entropy

Article

Reproducibility Probability Estimation and RP-Testing for Some Nonparametric Tests

Lucio De Capitani and Daniele De Martini *

Department of Statistics and Quantitative Methods, University of Milano-Bicocca, via Bicocca degli Arcimboldi, 8, Milano 20126, Italy; lucio.decapitani1@unimib.it

* Correspondence: daniele.demartini@unimib.it; Tel.: +39-02-6448-3130

Academic Editors: Julio Stern, Adriano Polpo and Kevin H. Knuth
Received: 13 January 2016; Accepted: 6 April 2016; Published: 16 April 2016

Abstract: Several reproducibility probability (RP)-estimators for the binomial, sign, Wilcoxon signed rank and Kendall tests are studied. Their behavior in terms of MSE is investigated, as well as their performances for RP-testing. Two classes of estimators are considered: the semi-parametric one, where RP-estimators are derived from the expression of the exact or approximated power function, and the non-parametric one, whose RP-estimators are obtained on the basis of the nonparametric plug-in principle. In order to evaluate the precision of RP-estimators for each test, the MSE is computed, and the best overall estimator turns out to belong to the semi-parametric class. Then, in order to evaluate the RP-testing performances provided by RP estimators for each test, the disagreement between the RP-testing decision rule, *i.e.*, "accept H_0 if the RP-estimate is lower than, or equal to, 1/2, and reject H_0 otherwise", and the classical one (based on the critical value or on the *p*-value) is obtained. It is shown that the RP-based testing decision for some semi-parametric RP estimators exactly replicates the classical one. In many situations, the RP-estimator replicating the classical decision rule also provides the best MSE.

Keywords: asymptotic power approximation; sign test; binomial test; Wilcoxon signed rank test; Kendall test; stability of test outcomes; reproducibility of tests outcomes

1. Introduction

Statistical tests are usually applied in almost all fields of science to evaluate experimental results. The reproducibility probability (RP) is the true power of a statistical test, and its estimation provides useful information to evaluate the stability of statistical test results. Indeed, when the Neyman–Pearson approach is adopted, that is the Type I error probability is fixed before starting the experiment, the statistical test turns out to be a Bernoullian random variable (*viz.* significant/non-significant), whose parameter is the RP. Therefore, looking at the RP estimate is the natural perspective for evaluating the stability of test results: the higher the estimated RP, the more stable the observed result is estimated to be; see [1]. RP estimation was applied, for example, in the context of clinical trials [2–6]. Moreover, RP-testing, that is the adoption of the RP estimate to evaluate the significance of statistical test results, can substitute the *p*-value testing [7,8]. In detail, the RP-testing decision rule, which sounds very intuitive, states: "accept H_0 if the RP-estimate is lower than, or equal to, 1/2, and reject H_0 otherwise". We argue that the RP-testing rule can be adopted in order to bypass the many, well-known criticisms raised by the *p*-value [9–13] In the context of nonparametric tests, RP estimation has not yet been widely studied. The only works in this field concern RP estimation and testing for the Wilcoxon rank sum test [14,15].

In this paper, some RP estimators for the most commonly-used nonparametric tests are introduced and studied. Specifically, the sign test, the binomial test, the Kendall test and the Wilcoxon signed rank test are considered. Both nonparametric and semi-parametric RP estimators are presented, for each test. Focus is placed on two features: (1) the behavior of different estimators for a given test and their consequent comparison, for example in terms of MSE; (2) the validity, exact or approximated, of the RP-testing rule based on the RP estimators presented here. For the first task, we resort to some simulation studies, whereas appropriate theoretical results are developed for the second one.

The theoretical framework of nonparametric RP estimation and testing is introduced in Section 2, where the problems that can be encountered are explained in depth; then, the class of semi-parametric estimators and that of nonparametric plug-in estimators are introduced, and some theoretical results on RP-testing are provided. In Section 3, the sign test and the binomial test are considered: semi-parametric RP estimation and testing for the binomial test are studied first; then, nonparametric estimation techniques are studied for the same aim; finally, the sign test is considered, showing that the results obtained for the binomial test hold true also for the sign test. RP estimation and testing for the Wilcoxon signed rank test is studied in Section 4, where semi-parametric and nonparametric plug-in estimators are considered and studied separately; then, the behavior of different estimators is compared through simulation. The last test considered (Section 5) is the Kendall test of monotonic association. As in the previous sections, semi-parametric and nonparametric estimators are studied separately, then a simulation is run to compare the behavior of different estimators, in terms of MSE and RP-testing performances. An example of the applications is shown in Section 6, and the conclusions are reported in Section 7.

2. RP-Estimation and Testing in the Nonparametric Framework

2.1. The General Nonparametric Framework

Let ${}_tF$ be the true cumulative distribution function of a study variable X. This distribution function is unknown and belongs to the class of distributions $\mathcal{F}$. Assume that starting from a random sample $\mathbf{X}_n = (X_1, \cdots, X_n)$ drawn from ${}_tF$, it is of interest to solve the testing problem:

$$H_0 : {}_tF \in \mathcal{F}_0 \quad vs \quad H_1 : {}_tF \in \mathcal{F} \backslash \mathcal{F}_0 \, , \tag{1}$$

where $\mathcal{F}_0 \subset \mathcal{F}$. Let $T_n = \mathcal{T}(\mathbf{X_n})$ be the test statistic used to solve (1). There are two typical cases that can be encountered when considering nonparametric tests:

(**A**) the exact and asymptotic distributions of T_n are known both under H_0 and H_1;
(**B**) the exact and asymptotic distributions of T_n are known under H_0. Under H_1, only the asymptotic distribution can be derived.

Case (**A**) is rather an exception. The binomial and sign tests are examples of tests under this case. Case (**B**) is the common situation: for almost all of the distribution-free tests, the exact null-distribution of T_n can be derived by using permutations, combinatorics and *ad hoc* algorithms (see, e.g., [16]). On the contrary, the non-null distribution can be derived only recurring to large-sample approximations. A few examples include the Wilcoxon signed rank test, the Wilcoxon rank sum test and the Kendall test.

Under both Cases (**A**) and (**B**), the knowledge of the exact null distribution of T_n allows the definition of the exact test:

$$\Psi_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad T_n \in R_{n,\alpha} \\ 0 & \text{if} \quad T_n \notin R_{n,\alpha} \end{cases} \, , \tag{2}$$

where, as usual, α denotes the Type I error probability and $R_{n,\alpha}$ is a level-α critical region corresponding to the sample size n. For example, if the testing problem (1) is one-sided, the critical region takes, without loss of generality, the form $R_{n,\alpha} = (t_{n,1-\alpha}, \infty)$, where $t_{n,1-\alpha}$ is the $(1-\alpha)$-quantile of the null distribution G_0 of T_n. Note that, if T_n is a discrete random variable, the critical region $R_{n,\alpha}$ is exact, but conservative, *i.e.*, its Type I error probability can be lower than α, since

$t_{n,1-\alpha} = \inf\{t : G_0(t) \geq 1-\alpha)\}$. In practice, if the sample size n is sufficiently high, an asymptotic test is usually preferred to avoid the computational effort needed to compute the exact distribution of T_n. In particular, the following test is used:

$$\widetilde{\Psi}_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad T_n \in \widetilde{R}_{n,\alpha} \\ 0 & \text{if} \quad T_n \notin \widetilde{R}_{n,\alpha} \end{cases}, \tag{3}$$

where $\widetilde{R}_{n,\alpha}$ is the level-α asymptotic critical region, which, considering the one-sided example mentioned above, takes the form $\widetilde{R}_{n,\alpha} = (\tilde{t}_{n,1-\alpha}, \infty)$, where $\tilde{t}_{n,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the large sample null distribution of T_n. Obviously, the tests Ψ_α and $\widetilde{\Psi}_\alpha$ become closer as the sample size n increases, and they are asymptotically equivalent. However, whatever the sample size is, there is a certain probability of disagreement between (2) and (3). To clearly explain the definition of the probability of disagreement, consider sets A_1, A_2 and A defined as follows: $A_1 = \left\{\mathbf{x}_n : \Psi_\alpha(\mathbf{x}_n) = 0 \text{ and } \widetilde{\Psi}_\alpha(\mathbf{x}_n) = 1\right\}$,$A_2 = \left\{\mathbf{x}_n : \Psi_\alpha(\mathbf{x}_n) = 1 \text{ and } \widetilde{\Psi}_\alpha(\mathbf{x}_n) = 0\right\}$, $A = A_1 \cup A_2$. Set A_2 collects the realizations $\mathbf{x}_n$ of $\mathbf{X}_n$ for which the null hypothesis is accepted by the asymptotic test and rejected by the exact one. Conversely, set A_1 collects the realizations $\mathbf{x}_n$ of $\mathbf{X}_n$ for which the null hypothesis is accepted by the exact test and rejected by the asymptotic one. Therefore, the probability of disagreement between (2) and (3) is:

$$D(\alpha, n, F) = P_F(A) = P_F(A_2) + P_F(A_1). \tag{4}$$

The differences between Cases (**A**) and (**B**) do not impact the definition of the statistical test to solve (1), but they determine the way the power of the test and, therefore, the RP can be evaluated: under Case (**A**), the power of the test can be exactly computed; under Case (**B**), the power can be evaluated only approximately. In detail, under Case (**A**), the exact power Ψ_α corresponding to the distribution $F \in \mathcal{F}$ can be computed as $\pi(n,\alpha,F) = P_F(T_n \in R_{n,\alpha}) = E_F[\Psi_\alpha(\mathbf{X}_n)]$. Consequently, the exact RP of the test (*i.e.*, the exact "true power" of the test) coincides with $RP = \pi(n,\alpha,{}_tF) = P_{{}_tF}(T_n \in R_{n,\alpha}) = E_{{}_tF}[\Psi_\alpha(\mathbf{X}_n)]$. Under Case (**B**), the exact power of Ψ_α can be approximated by $\tilde{\pi}(n,\alpha,F) = \widetilde{P}_F(T_n \in R_{n,\alpha}) = \widetilde{E}_F[\Psi_\alpha(\mathbf{X}_n)]$, where the symbols $\widetilde{P}_F$ and $\widetilde{E}$ emphasize that probability and expectation are computed according to the asymptotic distribution of T_n. In this case, the approximated RP is $\widetilde{RP} = \tilde{\pi}(n,\alpha,{}_tF)$. Analogously, under Case (**B**), the power of $\widetilde{\Psi}_\alpha$ can be approximated by $\pi_a(n,\alpha,F) = \widetilde{P}_F(T_n \in \widetilde{R}_{n,\alpha}) = \widetilde{E}_F[\widetilde{\Psi}_\alpha(\mathbf{X}_n)]$ and the approximate RP results $RP_a = \pi_a(n,\alpha,{}_tF)$. Obviously, the approximate power $\tilde{\pi}(n,\alpha,F)$ and $\pi_a(n,\alpha,F)$ and the approximate RP $\tilde{\pi}(n,\alpha,{}_tF)$ and $\pi_a(n,\alpha,{}_tF)$, can be computed under Case (**A**), as well. Moreover, in this latter case, it is also possible to compute the exact power of the approximate test. However, in practice, under Case (**A**), if the computational burden is acceptable, the exact test and power are usually computed. In the case of a huge computational cost, the asymptotic test and its approximate power are used. To summarize, in Table 1, the possible approaches to compute the power of a test are represented under the different scenarios that can arise under Cases (**A**) and (**B**). The background of the cell representing the approaches commonly employed in practice are colored in gray.

