Nonparametric predictive inference for test reproducibility by sampling future data orderings

Frank P.A. Coolen, *Durham University,*
Durham, UK.
Email: frank.coolen@durham.ac.uk
Filipe J. Marques, *Universidade Nova de Lisboa,*
Lisbon, Portugal.
Email: fjm@fct.unl.pt

Abstract

This paper considers nonparametric predictive inference (NPI) for reproducibility of likelihood ratio tests with the test criterion in terms of the sample mean. Given a sample of size n used for the actual test, the NPI approach provides lower and upper probabilities for the event that a repeat of the test, also with n observations, will lead to the same overall test conclusion, that is rejecting a null-hypothesis or not. This is achieved by considering all orderings of n future observations among the n data observations, which based on an exchangeability assumption are equally likely. However, exact lower and upper probabilities can only be derived for relatively small values of n due to computational limitations. Therefore, the main aim of this paper is to explore sampling of the orderings of the future data among the observed data in order to approximate the lower and upper reproducibility probabilities. The approach is applied for the Exponential and Normal distributions and the performance of the ordering sampling for approximation of the NPI lower and upper reproducibility probabilities is investigated. An application with real data of the methodology developed is provided.

AMS Subject Classification: 62A99, 62G99 and 62P30

Keywords: Exponential family; likelihood ratio test; lower and upper probabilities; nonparametric predictive inference; normal distribution; reproducibility probability; sampling orderings of future observations.

1 Introduction

The reproducibility probability (RP) of a statistical test measures how likely it is that if a statistical test were repeated under the same circumstances, it would lead to the same conclusion, that is the rejection or non-rejection of the null hypothesis. This is an important property which was first addressed by Goodman (1992) and then by Shao and Chow (2002); De Martini (2008); De Capitani and De Martini (2011); Shao and Chow (2002) who dealt with this issue as being an estimation of the power of a test problem.

In Coolen and Bin Himd (2014) a new perspective was presented using the nonparametric predictive inference (NPI) framework of frequentist statistical methods (Augustin and Coolen, 2004; Coolen, 2006, 2011). This NPI approach for the reproducibility probability of a test (NPI-RP) considers the test result for a predicted future sample of the same size as the original sample, this approach will be detailed in Section 2. The NPI approach for reproducibility of likelihood ratio tests was introduced in Marques et al. (2019a) and used in Marques et al. (2019b) to study the reprocibility of hypotheses testing between two Beta distributions. In Marques et al. (2019a) the authors considered only simple hypotheses and discussed the reproducibility property of some basic tests. In Marques et al. (2019a), only small samples of sizes 5 and 10 were considered, since for larger samples the methodology proposed is difficult to implement in computational terms. The number of orderings required for exact computation of the NPI lower and upper reproducibility probabilities is equal to $\binom{2n}{n}$, which is only feasible for small data sets. One way to overcome this problem is by using an NPI-based bootstrap method (Coolen and Bin Himd, 2020), which however has the disadvantage that imprecision is no longer present. To remain closer to the nature of NPI, using lower and upper probabilities to quantify uncertainty and reflecting the amount of information through imprecision, we propose and alternative computational method, namely by estimating the NPI lower and upper RP probabilities via sampling of the future orderings.

In this work, we study the reproducibility property of likelihood ratio tests for composite hypotheses on the mean value, such that the decision rule may be expressed in terms of the sample mean. Furthermore, a new sampling methodology is proposed, based on the sampling of future orderings, to overcome the computational limitations described in Marques et al. (2019a) to address scenarios with larger sample sizes.

Although, the sampling of future orderings technique is illustrated for likelihood ratio tests with test criterion based on the sample mean, which somehow limits the distributions that may be assumed in this testing procedure, this methodology may also be applied to other testing procedures for which the test criterion may be expressed in terms of the sample mean. Moreover, convergence theorems such as the Central Limit Theorem allow us to extend the results to other distributions and even in cases where the distributions are not known. We emphasize that the sampling of orderings to estimate NPI lower and upper RPs is far more widely applicable, and enables the NPI-RP approach to be implemented to a wide range of statistical tests and not only for likelihood ratio tests.

This paper is organized as follows; in Section 2 we present the methodology used to compute the lower and upper reproducibility probabilities for likelihood ratio tests on the mean value where decision rule is based on the mean sample. In Section 3, the computation of the reproducibility probabilities is illustrated for samples from the Normal and Exponential distributions. The new sampling methodology is introduced in Section 4 together with numerical studies which show the adequacy of the methodology proposed. Section 5 is dedicated to simulations. In Section 6, the procedure for the computation of lower and upper reproducibility probabilities, for two sided tests, is briefly illustrated. An application with real data is presented in Section 7. The discussions and conclusions are presented in Section 8. troduces NPI-RP to the important setting of likelihood ratio tests (LRT). These tests were introduced by Neyman and Pearson in 1928 and since then have been widely applied in the most different fields of statistics, for example, applications can be easily found in engineering, economics, medicine and

2 NPI-RP for LRT

As mentioned in Marques et al. (2019a), nonparametric predictive inference (Augustin and Coolen, 2004; Coolen, 2006, 2011) is a frequentist statistical method based on Hill's assumption $A_{(n)}$ (Hill, 1968). This assumption considers a single future real-valued observation X_{n+1}, given n data observations, with the assumption that there are no ties among the data (this assumption will also be made throughout this paper), and assigns probability $1/(n+1)$ for X_{n+1} to each open interval in the partition created by the n observations. We denote the n data observations by $x_1 < x_2 < \ldots < x_n$. For distributions with unlimited support we have to define bounds which we denote by $x_0 = L$ and $x_{n+1} = R$. No further assumptions are made, in particular not on the distribution of the probability $1/(n+1)$ within each interval.

This assumptions may be generalized for $m \geq 1$ future real-valued observations, based on n data observations, considering the sequential assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ (Arts et al., 2004), These assumptions lead to the following inferential method: given n data observations and m future observations, the $\binom{m+n}{m}$ different orderings of all these observations are all equally likely, with again no further assumptions on where future observations would be within intervals between consecutive data observations.

Considering $m = n$, we denote the $\binom{2n}{n}$ different orderings of the n future real-valued observations among the n data observations, by O_j for $j = 1, \ldots, \binom{2n}{n}$. Each ordering O_j can be represented by $(s_1^j, \ldots, s_{n+1}^j)$, where s_i^j is the number of future observations in the interval (x_{i-1}, x_i), according to ordering O_j. Here $s_i^j \geq 0$ and $\sum_{i=1}^{n+1} s_i^j = n$.

The general idea of the NPI-RP approach is as follows (Coolen and Bin Himd, 2014; Marques et al., 2019a). Given n real-valued observations for which the original test is performed, we consider the $\binom{2n}{n}$ different orderings of the n future observations among the n data observations; these orderings all have the same probability $\binom{2n}{n}^{-1}$ to occur. For each such future ordering O_j, we do not know precise values of the future data, but O_j specifies the number s_i^j of observations in the interval (x_{i-1}, x_i), for each $i = 1, \ldots, n+1$. For these future observations nothing more is assumed, so they can take on any value within the specific interval. We wish to perform the same test on the future data as was applied to the real data, hence we need to consider the mean of the n future observations for each ordering O_j.

In this work, we are considering likelihood ratio tests which result in the following test criterion involving the mean of the observed values. We will mainly consider a null hypothesis H_0 with a single-sided alternative hypothesis, leading (without loss of generality) to the test criterion that H_0 is rejected if and only if

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

with c dependent on the assumed statistical model and significance level for the test.

