# Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests

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

**Abstract**

This paper introduces the nonparametric predictive inference approach for reproducibility of likelihood ratio tests. The general idea of this approach is outlined for tests between two simple hypotheses, followed by an investigation of reproducibility for tests between two Beta distributions. The paper reports on the first steps of a wider research programme towards tests involving composite hypotheses and substantial computational challenges.

*AMS Subject Classification:* 60A99, 62G99 and 62P30

*Keywords:* Beta distribution; lower and upper probabilities; nonparametric predictive inference; likelihood ratio test; reproducibility probability.

## 1 Introduction

In recent years, reproducibility of statistical hypotheses tests has received increasing attention. The issue involves a straightforward question: if a statistical test were repeated, under the same circumstances, would it lead to the same conclusion with regard to rejection or non-rejection of the null-hypothesis? Or, more precisely, what would be the probability of the test conclusion for the repeated test to be the same as for the original test? This is called the reproducibility probability (RP). The issue was first raised by

Goodman (1992), who pointed out that among practitioners there appeared to be a misunderstanding about the meaning of the $p$-value. Senn (2002) provided an extensive discussion of Goodman's paper, from a statistical perspective, emphasizing the difference between RP and the $p$-value. Of course, as explained by Senn, upon rejection of a null-hypothesis, a smaller value of the $p$-value suggests a larger RP. But, remarkably, it was not clear at all how RP could be computed or estimated. Traditionally, a test is designed for a specified level of significance, and the power of the test for a precisely specified alternative hypothesis, also called a 'simple hypothesis', can also be taken into account for the sample size or more general aspects of the test design. But it remains somewhat vague how the concept of a repeat of a test, and hence reproducibility of the test results, fits in to the classical frequentist framework of statistics.

The power of a test is the probability that the null-hypothesis is rejected if a simple alternative hypothesis is true. Due to the typical test formulation where a strong indication is being sought in favour of the alternative hypothesis, it is the correct rejection of the null-hypothesis that is often considered to be the target of the test. For example, in development of new medication one may test its superior effect on patients compared to existing medication by formulating a null-hypothesis of there being no difference and an alternative hypothesis specifying a specific level of improvement for the patients by using the new medication. This led Shao and Chow (2002) to focus on the power of the test, and they suggested to call an estimate of the power, based on the data of the original test, the 'reproducibility' of the test. In this approach, if the hypotheses involve a value for a parameter for an assumed model, and the original null hypothesis is rejected, then the data are used to estimate the parameter value, and this estimated parameter is then considered to be the simple alternative hypothesis parameter value for which the power of the test is computed, hence overall this leads to an estimate for the power of the test which is interpreted as an estimate of RP. This approach was also followed by De Martini (2008), who in addition proposed to use the estimated RP to design tests, and following work by De Capitani and De Martini (2011). While it is of course necessary to base inference on RP on the data of the original test, the explicit focus on the power of the test, hence on the assumption that the simple alternative hypothesis is true, is somewhat restrictive. In this 'estimated power' approach, RP can be regarded as a 'within the model' concept in the sense that the data of the original are used to estimate a parameter of the model, which in turn is linked to the power and then interpreted as RP. While such a power estimate is of interest, we do not think that it is in line with the natural interpretation of test reproducibility, both because it only considers cases where the null-hypothesis is rejected and because it does not really consider repeat application of the test, which would lead to different data. We should emphasize that there have been quite a few further attempts to specify RP, but they are less convincing than the approach by Shao and Chow (2002), a short introduction was provided by Coolen and Bin Himd (2014).

Coolen and Bin Himd (2014) presented a different perspective on test reproducibility, using the nonparametric predictive inference (NPI) framework of frequentist statistical methods (Augustin and Coolen, 2004; Coolen, 2006, 2011). The main difference to the estimated power approach of Shao and Chow (2002) is that the NPI approach for the reproducibility probability of a test (NPI-RP) is explicitly predictive, so it considers the test result for a predicted future sample of the same size as the original sample.