Table 1. Possible approaches to compute the power of a test under the different scenarios related to Cases (**A**) and (**B**). The cells with gray background represent the possible approaches commonly employed in practice.

	Case (A)		**Case (B)**	
	Exact Test (Ψ_α)	**Asymptotic Test ($\widetilde{\Psi}_\alpha$)**	**Exact Test (Ψ_α)**	**Asymptotic Test ($\widetilde{\Psi}_\alpha$)**
Computation of the exact power	Possible Case (**A.1**)	Possible (not considered)	Not Possible	Not Possible
Computation of the approximated power	Possible (not considered)	Possible Case (**A.2**)	Possible Case (**B.1**)	Possible Case (**B.2**)

Under both Cases (**A**) and (**B**), it is possible to get an RP-estimator following several methodologies. These methodologies can be divided into two main subgroups: semi-parametric estimators and non-parametric estimators.

2.2. Semi-Parametric RP-Estimation and RP-Testing

As for the WRS test (see [14,15]), in common nonparametric tests, the asymptotic/exact distribution of T_n depends on a vector θ_t of parameters defined as particular functionals of ${}_tF$. In such cases, the asymptotic/exact power can be interpreted as a function of θ_t instead of a functional of ${}_tF$. Now, a semi-parametric RP-estimator can be obtained by plugging an appropriate point estimator $\hat{\theta}$ of θ_t into the expression of the exact/asymptotic power:

- under Case (**A.1**), the semi-parametric RP-estimator is $\hat{\pi} = \pi(n, \alpha, \hat{\theta})$;
- under Case (**A.2**) and Case (**B.2**), the semi-parametric RP-estimator is $\hat{\pi}_a = \pi_a(n, \alpha, \hat{\theta})$;
- under Case (**B.1**), the semi-parametric RP-estimator is $\hat{\tilde{\pi}}_a = \tilde{\pi}(n, \alpha, \hat{\theta})$.

As will be explained later, if the estimator $\hat{\theta}$ is appropriately chosen and the testing problem (1) is one sided, the semi-parametric RP-estimator $\hat{\pi}$ and $\hat{\pi}_a$ can be used to replicate the tests Ψ_α and $\tilde{\Psi}_\alpha$ through the RP-testing technique: "accept H_0 if the RP-estimate is lower or equal to 1/2 and reject H_0 otherwise". For several non-parametric tests (see [8] for the general parametric case), if the estimator $\hat{\theta}$ is appropriately chosen, it is possible to demonstrate that $T_n \in R_{n,\alpha} \Leftrightarrow \hat{\pi} > 1/2$ or $T_n \in \widetilde{R}_{n,\alpha} \Leftrightarrow \hat{\pi}_a > 1/2$. Then, the exact and asymptotic tests can be rewritten as

$$\Psi_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi} > 1/2 \\ 0 & \text{if} \quad \hat{\pi} \le 1/2 \end{cases} \quad \text{and} \quad \widetilde{\Psi}_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi}_a > 1/2 \\ 0 & \text{if} \quad \hat{\pi}_a > 1/2 \end{cases}.$$

The above identities cover Cases (**A.1**), (**A.2**) and (**B.2**). In Case (**B.1**), the exact test cannot generally be replicated through the RP-testing technique based on semi-parametric estimators. However, the following lemma (proved in the Supplementary Material) describes a case in which this is possible:

Lemma 1. *Assume that the testing problem (1) is one sided and that the exact test based on the test statistic T_n is $\Psi_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad T_n > t_{n,1-\alpha} \\ 0 & \text{if} \quad T_n \le t_{n,1-\alpha} \end{cases}$. Moreover, assume that $\frac{T_n - E[T_n]}{\sqrt{Var[T_n]}} \xrightarrow{d} \mathcal{N}(0,1)$, with $E[T_n] = e(\theta_t)$ and $Var[T_n] = v(\theta_t)$. If $\hat{\theta}$ is such that $T_n = e(\hat{\theta})$, then:*

1. *the RP-based decision rule defined by $\hat{\tilde{\pi}}_a = \tilde{\pi}(n, \alpha, \hat{\theta}) = 1 - \Phi\left(\frac{t_{n,1-\alpha} - e(\hat{\theta})}{\sqrt{v(\hat{\theta})}}\right)$ exactly replicates the exact test Ψ_α;*
2. *the RP-based decision rule defined by $\hat{\tilde{\pi}}_a^* = 1 - \Phi\left(\frac{t_{n,1-\alpha} - e(\hat{\theta})}{\sqrt{\hat{V}}}\right)$ with $\hat{V}$ any estimator for $Var[T_n]$, exactly replicates the exact test Ψ_α.*

2.3. Non-Parametric RP-Estimation and RP-Testing: The Non-Parametric Plug-In Approach

As pointed out in [17] and in [3,18], it is possible to estimate the RP by using a non-parametric plug-in estimator. Under Cases (**A.1**) and (**B.1**), it is possible to consider the plug-in estimators $\hat{\pi}_e^{PI} = P_{\hat{F}_n}(T_n > t_{n,1-\alpha}) = E_{\hat{F}_n}[\Psi_\alpha(\mathbf{X}_n)]$ where $\hat{F}_n$ denotes the empirical cumulative distribution function (ecdf). In practice, $\hat{\pi}_e^{PI}$ coincides with the rejection rate computed performing test Ψ_α over all n^n possible samples of size n that can be drawn from the ecdf: $\hat{\pi}_e^{PI} = \frac{1}{n^n} \sum_{\mathbf{x}_n^i \in \mathcal{X}(\mathbf{X}_n)} \Psi_\alpha(\mathbf{x}_n^i)$ where $\mathcal{X}(\mathbf{X}_n)$ denotes the set of all of the samples of size n that can be drawn with replacement from the ecdf corresponding to $\mathbf{X}_n$. Apart from some special cases, the analytical expression of $\hat{\pi}_e^{PI}$ cannot be derived. Consequently, it is usually approximated by the Monte-Carlo method: B samples of length

n are drawn from the ecdf. The test Ψ_α is then performed over all of the B samples, and the plug-in RP-estimate is computed as the rejection rate. In detail:

$$\hat{\pi}^{PI} = \frac{1}{B}\sum_{j=1}^{B} \Psi_\alpha(\mathbf{X}_n^j) \tag{5}$$

where $\mathbf{X}_n^j$ denotes the j-th re-sample drawn from the ecdf. Similarly, under Case (**A.2**) and Case (**B.2**), it is possible to define the plug-in RP-estimator starting from the asymptotic test obtaining $\hat{\pi}_{a,e}^{PI} = \dfrac{1}{n^n} \displaystyle\sum_{\mathbf{x}_n^i \in \mathcal{X}(\mathbf{X}_n)} \widetilde{\Psi}_\alpha(\mathbf{x}_n^i)$ and:

$$\hat{\pi}_a^{PI} = \frac{1}{B}\sum_{j=1}^{B} \widetilde{\Psi}_\alpha(\mathbf{X}_n^j) \ . \tag{6}$$

The plug-in RP-estimators introduced above can be used to define the RP-based test $\Psi_\alpha^{PI,e}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi}_e^{PI} > 1/2 \\ 0 & \text{if} \quad \hat{\pi}_e^{PI} > 1/2 \end{cases}$, $\widetilde{\Psi}_\alpha^{PI,e}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi}_{a,e}^{PI} > 1/2 \\ 0 & \text{if} \quad \hat{\pi}_{a,e}^{PI} > 1/2 \end{cases}$,
$\Psi_\alpha^{PI}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi}^{PI} > 1/2 \\ 0 & \text{if} \quad \hat{\pi}^{PI} > 1/2 \end{cases}$ and $\widetilde{\Psi}_\alpha^{PI}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{\pi}_a^{PI} > 1/2 \\ 0 & \text{if} \quad \hat{\pi}_a^{PI} > 1/2 \end{cases}$.

However, there are no general theoretical results assuring that $\Psi_\alpha^{PI,e}$ and $\widetilde{\Psi}_\alpha^{PI,e}$ or Ψ_α^{PI} and $\widetilde{\Psi}_\alpha^{PI}$ are level-α tests equivalent to Ψ_α and $\widetilde{\Psi}_\alpha$, respectively.