When considering a specific ordering O_j of the n future observations in the NPI approach, we cannot derive a precise value for their mean, as we do not assume precise values within the intervals (x_{i-1}, x_i). Hence, we can only derive the maximum lower bound and minimum upper bound for the mean corresponding to O_j, we denote these by $\underline{m}_j$ and $\overline{m}_j$, respectively. They are easily derived as

$$\begin{aligned} \underline{m}_j &= \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_{i-1} && (2) \\ \overline{m}_j &= \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_i && (3) \end{aligned}$$

where s_i^j is the number of future observations in the interval (x_{i-1}, x_i), according to ordering O_j. Suppose that the original data sample of size n led to rejection of H_0, so their mean exceeds c. Then, this test result is reproduced if the future sample, also of size n, also leads to rejection of H_0. For an ordering O_j, this occurs certainly if $\underline{m}_j > c$, while it certainly does not occur if $\overline{m}_j \leq c$. However, if $\underline{m}_j \leq c < \overline{m}_j$ then we cannot conclude whether the original test result is reproduced or not. Remembering that all orderings O_j are equally likely, we derive the NPI lower and upper probabilities for test reproducibility, for the case that H_0 was rejected for the original test data, as

$$\begin{aligned} \underline{RP} &= \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > c\} && (4) \\ \overline{RP} &= \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > c\} && (5) \end{aligned}$$

where the summations are taken over $j = 1, 2, \ldots, \binom{2n}{n}$ and $\mathbf{1}\{A\}$ is the indicator function which is equal to 1 if A is true and 0 if A is not true.

If the original data did not lead to rejection of H_0, then the reasoning is similar and the resulting NPI lower and upper probabilities for test reproducibility are

$$\begin{aligned} \underline{RP} &= \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j \leq c\} && (6) \\ \overline{RP} &= \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j \leq c\}\,. && (7) \end{aligned}$$

It is now easy to see why we must assume a finite range $[L, R]$ of possible values for the future observations, where L and R can depend on the actual data observations.

Suppose that we did not restrict the range, and allowed all real values in the method described above. Half of the $\binom{2n}{n}$ orderings have at least one observation in the interval $(-\infty, x_1)$. This follows easily from combinatorics, as it requires the smallest future observation to be in this interval while the $n-1$ further observations can be anywhere, hence there are $\binom{2n-1}{n-1}$ orderings with this property, which is precisely half of the $\binom{2n}{n}$ orderings. A further simple argument to see that indeed precisely half of the orderings have at least one future observation that is less than x_1 follows immediately from the assumed exchangeability of all $2n$ observations, which implies that the smallest of these $2n$ observations has equal probability to be among the n observations in the data set and to be among the n future observations. If we allow $x_0 = -\infty$, then for each of these orderings, $\underline{m}_j < c$ for all finite values of c. If we allow $x_{n+1} = \infty$, by the same reasoning, also half the orderings have $\overline{m}_j > c$ for all finite c. Hence, we would have $\underline{RP} \leq 0.5$ and $\overline{RP} \geq 0.5$ for all values of n and for any possible data set. Furthermore, there are $\binom{2n-2}{n-2}$ orderings which have both at least one future observation less than x_1 and at least one future observation greater than x_n. Note that this is about a quarter of all the orderings. For these orderings, $\underline{m}_j < c$ and $\overline{m}_j > c$ for all finite c, which implies that the imprecision in the reproducibility inferences, that is $\overline{RP} - \underline{RP}$, is at least (about) 0.25, for all values of n. As we will see later, by assuming finite values for L and R we will get inferences that have more attractive properties for increasing values of n, in particular reducing imprecision.

Of course, we need to assume values $L < x_1$ and $R > x_n$ such that the data observations are within $[L, R]$. For the specific choice of L and R, however, there do not appear to be compelling theoretic arguments, yet it does influence the reproducibility inference. In Section 5, we will explore a few heuristic arguments for specific choices of L and R and we will investigate the sensitivity of the inferences to the choice of these values.

In the next section, we will illustrate the NPI-RP methodology for a few scenarios with small sample sizes, so that these NPI lower and upper reproducibility probabilities can be computed exactly. Thereafter, using sampling of the orderings, we will consider larger samples sizes.

3 Applications for small sample sizes

We consider likelihood ratio tests for the mean value, for which the hypotheses may be formulated as

$$H_0 : \mu \leq \mu_0 \ \text{ vs } \ H_1 : \mu > \mu_0 \tag{8}$$

and such that the test criterion may be expressed in terms of the sample mean. The main focus of the paper is on likelihood ratio tests, but the methodology proposed can also be applied to other test procedures. We will apply the methodology introduced in Section 2 when the underlying population has an Exponential or Normal distribution. Note that the Normal distribution, for known variance, and the Exponential distribution belong to the regular Exponential family with one unknown parameter, for which the density function may be expressed as

$$f_X(x|\theta) = h(x)g(\theta)\exp\{t(x)w(\theta)\}$$

and that if $w(\theta)$ is an increasing function of θ, then we have a monotone likelihood ratio in $T(\underline{X}) = \sum_{i=1}^{n} t(X_i)$. Both the Exponential and Normal distribution have monotone likelihood ratios in the statistic $T(\underline{X}) = \sum_{i=1}^{n} X_i$. In this section we consider the method described in Section 2 for small sample sizes. In this case, all the $\binom{2n}{n}$ orderings are considered for the computation of the upper and lower RPs. Both for the Normal and Exponential distributions, a sample of size n, $X_1, \dots, X_n$, is considered to test the null hypothesis in (8). The decision rule for the likelihood ratio test may be expressed in terms of the sample mean $\overline{X}$ and the test criterion in (1) is to reject the null hypothesis, for a significance level α, if

$$\overline{X} > q_{1-\alpha} \tag{9}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of $\overline{X}$. It is well known that for independent and identically distributed $X_i \sim \mathrm{Exp}(\lambda)$, $i = 1, \dots, n$, the distribution of the mean is

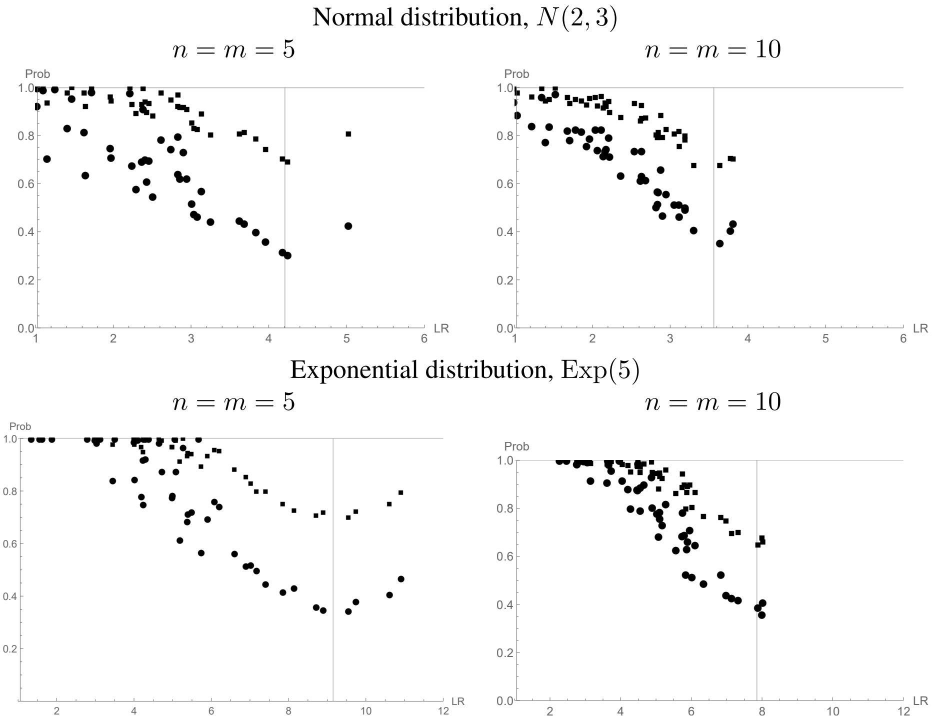
$$\overline{X} \sim \mathrm{Gamma}(n, \lambda/n)$$

and for $X_i \sim \mathrm{N}(\mu, \sigma)$, $i = 1, \dots, n$,

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

thus the quantile $q_{1-\alpha}$ can be easily determined in both cases. To apply the methodology introduced in Section 2, and since the distributions considered are not bounded, we need to define upper and lower limits, L and R. Our first heuristic approach is as follows. For n data observations, $x_1 < x_2 < \cdots < x_n$, L and R may be defined as $L = x_1 - \frac{x_n - x_1}{n-1}$ and $R = x_n + \frac{x_n - x_1}{n-1}$. In Section 5 we will discuss and analyze two other options for the choice of L and R. Note that when the underlying distribution is Exponential we use $L = 0$.