This approach seems to be well aligned to the nature of test reproducibility, which is more naturally considered as a prediction problem than as an estimation problem. Given the observed data from the original test, the NPI-RP approach first predicts future data sets, this is accomplished nonparametrically and without any consideration of the model assumed for the test. Then the same test as performed on the original data is considered for the random future data sets, and the proportion of these that lead to the same conclusion as the original test is investigated. Due the NPI being only based on few modelling assumptions, there is imprecision in this process as will be explained in more detail later within the context of the specific test scenario considered in this paper.

Coolen and Bin Himd (2014) introduced NPI for RP by considering some basic nonparametric tests, namely the sign test, Wilcoxon's signed rank test, and the two-sample rank sum test (Gibbons and Chakraborti, 2010). For these inferences NPI for Bernoulli quantities (Coolen, 1998) and for real-valued observations (Augustin and Coolen, 2004) were used. Recently, Coolen and Alqifari (2018) presented NPI-RP for two basic nonparametric tests based on order statistics, namely a quantile test (Gibbons and Chakraborti, 2010) and a precedence test (Balakrishnan and Ng, 2006), using NPI for future order statistics (Alqifari, 2017; Coolen et al, 2018). These basic tests all enabled analytical results for NPI-RP. To enable NPI for more complex test scenarios, the NPI-bootstrap method can be used, as introduced by Bin Himd (2014) who illustrates the use of NPI-bootstrap for NPI-RP for the Kolmogorov-Smirnov test. Computational aspects for more complex test scenarios are briefly commented on in this paper; they are an important topic for future research towards real-world implementation of NPI-RP.

This paper introduces 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 ecology (Chen et al, 2013; Nandakumar et al, 2008; Pirie et al, 2015; Zhang et al, 2010). Their good large sample properties and Wilks' theorem (Wilks, 1983) which states, for composite hypotheses, that the distribution of the logarithm of the LRT statistic can be approximated by a $\chi^2$ distribution allows the simple use of this testing procedure and makes it an attractive and commonly used tool. However, as already stated by several authors (see for example Johansen (2000) and Marques et al (2016)) this approximation, although simple and easy to use, does not provide precise results in many situations, for example in the multivariate setting, it often does not perform well in scenarios with large number of variables or for small sample sizes. For these more complex scenarios other more precise approximations can be considered, namely the so-called near-exact approximations (Coelho, 2004). In this paper, one will consider mainly the case of simple hypotheses but the case of composite hypotheses will be addressed in a possible follow up paper. For simple hypotheses, where all the parameters are specified, one will have, for a random sample $X_1, \ldots, X_n$ extracted from a population $X$, the null and alternative hypotheses specified in the following form

$$H_0 : X \sim f_0(x) \text{ vs } H_1 : X \sim f_1(x)$$

where $f_0$ and $f_1$ stand for the densities of the model considered under the null and

alternative hypotheses respectively. The LRT statistic is given by

$$LR = \prod_{i=1}^{n} \frac{f_0(x_i)}{f_1(x_i)}\,.$$

As already mentioned before, we will be interested in studying the RP of the LRT for some introductory examples. Along the way one will also derive the exact distribution of the LR statistics which will allow us to determine exact quantiles. Therefore, Section 2 of this paper introduces the general idea of NPI-RP for LRT through the case of a test between two simple hypotheses. This is then illustrated, first with a simple scenario, in Section 3, where one considers tests between a distribution with an increasing density in (0,1) versus the Uniform distribution, and then, in Section 4, for a test between two Beta distributions. Section 5 provides some concluding remarks and outlines important challenges for future research related to the generalization of this methodology to composite hypotheses, and the computational problems involved.