3. RP-Estimation and Testing for the Binomial and Sign Test

In this section, the performances of the semi-parametric and non-parametric *RP* estimators for binomial and sign tests are evaluated. At first, the binomial test is considered. Let $\mathbf{X}_n = (X_1, ..., X_n)$ be a random sample drawn from the Bernoulli distribution with unknown parameter p_t. The statistical hypotheses of interest are:

$$H_0 : p_t \leq p_0 \qquad \textit{versus} \qquad H_1 : p_t > p_0 \ . \tag{7}$$

The previous hypotheses can be tested by using the statistic $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The exact and asymptotic distribution of $\hat{P}$ is known both under H_0 and under H_1, and consequently, this test falls under Case (**A**). Specifically, $n\hat{P} \sim \mathit{Binomial}(n, p_t)$ and $\sqrt{n}\frac{\hat{P}-p_t}{\sqrt{p_t(1-p_t)}} \xrightarrow{d} \mathcal{N}(0,1)$. The exact test is then given by $\Psi_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{P} > c_\alpha \\ 0 & \text{if} \quad \hat{P} \leq c_\alpha \end{cases}$ where $c_\alpha = \frac{b_{(1-\alpha;n,p_0)}}{n}$ and $b_{(q;n,p)}$ is the q-quantile of the binomial distribution with parameters n and p (the test so-defined is conservative). The asymptotic test results $\widetilde{\Psi}_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad n\hat{P} > \lfloor n\tilde{c}_\alpha \rfloor \\ 0 & \text{if} \quad n\hat{P} \leq \lfloor n\tilde{c}_\alpha \rfloor \end{cases}$ where $\tilde{c}_\alpha = p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$, z_q is the q-quantile of the standard normal distribution and $\lfloor \cdot \rfloor$ denotes the floor function. Obviously, the exact and the asymptotic critical regions are not equivalent. Their disagreement can be evaluated using Expression (4), which, in this case,can be exactly evaluated. In Table S1 of the Supplementary Material, the values of $D(p_t, n, \alpha, p_0)$ are computed by fixing $\alpha = 0.05$ for some values of n, p_0 and p_t. From this table, it emerges that the probability of disagreement between the tests $\Psi_\alpha(\mathbf{X}_n)$ and $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ can be very high for some combinations of n, p_0 and p_t: it is often higher than 10% (up to 20%) with sample size $n = 15$, and it remains higher than 10%, for just a few cases, even with $n = 30$.

3.1. Semi-Parametric RP-Estimation and Testing for the Binomial Test

The exact power function and the exact RP of Ψ_α (Case (**A.1**)) are $\pi(n, \alpha, p) = 1 - B(nc_\alpha; n, p)$ and $RP = \pi(n, \alpha, p_t)$, where $B(\cdot; n, p)$ is the binomial cumulative distribution function with parameters n and p. Following [8], the semi-parametric RP-estimator based on the exact power is obtained by plugging the median estimator for p_t into the expression of $\pi(n,\alpha,p)$. The median estimator $\hat{P}^{\bullet}$ is defined as the solution of the equation $B(n\hat{p};n,\hat{P}^{\bullet}) = 1/2$, and the resulting RP-estimator is $\hat{\pi} = 1 - B(nc_{\alpha};n,\hat{P}^{\bullet})$. Similarly, the approximate power function of $\widetilde{\Psi}_{\alpha}$ (Case (**A.2**)) is $\pi_a(n,\alpha,p) = 1 - \Phi\left(\sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{p(1-p)}}\right)$, where $\Phi(\cdot)$ is the standard normal cdf, and the approximate RP results $RP = \pi_a(n,\alpha,p_t)$. The corresponding RP-estimator is then $\hat{\pi}_a = \pi_a(n,\alpha,\hat{P})$. Note that, in this case, the probability distribution of $\hat{\pi}$ and $\hat{\pi}_a$ can be obtained analytically. In particular, the support of $\hat{\pi}$ is given by the values $\hat{\pi}(s) = 1 - B(nc_{\alpha};n,\hat{p}^{\bullet}_s)$, $s = 0,1,...,n$, where $\hat{p}^{\bullet}_s$ is the solution of $B(s;n,\hat{p}^{\bullet}) = 1/2$. The probability function of $\hat{\pi}$ is given by $P\left(\hat{\pi} = \hat{\pi}(s)\right) = \binom{n}{s}p_t^s(1-p_t)^{n-s}$, $s = 0,1,...,n$. Analogously, the support of $\hat{\pi}_a$ is $\hat{\pi}_a(s) = 1 - \Phi\left(\sqrt{n}\frac{p_0-s/n}{\sqrt{s/n(1-s/n)}} + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{s/n(1-s/n)}}\right)$, $s = 0,1,...,n$, and $P\left(\hat{\pi}_a = \hat{\pi}_a(s)\right) = P\left(\hat{\pi} = \hat{\pi}(s)\right) = \binom{n}{s}p_t^s(1-p_t)^{n-s}$, $s = 0,1,...,n$.

Now, both the asymptotic and the exact tests can be replicated by using the RP-estimators defined above. Specifically, thanks to the results in [8] (which require the adoption of the median estimator $\hat{P}^{\bullet}$ in the definition of the RP-estimator $\hat{\pi}$), it results that:

$$\Psi_{\alpha}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{P} > c_{\alpha} \\ 0 & \text{if} \quad \hat{P} \le c_{\alpha} \end{cases} = \begin{cases} 1 & \text{if} \quad \hat{\pi} > 1/2 \\ 0 & \text{if} \quad \hat{\pi} \le 1/2 \end{cases}.$$

Similarly, it is easy to verify that: $\widetilde{\Psi}_{\alpha}(\mathbf{X}_n) = \begin{cases} 1 & \text{if} \quad \hat{P} > \tilde{c}_{\alpha} \\ 0 & \text{if} \quad \hat{P} \le \tilde{c}_{\alpha} \end{cases} = \begin{cases} 1 & \text{if} \quad \hat{\pi}_a > 1/2 \\ 0 & \text{if} \quad \hat{\pi}_a \le 1/2 \end{cases}.$

Note that, also for the validity of this last identity, the use of the point estimator $\hat{P}$ in the definition of $\hat{\pi}_a$ is fundamental.

3.2. Non-Parametric RP-Estimation and Testing for the Binomial Test

The case of the binomial test is particularly interesting when studying the features of the plug-in RP-estimators, since, in this context, the probability function of the estimators $\hat{\pi}_e^{PI}$ and $\hat{\pi}_{a,e}^{PI}$ can be analytically derived, and the RP-based decision rules based on the latter can be analytically studied. Lemma 2 below describes the analytical expression of the non-parametric plug-in RP-estimator for the exact binomial test (Point *1*); provides the probability distribution of this estimator (Point *2*); establishes the equivalence between the exact binomial test and the RP-based decision rule derived by the non-parametric plug-in estimator (Point *3*). Similar results concerning the asymptotic binomial test are provided in Lemma 3.

Lemma 2. *Let $\mathbf{X}_n = (X_1,...,X_n)$ be a random sample drawn from the Bernoulli distribution with unknown parameter p_t in order to test hypotheses (7). It results that:*

1. $\hat{\pi}_e^{PI} = \dfrac{1}{n^n}\displaystyle\sum_{\mathbf{x}_n^i \in \mathcal{X}(\mathbf{X}_n)} \Psi_{\alpha}(\mathbf{x}_n^i) = 1 - B\left(nc_{\alpha};n,\widehat{P}\right)$;
2. *the support of $\hat{\pi}_e^{PI}$ is $\hat{\pi}_e^{PI}(s) = 1 - B\left(nc_{\alpha};n,\frac{s}{n}\right)$, $s = 0,1,...,n$, and* $P(\hat{\pi}_e^{PI} = \hat{\pi}_e^{PI}(s)) = \binom{n}{s}p_t^s(1-p_t)^{n-s}$.
3. *the decision rule based on the RP-estimator $\hat{\pi}_e^{PI}$ exactly replicates the exact Binomial test Ψ_{α}.*

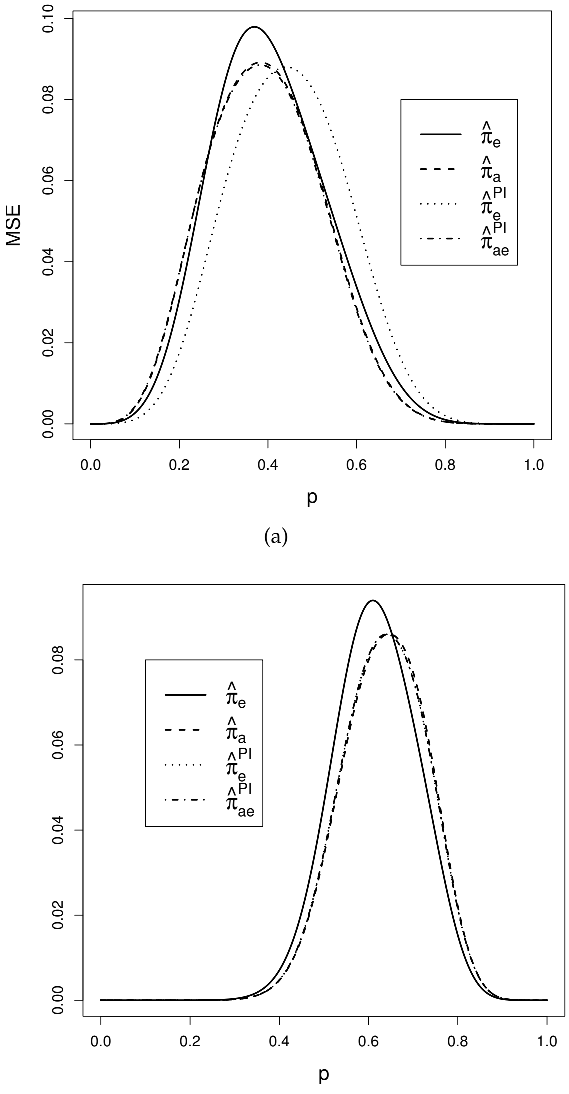
Lemma 3. *Let $\mathbf{X}_n = (X_1,...,X_n)$ be a random sample drawn from the Bernoulli distribution with unknown parameter p_t in order to test Hypotheses (7). It results that:*

1. $\hat{\pi}_{a,e}^{PI} = \dfrac{1}{n^n}\displaystyle\sum_{\mathbf{x}_n^i \in \mathcal{X}(\mathbf{X}_n)} \widetilde{\Psi}_{\alpha}(\mathbf{x}_n^i) = 1 - B\left(\lfloor n\tilde{c}_{\alpha}\rfloor;n,\widehat{P}\right)$;
2. *the support of $\hat{\pi}_{a,e}^{PI}$ is $\hat{\pi}_{a,e}^{PI}(s) = 1 - B\left(\lfloor n\tilde{c}_{\alpha}\rfloor;n,\frac{s}{n}\right)$, $s = 0,1,...,n$, and* $P(\hat{\pi}_{a,e}^{PI} = \hat{\pi}_{a,e}^{PI}(s)) = \binom{n}{s}p_t^s(1-p_t)^{n-s}$.
3. *the decision rule based on the RP-estimator $\hat{\pi}_{a,e}^{PI}$ exactly replicates the asymptotic Binomial test $\widetilde{\Psi}_{\alpha}$.*

The proofs of Lemma 2 and Lemma 3 are reported in the Supplementary Material.