In Figure 1, we simulated 50 replicates of samples of sizes $n = 5$ and $n = 10$ extracted from the Normal distribution, $\mathrm{N}(2, 3)$, and also from the Exponential distribution with expected value 5, $\mathrm{Exp}(5)$, that is we simulated under the null hypothesis in (8), considering $\mu_0 = 2$ for the Normal distribution and $\mu_0 = 5$ for the Exponential distribution. The decision rule is given in (9). For original samples of sizes $n = 5$ and $n = 10$, and for future samples of the same size there are respectively 252 and 184756 possible orderings. All the orderings were considered in the computation of the upper and lower RPs. The observed likelihood ratio statistic and the upper and lower RPs were determined for each of the 50 replications. In Figure 1, the vertical line indicates the $q_{0.95}$ quantile, the black circles and squares are respectively the lower and upper RPs. From Figure 1, it is possible to observe similar patterns to the ones described in the paper that introduced this topic (Marques et al., 2019a). The upper and lower RPs tend to increase when $|LR_{obs} - q_{0.95}|$ increases and it seems that there is some oscillation of the values of the RPs. This may be due to the method used to define L and R or to the definition of the lower and upper RPs. These properties are also present in Table 1 where three samples of sizes 5 and 10 from the $N(2, 3)$ and $Exp(5)$ are considered. In this table, for each sample, the observed sample mean, the test threshold and lower and upper RPs are presented.


Normal distribution, $N(2,3)$

$n = m = 5$

$n = m = 10$

Exponential distribution, $\mathrm{Exp}(5)$

$n = m = 5$

$n = m = 10$

Figure 1: Values of the upper (squares) and lower (circles) RPs, for 50 replications of samples of sizes 5 and 10 from the Normal distribution with mean 2 and standard deviation 3 and also from the Exponential distribution with mean 5. The vertical line indicates the $q_{0.95}$ quantile.

4 Sampling of orderings for larger sample sizes

As there are $\binom{2n}{n}$ orderings O_j of n future observations among n data observations, exact computation of the NPI lower and upper reproducibility probabilities, as presented in the previous section, is only possible for small values of n. One possibility to overcome this is by sampling the orderings, and for each sampled ordering determine $\underline{m}_j$ and $\overline{m}_j$, which enables estimation of $\underline{RP}$ and $\overline{RP}$ following the standard theory of estimation of proportions, as long as the sampling process of the orderings satisfies the conditions for 'simple random sampling' (SRS). There are several possible ways to achieve this, one that is easy to implement is explained below and implemented in the computations in this paper. Note that the standard theory of estimation of proportions also enables us to determine a suitable size for the sample of orderings, depending on a required accuracy of the estimates.

For SRS, at each selection of an ordering to be included in the sample, each ordering must have the same probability of being selected, and the selection of an ordering should be independent of the other selections. It is important to emphasize that, once n

Normal distribution with $\mu = 2$ and $\sigma = 3$					
		Sample mean	Test threshold	$\underline{RP}$	$\overline{RP}$
$n = m = 5$	Sample 1	4.680	4.207	0.36	0.74
	Sample 2	-0.128	4.207	0.94	1.00
	Sample 3	1.824	4.207	0.70	0.94
$n = m = 10$	Sample 1	2.012	3.560	0.90	0.97
	Sample 2	1.917	3.560	0.72	0.92
	Sample 3	2.201	3.560	0.78	0.93
Exponential distribution with parameter $\lambda = 5$					
		Sample mean	Test threshold	$\underline{RP}$	$\overline{RP}$
$n = m = 5$	Sample 1	6.919	9.154	0.49	0.81
	Sample 2	3.661	9.154	0.98	1.00
	Sample 3	4.826	9.154	0.90	0.99
$n = m = 10$	Sample 1	3.348	7.853	0.97	1.00
	Sample 2	6.986	7.853	0.46	0.75
	Sample 3	4.971	7.853	0.73	0.93

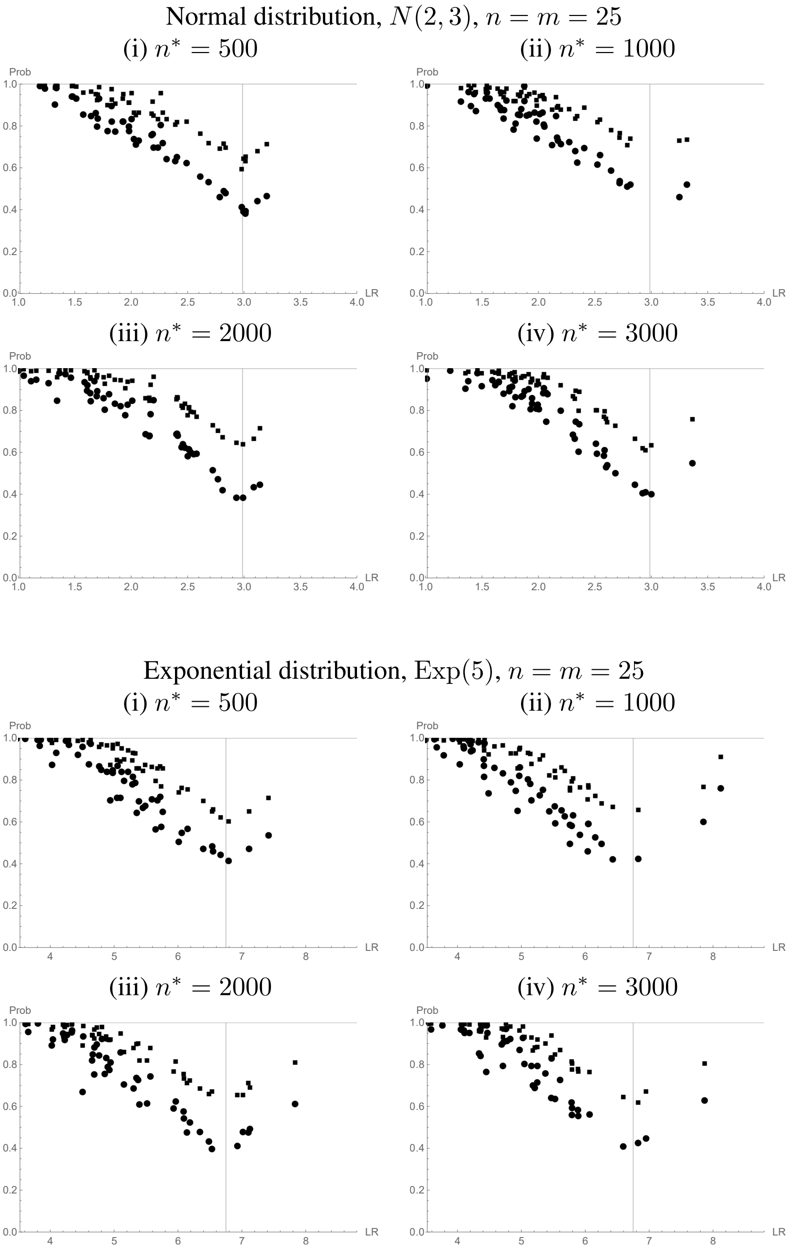
Table 1: Upper and lower RPs for three observed samples of sizes $n = 5$ and $n = 10$

is not very small, the total number of orderings is large enough in order to neglect any possible differences between sampling with or without replacement, for simplicity we will sample with replacement throughout this paper.

