On stochastic orderings of the Wilcoxon Rank Sum test statistic–with applications to reproducibility probability estimation testing

L. de Capitani, D. de Martini

▶ **To cite this version:**

L. de Capitani, D. de Martini. On stochastic orderings of the Wilcoxon Rank Sum test statistic–with applications to reproducibility probability estimation testing. Statistics and Probability Letters, 2011, 81 (8), pp.937. 10.1016/j.spl.2011.04.001 . hal-00753939

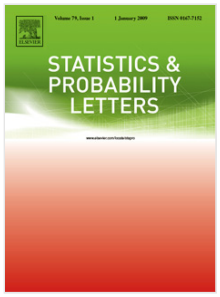
HAL Id: hal-00753939
https://hal.science/hal-00753939

Submitted on 20 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



On stochastic orderings of the Wilcoxon Rank Sum test statistic—with applications to reproducibility probability estimation testing

L. De Capitani, D. De Martini

PII: S0167-7152(11)00132-5
DOI: 10.1016/j.spl.2011.04.001
Reference: STAPRO 5976

To appear in: *Statistics and Probability Letters*

Received date: 23 February 2010
Revised date: 31 March 2011
Accepted date: 1 April 2011

Please cite this article as: De Capitani, L., De Martini, D., On stochastic orderings of the Wilcoxon Rank Sum test statistic—with applications to reproducibility probability estimation testing. *Statistics and Probability Letters* (2011), doi:10.1016/j.spl.2011.04.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

*Manuscript
Click here to view linked References
ACCEPTED MANUSCRIPT

On stochastic orderings of the Wilcoxon Rank Sum test statistic - with applications to reproducibility probability estimation testing

De Capitani L.[a,*], De Martini D.[a]

[a]*Department of Quantitative Methods for Economics and Business, University of Milan-Bicocca, via Bicocca degli Arcimboldi 8, 20126, Italy.*

Abstract

Recently, the possibility of testing statistical hypotheses through the estimate of the reproducibility probability (i.e. the estimate of the power of the statistical test) in a general parametric framework has been introduced. In this paper, we provide some results on the stochastic orderings of the Wilcoxon Rank Sum (WRS) statistic, implying, for example, that the related test is strictly unbiased. Moreover, under some regularity conditions, we show that it is possible to define a continuous and strictly monotone power function of the WRS test. This last result is useful in order to obtain a point estimator and lower bounds for the power of the WRS test. In analogy with the parametric setting, we show that these power estimators, alias reproducibility probability estimators, can be used as test statistic, i.e. it is possible to refer directly to the estimate of the reproducibility probability to perform the WRS test. Some reproducibility probability estimators based on asymptotic approximations of the power are provided. A brief simulation shows a very high agreement between the approximated reproducibility probability based tests and the classical one.

Keywords: Wilcoxon Rank Sum Test, Reproducibility Probability Estimation, Reproducibility Probability Estimation testing

1. Introduction

In the context of clinical trials, the power of a test is sometimes referred to as Reproducibility Probability (RP). This terminology was introduced by Goodman (1992) and it is due to the fact that the power is estimated only after a statistical test has been performed, in order to evaluate the reproducibility of the test result. Roughly speaking, once a statistical test is computed referring to data from a particular experiment, the RP is the probability of obtaining the same test result in a second, identical experiment. In detail, if we accept H_0 in the first experiment, 1-RP is the probability to accept H_0 even in the second experiment. Otherwise, if we reject H_0 in the first experiment, the probability of a further rejection is the RP itself. Then, the RP is an indicator of the stability of the test result and its estimate can be used to measure the reproducibility of the outcome of an experiment. The evaluation of the reproducibility of the outcome of an experiment is a pillar of the experimental method (alias Galilean method). For this reason the RP-Estimation (RPE) is a very important tool whereas the experimental method is applied. For example, the evaluation of the reproducibility is very important in the context of clinical trial. For a discussion on this topic see Goodman (1992) and Shao and Chow (2002).