3.3. Evaluating the Performances of the RP-Estimators for the Binomial Test

In the case of the binomial test, it is possible to compute the exact expectation and the MSE of $\hat{\pi}$ and $\hat{\pi}_a$, $\hat{\pi}_e^{PI}$ and $\hat{\pi}_{a,e}^{PI}$. In order to make a comparison among these estimators, their exact bias and MSE are represented in Figure S1 and Figure S2 of the Supplementary Material. Here, in Figure 1, only the MSE curves with $n = 15$, $\alpha = 0.05$, and $p = 0.2, 0.5$, are given. From these figures, it emerges that there is no RP-estimator that uniformly performs best. Concerning the estimators for the power of $\Psi_\alpha(\mathbf{X}_n)$, there is a tangible difference between the performance of $\hat{\pi}$ and $\hat{\pi}_e^{PI}$. For a wide range of small values of p_t, $\hat{\pi}$ has a bias and MSE, which is greater than the one of $\hat{\pi}_e^{PI}$; for large values of p_t, $\hat{\pi}$ generally performs better than $\hat{\pi}_e^{PI}$; whereas, the performances of $\hat{\pi}_a$ and $\hat{\pi}_a^{PI}$ for the power of $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ are very similar. Regarding RP-testing, we recall that there is no disagreement between classical binomial tests (exact or approximated) and their RP-based version. The results obtained here for the binomial test still hold for the sign test. The interested reader is referred to the Supplementary Material where the connection between these tests is explained in depth.



(a)

(b)

Figure 1. MSE curves of the reproducibility probability (RP) estimators $\hat{\pi}_e$ (solid), $\hat{\pi}_a$ (dashed), $\hat{\pi}_e^{PI}$ (dotted) and $\hat{\pi}_{a,e}^{PI}$ (dot-dashed). The MSE curves are computed considering the testing problem (7) with $\alpha = 0.05$, $p_0 = 0.2, 0.5$ and $n = 15$. (**a**) $p_0 = 0.2$, $n = 15$; (**b**) $p_0 = 0.5$, $n = 30$.

4. RP-Estimation and Testing for the Wilcoxon Signed Rank Test

Let $\mathbf{X}_n = (X_1, ..., X_n)$ be a random sample from a continuous and symmetric cdf F_{θ_t} with median θ_t. In order to test $H_0 : \theta_t \leq \theta_0$ vs $H_1 : \theta_t > \theta_0$, it is possible to apply the Wilcoxon signed rank (WSR) test, which is based on the statistic $W = \sum_{i=1}^{n} I_{ii} R_i = \sum_{i=1}^{n} \sum_{j=i}^{n} I_{ij}$ where $Z_i = X_i - \theta_0$, $R_i = \text{rank}(|Z_i|)$ and:

$$I_{ij} = \begin{cases} 1 & \text{if } Z_i + Z_j > 0 \\ 0 & \text{if } Z_i + Z_j < 0 \end{cases} . \tag{8}$$

Following the classification proposed in Section 2, the WSR test falls under Case (**B**), since the exact distribution of W can be derived by enumeration (see [19] on p. 126) under H_0, but, under H_1, it can only be approximated by using a central limit theorem. In particular, it is well known (see [19] on p. 166) that $\frac{W - E_{F_{\theta_t}}[W]}{\sqrt{Var_{F_{\theta_t}}(W)}} \xrightarrow{d} \mathcal{N}(0,1)$ where:

$$E_{F_{\theta_t}}[W] = e(p, p_1) = \frac{n(n-1)}{2} p_1 + np \, , \tag{9}$$

$$\begin{aligned} Var_{F_{\theta_t}}[W] &= v(p, p_1, p_2) \\ &= n(n-1)(n-2)(p_2 - p_1^2) + \frac{n(n-1)}{2}\left[2(p-p_1)^2 + 3p_1(1-p_1)\right] + np(1-p) \, , \end{aligned} \tag{10}$$

with:

$$p = P_{F_{\theta_t}}(Z > 0) \, , \quad p_1 = P_{F_{\theta_t}}(Z + Z' > 0) \, , \quad p_2 = p_1 = P_{F_{\theta_t}}(Z + Z' > 0 \text{ and } Z + Z'' > 0) \, , \tag{11}$$

being $Z = X - \theta_0$, Z' and Z'' i.i.d. to Z.

Note that, under H_0, $p = p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{3}$. These results allow the use of W in order to define the exact and asymptotic tests

$$\Psi_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if } \; W > w_\alpha \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \widetilde{\Psi}_\alpha(\mathbf{X}_n) = \begin{cases} 1 & \text{if } \; W > \tilde{w}_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where w_α denotes the $(1-\alpha)$-quantile of the exact null distribution of W and $\tilde{w}_\alpha = \frac{n(n+1)}{4} + z_{1-\alpha}\sqrt{\frac{n(n+1)(2n+1)}{24}}$. Obviously, the exact and asymptotic tests are not equivalent, and their disagreement is evaluated using Expression (4). In Table S2 of the Supplementary Material, the values of $D(\alpha, n, F_{\theta_t}, \theta_0)$ are computed by fixing $\alpha = 0.05$ and $\theta_0 = 0$ for some values of n and θ_t and by considering $X \sim \mathcal{N}(\theta_t, 1)$ (light tails) and $X \sim \text{Cauchy}(\theta_t)$ (fat tails).

4.1. Semi-Parametric RP-Estimation and Testing for the WSR Test

As mentioned above, the WSR is classified under Case (**B**). Therefore, its exact power function cannot be generally determined. However, it can be approximated thanks to the asymptotic normality of W. The approximation of the power function of the exact test $\Psi_\alpha(\mathbf{X}_n)$ is $\tilde{\pi}(n, \alpha, F_\theta, \theta_0) \approx 1 - \Phi\left(\frac{w_\alpha - E_{F_\theta}[W]}{\sqrt{Var_{F_\theta}[W]}}\right)$. Analogously, the approximation of the power function of the asymptotic test $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ is $\pi_a(n, \alpha, F_\theta, \theta_0) \approx 1 - \Phi\left(\frac{\tilde{w}_\alpha - E_{F_\theta}[W]}{\sqrt{Var_{F_\theta}[W]}}\right)$. Now, in order to define some semi-parametric RP-estimators starting from the approximated power function reported above, it is necessary to derive the estimators for $E_{F_\theta}[W]$ and $Var_{F_\theta}[W]$. They can be obtained by plugging into Expressions (9) and (10) the estimators for the parameters p, p_1 and p_2, defined in (11). Below, two different estimators for these parameters are considered.

- **Analogic estimators**:
  $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I_{ii}$, $\hat{p}_1 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} I_{ij} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} I_{ij}$,
  $\hat{p}_2 = \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j\neq k}}^{n}\sum_{k=1}^{n} I_{ij}I_{ik} = \frac{2}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{\substack{k=j+1\\ i\neq k}}^{n} I_{ij}I_{ik}$.
- **Plug-in estimators**. In order to introduce the plug-in estimators for p, p_1 and p_2, let $G_{\theta_t}(z) = F_{\theta_t}(z+\theta_0)$ and $g_{\theta_t}(z)$ be the cumulative distribution function and the density function of $Z = X - \theta_0$. By using this notation, it is easy to note that $p = G_{\theta_t}(0)$, $p_1 == 1 - E_{G_{\theta_t}}[G_\theta(-Z)]$, and $p_2 = 1 - 2E_{G_{\theta_t}}[G_\theta(-Z)] + E_{G_{\theta_t}}[G_\theta(-Z)^2]$. Let G_n be the empirical distribution function of the Z_i's (*i.e.*, of the $X_i - \theta_0$'s). By plugging G_n into the above expressions, the following estimators are obtained:

$$\tilde{p} = 1 - G_n(0) \equiv \hat{p}\ , \quad \tilde{p}_1 = 1 - \frac{1}{n}\sum_{i=1}^{n} G_n(-Z_i)\ , \quad \tilde{p}_2 = 1 - 2\frac{1}{n}\sum_{i=1}^{n} G_n(-Z_i) + \frac{1}{n}\sum_{i=1}^{n} G_n^2(-Z_i)\ .$$

Now, the following RP-estimators for the exact test can be introduced: $\hat{\pi}_1 = 1 - \Phi\left(\frac{w_\alpha - \hat{E}}{\sqrt{\hat{V}}}\right)$ and $\hat{\pi}_2 = 1 - \Phi\left(\frac{w_\alpha - \tilde{E}}{\sqrt{\tilde{V}}}\right)$, where $\hat{E} = e(\hat{p}, \hat{p}_1)$, $\hat{V} = v(\hat{p}, \hat{p}_1, \hat{p}_2)$, $\tilde{E} = e(\tilde{p}, \tilde{p}_1)$ and $\tilde{V} = v(\tilde{p}, \tilde{p}_1, \tilde{p}_2)$. Analogously, the following RP-estimators for the asymptotic test can be introduced: $\hat{\pi}_{a1} = 1 - \Phi\left(\frac{\tilde{w}_\alpha - \hat{E}}{\sqrt{\hat{V}}}\right)$ and $\hat{\pi}_{a2} = 1 - \Phi\left(\frac{\tilde{w}_\alpha - \tilde{E}}{\sqrt{\tilde{V}}}\right)$.

Following the idea in [20], the approximated power of nonparametric tests can be simplified by assuming that the variance of the test statistic is close to its value under H_0 (see [19] on pp. 72 and 167, for other applications of Noether's approach). In that case, the approximated and simplified power functions of $\Psi_\alpha(\mathbf{X}_n)$ and $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ result: $\widetilde{\pi}(n, \alpha, F_\theta, \theta_0) \approx 1 - \Phi\left(\frac{w_\alpha - E_{F_{\theta_t}}[W]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$, and $\pi_a(n, \alpha, F_\theta, \theta_0) \approx 1 - \Phi\left(\frac{\tilde{w}_\alpha - E_{F_{\theta_t}}[W]}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$. These expressions give rise to the following additional RP-estimators:

- RP-estimators for the exact test:
  $\hat{\pi}_{1S} = 1 - \Phi\left(\frac{w_\alpha - \hat{E}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$ and $\hat{\pi}_{2S} = 1 - \Phi\left(\frac{w_\alpha - \tilde{E}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$;
- RP-estimators for the asymptotic test:
  $\hat{\pi}_{aS1} = 1 - \Phi\left(\frac{\tilde{w}_\alpha - \hat{E}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$ and $\hat{\pi}_{aS2} = 1 - \Phi\left(\frac{\tilde{w}_\alpha - \tilde{E}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$.