Each ordering O_j, for $j = 1, \ldots, \binom{2n}{n}$, of n future observations is characterized by $(s_1^j, \ldots, s_{n+1}^j)$, where s_i^j is the number of future observations in the interval (x_{i-1}, x_i). Hence, for SRS of the orderings we must randomly select such vectors. An easy way to do this is by simple random sampling of a vector of integers $(r_1, \ldots, r_n)$, with $r_1 \geq 1$, $r_l > r_{l-1}$ for all $l = 2, \ldots, n$ and $r_n \leq 2n$. Then r_l is considered to be the rank of the j-th ordered data observation among the $2n$ combined data and future observations. Defining $r_0 = 0$ and $r_{n+1} = 2n+1$, and with a sampled vector $(r_1, \ldots, r_n)$, we define $s_l^j = r_l - r_{l-1} - 1$ for $l = 1, \ldots, n+1$, thus creating the j-th sampled future ordering in the SRS process. It is easy to verify that this process ensures that, at each selection, each of the possible orderings is equally likely to be selected, and as each selection is executed independently of the other selections, the conditions for SRS are satisfied.

To illustrate this SRS methodology, 50 replications of samples of size 25 were considered for the Normal distribution $N(2, 3)$ and for the Exponential distribution $\text{Exp}(5)$. The computation of the lower and upper RPs was achieved by sampling orderings of sizes $n^* = 500, 1000, 2000, 3000$.

Figure 2, shows that there are no substantial differences on the patterns for different values of n^*. In Tables 2, 3 and 4, we computed the lower and upper RPs and the corresponding 95% confidence intervals, for three samples of sizes 25, 50 and 100, and for different numbers of orderings sampled. From these tables we may see that reasonable results, for the approximating values of the lower and upper RPs, may be obtained by considering the number orderings sampled equal or greater than 2000 which is a quite small number when compared with the number of all possible orderings. This is an ex-



Normal distribution, $N(2,3)$, $n = m = 25$

(i) $n^* = 500$ (ii) $n^* = 1000$

(iii) $n^* = 2000$ (iv) $n^* = 3000$

Exponential distribution, $\mathrm{Exp}(5)$, $n = m = 25$

(i) $n^* = 500$ (ii) $n^* = 1000$

(iii) $n^* = 2000$ (iv) $n^* = 3000$

Figure 2: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of the original sample of size $n = 25$ from the Normal distribution with mean 2 and standard deviation 3, and also from the Exponential distribution with parameter 5, when $n = m$ and $n^* = 500, 1000, 2000, 3000$. The vertical line indicates the $q_{0.95}$ quantile of $\overline{X}$.

Normal distribution with $\mu = 2$ and $\sigma = 3$

n^*	lower	CI(0.95)	upper	CI(0.95)
100	0.63000	(0.53537,0.72463)	0.85000	(0.78002,0.91999)
500	0.59400	(0.55096,0.63705)	0.81000	(0.77561,0.84439)
1000	0.62400	(0.59398,0.65402)	0.81400	(0.78988,0.83812)
2000	0.64250	(0.62150,0.66350)	0.82300	(0.80627,0.83973)
5000	0.63320	(0.61984,0.64656)	0.81500	(0.80424,0.82576)
10000	0.62070	(0.61119,0.63021)	0.80540	(0.79764,0.81316)
20000	0.62705	(0.62035,0.63375)	0.81340	(0.80800,0.81880)
50000	0.62418	(0.61994,0.62843)	0.81164	(0.80821,0.81507)
100000	0.62756	(0.62456,0.63056)	0.81265	(0.81023,0.81507)

Exponential distribution with parameter $\lambda = 5$

100	0.79000	(0.71017,0.86983)	0.92000	(0.86683,0.97317)
500	0.76600	(0.72889,0.80311)	0.91000	(0.88492,0.93508)
1000	0.72100	(0.69320,0.74880)	0.88200	(0.86200,0.90200)
2000	0.73350	(0.71412,0.75288)	0.88050	(0.86628,0.89472)
5000	0.74460	(0.73251,0.75669)	0.88680	(0.87802,0.89558)
10000	0.74660	(0.73807,0.75513)	0.88490	(0.87864,0.89116)
20000	0.74165	(0.73558,0.74772)	0.88515	(0.88073,0.88957)
50000	0.73368	(0.72981,0.73755)	0.87696	(0.87408,0.87984)
100000	0.73889	(0.73617,0.74161)	0.88308	(0.88109,0.88507)

Table 2: Upper and lower RPs for an observed sample of size $n = 25$ and for increasing values of the number of the orderings sampled

cellent property of the SRS methodology that allows us to overcome the computational difficulties, identified in Marques et al. (2019a), in the computation of the lower and upper RPs, for samples larger than 10. Of course, if one wants to obtain more accurate approximations for the RPs, then one can increase the number of orderings sampled.

In order to assess the precision of the SRS methodology for the computation of the lower and upper RPs when the orders are sampled, we computed the exact lower and upper RPs, for $n = m = 14$, considering the 40 116 600 possible orderings and assuming as underlying distributions the Normal and Exponential distributions. Then, using the SRS technique, we performed 100 simulations with n^* equal to 100 and 1000, and in each simulation the 95% confidence interval was computed, for both the lower and upper RPs. The coverage probabilities of the 95% confidence intervals, assuming the Normal distribution with $\mu = 2$ and $\sigma = 3$, based on an observed sample of size 14, for the exact values $\overline{RP} = 0.87$ and $\underline{RP} = 0.70$, were respectively 0.96 and 0.95 for $n^* = 100$ and 0.99 and 1.00 for $n^* = 1000$. For the Exponential distribution with parameter, $\lambda = 5$, the coverage probabilities of the 95% confidence intervals for the exact values $\overline{RP} = 0.72$ and exact $\underline{RP} = 0.49$, were respectively 0.96 and 0.98 for $n^* = 100$ and 0.99 and 1.00 for $n^* = 1000$. For the computation of these confidence intervals for proportions, p, we used the standard result based on the normal

Normal distribution with $\mu = 2$ and $\sigma = 3$

		$n = 50$		
n^*	lower	CI(0.95)	upper	CI(0.95)
100	0.68000	(0.58857,0.77143)	0.82000	(0.74470,0.89530)
500	0.70600	(0.66607,0.74593)	0.84000	(0.80787,0.87213)
1000	0.70400	(0.67571,0.73229)	0.83000	(0.80672,0.85328)
2000	0.71050	(0.69062,0.73038)	0.83950	(0.82341,0.85559)
5000	0.71340	(0.70087,0.72593)	0.83620	(0.82594,0.84646)
10000	0.71540	(0.70656,0.72424)	0.83930	(0.83210,0.84650)
20000	0.71550	(0.70925,0.72175)	0.83970	(0.83462,0.84478)
50000	0.71628	(0.71233,0.72023)	0.84118	(0.83798,0.84438)
100000	0.71804	(0.71525,0.72083)	0.84131	(0.83905,0.84357)

Exponential distribution with parameter $\lambda = 5$

100	0.57000	(0.47297,0.66703)	0.73000	(0.64299,0.81701)
500	0.61000	(0.56725,0.65275)	0.75400	(0.71625,0.79175)
1000	0.63600	(0.60618,0.66582)	0.79100	(0.76580,0.81620)
2000	0.61600	(0.59468,0.63732)	0.76200	(0.74334,0.78066)
5000	0.63000	(0.61662,0.64338)	0.78420	(0.77280,0.79560)
10000	0.62700	(0.61752,0.63648)	0.77120	(0.76297,0.77943)
20000	0.63075	(0.62406,0.63744)	0.77430	(0.76851,0.78009)
50000	0.62982	(0.62559,0.63405)	0.77796	(0.77432,0.78160)
100000	0.62658	(0.62358,0.62958)	0.77482	(0.77223,0.77741)

Table 3: Upper and lower RPs for an observed sample of size 50 and for increasing values of the number of the orderings sampled

approximation

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

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution, and $\hat{p}$ is the estimated value of the lower or upper RP. In the literature (Brown et al., 2001; Mantalos and Zografos, 2008), it is well documented that some problems may arise when using these intervals for proportions; this may explain the high values for the coverage probabilities when $n^* = 1000$. However, our interest is only to document the reliability of the SRS method in the computation of approximating values for the lower and upper RPs and this is well stated in the results presented.