Recently, De Martini (2008) showed that RPE can also be used for testing parametric statistical hypotheses. In particular, it is shown that the point estimate of the RP is greater than $1/2$ if and only if the null hypotheses is rejected. This leads to the simple and intuitive decision rule *"accept H_0 if the estimate of the RP is lower or equal to* $1/2$ *and reject H_0 otherwise"* which is equivalent to the commonly used rules based on the p-value and on the direct evaluation of the test statistic. In general, the rule based on the p-value is preferred to the one based on test statistic. This is due to the fact that the p-value can be used not only to reject/accept H_0 but it can also be viewed as a measure of the degree to which the data support or contradict H_0 (i.e. p-values measures the evidence against or in favor to H_0).

*Corresponding author
Email addresses: l.decapitani@campus.unimib.it (De Capitani L.), daniele.demartini@unimib.it (De Martini D.)

ACCEPTED MANUSCRIPT

The evaluation of the evidence is important since it allows the researcher to understand how much stable the decision taken is, shedding light on the reproducibility of the test result. However, there are some arguments suggesting that the rule based on the RP is better than those based on the p-value. The first argument stems from the work of Goodman (1992) who showed that a small p-value can overstate the evidence against the null hypotheses and it can thus lead to an underestimation of the variability of the test result (i.e. to an overestimation of the reproducibility). Indeed, unlike the RP, the p-value is not a direct measure of reproducibility. As a further argument, it should be noted that the p-value is a very misinterpreted measure. In particular, as pointed out by several authors (see, e.g., Goodman (1992), Berger and Sellke (1987) and Hubbard and Bayarri (2003)) the p-value is often confused with the Type I error rate and erroneously interpreted as the "observed α". On the contrary, in our opinion, the use of the RP avoid all these misunderstandings since its interpretation is easier and more direct.

The technical result obtained in De Martini (2008) holds for a large class of parametric tests (e.g. those whose test statistic has Gaussian, or χ^2, or t, or F distribution) and it is based on the key assumption that the test statistics is stochastically strictly ordered by the parameter under testing. Consequently, in the case of one-sided hypotheses, the power function of the test is strictly monotone in the parameter under testing and the test is strictly unbiased. These are the necessary properties that assure the possibility of making "RP-testing", i.e. they assure that the RPE can be used for testing statistical hypotheses.

In this paper we focus on the Wilcoxon Rank Sum (WRS) test in order to provide the regularity conditions that make RP-testing possible. The problem arising in the nonparametric context of the WRS test consists in the absence of the parameter under testing which, in this case, is a distribution. As a consequence, the power is not a function but a functional, and the extension of the results given in De Martini (2008) is not immediate. Then, we show that the WRS statistics is stochastically strictly ordered for the stochastically strictly ordered distributions, and we use this result to show that the WRS test is strictly unbiased. Moreover, we find the regularity conditions under which the WRS test has a continuous and strictly monotone power function. In doing that, we analyze first the case of the well known location shift model and later the more general context of the strictly stochastically ordered alternative hypothesis. Finally we apply these results to RP-testing for the WRS test. We remark that the above results concerning the WRS statistic and the WRS test are of general theoretical interest even if they were mainly derived in order to study the RP-testing.

The paper is organized as follows. Section 2 recalls the WRS test and the asymptotic distribution of its test statistic. In section 3 some general qualitative features of the WRS test statistic are stated and proved. These qualitative features concerns the stochastic orderings of WRS statistics, the power and the unbiasedness of the WRS test. In Section 4 the RP-testing result given in De Martini (2008) is extended to the nonparametric context of the WRS test. In section 5 an application of the results of section 4 and a brief simulation are provided.