## 2 NPI-RP for LRT with simple hypotheses

Nonparametric predictive inference (NPI) (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 is made throughout this paper), and assigns probability $1/(n+1)$ for $X_{n+1}$ to each open interval between consecutive data observations (and $-\infty$ and $\infty$ for the left- and right-most intervals). We denote the $n$ data observations by $x_1 < x_2 < \ldots < x_n$ and for ease of notation we define $x_0 = -\infty$ and $x_{n+1} = \infty$. Of course, if finite bounds are known for the observation values then we can use these bounds as $x_0$ and $x_{n+1}$. It should be emphasized that no further assumptions are made, in particular not on the distribution of the probability $1/(n+1)$ within each interval. As a generalization, NPI for $m \geq 1$ future real-valued observations, based on $n \geq 1$ data observations, uses the sequential assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ (Arts et al, 2004), and by doing so it explicitly takes the interdependence of the future observations into account. 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. In this paper, we restrict attention to the case $m = n$, as this is most logical for studying reproducibility of a test based on $n$ observations.

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. 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 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 than was applied to the real data, and hence we wish to compute the likelihood ratio based on the future data, for each given ordering $O_j$. This is not possible, but we can find bounds for the likelihood ratio by minimizing and maximizing it over the ranges of values that the observations can have, given the specific ordering. This leads to three groups of orderings. First, orderings for which we certainly do not reject $H_0$, so for all possible locations of the $n$ future observations within the respective intervals, the resulting value of the likelihood ratio leads to non-rejection of $H_0$. Secondly, and following from similar arguments as for the first case, there are orderings for which we certainly reject $H_0$. Thirdly, orderings for which the minimum and maximum values of the LR lead to different conclusions with regard to rejection of $H_0$. All the orderings $O_j$ are equally likely, so to calculate the NPI lower RP, if for the original data we do not reject $H_0$ then we count the number of orderings in the first group, and to calculate the corresponding NPI upper RP in this case we count the number of orderings in the first and third groups. Similarly, if for the original data we reject $H_0$ then we count the number of orderings in the second group to calculate the NPI lower RP, and to calculate the corresponding NPI upper RP in this case we count the number of orderings in the second and third groups.

Clearly, imprecision in NPI-RP results from future orderings for which it is both possible that $H_0$ would be rejected or not rejected, given the ranges of values the observations can have within the intervals created by the original data. The main challenge for the NPI-RP approach to LRT is the derivation of the minimum and maximum values of the LR for each ordering of future observations. We start exploring this method, in Section 3, by considering a basic scenario for NPI-RP for the likelihood ratio test with two simple hypotheses. Suppose we have independent and identically distributed random quantities $X_i$ on $[0, 1]$ and wish to test $H_0 : X_i \sim f(x)$ versus $H_1 : X_i \sim U[0, 1]$, with probability density function (pdf) $f(x) > 0$ increasing on $[0, 1]$ and $U[0, 1]$ denoting the uniform distribution on $[0, 1]$. Due to the specific choice of $H_1$, the likelihood ratio based on data $x_1 < x_2 < \ldots < x_n$ is

$$LR(x) = \prod_{i=1}^{n} f(x_i)$$

The LRT is such that $H_0$ is not rejected if $LR(x) > K$, for some $K$ depending on the chosen level of significance for the test, and $H_0$ is rejected if $LR(x) \leq K$. With this specific support for the probability distribution of the random quantities of interest, we define $x_0 = 0$ and $x_{n+1} = 1$.