5 Simulations

In this section we will (i) analyze the impact of the increase of the sample size in the patterns of the lower and upper RPs, (ii) compare the results obtained for the lower and upper RPs considering three different methods to determine L and R, and (iii) illustrate, through simulations, the changes on the pattern of the upper and lower RPs when the simulations are performed by sampling under the alternative hypothesis.

Normal distribution with $\mu = 2$ and $\sigma = 3$

		$n = 100$		
n^*	lower	CI(0.95)	upper	CI(0.95)
100	0.88000	(0.81631,0.94369)	0.93000	(0.87999,0.98001)
500	0.79800	(0.76281,0.83319)	0.88000	(0.85152,0.90848)
1000	0.81400	(0.78988,0.83812)	0.87800	(0.85771,0.89829)
2000	0.82650	(0.80990,0.84310)	0.90050	(0.88738,0.91362)
5000	0.81280	(0.80199,0.82361)	0.89200	(0.88340,0.90060)
10000	0.82110	(0.81359,0.82861)	0.89310	(0.88704,0.89916)
20000	0.81910	(0.81377,0.82443)	0.89435	(0.89009,0.89861)
50000	0.81872	(0.81534,0.82210)	0.89284	(0.89013,0.89555)
100000	0.81824	(0.81585,0.82063)	0.89200	(0.89008,0.89392)

Exponential distribution with parameter $\lambda = 5$

100	0.80000	(0.72160,0.87840)	0.89000	(0.82867,0.95133)
500	0.80800	(0.77348,0.84252)	0.91000	(0.88492,0.93508)
1000	0.85000	(0.82787,0.87213)	0.91300	(0.89553,0.93047)
2000	0.82750	(0.81094,0.84406)	0.90400	(0.89109,0.91691)
5000	0.83340	(0.82307,0.84373)	0.91480	(0.90706,0.92254)
10000	0.82630	(0.81887,0.83373)	0.90680	(0.90110,0.91250)
20000	0.83170	(0.82651,0.83689)	0.90640	(0.90236,0.91044)
50000	0.82694	(0.82362,0.83026)	0.90514	(0.90257,0.90771)
100000	0.82877	(0.82644,0.83110)	0.90896	(0.90718,0.91074)

Table 4: Upper and lower RPs for an observed sample of size 100 and for increasing values of the number of the orderings sampled

5.1 Sample size effect on the lower and upper RPs

To study the impact of the sample size on the upper and lower RPs, we consider the same number of orderings sampled, $n^* = 2000$, and different sample sizes $n = 25, 50, 100, 200$. For each sample size we considered 50 replications from the Normal distribution, $N(2, 3)$ and from the Exponential distribution, $\mathrm{Exp}(5)$. The results are presented in Figures 3 and 4.

Normal distribution, $N(2,3)$, $n^* = 2000$

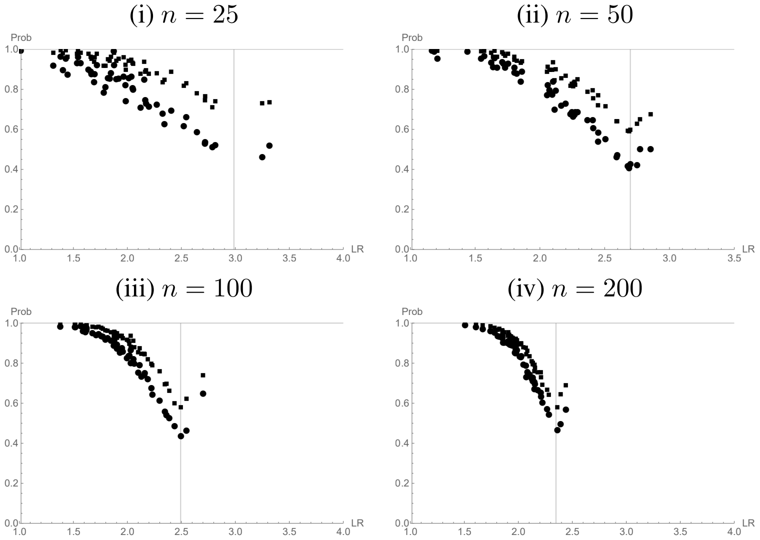
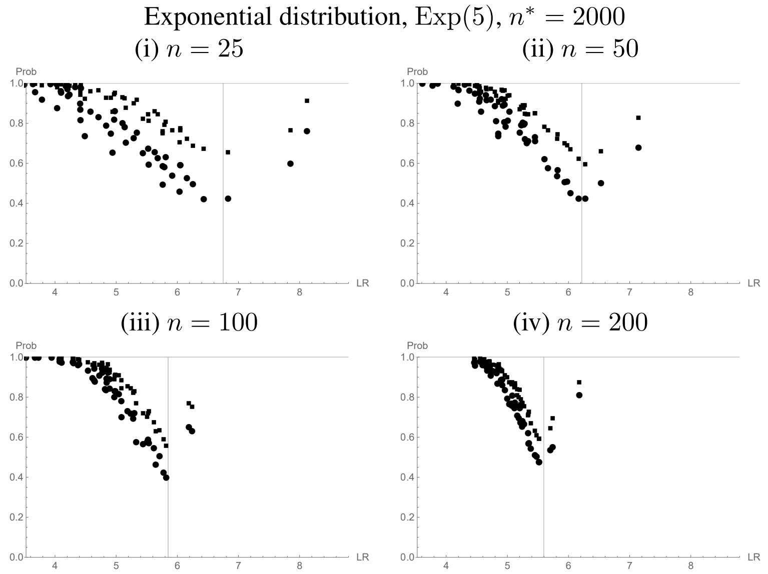


Figure 3: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of different sizes, n, of the original sample from the Normal distribution with mean 2 and standard deviation 3 when $n = m$ and $n^* = 2000$. The vertical line indicates the $q_{0.95}$ quantile of $\overline{X}$.



Exponential distribution, $\mathrm{Exp}(5)$, $n^* = 2000$

(i) $n = 25$

(ii) $n = 50$

(iii) $n = 100$

(iv) $n = 200$

Figure 4: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of different sizes, n, of the original sample from the Exponential distribution with parameter 5, when $n = m$ and $n^* = 2000$. The vertical line indicates the $q_{0.95}$ quantile of $\overline{X}$.

In Figures 3 and 4, it is interesting to note that, when n increases, the lower RP value seems to tend to 0.5 when the observed likelihood ratios are close to the quantile and the range between the lower and upper RPs decreases.