2. Recalling WRS test

Let X and Y be absolutely continuous random variables with distribution functions F and G, respectively. Further, let $\mathbf{X}_m = (X_1, ..., X_m)$ and $\mathbf{Y}_n = (Y_1, ..., Y_n)$ be independent random samples from F and G, respectively. We want to test the null hypothesis that X and Y are equal in distribution against the alternative that Y is stochastically strictly greater than X:

$$H_0 : \ Y \stackrel{d}{=} X \qquad vs \qquad H_1 : \ Y >_{st} X \tag{1}$$

To test these hypotheses it is possible to use the WRS test, which is based on the following statistic (written in the Mann-Whtiney form):

$$W_{XY} = \#\left\{(X_i, Y_j) : X_i < Y_j \ ; i = 1, ..., m \ \ j = 1, ..., n\right\} \ . \tag{2}$$

For a given level of the type I error (α), the WRS test is

$$\Phi_\alpha(W_{XY}) = \begin{cases} 1 & \text{if} \quad W_{XY} > w_{1-\alpha} \\ 0 & \text{if} \quad W_{XY} \leq w_{1-\alpha} \end{cases} \tag{3}$$

where w_q is the q-quantile of the null distribution of W_{XY} (hereafter denoted by $\mathcal{W}_0$). The null distribution $\mathcal{W}_0$ depends only on the sample sizes m and n and it has been widely studied. In particular, $\mathcal{W}_0$ can be exactly computed using the recurrence relations known in literature (see Di Bucchianico (1999) and references therein) and implemented,

ACCEPTED MANUSCRIPT

for example, in **R**. Approximated values of $\mathcal{W}_0$ can be obtained through the well known normal approximation

$$\frac{W_{XY} - mn/2}{(mn(m+n+1)/12)^{1/2}} \overset{a}{\sim} \mathcal{N}(0,1) \ . \tag{4}$$

Further, in Bean *et al.* (2004) it is shown that the previous normal approximation can be improved by an Edgeworth approximation or a saddlepoint one.

Unlike $\mathcal{W}_0$, the distribution of W_{XY} under the alternative hypothesis, denoted by $\mathcal{W}_{FG}$, depends not only on the sample sizes m and n, but also on F and G. For this reason, it is more difficult to study $\mathcal{W}_{FG}$ than $\mathcal{W}_0$. In detail, due to analytical difficulties, the exact expression of $\mathcal{W}_{FG}$ has been derived just in some particular cases. For example, Haynam and Govindarajulu (1966) derived the exact expression of $\mathcal{W}_{FG}$ for the exponential and uniform shift alternatives. A similar result is given in Lehmann (1953) for the so called Lehmann's alternatives. For some other references on this subject see Lehmann (1998), pp. 98-99. However it is well known that W_{XY} is asymptotically normally distributed also under H_1 (see Lehmann (1951)). In particular, when m and n diverge, and $0 < p_1 = P(X < Y) < 1$, we have that

$$\frac{W_{XY} - mnp_1}{\sqrt{V(p_1, p_2, p_3)}} \overset{a}{\sim} \mathcal{N}(0,1) \tag{5}$$

where $V(p_1, p_2, p_3) = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$, with $p_2 = P(X < Y \wedge X < Y')$ (Y' and Y i.i.d.), and $p_3 = P(X < Y \wedge X' < Y)$ (X' and X i.i.d.). This last approximation can be improved by the Edgeworth approximation given in Witting (1960) and it is very useful in order to study the power of the WRS test. In particular, let $\mathbf{n}$ denotes the couple of sample sizes (m, n) and let $\pi_{\mathbf{n},\alpha}(G, F) = P(W_{XY} > w_{1-\alpha}) = 1 - \mathcal{W}_{FG}(w_{1-\alpha})$ denotes the power of the test (3). From approximation (5) it follows that (see Lehmann (1998), p. 71):

$$\pi_{\mathbf{n},\alpha}(F, G) \cong \Phi\left[\frac{mn\left(p_1 - \frac{1}{2}\right) - z_{1-\alpha}\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V(p_1, p_2, p_3)}}\right] \ . \tag{6}$$

In the following, we will refer to the power $\pi_{\mathbf{n},\alpha}(F, G)$ as to the RP of the test (3) (see De Martini , 2008, Definition 1).