Due to the assumption that the pdf $f$ is increasing on $[0, 1]$, the NPI lower and upper reproducibility probabilities are relatively straightforward to derive. For ordering $O_j$, the likelihood ratio $LR(O_j)$ has minimum possible value $\underline{LR}(O_j)$ and maximum

possible value $\overline{LR}(O_j)$ derived by

$$\underline{LR}(O_j) = \prod_{l=1}^{n+1} f(x_{l-1})^{s_l^j} \quad (1)$$

$$\overline{LR}(O_j) = \prod_{l=1}^{n+1} f(x_l)^{s_l^j} . \quad (2)$$

Suppose that the LRT for the original data $x_1, \ldots, x_n$ leads to non-rejection of $H_0$, so $LR(x) > K$. Then the NPI lower RP is derived by counting all of the $\binom{2n}{n}$ orderings $O_j$ for which $\underline{LR}(O_j) > K$, while the corresponding NPI upper RP is derived by counting all orderings for which $\overline{LR}(O_j) > K$. Similarly, if the LRT for the original data leads to rejection of $H_0$, so $LR(x) \leq K$, then the NPI lower RP is derived by counting all of the $\binom{2n}{n}$ orderings $O_j$ for which $\overline{LR}(O_j) \leq K$, while the corresponding NPI upper RP is derived by counting all orderings for which $\underline{LR}(O_j) \leq K$. In this methodology the computation of all possible orderings, $O_j$, may be time consuming and an obstacle to its implementation when large samples are considered. However, this problem may be overcome by using bootstrap techniques (Bin Himd, 2014) or by sampling the orderings. The authors intend to extend and apply these techniques to NPI-RP for LRTs in future works.

In Section 4, one considers a more general setting where the densities may not be increasing functions and may assume the value zero in the extremes of their support. In this case, for independent and identically distributed random quantities $X_i$, $i = 1, \ldots, n$, on $[0, 1]$, we wish to test $H_0 : X_i \sim f_0(x)$ versus $H_1 : X_i \sim f_1(x)$. In the example provided in Section 4, $f_0$ and $f_1$ are the densities of Beta distributions. The LR based on the observed data $x_1 < x_2 < \ldots < x_n$ is

$$LR = \prod_{i=1}^{n} \frac{f_0(x_i)}{f_1(x_i)} = \prod_{i=1}^{n} f(x_i)$$

with $f(x_i) = f_0(x_i)/f_1(x_i)$. Then, if $f$ is a monotone function, for an ordering $O_j$, the likelihood ratio $LR(O_j)$ will have minimum possible value $\underline{LR}(O_j)$ and maximum possible value $\overline{LR}(O_j)$ given respectively by

$$\underline{LR}(O_j) = \prod_{l=1}^{n+1} f(x_l^-)^{s_l^j} \quad (3)$$

$$\overline{LR}(O_j) = \prod_{l=1}^{n+1} f(x_l^+)^{s_l^j} \quad (4)$$

where, for a given $l$, with $l = 1, \ldots, n+1$

$$x_l^- = \begin{cases} x_{l-1} & if \quad f(x_{l-1}) < f(x_l) \\ x_l & if \quad f(x_{l-1}) \geq f(x_l) \end{cases}$$

and

$$x_l^+ = \begin{cases} x_{l-1} & if \quad f(x_{l-1}) > f(x_l) \\ x_l & if \quad f(x_{l-1}) \leq f(x_l) . \end{cases}$$

For the general case with any two likelihood functions, we need to find $\underline{lr}_l$ and $\overline{lr}_l$, the infimum and supremum, respectively, of the (likelihood) ratio of the probability density functions over $(x_{l-1}, x_l)$; while this may not be trivial, it only needs to be done once for a given data set, and perhaps some approximate results may be possible, note that for large data sets it is likely that for most of these intervals the ratio of the pdf values within it do not change much. These $\underline{lr}_l$ and $\overline{lr}_l$ then replace the pdfs in Equations (1) and (2) above.

To avoid possible issues related with the fact that the densities functions of the models under the null and alternative hypotheses may assume the value zero in the extremes of interval (0,1) we propose a basic and initial approach to this problem, which is to consider $x_0 = \frac{0+x_1}{2}$ and $x_{n+1} = \frac{x_n+1}{2}$. Other possible techniques as well as the possible effects of this choice will be studied in future works.