Finally, the estimators based on the following Noether's power approximation $\widetilde{\pi}(n, \alpha, F_\theta, \theta_0) \approx \pi_a(n, \alpha, F_\theta, \theta_0) \approx 1 - \Phi\left(z_{1-\alpha} - \sqrt{3n}\left(p_1 - \frac{1}{2}\right)\right)$ are also considered here. In particular, the estimators $\hat{\pi}_{N1} = 1 - \Phi\left(z_{1-\alpha} - \sqrt{3n}\left(\hat{p}_1 - \frac{1}{2}\right)\right)$ and $\hat{\pi}_{N2} = 1 - 1 - \Phi\left(z_{1-\alpha} - \sqrt{3n}\left(\tilde{p}_1 - \frac{1}{2}\right)\right)$ are applied to estimate the RP of both the exact and asymptotic WSR tests.

Concerning the RP-based version of the WSR test based on the introduced semi-parametric RP-estimators, the following corollary (proven in the Supplementary Material) can be stated:

Corollary 4. *The decision rules based on the RP-estimators $\hat{\pi}_1$ and $\hat{\pi}_{1S}$ exactly replicate the exact WSR test Ψ_α. Analogously, the decision rules based on the RP-estimators $\hat{\pi}_{a1}$ and $\hat{\pi}_{aS1}$ exactly replicate the asymptotic WSR test $\widetilde{\Psi}_\alpha$.*

Concerning the RP-based decision rules stemming from the remaining semi-parametric RP-estimators (*i.e.*, $\hat{\pi}_2$, $\hat{\pi}_{2S}$, $\hat{\pi}_{a2}$, $\hat{\pi}_{aS2}$, $\hat{\pi}_{N1}$ and $\hat{\pi}_{N2}$), they do not replicate the exact/asymptotic WSR tests, and their disagreement probabilities will be evaluated in Section 4.3.

4.2. Non-Parametric RP-Estimation and Testing for the WSR Test

As explained, in Section 2, the RP of the exact and asymptotic WSR test can be estimated by using (5) and (6), respectively. Here, we consider the non-parametric RP-estimators $\hat{\pi}_5^{PI}$, $\hat{\pi}_{10}^{PI}$, $\hat{\pi}_{20}^{PI}$, $\hat{\pi}_{a5}^{PI}$, $\hat{\pi}_{a10}^{PI}$ and $\hat{\pi}_{a20}^{PI}$. The first three estimators coincide with (5) with $B = 500$, $B = 1000$, $B = 2000$. The last three estimators coincide with (6) with $B = 500$, $B = 1000$, $B = 2000$. As mentioned above, the RP-based decision rules based on these estimators do not replicate the exact and asymptotic WSR tests, respectively, and their disagreement probabilities will be evaluated in Section 4.3.

4.3. Evaluating the Performances of the RP-Estimators for the WSR Test

In order to evaluate the performances of the several RP-estimators introduced above for the exact and asymptotic WSR test, a simulation study is built. The scenarios considered in the simulation study regard the testing problem $H_0 : \theta_t \leq 0$ *vs.* $H_0 : \theta_t > 0$ with $\alpha = 0.05$. The considered sample sizes are $n = 15, 30, 60, 120, 240$. Data are drawn from normal distribution with unit variance and mean (median) θ_t and shifted Cauchy with median θ_t. For each one of the considered sample sizes and distributions (normal or Cauchy), 19 values for θ_t have been considered. These values have been obtained by simulation and have been chosen in order to provide the following prefixed values for the power of the exact/asymptotic test: (α, 0.1, 0.15, 0.20, 0.25, ..., 0.85, 0.9, 0.95). In each simulation, 10^4 replications are considered.

The results of the simulation study are summarized in Tables S3 and S4 in the Supplementary Material, where the averages (computed over the 19 different values of θ_t) of the simulated MSE, simulated bias and disagreement rate are provided. Here, in Table 2, only the simulated MSE and disagreement rate related to the Cauchy distribution are provided.

Table 2. Averaged MSE and disagreement rate for the asymptotic and exact Wilcoxon signed rank (WSR) test when sampling from the Cauchy distribution. The averages are computed over the 19 different values of θ considered in the simulation study. The smallest values for the averaged MSE and disagreement are highlighted in bold.

RP-estimation and Testing for the Asymptotic WSR Test										
	$n = 15$		$n = 30$		$n = 60$		$n = 120$		$n = 240$	
RP-est.	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**
$\hat{\pi}_{N1}$	0.0684	0.0286	0.0678	0.0073	0.0686	0.0045	0.0688	0.0018	0.0683	0.0012
$\hat{\pi}_{N2}$	0.0664	0.0131	0.0669	0.0057	0.0682	0.0025	0.0686	0.0010	0.0682	0.0008
$\hat{\pi}_{aS1}$	0.0652	**0.0000**	0.0664	**0.0000**	0.0680	**0.0000**	0.0685	**0.0000**	0.0681	**0.0000**
$\hat{\pi}_{aS2}$	**0.0636**	0.0122	**0.0656**	0.0034	**0.0676**	0.0019	**0.0683**	0.0009	**0.0680**	0.0004
$\hat{\pi}_{a1}$	0.0793	**0.0000**	0.0734	**0.0000**	0.0713	**0.0000**	0.0702	**0.0000**	0.0690	**0.0000**
$\hat{\pi}_{a2}$	0.0735	0.0122	0.0702	0.0034	0.0697	0.0019	0.0693	0.0009	0.0685	0.0004
$\hat{\pi}_{a5}^{PI}$	0.0718	0.0180	0.0698	0.0142	0.0696	0.0136	0.0695	0.0133	0.0687	0.0129
$\hat{\pi}_{a10}^{PI}$	0.0717	0.0164	0.0696	0.0111	0.0694	0.0097	0.0693	0.0093	0.0686	0.0091
$\hat{\pi}_{a20}^{PI}$	0.0716	0.0156	0.0695	0.0091	0.0694	0.0072	0.0692	0.0068	0.0685	0.0065
RP-estimation and Testing for the Exact WSR Test										
	$n = 15$		$n = 30$		$n = 60$		$n = 120$		$n = 240$	
RP-est.	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**
$\hat{\pi}_{N1}$	0.0677	0.0200	0.0678	0.0082	0.0686	0.0040	0.0688	0.0018	0.0683	0.0011
$\hat{\pi}_{N2}$	0.0657	0.0045	0.0668	0.0033	0.0682	0.0020	0.0686	0.0010	0.0682	0.0006
$\hat{\pi}_{S1}$	0.0647	**0.0000**	0.0664	**0.0000**	0.0680	**0.0000**	0.0685	**0.0000**	0.0681	**0.0000**
$\hat{\pi}_{S2}$	**0.0631**	0.0066	**0.0655**	0.0036	**0.0675**	0.0018	**0.0683**	0.0009	**0.0680**	0.0005
$\hat{\pi}_1$	0.0797	**0.0000**	0.0734	**0.0000**	0.0713	**0.0000**	0.0702	**0.0000**	0.0690	**0.0000**
$\hat{\pi}_2$	0.0739	0.0066	0.0703	0.0036	0.0697	0.0018	0.0693	0.0009	0.0685	0.0005
$\hat{\pi}_5^{PI}$	0.0722	0.0211	0.0698	0.0143	0.0696	0.0137	0.0695	0.0133	0.0687	0.0130
$\hat{\pi}_{10}^{PI}$	0.0721	0.0200	0.0697	0.0113	0.0694	0.0099	0.0693	0.0093	0.0686	0.0092
$\hat{\pi}_{20}^{PI}$	0.0720	0.0193	0.0696	0.0091	0.0694	0.0073	0.0692	0.0068	0.0685	0.0066

Note that the disagreement between the exact Wilcoxon signed rank test and its approximated versions is often higher than 2% with $n = 15$ (up to 2.5%) and can reach 0.8% with $n = 30$ (see Table S2 in the Supplementary Material). Rather, the averaged disagreement between classical tests and their RP-based version, with $n = 15$, surpasses 2% just with two estimators, whereas for some of them, no disagreement is shown; with $n = 30$, some RP estimators provide a disagreement between the classical test and the RP-based one resulting in a little higher than 1%, but no disagreement is shown for two of them.

Regarding RP estimation, the estimators that globally have the lowest MSE are $\hat{\pi}_{aS2}$ for the approximated test and $\hat{\pi}_{S2}$ for the exact test. However, these estimators do not exactly replicate the corresponding classical test. By considering both the estimation performance and the disagreement probability, we suggest using the estimators $\hat{\pi}_{aS1}$ for the approximated test and $\hat{\pi}_{S1}$ for the exact test, since their MSE is very similar to the ones of $\hat{\pi}_{aS2}$ and $\hat{\pi}_{S2}$, but their disagreement probability is null. As a final remark, note the good performance of the non-parametric plug-in estimators, which is not far from the one of the semi-parametric ones, even if they are not *ad hoc* estimators.