3. Main results

First, we introduce some notations. Let $\mathcal{F}$ be the set of all absolutely continuous distribution functions. $\mathcal{S}_L$ is the support of a distribution function $L \in \mathcal{F}$. Let K and T be two random variables with distribution functions $L \in \mathcal{F}$ and $M \in \mathcal{F}$. W_{KT} denotes the Mann-Whitney statistic defined, according to (2), on the basis of two independent samples $\mathbf{K}_m$ and $\mathbf{T}_n$ from L and M, respectively. We denote with the lower case w_{kt} the realization of W_{KT} corresponding to the realizations $\mathbf{k}_m$ and $\mathbf{t}_n$ of $\mathbf{K}_m$ and $\mathbf{T}_n$, respectively. $\mathcal{W}_{LM}$ is the distribution function of W_{KT}. $\mathcal{G}_L$ denotes the set of the distributions of all the absolutely continuous random variables strictly dominating K in the sense of the usual stochastic order, that is: $\mathcal{G}_L = \{M \in \mathcal{F} \ : \ M(x) \leq L(x) \ \ \forall\, x \in \mathbb{R} \ \ \wedge \ \ \exists\, x : M(x) < L(x)\}$. $\tilde{\mathcal{G}}_L$ denotes the set of the distributions of all the absolutely continuous random variables strictly dominated by K in the sense of the usual stochastic order, that is: $\tilde{\mathcal{G}}_L = \{M \in \mathcal{F} \ : \ M(x) \geq L(x) \ \ \forall\, x \in \mathbb{R} \ \ \wedge \ \ \exists\, x : M(x) > L(x)\}$. Finally, hereafter we rewrite hypotheses (1) as follows:

$$H_0 : \ G = F \qquad vs \qquad H_1 : \ G \in \mathcal{G}_F \ . \tag{7}$$

We can now state the following theorem.

Theorem 1. *Let X,Y and Z be absolutely continuous random variables with distribution functions respectively given by F, $G \in \{\mathcal{G}_F \cup F\}$ and $H \in \mathcal{G}_G$. Suppose that Y and Z are independent on X and let $\mathbf{X}_m$, $\mathbf{Y}_n$, and $\mathbf{Z}_n$ be random samples from F, G, and H, respectively. Let $W_{XY} \sim \mathcal{W}_{FG}$ be the WRS statistic defined, according to (2), on the basis of $\mathbf{X}_m$ and $\mathbf{Y}_n$. Similarly, $W_{XZ} \sim \mathcal{W}_{FH}$ is the WRS statistic defined on the basis of $\mathbf{X}_m$ and $\mathbf{Z}_n$.*
If the following condition holds

$$\exists\, x \in \mathcal{S}_F \ : \ G(x) > H(x) \ , \tag{8}$$

then $\mathcal{W}_{FG}(w) > \mathcal{W}_{FH}(w)$ for all $w \in [0, mn)$.

ACCEPTED MANUSCRIPT

P . Let $\mathbf{Z}_n^*$ be the random vector obtained from $\mathbf{Y}_n$ through the transformation

$$Z_j^* = H^{-1}\left(G(Y_j)\right) \qquad j = 1, ..., n \tag{9}$$

where H^{-1} indicates the generalized inverse of H. The components of $\mathbf{Z}_n^*$ are i.i.d. with distribution H. Further, $\mathbf{Z}_n^*$ is independent on $\mathbf{X}_m$ thanks to the independence of X and Y. Then the theorem can be proved by referring to the distribution of $W_{XZ^*} = \#\left\{(X_i, Z_j^*) : X_i < Z_j^* \; ; i = 1, ..., m \;\; j = 1, ..., n\right\}$ which is equal to the distribution of W_{XZ}, i.e. $\mathcal{W}_{FH}$. Now, consider the realizations $\mathbf{x}_m$ and $\mathbf{y}_n$ and the realization $\mathbf{z}_n^*$ obtained from $\mathbf{y}_n$ through the transformation (9). Further, Let w_{xy} and w_{xz^*} be the values of W_{XY} and W_{XZ^*} corresponding to $\mathbf{x}_m$,$\mathbf{y}_n$ and $\mathbf{z}_n^*$. By construction, $z_j^* \geq y_j$ for every $j = 1, ...n$ and $w_{xy} \leq w_{xz^*}$ for all the $\mathbf{x}_m$ and $\mathbf{y}_n$ in the sample space. As a consequence $P[W_{XY} \leq W_{XZ^*}] = 1$ and