## 3 A first simple example

In our first example we consider the following hypotheses

$$H_0 : X_i \sim f_0(x) \text{ vs } H_1 : X_i \sim U[0,1]$$

with

$$f_0(x) = \begin{cases} x + \frac{1}{2} & , \quad x \in [0,1] \\ 0 & , \quad \text{otherwise} \end{cases} \tag{5}$$

the LR, for an observed sample of size $n$, is given by

$$LR = \prod_{i=1}^{n} f_0(x_i)\,.$$

The exact distribution of the LR statistic, under the null hypothesis, is given in the following theorem.

**Theorem 3.1** *Let $X_1, \ldots, X_n$ be independent and identically distributed with density given in (5). Then, the cumulative distribution function of $LR = \prod_{i=1}^{n}(X_i + 1/2)$ is given by*

$$1 - \sum_{k=0}^{n} \frac{(-1)^k 3^{2(n-k)} \binom{n}{k}}{2^{3n}} F_{\Gamma(n,2)} \left[-(n\log(2) - \log(3)(n-k)) - \log(x)\right]$$

*where $F_{\Gamma(n,2)}(.)$ represents the cumulative distribution function of Gamma distribution with shape parameter $n$, rate parameter $2$.*

**Proof** It is easy to note that $Y_i = X_i + 1/2$ has density given by

$$f_{Y_i}(y) = \begin{cases} y & , \quad y \in [1/2, 3/2] \\ 0 & , \quad \text{otherwise} \end{cases} \quad .$$

and that the $h$-th moment of $Y$ is given by

$$E[Y_i^h] = \frac{2^{-h-2}\left(3^{h+2} - 1\right)}{h+2}\,.$$

Using the equality $E[\mathrm{e}^{-\mathrm{i}t\log(Y_i)}] = E[Y_i^{-\mathrm{i}t}]$ it is easy to obtain the characteristic function of $-\log(Y_i)$ as

$$\Phi_{-\log(Y_i)}(t) = \frac{2^{-2+\mathrm{i}t}\left(-1+3^{2-\mathrm{i}t}\right)}{2-\mathrm{i}t}\,.$$

If we consider the random variable $W = -\log(\prod_{i=1}^{n} Y_i) = \sum_{i=1}^{n} -\log(Y_i)$, given the properties of characteristic functions, we have that the characteristic function of $W$ is given by

$$\Phi_W(t) = \left(\frac{2^{-2+\mathrm{i}t}\left(-1+3^{2-\mathrm{i}t}\right)}{2-\mathrm{i}t}\right)^n$$

which after some algebraic manipulation and using the binomial expansion it is possible to write as

$$\Phi_W(t) = \sum_{k=0}^{n} \frac{(-1)^k 3^{2(n-k)} 2^{\mathrm{i}nt}\left(\frac{2}{2-\mathrm{i}t}\right)^n \binom{n}{k} 3^{-\mathrm{i}t(n-k)}}{2^{3n}}\,.$$

The previous expression, is the characteristic function of a mixture of shifted Gamma distributions. After the necessary transformations the cumulative distribution function of the LR can be written as

$$1 - \sum_{k=0}^{n} \frac{(-1)^k 3^{2(n-k)}\binom{n}{k}}{2^{3n}} F_{\Gamma(n,2)}\left[-(n\log(2) - \log(3)(n-k)) - \log(x)\right]$$

where $F_{\Gamma(n,2)}(.)$ represents the cumulative distribution function of Gamma distribution with shape parameter $n$, rate parameter 2. □

Using the previous results one may determine the exact quantiles for the LR statistic and develop numerical simulations to study the RP of this test. In the following simulations we consider samples of sizes $n = 5$ and $n = 10$ and for each case we consider 15 and 50 replications simulated under $H_0$. The 0.2 exact quantiles were determined using the cumulative distribution function given in Theorem 3.1, and are equal to 0.7267 and 0.7240 respectively for $n = 5$ and $n = 10$. We have considered the 0.2 significance level in order to make it easier to analyse the figures. In Figure 1, the filled circle dots are the upper RP and the filled square dots are the lower RP evaluated for each simulated value of the LR statistic. The vertical line marks the value of the exact quantile. From Figure 1 one may observe that the values of the RP tend to increase when the simulated value of the LR statistic moves away from the quantile considered, this was already expected and it is also observed in the next examples.