5. RP-Estimation and Testing for the Kendall Test of Monotonic Association

Let (X, Y) be a bivariate continuous random variable with joint distribution $\mathbf{F}_t$ and margins ${}_tF_X$ and ${}_tF_Y$. Let $(\mathbf{X}, \mathbf{Y})_n = \{(X_i, Y_i), i = 1, ..., n\}$ be a random sample drawn from $\mathbf{F}_t$. To test the presence of positive or negative monotone association between X and Y, the Kendall test can be adopted. Without loss of generality, consider the alternative hypothesis of positive monotone association. In that case, the testing problem of interest is $H_0 : \tau_t \leq 0$ *vs.* $H_0 : \tau_t > 0$, where τ_t is the Kendall rank-correlation coefficient, which, under the assumption of absolute continuity of $\mathbf{F}_t$, is defined as the difference between the probability of concordance p_1 and the probability of discordance p_1': $\tau_t = p_1 - p_1' = 2p_1 - 1$ with $p_1 = P_{\mathbf{F}_t}\left((X - X')(Y - Y') > 0\right)$, $p_1' = P_{\mathbf{F}_t}\left((X - X')(Y - Y') < 0\right) = 1 - p_1$ and (X', Y') i.i.d as (X, Y). The test statistics is $\hat{\tau} = \frac{2K}{n(n-1)}$ where $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\text{sign}(X_i - X_j) \cdot \text{sign}(Y_i - Y_j)\right)$ and $sign(a) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$. As for the WSR test, the Kendall test falls under Case (**B**). The exact distribution of $\hat{\tau}$ can be derived, under H_0, by enumeration or by using a recurrence relation (see [21]), but generally, it can only be approximated through a central limit theorem under H_1. In particular, it is well known (see [22]) that $\frac{\hat{\tau} - E_{\mathbf{F}_t}[\hat{\tau}]}{\sqrt{Var_{\mathbf{F}_t}[\hat{\tau}]}} \xrightarrow{d} \mathcal{N}(0, 1)$ where:

$$E_{\mathbf{F}_t}[\hat{\tau}] = \tau_t \tag{12}$$

$$\text{and} \qquad Var_{\mathbf{F}_t}[\hat{\tau}] = u(\tau_t, p_2) = \frac{2}{n(n-1)}(1 - \tau_t^2) + \frac{4(n-2)}{n(n-1)}(2p_2 - 1 - \tau_t^2) \tag{13}$$

$$\text{with} \qquad p_2 = P_{\mathbf{F}_t}[(X - X')(Y - Y')(X - X'')(Y - Y'') > 0] \tag{14}$$

being (X', Y') and (X'', Y'') i.i.d as (X, Y).

Note that, under H_0, $p_2 = \frac{5}{9}$, and consequently: $Var_0[\hat{\tau}] = u\left(0, \frac{5}{9}\right) = \frac{2(2n+5)}{9n(n-1)}$. These results allow the introduction of the exact and asymptotic Kendall tests $\Psi_\alpha((\mathbf{X}, \mathbf{Y})_n) = \begin{cases} 1 & \text{if } \hat{\tau} > t_\alpha \\ 0 & \text{otherwise} \end{cases}$ and $\widetilde{\Psi}_\alpha((\mathbf{X}, \mathbf{Y})_n) = \begin{cases} 1 & \text{if } \hat{\tau} > \tilde{t}_\alpha \\ 0 & \text{otherwise} \end{cases}$ where t_α denotes the $(1 - \alpha)$-quantile of the exact null distribution of $\hat{\tau}$ and $\tilde{t}_\alpha = z_{1-\alpha}\sqrt{\frac{2(2n+5)}{9n(n-1)}}$.

Note that the computational burden necessary to compute the exact null distribution of $\hat{\tau}$ increases rapidly with n. From a practical point of view, the exact test can be performed, if $n < 9$, by computing the exact $(1 - \alpha)$-quantile from the null distribution of $\hat{\tau}$ using, for example, the software *R* [23] function *qKendall* of package SuppDists [24]. If $n > 9$, the asymptotic test $\widetilde{\Psi}_\alpha$ is generally performed or an Edgeworth expansion (see [25]) is used to obtain a better approximation of t_α. The $(1 - \alpha)$-quantile from the null distribution of $\hat{\tau}$ approximated by the Edgeworth expansion is also computed by *qKendall*. When $n > 9$, it is common practice to refer to the test based on the Edgeworth expansion as the "exact" Kendall test, even if it is actually an approximated test. From here onwards, this commonly-used terminology will be adopted.

Obviously, the exact and asymptotic tests are not equivalent, and their disagreement is evaluated, again, using Expression (4). In Table S5 of the Supplementary Material, the probabilities of disagreement between the asymptotic and exact (based on Edgeworth expansion) tests are computed by fixing $\alpha = 0.05$ for some values of n and τ_t when sampling from the bivariate normal distribution with correlation coefficient ρ and from the bivariate Student's t distribution with three degrees of freedom (df) and correlation coefficient ρ.

5.1. Semi-Parametric RP-Estimation and Testing for the Kendall Test

The exact power function of the Kendall test cannot be generally determined, but it can be approximated thanks to the asymptotic normality of $\hat{\tau}$. In particular, the approximation of the power function of the exact test $\Psi_\alpha((\mathbf{X},\mathbf{Y})_n)$ is $\widetilde{\pi}(n,\alpha,\mathbf{F}_t) \approx 1 - \Phi\left(\frac{t_\alpha - E_{\mathbf{F}_t}[\hat{\tau}]}{\sqrt{Var_{\mathbf{F}_t}[\hat{\tau}]}}\right)$. Analogously, the approximation of the power function of the asymptotic test $\widetilde{\Psi}_\alpha((\mathbf{X},\mathbf{Y})_n)$ is $\pi_a(n,\alpha,\mathbf{F}_\tau) \approx 1 - \Phi\left(\frac{\tilde{t}_\alpha - E_{\mathbf{F}_t}[\hat{\tau}]}{\sqrt{Var_{\mathbf{F}_t}[\hat{\tau}]}}\right)$.

Now, in order to define semi-parametric RP-estimators starting from the approximated power function reported above, it is necessary to derive estimators for $E_{\mathbf{F}_t}[\hat{\tau}]$ and $Var_{\mathbf{F}_t}[\hat{\tau}]$. From Expressions (12) and (13), it follows that $E_{\mathbf{F}_t}[\hat{\tau}]$ can be estimated by $\hat{\tau}$, while an estimator for $Var_{\mathbf{F}_t}[\hat{\tau}]$ can be introduced once an estimator for p_2 has been defined. Two different estimators for p_2 are considered here:

- **Analogic estimators**: Remembering Expression (14), the analogic estimator for p_2 results: $\hat{p}_2 = \frac{1}{n(n-1)(n-2)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j \neq k}}^n \sum_{k=1}^n I_{ijk}$ where
$$I_{ijk} = \begin{cases} 1 & \text{if } (x_i - x_j)(x_i - x_k)(y_i - y_j)(y_i - y_k) > 0 \\ 0 & \text{otherwise} \end{cases}.$$
- **Plug-in estimators**: In order to introduce these estimators, the following alternative expression for p_2 is useful: $p_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[p(x,y)^2 + (1-p(x,y))^2\right] \mathbf{f}_t(x,y)dxdy = E_{\mathbf{F}_t}\left[p(X,Y)^2 + (1-p(X,Y)^2)\right]$ where $p(x,y) = P_{\mathbf{F}_t}[X \leq x \cap Y \leq y] + P_{\mathbf{F}_t}[X > x \cap Y > y] = 1 - {}_tF_X(x) - {}_tF_Y(y) + 2\mathbf{F}_t(x,y)$.

  Now, let $\widehat{F}_{nX}$, $\widehat{F}_{nY}$ and $\widehat{\mathbf{F}}_n$ be the ecdfs of X, Y and (X,Y), respectively. The plug-in estimators for p_2 results: $\tilde{p}_2 = \frac{1}{n}\sum_{i=1}^n \left[\tilde{p}(x_i,y_i)^2 + (1-\tilde{p}(x_i,y_i))^2\right]$ where $\tilde{p}(x,y) = 1 - \widehat{F}_{nX}(x) - \widehat{F}_{nY}(y) + 2\widehat{\mathbf{F}}_n(x,y)$.

Now, the following RP-estimators for the exact test can be introduced: $\hat{\pi}_1 = 1 - \Phi\left(\frac{t_\alpha - \hat{\tau}}{\sqrt{\hat{U}}}\right)$ and $\hat{\pi}_2 = 1 - \Phi\left(\frac{t_\alpha - \hat{\tau}}{\sqrt{\tilde{U}}}\right)$, where $\hat{U} = u(\hat{\tau},\hat{p}_2)$, $\tilde{U} = u(\hat{\tau},\tilde{p}_2)$. Analogously, the following RP-estimators for the asymptotic test can be introduced: $\hat{\pi}_{a1} = 1 - \Phi\left(\frac{\tilde{t}_\alpha - \hat{\tau}}{\sqrt{\hat{U}}}\right)$ and $\hat{\pi}_{a2} = 1 - \Phi\left(\frac{\tilde{t}_\alpha - \hat{\tau}}{\sqrt{\tilde{U}}}\right)$.

As for the WSR test, the approximated power of nonparametric tests can be simplified following Noether's idea by assuming that the variance of the test statistic is close to the value it assumes under H_0, obtaining the estimators $\widehat{\pi}_S = 1 - \Phi\left(\frac{t_\alpha - \hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}\right)$ and $\widehat{\pi}_{aS} = 1 - \Phi\left(\frac{\tilde{t}_\alpha - \hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}\right)$ for the exact and asymptotic tests, respectively. Note that the above estimators are very simple, since they do not require the estimation of p_2. Another approach that can be followed in order to introduce an approximation of the power function $\widetilde{\pi}(n,\alpha,\mathbf{F}_t)$ and $\pi_a(n,\alpha,\mathbf{F}_t)$ is described below. From Expression (14), it is clear that p_2 is not independent from τ_t, and there is no unique function describing the behavior of p_2 as a function of τ_t since the relation between p_2 and τ_t depends on the entire shape of $\mathbf{F}_t$. However, if $\tau_t = \pm 1$, then $p_2 = 1$, while if $\tau_t = 0$, then $p_2 = 5/9$. Then, the relation between τ_t and p_2 can be intuitively represented by the parabola passing through the points $(-1,1)$, $(0,5/9)$ and $(1,1)$: $p_2 = 1 - \frac{4}{9}(1-\tau_t^2)$. By substituting this expression into (13) one obtains the RP-estimators $\widehat{\pi}_L = 1 - \Phi\left(\frac{t_\alpha - \hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}(1-\hat{\tau}^2)}}\right)$ and $\widehat{\pi}_{aL} = 1 - \Phi\left(\frac{\tilde{t}_\alpha - \hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}(1-\hat{\tau}^2)}}\right)$ for the exact and asymptotic tests, respectively.

For completeness, the [20] estimator is also considered and applied both to the exact and asymptotic tests: $\widehat{\pi}_N = 1 - \Phi\left(z_{1-\alpha} - \frac{3}{2}\sqrt{n}\hat{\tau}\right)$. Finally, the estimators deduced from the power approximation provided in [2] are considered: $\widehat{\pi}_{C1} = 1 - \Phi\left(\frac{z_{1-\alpha}3 - \frac{\sqrt{n}}{2}\hat{\tau}}{\sqrt{2\hat{p}_2 - 1 - \hat{\tau}^2}}\right)$, $\widehat{\pi}_{C2} = 1 - \Phi\left(\frac{z_{1-\alpha}3 - \frac{\sqrt{n}}{2}\hat{\tau}}{\sqrt{2\tilde{p}_2 - 1 - \hat{\tau}^2}}\right)$.