$$P[W_{XY} \leq w] - P[W_{XZ^*} \leq w] = \sum_{i=0}^{w} \sum_{j=w+1}^{mn} P[W_{XY} = i \cap W_{XZ^*} = j] \geq P[W_{XY} = 0 \cap W_{XZ^*} = mn] \qquad \forall w \in [0, 1) \;\; . \tag{10}$$

From expression (10) it follows that the statement of Theorem 1 can be proved by showing that $P[W_{XY} = 0 \cap W_{XZ^*} = mn] > 0$. To do this, let us first note that $w_{xy} < w_{xz^*}$ if and only if $y_j < x_i < z_j^*$ for at least one couple (i, j). The previous chain of inequalities holds true if $y_j < x_i$ and $x_i < z_j^*$. The inequality $x_i < z_j^*$ coincide with $x_i < H^{-1}(G(y_j))$ which is equivalent to $y_j > G^{-1}(H(x_i))$. To see why, we recall that if Y is absolutely continuous then the generalized inverse G^{-1} is strictly increasing and satisfies the following relations: $G(G^{-1}(q)) = q$ for all $q \in (0, 1)$; $G^{-1}(G(y)) \leq y$ for $y \in \mathbb{R}$ with equality if $y \in \mathcal{S}_G$; $G(y) < q$ iff $y < G^{-1}(q)$, $q \in (0, 1)$ (similar relations hold for H^{-1}). Then, being $y_j \in \mathcal{S}_G$, we have $x_i < H^{-1}(G(y_j)) \Leftrightarrow H(x_i) < G(y_j) \Leftrightarrow y_j > G^{-1}(H(x_i))$. Putting together the inequalities $y_j > G^{-1}(H(x_i))$ and $y_j < x_i$ we obtain that $w_{xy} < w_{xz^*}$ if and only if $y_j \in (G^{-1}(H(x_i)), x_i)$ for at least one couple (i, j). Now, let us introduce the set $E = \{x \in \mathcal{S}_F : G(x) > H(x)\}$. Thanks to condition (8), the set E is nonempty. Moreover, the continuity of G and H (assured by the absolute continuity of Y and Z) ensures that the set E contains at least one interval. Without loss of generality, let us assume that E is an interval and let $x^* \in E$. Then x^* is an interior point of $\mathcal{S}_F$ and $(G^{-1}(H(x^*)), x^*) \cap \mathcal{S}_F = (\tilde{x}, x^*)$ where $\tilde{x} \geq G^{-1}(H(x^*))$. Analogously, the set $(G^{-1}(H(x^*)), x^*) \cap \mathcal{S}_G$ contains at least an interval, say (y_l, y_u), since

$$P[Y \in (G^{-1}(H(x^*)), x^*)] = G(x^*) - G(G^{-1}(H(x^*))) = G(x^*) - H(x^*) > 0 \;\; . \tag{11}$$

Now, assume that $(y_l, y_u) \cap (\tilde{x}, x^*) = \emptyset$. In that case $y_u < \tilde{x}$. Furthermore, if $Y_j \in (y_l, y_u)$ for all $j = 1, ..., n$ then $Z_j^* > x^*$ for all $j = 1, ..., n$. From expression (11) it follows that $P[Y_j \in (y_l, y_u) \;\; \forall \;\; j = 1, ..., n] = P[Y \in (y_l, y_u)]^n > 0$. Analogously, $P[X_i \in (\tilde{x}, x^*) \;\; \forall \;\; i = 1, ..., m] = P[X \in (\tilde{x}, x^*)]^n > 0$. Consequently

$$P[Y_j \in (y_l, y_u) \cap X_i \in (\tilde{x}, x^*) \; \forall \; i, j] = P[X_i \in (\tilde{x}, x^*) \; \forall \; i] P[Y_j \in (y_l, y_u) \; \forall \; j] > 0 \;\; .$$