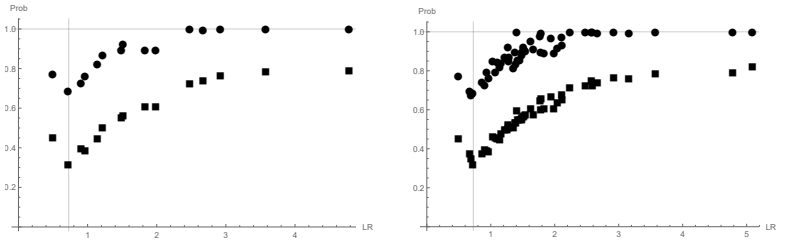
## 4 A test between two Beta distributions

In this section we will consider a more complex case using the procedure illustrated in Section 2. We are interested in testing

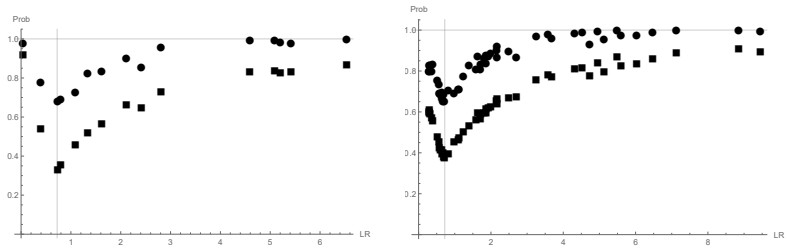
$$H_0 : X_i \sim \mathrm{Beta}(a,1) \;\text{ vs }\; H_1 : X_i \sim \mathrm{Beta}(b,1)$$

Figure 1: Simulated values of the RP

15 replications and $n = m = 5$ 50 replications and $n = m = 5$




15 replications and $n = m = 10$ 50 replications and $n = m = 10$




for $a \neq b$, the LR, for a sample of size $n$, is given by

$$LR = \prod_{i=1}^{n} \frac{f_0(x_i)}{f_1(x_i)}$$

with

$$f_0(x) = \frac{x^{a-1}}{B(a,1)} \text{ and } f_1(x) = \frac{x^{b-1}}{B(b,1)}$$

thus

$$LR = \prod_{i=1}^{n} \frac{a}{b}\, x_i^{a-b}\,.$$

The following theorem specifies the distribution of the LR statistic. In this theorem one will consider just the case $a > b$, however the case $a < b$ can be considered using a similar procedure.

**Theorem 4.1** *For a sample $X_1, \ldots, X_n$, independent and identically distributed, with $X_i \sim Beta(a,1)$, the cumulative distribution function of $LR = \left(\frac{a}{b}\right)^n \prod_{i=1}^{n} X_i^{a-b}$ with $a > b$ is given*

$$1 - F_{\Gamma\left(n, \frac{a}{a-b}\right)}\left(-\log\left(\frac{x}{(a/b)^n}\right)\right)$$

*where $F_{\Gamma\left(n, \frac{a}{a-b}\right)}(.)$ is the cumulative distribution function of a Gamma distribution with shape parameter $n$ and rate parameter $\frac{a}{a-b}$.*