5.2 Choice of L and R

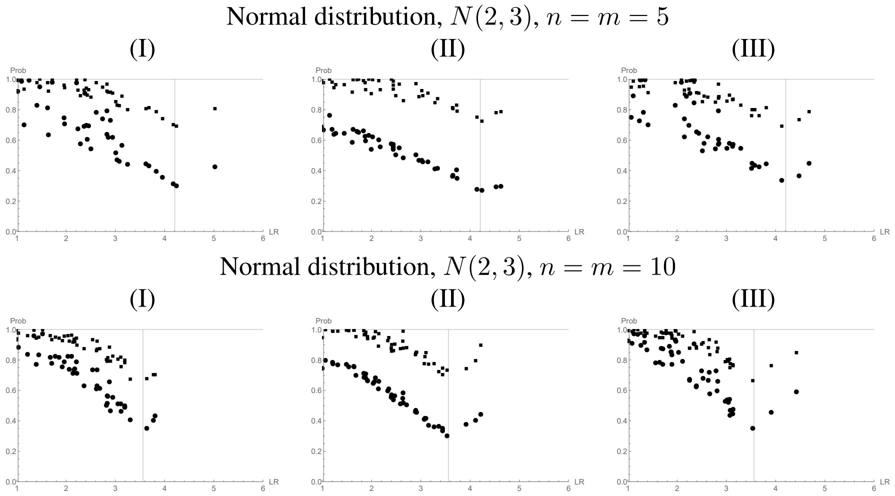
In this section we consider three heuristic methods to determine the values of L and R, and the results obtained in the computation of the lower and upper RPs are compared. For n data observations, $x_1 < x_2 < \cdots < x_n$, L and R may be defined as follows

(I) the method already introduced in Section 3 and used throughout this work, with $L = x_1 - \frac{x_n - x_1}{n-1}$ and $R = x_n + \frac{x_n - x_1}{n-1}$;

(II) $L = q_\alpha$ and $R = q_{1-\alpha}$, for $\alpha = 0.0001$, where q_α and $q_{1-\alpha}$ are the α and $1 - \alpha$ quantiles of the distribution assumed under H_0, provided that all data observations are within (L, R);

(III) $L = \frac{x_1 + x_n}{2} - y$ and $R = \frac{x_1 + x_n}{2} + y$; y is defined, for the Normal case, as the solution of the system of equations

$$\begin{cases} P\left(W_1 \geq \dfrac{x_1 + x_n}{2} + y\right) = \dfrac{1}{2n+1} & \text{with } W_1 \sim \mathrm{N}\left(\frac{x_1+x_n}{2}, \sigma\right) \\ P\left(W_1 \geq x_n\right) = \dfrac{1}{n+1} & \text{with } W_1 \sim \mathrm{N}\left(\frac{x_1+x_n}{2}, \sigma\right). \end{cases}$$


Normal distribution, $N(2,3)$, $n=m=5$

(I) (II) (III)

Normal distribution, $N(2,3)$, $n=m=10$

(I) (II) (III)

Figure 5: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of samples of sizes 5 and 10 from the Normal distribution with mean 2 and standard deviation 3, using methods (I), (II) and (III). The vertical line indicates the $q_{0.95}$ quantile.

For the Exponential case, using a similar method, R is defined as

$$R = \frac{\log(2n+1)}{\log(n+1)} X_n \, .$$

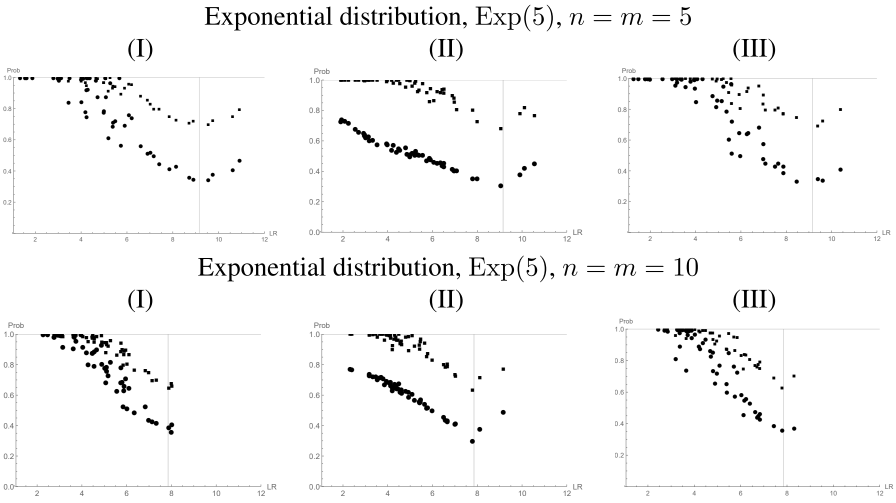
We should point out that for the Exponential distribution, in the previous three methods and throughout the paper, L is considered to be equal to 0.

In Figures 5 and 6, we simulated 50 replicates of samples of sizes $n = 5$ and $n = 10$ extracted from the Normal distribution, $\mathrm{N}(2,3)$, and also from the Exponential distribution, $\mathrm{Exp}(5)$. Since the samples are small, for each replications, we computed the exact lower and upper RPs.

From Figures 5 and 6 we may see that the patterns are similar using methods (I), (II) or (III). Although, it seems that with method (II) there is less oscillation of the values of $\underline{RP}$ and $\overline{RP}$, but there is more imprecision away from the test threshold. In Table 5, we considered three samples and computed the values of the $\underline{RP}$ and $\overline{RP}$ using the three methods. From this table we may see that, for the same observed sample, the upper and lower RPs may differ substantially, mainly between method (II) and methods (I) and (III). This may be explained by the fact that, in method (II), using the 0.0001 upper or lower quantiles we may be considering the first or last intervals with a larger length than the ones obtained using methods (I) and (III).

5.3 Simulations under the alternative hypothesis

This subsection illustrates, through simulations, the changes on the pattern of the upper and lower RPs when the simulations are performed assuming the underlying popula-



Exponential distribution, $\mathrm{Exp}(5)$, $n = m = 5$

(I) (II) (III)

Exponential distribution, $\mathrm{Exp}(5)$, $n = m = 10$

(I) (II) (III)

Figure 6: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of samples of sizes 5 and 10 from the Exponential distribution, $\mathrm{Exp}(5)$, using methods (I), (II) and (III). The vertical line indicates the $q_{0.95}$ quantile.

tions under the alternative hypothesis. For the hypotheses in (8), and for the Normal case, we consider samples of size 25 from the following distributions (i) $N(2, 3)$ (ii) $N(\delta_1, 3)$ and (iii) $N(2\delta_1 - 2, 3)$ where $\delta_1 = q_{1-\alpha}$ is the threshold considered in the test criterion and marked with a vertical line in the plots of Figure 7. For the Exponential distribution, we also considered samples of size 25 from (i) $\mathrm{Exp}(5)$, (ii) $\mathrm{Exp}(\delta_2)$ and (iii) $\mathrm{Exp}(2\delta_2 - 5)$, where $\delta_2 = q_{1-\alpha}$ is also the threshold considered in the test criterion and is also marked with a vertical line in the plots. We considered 50 replications, and in all cases we defined $n^* = 2000$ and used the method (I), introduced in Section 3, to define L and R. In the results presented in Figure 7 we may observe a change on the pattern of the RPs obtained through simulations, together with the change of "side" of the observed mean values in relation to the threshold considered.