Moreover, if $Y_j \in (y_l, y_u) \cap X_i \in (\tilde{x}, x^*) \; \forall \; i, j$ then $W_{XY} = 0$ but $W_{XZ^*} = mn$. If the intervals (y_l, y_u) and $(\tilde{x}, x^*)$ intersect and $y_l < \tilde{x}$ the last passages can be made by referring to the intervals $(y_l, \tilde{x})$ and $(\tilde{x}, x^*)$. Finally, if the intervals (y_l, y_u) and $(\tilde{x}, x^*)$ intersect and $y_l > \tilde{x}$ the last passages can be made by referring to the intervals $\left(y_l, \frac{y_l + x^*}{2}\right)$ and $\left(\frac{y_l + x^*}{2}, x^*\right)$. Then $P[W_{XY} = 0 \cap W_{XZ^*} = mn] > 0$ and this completes the proof.

Remark 1. Roughly speaking, Theorem 1 shows that, being F set, it is sufficient that G is greater than H over an interval contained in $\mathcal{S}_F$ to assure that $\mathcal{W}_{FG}$ is always greater than $\mathcal{W}_{FH}$.

Remark 2. If the distributions of the random variables X, Y and Z have common support, condition (8) is superfluous thanks to the definition of $\mathcal{G}_F$ and $\mathcal{G}_G$. In the other cases, condition (8) is necessary in order to guarantee the validity of Theorem 1. This is obvious if $\sup\{\mathcal{S}_F\} \leq \inf\{\mathcal{S}_G\}$. In fact, in that case we have that $\mathcal{W}_{FG}(w) = \mathcal{W}_{FH}(w) = 0$ for all $x \in [0, mn)$. Furthermore, if $G(x) = H(x)$ for every $x \in \mathcal{S}_F$ then $\mathcal{W}_{FG}(w) = \mathcal{W}_{FH}(w)$ for all $w \in [0, mn]$.

Remark 3. If the sample size for Z (say n_Z) is lower than the sample size for Y (say $n_Y > n_Z$), Theorem 1 does not hold because $\mathcal{W}_{FH}(mn_Z) = 1$ while $\mathcal{W}_{FG}(mn_Z) < 1$. On the contrary, if $n_Z > n_Y$ then Theorem 1 still holds since, keeping fixed F, H and m, the random variable W_{XZ} is stochastically increasing in n_Z.

ACCEPTED MANUSCRIPT

We shall now show a first application of Theorem 1. In Lehmann (1951) it is shown that the WRS test is unbiased against the alternatives (7). Namely: $\pi_{\mathbf{n},\alpha}(F,H) \geq \alpha$ for all $H \in \mathcal{G}_F$. Thanks to Theorem 1 it is now possible to refine this latter result.

Corollary 1. *The test (3) for testing hypothesis (7) is strictly unbiased : $\pi_{\mathbf{n},\alpha}(F,H) > \alpha$ for all $H \in \mathcal{G}_F$.*

P . The corollary follows directly from the Theorem 1 specialized to the case $F = G$. In more detail, if $F = G$, Theorem 1 states that $\mathcal{W}_0(w) > \mathcal{W}_{FH}(w)$ for all $w \in [0, mn)$. Consequently $1 - \mathcal{W}_0(w_{1-\alpha}) < 1 - \mathcal{W}_{FH}(w_{1-\alpha})$. The corollary now follows observing that $\alpha \leq 1 - \mathcal{W}_0(w_{1-\alpha})$ and remembering the definition of $\pi_{\mathbf{n},\alpha}(F,H)$.

For completeness, we state the following theorem concerning, in a sense, the opposite situation of Theorem 1.