**Proof** Following a similar procedure to the one used in two previous cases, we will use the random variable $W = -\log(\prod_{i=1}^{n} X_i^{a-b}) = \sum_{i=1}^{n} -(a-b)\log(X_i)$. Since we know that $-\log(X_i)$ has an Exponential distribution with parameter $a$ and that the $h$th moment of $X_i^{a-b}$ is given by

$$E\left[X_i^{(a-b)h}\right] = \frac{a}{a + h(a-b)}$$

the expression of the characteristic function of $-(a-b)\log(X_i)$ will be given by

$$\Phi_{-(a-b)\log(X_i)}(t) = \frac{a}{a - \mathrm{i}t(a-b)}$$

and the characteristic function of $W$ by

$$\Phi_W(t) = \left(\frac{\frac{a}{a-b}}{\frac{a}{a-b} - \mathrm{i}t}\right)^n$$

which is the characteristic function of a Gamma distribution with shape parameter $n$ and rate parameter $\frac{a}{a-b}$. Therefore it is, again, straightforward to determine, with the necessary transformations, the cumulative distribution function of the LR which is given by

$$1 - F_{\Gamma\left(n, \frac{a}{a-b}\right)}\left(-\log\left(\frac{x}{(a/b)^n}\right)\right)$$

where $F_{\Gamma\left(n, \frac{a}{a-b}\right)}(.)$ is the cumulative distribution function of a Gamma distribution with shape parameter $n$ and rate parameter $\frac{a}{a-b}$. □

For example, when $a = 5$ and $b = 1/5$, for samples of sizes $n = 5$ and $n = 10$, the 0.2 exact quantiles, were determined using the cumulative distribution function given in Theorem 4.1, and are respectively 15401.8 and $5.755 \times 10^8$ . The simulations were performed considering 15 and 50 replications of data generated under $H_0$.

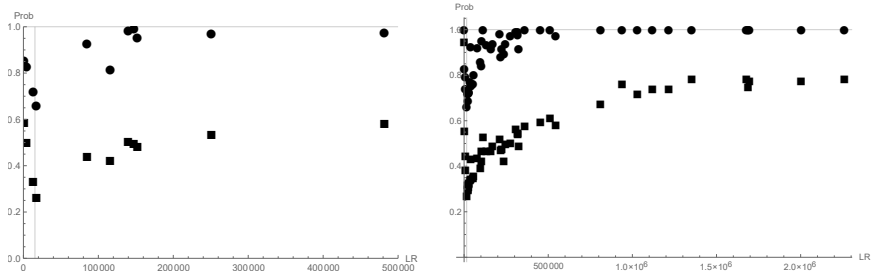
In Figure 2 we observe similar features to the ones already described in Figure 1; (i) the values of the lower and upper RPs tend to increase with increasing distance between the observed LR and the quantiles considered, (ii) if the observed values of the LR are close to the quantile, the lower RP decreases substantially and may even assume values below 0.5, (iii) these figures show some oscillation of the values of the RPs which is, essentially, due to randomness and to the products involved in the expression of the LR which are reflected in the process for computing the minimum and maximum possible values of the LR.

## 5 Concluding remarks

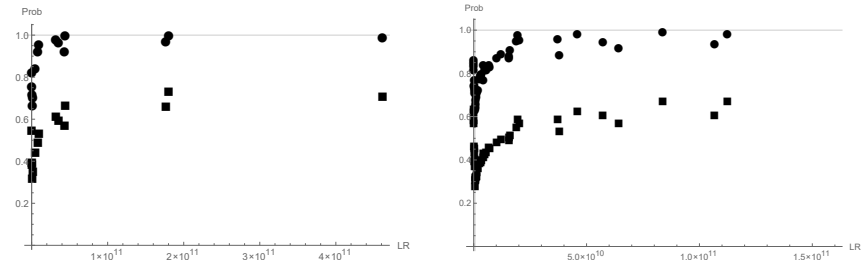
This paper introduced nonparametric predictive inference methods for reproducibility of likelihood ratio tests. The main idea is exemplified with two examples of testing procedures. The simulations carried out show the increasing trend of the lower and upper

Figure 2: Simulated values of the RP

5 replications and $n = m = 5$ 50 replications and $n = m = 5$




15 replications and $n = m = 10$ 50 replications and $n = m = 10$




RPs together with the decreasing trend of the difference between these probabilities, for increasing values of the distance between the observed LRs and the quantiles. The simulations were performed for small samples due to the computational time required for the computation of the number of possible orderings. However, this difficulty may be overcome by using bootstrap techniques (Bin Himd, 2014) or sampling of orderings. Further research challenges include discussion of significance level (p-value), power and RP to be taken into account for test designs, and the development of a general set up for composite hypotheses.