6 Two sided test

This work focused on the right sided test, however a similar procedure may be implemented for the left sided test or even for the two sided tests. In this section, we illustrate briefly, the procedure for the computation of the upper and lower RPs for the two sided test. Suppose we want to test the hypotheses

$$H_0 : \mu = \mu_0 \ \text{ vs } \ H_1 : \mu \neq \mu_0 \,. \tag{10}$$

The decision rule for this test may be expressed in terms of the sample mean $\overline{X}$ and the test criterion is to reject the null hypothesis, for a significance level α, if

$$\overline{X} < q_{\alpha/2} \ \vee \ \overline{X} > q_{1-\alpha/2} \tag{11}$$

Normal distribution with $\mu = 2$ and $\sigma = 3$							
		Sample 1		Sample 2		Sample 3	
$n = m = 5$	(I)	0.36	0.74	0.94	1.00	0.70	0.94
	(II)	0.33	0.72	0.77	1.00	0.60	0.94
	(III)	0.37	0.73	0.96	1.00	0.73	0.94
		Sample 1		Sample 2		Sample 3	
$n = m = 10$	(I)	0.90	0.97	0.72	0.92	0.78	0.93
	(II)	0.68	0.98	0.65	0.93	0.63	0.94
	(III)	0.89	0.97	0.72	0.92	0.78	0.93
Exponential distribution with parameter $\lambda = 5$							
		Sample 1		Sample 2		Sample 3	
$n = m = 5$	(I)	0.49	0.81	0.98	1.00	0.90	0.99
	(II)	0.40	0.81	0.59	1.00	0.54	0.99
	(III)	0.48	0.81	0.97	1.00	0.87	0.99
		Sample 1		Sample 2		Sample 3	
$n = m = 10$	(I)	0.97	1.00	0.46	0.75	0.73	0.93
	(II)	0.73	1.00	0.40	0.75	0.60	0.93
	(III)	0.95	1.00	0.43	0.75	0.70	0.93

Table 5: Upper and lower RPs for three observed samples of sizes $n = 5$ and $n = 10$ using the three methods

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

For a given order O_j, the maximum lower bound, $\overline{m}_j$, and minimum upper bound, $\underline{m}_j$, for the mean are still determined as in equations (3) and (2) respectively.

However, the procedure for the computation of $\underline{RP}$ and $\overline{RP}$ is different since it needs to account for the two rejection regions. Thus, for the two sided test, if the decision for the initial observed sample is the rejection of the null hypothesis, we have

$$\underline{RP} = \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > q_{1-\alpha/2} \;\vee\; \overline{m}_j < q_{\alpha/2}\} \tag{12}$$

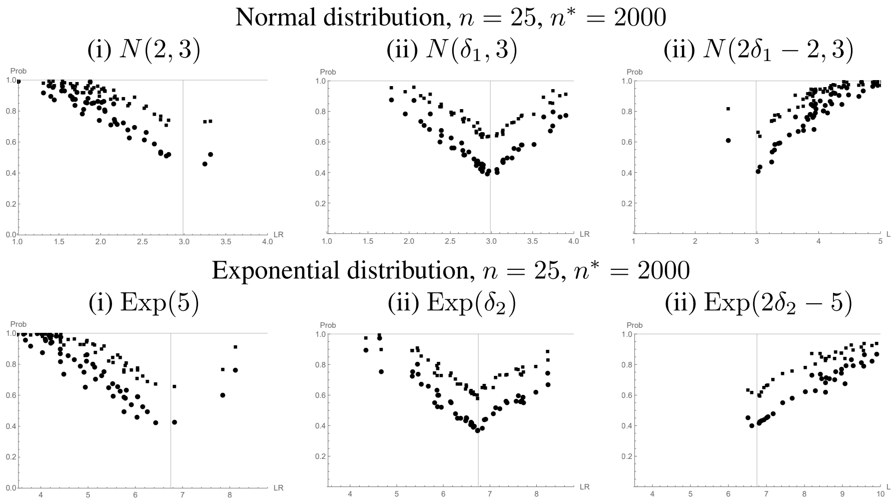
$$\overline{RP} = \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > q_{1-\alpha/2} \;\vee\; \underline{m}_j < q_{\alpha/2}\}\,. \tag{13}$$

If the original data did not lead to rejection of H_0, then the lower and upper RPs are determined as follows

$$\underline{RP} = \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > q_{\alpha/2} \;\wedge\; \overline{m}_j < q_{1-\alpha/2}\} \tag{14}$$

$$\overline{RP} = \binom{2n}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > q_{\alpha/2} \;\wedge\; \underline{m}_j < q_{1-\alpha/2}\}\,. \tag{15}$$

In Figure 8, we have simulated from the Normal and Exponential distributions 50 samples of sizes 50, 100 and 200 and, in each case, we computed the lower and upper RPs.



Normal distribution, $n = 25$, $n^* = 2000$

(i) $N(2,3)$ (ii) $N(\delta_1,3)$ (ii) $N(2\delta_1 - 2, 3)$

Exponential distribution, $n = 25$, $n^* = 2000$

(i) $\mathrm{Exp}(5)$ (ii) $\mathrm{Exp}(\delta_2)$ (ii) $\mathrm{Exp}(2\delta_2 - 5)$

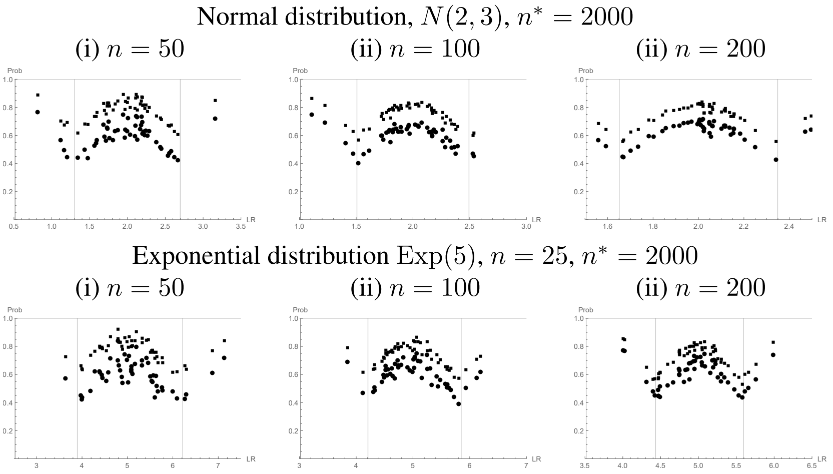
Figure 7: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of different sizes of the original sample $n = 25$, when $n = m$ and $n^* = 2000$. The vertical line indicates the $q_{0.95}$ quantile of $\overline{X}$.

Consistent with the methodology used throughout this work, we used method (I) to determine the values for the range $[L, R]$; remember that, for the Exponential distribution we used $L = 0$. In these computations we used the SRS methodology taking $n^* = 2000$.

Figure 8 shows that reproducibility for the two sided test is, of course, worst if the actual test results lead to mean values close to a test threshold. A main observation is that the maximum values of the reproducibility for cases where the null hypothesis is not rejected, so in between the two thresholds, may perhaps be lower than many would expect, with even the NPI upper RP value for some cases not going much above 0.8. For larger values of n the differences between upper and lower RP values decrease and randomness of the values is reduced, this is in line with our earlier observations for single-sided test.

7 An application

In this section we provide an illustration of a possible practical use of the results presented in this work. We consider the data available in *https://dasl.datadescription.com/datafile/farmed-salmon/* and in Hites (2004). From this data set, about organochlorine contaminants in farmed salmon, we selected the values of Mirex measured in farmed salmon in two countries, Chile and Scotland. It is known that high levels of concentration of Mirex may be dangerous for public health. In the American Cancer Society web-page the substance Mirex is listed as to be “reasonably anticipated to be human



Normal distribution, $N(2,3)$, $n^* = 2000$

(i) $n = 50$ (ii) $n = 100$ (ii) $n = 200$

Exponential distribution $\text{Exp}(5)$, $n = 25$, $n^* = 2000$

(i) $n = 50$ (ii) $n = 100$ (ii) $n = 200$

Figure 8: Simulated values of the upper (squares) and lower (circles) RPs, for 50 replications of different sizes of the original sample $n = 25$, when $n = m$ and $n^* = 2000$. The vertical line indicates the $q_{0.05}$ and $q_{0.95}$ quantiles of $\overline{X}$.

carcinogens". Thus, we may be interested in testing the hypotheses

$$H_0 : \mu \leq \ell \text{ vs } H_1 : \mu > \ell \tag{16}$$

where ℓ stands for a given limit above which the levels of Mirex may be considered dangerous. Instead of considering a specific value for ℓ we considered a range of values. For each value of ℓ we tested the null hypotheses in (16) and then we computed the lower and upper RPs of the test, which give us a prediction of how likely is to obtain the same decision if we repeat the test. For the computations of the lower and upper RPs we consider $n^* = 2000$. We have two samples of size 30, one from Chile and another from Scotland, of measurements of Mirex in farmed salmons. In the data set it is possible to find data from other countries and also measurements of other substances, however for this simple illustration we decided to consider only the data from Chile and Scotland of measurement of Mirex in farmed salmons.