5.2. Non-Parametric RP-Estimation and Testing for the Kendall Test

As for the WSR test, the RP of the exact and asymptotic Kendall test can be estimated by using the non-parametric RP-estimators $\hat{\pi}_5^{PI}$, $\hat{\pi}_{10}^{PI}$, $\hat{\pi}_{20}^{PI}$, $\hat{\pi}_{a5}^{PI}$, $\hat{\pi}_{a10}^{PI}$ and $\hat{\pi}_{a20}^{PI}$. It is recalled once again that the RP-based decision rules based on these estimators do not replicate the exact and asymptotic Kendall tests, respectively, and their disagreement probabilities will be evaluated next.

5.3. Evaluating the Performances of the RP-Estimators for the Kendall Test

In order to evaluate the performances of the several RP-estimators introduced above for the exact and asymptotic Kendall test, a simulation study is built. The scenarios considered in the simulation study regards the testing problem $H_0 : \tau_t \leq 0$ *vs.* $H_1 : \tau_t > 0$. The considered sample sizes are $n = 15, 30, 60, 120$. Data are drawn from the bivariate standard normal distribution with correlation coefficient ρ and from the bivariate Student's t distribution with three df and correlation coefficient ρ. For each one of the considered sample sizes and distributions, 19 values for ρ have been considered. These values have been obtained by simulation and have been chosen in order to provide the following prefixed values for the power of the exact/asymptotic test: $(\alpha, 0.1, 0.15, 0.20, 0.25, ..., 0.85, 0.9, 0.95)$. In each simulation, 10^4 replications are considered. The results of the simulation study are summarized in Table S6 and Table S7 of the Supplementary Materials. In these tables, the averages (computed over the 19 different values of θ) of the simulated MSE, simulated bias and disagreement rate are provided. Here (see Table 3), only the simulated MSE and disagreement rate obtained under the bivariate Student's t distribution with three df are reported. As for the binomial and sign tests, the disagreement between the exact Kendall test and its approximated versions is quite high: often higher than 5% and in some cases higher than 10%, both with $n = 15$ and $n = 30$. The disagreement is still remarkable even with $n = 120$. On the contrary, the averaged disagreement between the classical asymptotic test and its RP-based version is between 3% and 7% for just three estimators, for each sample size, whereas the other estimators provide a disagreement often lower than 1%. The disagreement between the classical exact test and the RP-based one results in being a little higher than the previous case, but still lower than the disagreements between classical tests.

Regarding RP estimation, the simulation results suggest that the best estimators for the approximated and exact tests are $\hat{\pi}_{a2}$ and $\hat{\pi}_2$, respectively. Indeed, these two estimators generally have the least MSE and a null probability of disagreement. Also in these cases, the good performance of the general non-parametric plug-in estimators should be noted.

Table 3. Averaged MSE and disagreement rate for the asymptotic and exact Kendall's test when sampling from the t copula with 3 degrees of freedom. The averages are computed over the 19 different values of ρ considered in the simulation study. The least values for the averaged MSE and disagreement are highlighted in bold.

RP-estimation and Testing for the Asymptotic Kendall's Test								
	$n = 15$		$n = 30$		$n = 60$		$n = 120$	
RP-est.	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**
$\hat{\pi}_N$	0.0759	0.0551	0.0751	0.0281	0.0752	0.0133	0.0748	0.0070
$\hat{\pi}_{aS}$	0.0689	**0.0000**	0.0721	**0.0000**	0.0738	**0.0000**	0.0741	**0.0000**
$\hat{\pi}_{a1}$	0.0778	**0.0000**	0.0735	**0.0000**	0.0709	**0.0000**	0.0693	**0.0000**
$\hat{\pi}_{a2}$	**0.0591**	**0.0000**	**0.0581**	**0.0000**	**0.0577**	**0.0000**	**0.0570**	**0.0000**
$\hat{\pi}_{C1}$	0.0881	0.0551	0.0765	0.0281	0.0720	0.0133	0.0697	0.0070
$\hat{\pi}_{C2}$	0.0617	0.0551	0.0589	0.0281	0.0579	0.0133	0.0571	0.0070
$\hat{\pi}_{aL}$	0.0736	**0.0000**	0.0741	**0.0000**	0.0747	**0.0000**	0.0745	**0.0000**
$\hat{\pi}^{PI}_{a5}$	0.0695	0.0769	0.0687	0.0595	0.0685	0.0448	0.0680	0.0329
$\hat{\pi}^{PI}_{a10}$	0.0694	0.0775	0.0685	0.0597	0.0683	0.0451	0.0679	0.0327
$\hat{\pi}^{PI}_{a20}$	0.0693	0.0779	0.0684	0.0596	0.0683	0.0455	0.0678	0.0329
RP-estimation and Testing for the Exact Kendall's Test								
	$n = 15$		$n = 30$		$n = 60$		$n = 120$	
RP-est.	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**	**MSE**	**D**
$\hat{\pi}_N$	0.0906	0.1411	0.0847	0.1063	0.0829	0.0955	0.0819	0.0882
$\hat{\pi}_S$	0.0672	**0.0000**	0.0713	**0.0000**	0.0734	**0.0000**	0.0739	**0.0000**
$\hat{\pi}_1$	0.0780	**0.0000**	0.0738	**0.0000**	0.0711	**0.0000**	0.0693	**0.0000**
$\hat{\pi}_2$	**0.0590**	**0.0000**	**0.0580**	**0.0000**	**0.0577**	**0.0000**	**0.0570**	**0.0000**
$\hat{\pi}_{C1}$	0.1064	0.1411	0.0876	0.1063	0.0803	0.0955	0.0769	0.0882
$\hat{\pi}_{C2}$	0.0764	0.1411	0.0680	0.1063	0.0651	0.0955	0.0634	0.0882
$\hat{\pi}_L$	0.0732	**0.0000**	0.0739	**0.0000**	0.0746	**0.0000**	0.0745	**0.0000**
$\hat{\pi}^{PI}_5$	0.0738	0.0781	0.0715	0.0601	0.0707	0.0457	0.0696	0.0324
$\hat{\pi}^{PI}_{10}$	0.0737	0.0784	0.0714	0.0604	0.0706	0.0461	0.0695	0.0326
$\hat{\pi}^{PI}_{20}$	0.0736	0.0791	0.0713	0.0607	0.0705	0.0463	0.0694	0.0323

6. Example of Applications

Let us consider the data reported in Table 4 (see [26], p.38), concerning the Hamilton depression scale factor (HDSF) in nine patients with mixed anxiety and depression, observed at a first visit before the initiation of a therapy (X) and at a second visit after administration of a tranquilizer (Y). An improvement due to the tranquilizer corresponds to a reduction of the HDSF. Six patients out of nine showed a reduction; one was almost invariant; and two gave small increments. The sign test and the WSR test have been applied to evaluate HDSF reduction and the Kendall test to evaluate the association between X and Y.

For each test, the RP estimates given by the best semiparametric estimator (among those studied above) and by the nonparametric $\hat{\pi}^{PI}_{20}$ are computed, at three levels of α: 0.01, 0.05, 0.1 (see Table 5). First, note that RP estimates decrease as tests become stricter, *i.e.*, as α decreases. Second, RP estimates fulfill RP-testing. Third, RP estimates might differ from one technique to another: the nonparametric technique is not the most reliable, but is a general one, whereas the best RP estimation technique should be customized for each test.

As concerns the interpretation of the results, RP estimates highlight that significant outcomes are often less reproducible than one may think. For example, when $\alpha = 0.05$, the significance threshold for the WSR test with $n = 9$ data is $w_{0.05} = 36$, and the significant result $W_{ob} = 40$, although providing a p-value that is quite small (*i.e.*, $\simeq 0.02$), gave an RP estimate of about 2/3: this means that, it is estimated that, under the same experimental conditions, about one out of three test replications will not show significance.

On the other hand, non-significant outcomes might be highly variable, and significance can be found with non-negligible probability when replicating the experiment. Continuing the example above, and assuming that α was 0.01, the observed test statistic provides a non-significant p-value of about two-times α, but also gives an RP estimate not far from 50%.

Table 4. Hamilton depression scale factor (HDSF) data: first visit (X) and second visit (Y).

x_i	y_i
1.83	0.878
0.50	0.647
1.62	0.598
2.48	2.050
1.68	1.060
1.88	1.290
1.55	1.060
3.06	3.140
1.30	1.290

Table 5. RP estimates for the example data.

Sign Test				
standard results	α	c_α	$\hat{\pi}$	$\hat{\pi}_e^{PI}$
$n = 9$	0.1	6	0.7905	0.6781
$B_{ob} = 7$	0.05	7	0.5	0.3719
$p\text{-}value = 0.0898$	0.01	8	0.1683	0.1042
WSR Test				
standard results	α	w_α	$\hat{\pi}_{1S}$	$\hat{\pi}_{20}^{PI}$
$n = 9$	0.1	34	0.7614	0.8835
$W_{ob} = 40$	0.05	36	0.6822	0.7435
$p\text{-}value = 0.0195$	0.01	41	0.4528	0.4505
Kendall Test				
standard results	α	t_α	$\hat{\pi}_2$	$\hat{\pi}_{20}^{PI}$
$n = 9$	0.1	0.3333	0.6979	0.6930
$\hat{\tau}_{ob} = 0.5$	0.05	0.4444	0.5685	0.5495
$p\text{-}value = 0.2231$	0.01	0.6111	0.3648	0.2615

Finally, we remark that even when p-values are quite a bit smaller than α, RP estimates may not be high, that is the test results (*viz.* significances) are estimated to be quite variable. For example, when p-values result in being about one order of magnitude smaller than α, RP estimates are still close to 80%.

7. Conclusions

Several results have been obtained, concerning both the precision of RP estimators and of RP-testing in the cases of the binomial, sign, Wilcoxon signed rank and Kendall tests.

For both the binomial and sign tests, the RP-testing rule holds exactly, also when nonparametric estimators of RP are adopted. In terms of estimation performances, semi-parametric and nonparametric estimators behave similarly.