**Acknowledgements**

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

**References**

Alqifari, H.N., 2017. *Nonparametric Predictive Inference for Future Order Statistics*. PhD Thesis, Durham University (*www.npi-statistics.com*).

Arts G.R.J., Coolen F.P.A., van der Laan P., 2004. Nonparametric predictive inference

in statistical process control. *Quality Technology and Quantitative Management*, 1, 201-216.

Arts G.R.J., Coolen F.P.A., 2008. Two nonparametric predictive control charts. *Journal of Statistical Theory and Practice*, 2, 499-512.

Atmanspacher H., Maasen S., 2016. *Reproducibility: Principles, Problems, Practices, and Prospects*. Wiley, Hoboken, New Jersey.

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

Augustin T., Coolen F.P.A., de Cooman G., Troffaes M.C.M. (Eds.), 2014. *Introduction to Imprecise Probabilities*. Wiley, Chichester.

Balakrishnan, N., Ng, H.K.T., 2006. *Precedence-Type Tests and Applications*. Wiley, Hoboken, New Jersey.

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

Chen, P., Dong, L., Chen, W. and Lin, J.-G., 2013. Outlier Detection in Adaptive Functional-Coefficient Autoregressive Models Based on Extreme Value Theory. *Mathematical Problems in Engineering*, Article ID 910828, 9 pp..

Coelho, C.A., 2004. The generalized near-integer gamma distribution: a basis for near-exact approximations to the distribution of statistics which are the product of an odd number of independent beta random variables. *Journal of Multivariate Analysis*, 89, 191-218.

Coolen F.P.A., 1998. Low structure imprecise predictive inference for Bayes' problem. *Statistics and Probability Letters*, 36, 349-357.

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., 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., Coolen-Maturi, T., Alqifari, H.N., 2018. Nonparametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, 47, 2527-2548.

Coolen, F.P.A., Alqifari, H.N., 2018. Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT - Statistical Journal*, 16, 167-185.

De Capitani L., 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.

Gibbons, J.D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference* (5th ed.). Chapman & Hall, Boca Raton, Florida.

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.

Johansen S., 2000. A Bartlett correction factor for tests on the cointegrating relations. *Econometric Theory*, 16, 740-778.

Lawless J.F., Fredette M., 2005. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92, 529–542.

Marques, F.J., Coelho, C.A. and Rodrigues, P.C., 2016. Testing the equality of several linear regression models. *Computational Statistics*, 32, 1453-1480.

Nandakumar, K., Chen, Y., Dass, S.C. and Jain, A., 2008. Likelihood Ratio-Based Biometric Score Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 2.

Pirie, A., Wood, A., Lush, M., Tyrer, J. and Pharoah, P.D., 2015. The effect of rare variants on inflation of the test statistics in case-control analyses. *BMC Bioinformatics*, 16, 53, 5 pp.

Senn S., 2002. Comment on 'A comment on replication, p-values and evidence', by S.N. Goodman (Letter to the editor). *Statistics in Medicine*, 21, 2437-2444. With author's reply, pp. 245–247.

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

Zhang, J., Zou, C. and Wang, Z., 2010. A control chart based on likelihood ratio test for monitoring process mean and variability. *Quality and Reliability Engineering International*, 26, 1, 63-73.

Wilks, S.S., 1983. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9, 60-62.