The obtained values are presented in Tables 6 and 7. From these tables we may observe that, for the data from Chile, we have $\underline{RP} > 0.8$ for $\ell \leq 0.05$, thus in these cases, if the test were repeated in the same circumstances, we would expect the result to be rejection of the null hypothesis, and since $\underline{RP} > 0.8$ for $\ell \geq 0.06$ we would expect a non-rejection of the null hypothesis in a repetition of the test. Similar conclusions can be drawn from the data from Scotland, but for different values of ℓ, thus if the test were repeated in the same circumstances, we would expect the same decision for $\ell \leq 0.115$ –rejection– or for $\ell \geq 0.135$ – non-rejection of the null hypothesis.

Data from Chile

ℓ	Rejection	$\underline{RP}$	CI(0.95)	$\overline{RP}$	CI(0.95)
0.040	Yes	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.042	Yes	0.999	(0.999,1.000)	1.000	(1.000,1.000)
0.044	Yes	0.998	(0.996,1.000)	1.000	(0.999,1.000)
0.046	Yes	0.988	(0.983,0.993)	1.000	(0.999,1.000)
0.048	Yes	0.949	(0.939,0.958)	0.989	(0.984,0.994)
0.050	Yes	0.868	(0.853,0.883)	0.954	(0.944,0.963)
0.052	Yes	0.689	(0.668,0.709)	0.851	(0.835,0.867)
0.054	Yes	0.498	(0.476,0.519)	0.695	(0.675,0.715)
0.056	No	0.498	(0.476,0.520)	0.687	(0.667,0.707)
0.058	No	0.677	(0.656,0.697)	0.822	(0.805,0.838)
0.060	No	0.825	(0.808,0.841)	0.916	(0.903,0.928)
0.062	No	0.905	(0.892,0.918)	0.965	(0.956,0.973)
0.064	No	0.957	(0.948,0.965)	0.984	(0.979,0.989)
0.066	No	0.983	(0.977,0.989)	0.994	(0.991,0.997)
0.068	No	0.995	(0.991,0.998)	1.000	(0.999,1.000)
0.070	No	0.999	(0.998,1.000)	1.000	(0.999,1.000)
0.072	No	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.074	No	1.000	(1.000,1.000)	1.0000	(1.000,1.000)

Table 6: Upper and lower RPs for different values of ℓ.

8 Concluding remarks

In this paper we have investigated the use of sampling of orderings of future data among observed data for inference on reproducibility of tests. Due to the very large number of such orderings, for all but very small data sample sizes, one cannot consider all the orderings, and it was shown that simple random sampling leads to excellent approximations, already based on quite small numbers of orderings being sampled. In fact, both to approximate the NPI lower and upper reproducibility probabilities, the scenario is just regular estimation of a proportion, hence required numbers of orderings in order to achieve an accuracy requirement follow from basic theory of statistics. For most practical applications, sampling 2 000 orderings provides a good impression of reproducibility, while sampling 100 000 orderings ensures excellent approximations of the NPI lower and upper reproducibility probabilities and the numerical computations for such numbers of orderings require little time.

There are alternative approaches for approximating the NPI lower and upper RP. In particular, Bin Himd (2014) developed a bootstrap-based approach linked to NPI and explored its use for reproducibility computations from NPI perspective. While that approach does enable application of the NPI-RP idea in case of larger sample sizes, it does not provide estimates for the NPI lower and upper RPs but single values in between these. It may also be possible to quickly identify a proportion of all orderings for which one knows with certainty that the test necessarily leads to rejection or to non-

Data from Scotland

ℓ	Rejection	$\underline{RP}$	CI(0.95)	$\overline{RP}$	CI(0.95)
0.080	Yes	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.085	Yes	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.090	Yes	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.095	Yes	1.000	(0.999,1.000)	1.000	(1.000,1.000)
0.100	Yes	0.999	(0.998,1.000)	1.000	(1.000,1.000)
0.105	Yes	0.995	(0.992,0.998)	1.000	(0.999,1.000)
0.110	Yes	0.984	(0.978,0.989)	0.995	(0.991,0.998)
0.115	Yes	0.932	(0.921,0.943)	0.978	(0.971,0.984)
0.120	Yes	0.795	(0.777,0.813)	0.918	(0.905,0.930)
0.125	Yes	0.567	(0.545,0.589)	0.759	(0.740,0.778)
0.130	No	0.521	(0.499,0.543)	0.729	(0.710,0.748)
0.135	No	0.837	(0.821,0.853)	0.929	(0.917,0.940)
0.140	No	0.977	(0.970,0.984)	0.995	(0.991,0.998)
0.145	No	1.000	(0.999,1.000)	1.000	(1.000,1.000)
0.150	No	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.155	No	1.000	(1.000,1.000)	1.000	(1.000,1.000)
0.160	No	1.000	(1.000,1.000)	1.000	(1.000,1.000)

Table 7: Upper and lower RPs for different values of ℓ.

rejection of the null-hypothesis. For example, for the one-sided tests considered in this paper, using the sample mean in the test criterion, samples with all future observations guaranteed to be less than the test threshold value, necessarily lead to the future sample mean being less than that threshold, hence such samples could be excluded from the sampling. Similarly, future samples with all observations necessarily greater than the test threshold can be excluded. This would, however, only affect the total number of orderings to be considered for sampling quite marginally, and as we have seen that the sampling of orderings functions very well we consider such possible refinements of the procedure as being only of little practical interest. Of course, it may well be that for more complicated test scenarios one could benefit more from exploring the space of all possible future orderings in more detail, to simplify overall computational burden, this could be a topic of future research interest if such more complicated test scenarios are being investigated.

Acknowledgements

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UIDB/00297/2020 (Centro de Matemática e Aplicações).

References

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.

Augustin T. and Coolen F.P.A., 2004. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124, 251-272.

Bin Himd, S., 2014. *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility*. PhD Thesis, Durham University (*www.npi-statistics.com*).

Brown, L.D., Cai, T.T. and DasGupta, A. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.*, 16 , 101-133.

Coolen F.P.A., 2006. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15, 21-47.

Coolen F.P.A., 2011. Nonparametric predictive inference. In: *International Encyclopedia of Statistical Science*, Lovric M. (Ed.). Springer, Berlin, pp. 968-970.

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.

Coolen, F.P.A. and Bin Himd, S. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov-Smirnov test. Journal of Statistical Theory and Practice 14(2): 26.

De Capitani L. and De Martini D., 2011. On stochastic orderings of the Wilcoxon rank sum test statistic - with applications to reproducibility probability estimation testing. *Statistics and Probability Letters*, 81, 937-946.

De Martini D., 2008. Reproducibility probability estimation for testing statistical hypotheses. *Statistics and Probability Letters*, 78, 1056-1061.

Goodman S.N., 1992. A comment on replication, p-values and evidence. *Statistics in Medicine*, 11, 875-879.

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

Hites, R.A., Foran, J.A., Carpenter, D.O., Hamilton, M.C., Knuth, B.A. and Schwager, S.J., 2004. Global Assessment of Organic Contaminants in Farmed Salmon. *Science* 303, 226-229.

Mantalos, P. and Zografos, K. 2008. Interval estimation for a binomial proportion: a bootstrap approach. *Journal of Statistical Computation and Simulation*, 78, 1251-1265.

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

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.

Shao J. and Chow S.C., 2002. Reproducibility probability in clinical trials. *Statistics in Medicine*, 21, 1727-1742.