For the WSR and Kendall tests, the RP-testing rule holds exactly for just some RP estimators, and for the remaining ones, the disagreement is very small. It is worth noting that the disagreement between these two classical exact tests and their respective approximated version is often higher than the disagreement between the classical tests (exact or approximated) and their RP-based version.

In general, the disagreement between the several classes of tests decreases when the sample size increases.

Concerning the variability of RP estimators, there is not an overall best performer for the WSR and Kendall tests. Nevertheless, the estimators showing good estimation performances also present slow or null disagreement and, mainly, belong to the semi-parametric family. Nonparametric estimators present a slightly higher variability and disagreement with respect to the best semi-parametric ones, but have the advantage of being general, since they can be adopted even when the power functional has not been studied and parametrized.

To conclude, many useful and actually applicable solutions to estimate the RP and to perform RP-testing for the most commonly-used nonparametric tests, exact or approximated, are provided. The RP-testing rule is shown to be easily extended to these nonparametric tests. Further development in RP estimation might concern the application of Bayesian techniques in the nonparametric context, since in the parametric one, they showed promising improvement when uninformative priors have been adopted [27]. Furthermore, prediction intervals may be considered for nonparametric RP estimation (see [28]); in particular, it would be interesting to link prediction intervals, which provide likely results of future RP estimators, once experimental data have been observed, to the RP-testing rule.

Supplementary Materials: The following are available online at www.mdpi.com/1099-4300/18/4/142/s1. Figure S1. Bias (left) and MSE (right) of the RP estimators $\hat{\pi}_e$ (solid), $\hat{\pi}_a$ (dashed), $\hat{\pi}_e^{PI}$ (dotted), and $\hat{\pi}_{a,e}^{PI}$ (dot-dashed). The MSE and the Bias are computed setting $\alpha = 0.05$ and considering the testing problem (3.1) by setting $\alpha = 0.05$, $p_0 = 0.2$ and $n = (5, 10, 15)$. (a) $p_0 = 0.2$, $n = 5$; (b) $p_0 = 0.2$, $n = 5$; (c) $p_0 = 0.2$, $n = 15$; (d) $p_0 = 0.2$, $n = 15$; (e) $p_0 = 0.2$, $n = 30$; (f) $p_0 = 0.2$, $n = 30$; Figure S2. Bias (left) and MSE (right) of the RP estimators $\hat{\pi}_e$ (solid), $\hat{\pi}_a$ (dashed), $\hat{\pi}_e^{PI}$ (dotted), and $\hat{\pi}_{a,e}^{PI}$ (dot-dashed). The MSE and the Bias are computed setting $\alpha = 0.05$ and considering the testing problem (3.1) by setting $\alpha = 0.05$, $p_0 = 0.5$ and $n = (5, 10, 15)$. (a) $p_0 = 0.5$, $n = 5$; (b) $p_0 = 0.5$, $n = 5$; (c) $p_0 = 0.5$, $n = 15$; (d) $p_0 = 0.5$, $n = 15$; (e) $p_0 = 0.5$, $n = 30$; (f) $p_0 = 0.5$, $n = 30$; Table S1. Probability of disagreement $D(p_t, n, \alpha, p_0)$ between the tests $\Psi_\alpha(\mathbf{X}_n)$ and $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ evaluated for $\alpha = 0.05$, $n = (5, 15, 30)$, $p_0 = (0.2, 0.5)$ and $p_t = (0.1, 0.2, ..., 0.9)$.; Table S2. Probability of disagreement $D(\alpha, n, F_{\theta_t}, \theta_0)$ between the tests $\Psi_\alpha(\mathbf{X}_n)$ and $\widetilde{\Psi}_\alpha(\mathbf{X}_n)$ evaluated for $\alpha = 0.05$, $n = (15, 30, 60, 120, 240)$, $\theta_0 = 0$, and assuming that $F_{\theta_t} \equiv \mathcal{N}(\theta_t, 1)$ or $F_{\theta_t} \equiv \text{Cauchy}(\theta_t)$ with $\theta_t = (0.0, 0.1, 0.2, ..., 0.9)$; Table S3. Averaged MSE, Bias and Disagreement rate for the asymptotic and exact WSR test when sampling from the Normal distribution. The averages are computed over the 19 different values of θ considered in the simulation study. The smallest values for the averaged MSE, Bias and disagreement are highlighted in bold; Table S4. Averaged MSE, Bias and Disagreement rate for the asymptotic and exact WSR test when sampling from the Chaucy distribution. The averages are computed over the 19 different values of θ considered in the simulation study. The smallest values for the averaged MSE, Bias and disagreement are highlighted in bold; Table S5. Probability of disagreement between $\widetilde{\Psi}_\alpha((\mathbf{X}, \mathbf{Y})_n)$ and $\Psi_\alpha((\mathbf{X}, \mathbf{Y})_n)$ when $\alpha = 0.05$, $n = (15, 30, 60, 120)$, and assuming that $\mathbf{F}_t$ is the Bivariate Normal or Bivariate Student's t (3 df) with correlation coefficient $\rho = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$. The values of τ_t corresponding to the considered values of ρ are reported in the second column of the table. They have been obtained by using the relation $\tau_t = \frac{2}{\pi} \arcsin(\rho)$ which holds for all absolutely continuous elliptical distributions; Table S6. Averaged MSE, Bias and Disagreement rate for the asymptotic and exact Kendall's test when sampling from the Gaussian copula. The averages are computed over the 19 different values of ρ considered in the simulation study. The least values for the averaged MSE, Bias and disagreement are highlighted in bold; Table S7. Averaged MSE, Bias and Disagreement rate for the asymptotic and exact Kendall's test when sampling from the t copula with 3 degrees of freedom. The averages are computed over the 19 different values of ρ considered in the simulation study. The least values for the averaged MSE, Bias and disagreement are highlighted in bold.

Acknowledgments: We thank the Editors, J. Stern and A. Polpo, for the good timing of this Special Issue, and also thank two anonymous reviewers for their constructive comments which helped us to improve the manuscript.

Author Contributions: Lucio De Capitani and Daniele De Martini wrote together Sections: 2, 4.2, 4.3, 5.2, 5.3 and 7, and also conceived and designed numerical experiments. Lucio De Capitani performed numerical computations and wrote Sections: 3, 4.1, 5.1, 6 and supplementary material. Daniele De Martini wrote Section 1. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodman, S.N. A comment on replication, *p*-values and evidence. *Stat. Med.* **1992**, *11*, 875–879.
2. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*; Marcel Dekker: New York, NY, USA, 2003.
3. De Martini, D. Stability Criteria for the Outcomes of Statistical Tests to Assess Drug Effectiveness with a Single Study. *Pharm. Stat.* **2012**, *11*, 273–279.
4. De Martini, D. *Success Probability Estimation with Applications to Clinical Trials*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
5. Hsieh, T.C.; Chow, S.C.; Yang, L.Y.; Chi, E. The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Stat. Med.* **2013**, *32*, 406–414.
6. Shao, J.; Chow, S.C. Reproducibility Probability in Clinical Trials. *Stat. Med.* **2002**, *21*, 1727–1742.
7. De Capitani, L. An introduction to RP-testing. *Epidemiol. Biostat. Public Health* **2013**, *10*, doi:10.2427/8756.
8. De Martini, D. Reproducibility Probability Estimation for Testing Statistical Hypotheses. *Stat. Probab. Lett.* **2008**, *78*, 1056–1061.
9. Berger, J. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Stat. Sci.* **2003**, *18*, 1–12.
10. Berger, J.; Sellke, T. Testing a point null hypothesis: The irreconcilability of *P*-values and evidence. *J. Am. Stat. Assoc.* **1987**, *82*, 112–122.
11. Hubbard, R.; Bayarri, M.J. Confusion over measures of evidence (*p*s) *versus* errors (αs) in classical statistical testing. *Am. Stat.* **2003**, *57*, 171–178.
12. Hubbard, R.; Lindsay, R.M. Why *p*-values are not a useful measure of evidence in statistical significance testing. *Theory Psychol.* **2008**, *18*, 69–88.
13. Schervish, M.J. *p*-values: What they are and what they are not. *Am. Stat.* **1996**, *50*, 203–206.
14. De Capitani, L.; De Martini, D. On stochastic orderings of the Wilcoxon Rank Sum test statistic—with applications to reproducibility probability estimation testing. *Stat. Probab. Lett.* **2011**, *81*, 937–946.
15. De Capitani, L.; De Martini, D. Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *J. Stat. Comput. Simul.* **2015**, *85*, 468–493.
16. Van de Wiel, M.A.; Di Bucchianico, A.; van der Laan, A. Exact distributions of nonparametric test statistics and computer algebra. *J. R. Stat. Soc. Series D* **1999**, *48*, 507–551.
17. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993.
18. De Martini, D. Conservative Sample Size Estimation in Nonparametrics. *J. Biopharm. Stat.* **2011**, *21*, 24–41.
19. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Chapman & Hall: New York, NY, USA, 1998.
20. Noether, G.E. Sample size determination for some common non-parametric tests. *J. Am. Stat. Assoc.* **1987**, *82*, 645–647.
21. Kaarsemaker, L.; van Wijngaarden, A. Tables for use in rank correlation. *Stat. Neerlandica* **1953**, *7*, 41–54.
22. Gibbons, J.D.; Chakraborti, S. *Nonparametric Statistical Inference*; Dekker: New York, NY, USA, 2003.
23. The R Project for Statistical Computing. Available online: http://www.R-project.org/ (accessed on 8 April 2016).
24. Wheeler, B. SuppDists: Supplementary Distributions. Available online: http://CRAN.R-project.org/package=SuppDists (accessed on 8 April 2016).
25. Best, D.J.; Gipps, P.G. Algorithm AS 71: The Upper Tail Probabilities of Kendall's Tau. *J. R. Stat. Soc.* **1974**, *23*, 98–100.
26. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*, 2nd ed.; Wiley: Weinheim, Germany, 1999.
27. De Capitani, L.; De Martini, D. Improving Reproducibility Probability estimation, preserving RP-testing. Available online: https://boa.unimib.it/handle/10281/105595 (accessed on 14 April 2016).
28. Coolen, F.P.; Bin Himd, S. Nonparametric predictive inference for reproducibility of basic nonparametric tests. *J. Stat. Theory Pract.* **2014**, *8*, 591–618.