Theorem 2. *Let X,Y and Z be absolutely continuous random variables with distribution functions respectively given by F, $G \in \{\tilde{\mathcal{G}}_F \cup F\}$ and $H \in \tilde{\mathcal{G}}_G$. Suppose that Y and Z are independent on X and let $\mathbf{X}_m$, $\mathbf{Y}_n$, and $\mathbf{Z}_n$ be random samples from F, G, and H, respectively. Let $W_{XY} \sim \mathcal{W}_{FG}$ be the WRS statistic defined on the basis of $\mathbf{X}_m$ and $\mathbf{Y}_n$. Similarly, $W_{XZ} \sim \mathcal{W}_{FH}$ is the WRS statistic defined on the basis of $\mathbf{X}_m$ and $\mathbf{Z}_n$.*
If the following condition holds

$$\exists\, x \in \mathcal{S}_F \ : \ G(x) < H(x) \ , \tag{12}$$

then $\mathcal{W}_{FG}(w) < \mathcal{W}_{FH}(w)$ for all $w \in [0, mn)$.

The proof of Theorem 2 is omitted because it is quite similar to the proof of Theorem 1.

Theorems 1 and 2 can be used as starting points for a more detailed study on the features of $\mathcal{W}_{FG}$ when G belongs to some particular subset of $\mathcal{G}_F$. For example, consider the so called *location shift model* (*lsm*), where it is assumed that $G \in \mathcal{G}_F^{\Delta} = \{G \in \mathcal{G}_F \ : \ G(x) = F(x - \Delta) \ \forall\, x \in \mathbb{R}; \ \ \Delta > 0\}$. The testing problem then becomes:

$$H_0 : \ G = F \qquad vs \qquad H_1 : \ G \in \mathcal{G}_F^{\Delta} \ . \tag{13}$$

The amount of shift Δ can be used as a parameter for the distribution $\mathcal{W}_{FG}$. Namely, keeping fixed the distribution F, $\mathcal{W}_{FG}$ can be seen as a function of w and Δ. For this reason, in the context of the *lsm*, we write $\mathcal{W}_F(w;\, \Delta)$ rather than $\mathcal{W}_{FG}(w)$. The following Corollary describes the features of $\mathcal{W}_F(w;\, \Delta)$ as a function of Δ.

Corollary 2. *For a given $w \in [0, mn)$ and $F \in \mathcal{F}$, the function $\mathcal{W}_F(w;\Delta)$ is continuous and strictly decreasing in Δ over the range $-r < \Delta < r$, where $r = \sup\{\mathcal{S}_F\} - \inf\{\mathcal{S}_F\}$ if $\mathcal{S}_F$ is bounded and $r = \infty$ otherwise.*

P . Thanks to Theorems 1 and 2, keeping fixed F, the function $\mathcal{W}_F(w;\Delta)$ is strictly decreasing in Δ over the range $-r < \Delta < r$ for every $w \in [0, mn)$. To prove the continuity, note first that the monotonicity of $\mathcal{W}_F(w;\Delta)$ in Δ implies that the following limits exist for every $w \in [0, mn)$:

$$\lim_{\delta \to \Delta^-} \mathcal{W}_F(w;\delta) = a_w \ , \qquad \lim_{\delta \to \Delta^+} \mathcal{W}_F(w;\delta) = b_w \ .$$

For a given, arbitrary small $\epsilon > 0$, let $\{\Delta_h^-\}_{h\in\mathbb{N}}$ be a monotone increasing succession of values in $(\Delta - \epsilon, \Delta) \subset (-r, r)$. Further, let $\{\Delta_h^+\}_{h\in\mathbb{N}}$ be a monotone decreasing succession of values in $(\Delta, \Delta + \epsilon) \subset (-r, r)$. Suppose that $\lim_{h\to\infty} \Delta_h^- = \Delta = \lim_{h\to\infty} \Delta_h^+$. We prove the continuity of $\mathcal{W}_F(w;\Delta)$ by showing that $\lim_{h\to\infty} \mathcal{W}_F(w;\Delta_h^-) = \lim_{h\to\infty} \mathcal{W}_F(w;\Delta_h^+)$. For that purpose, the following notation is introduced:

$$G_{h-}(x) = F(x - \Delta_h^-) \ , \quad G_{h+}(x) = F(x - \Delta_h^+) \ .$$

Consider the random samples $\mathbf{X}_m$ and $\mathbf{Y}_n$ drawn from F and $G(x) = F(x - \Delta)$, respectively. Further, let $\mathbf{Y}_{n,h}^-$ and $\mathbf{Y}_{n,h}^+$ be the random samples obtained from $\mathbf{Y}_n$ through the transformations

$$Y_{j,h}^- = G_{h-}^{-1}(F(Y_j)) \qquad \text{and} \qquad Y_{j,h}^+ = G_{h+}^{-1}(F(Y_j)) \ , \qquad j = 1, ..., n \ . \tag{14}$$

By definition, the distribution of

$$W_{X,Y_h^-} = \#\left\{(X_i, Y_{j,h}^-) : X_i < Y_{j,h}^- \ ; i = 1, ..., m \ \ j = 1, ..., n\right\}$$

ACCEPTED MANUSCRIPT

is $\mathcal{W}_F(\cdot;\Delta_h^-)$ for all $h \in \mathbb{N}$. Analogously, the distribution of

$$W_{X,Y_h^+} = \#\left\{(X_i, Y_{j,h}^+) : X_i < Y_{j,h}^+ \; ; i = 1, ..., m \;\; j = 1, ..., n\right\}$$

is $\mathcal{W}_F(\cdot;\Delta_h^+)$ for all $h \in \mathbb{N}$. Now, consider the realizations $\mathbf{x}_m$ and $\mathbf{y}_n$ and the sequences of realizations $\left\{\mathbf{y}_{n,h}^-\right\}_{h\in\mathbb{N}}$ and $\left\{\mathbf{y}_{n,h}^+\right\}_{h\in\mathbb{N}}$ obtained from $\mathbf{y}_n$ through the transformations (14). Further, let w_{xy}, $\left\{w_{xy_h^+}\right\}_{h\in\mathbb{N}}$ and $\left\{w_{xy_h^-}\right\}_{h\in\mathbb{N}}$ the corresponding realizations of the WRS statistics. Finally, let us introduce the following subsets of the sample space:

$$\mathcal{E}_{XY}(w) = \left\{(\mathbf{x}_m, \mathbf{y}_n) \in \mathcal{S}_F^m \times \mathcal{S}_G^n \; : \; w_{xy} > w\right\} \;\; ,$$

$$\mathcal{E}_{XY_h^-}(w) = \left\{(\mathbf{x}_m, \mathbf{y}_n) \in \mathcal{S}_F^m \times \mathcal{S}_G^n \; : \; w_{xy_h^-} > w\right\} \;\; ,$$

$$\mathcal{E}_{XY_h^+}(w) = \left\{(\mathbf{x}_m, \mathbf{y}_n) \in \mathcal{S}_F^m \times \mathcal{S}_G^n \; : \; w_{xy_h^+} > w\right\} \;\; .$$

By construction $w_{xy_h^+} \leq w_{xy} \leq w_{xy_h^-}$ for all $(\mathbf{x}_m, \mathbf{y}_n) \in \mathcal{S}_F^m \times \mathcal{S}_G^n$ and for all $h \in \mathbb{N}$. In addition, $w_{xy_h^+} \leq w_{xy_{h+1}^+}$ and $w_{xy_h^-} \leq w_{xy_{h+1}^-}$ for all $(\mathbf{x}_m, \mathbf{y}_n) \in \mathcal{S}_F^m \times \mathcal{S}_G^n$ and for all $h \in \mathbb{N}$. Consequently, we have that $\mathcal{E}_{XY_1^-}(w) \supseteq \mathcal{E}_{XY_2^-}(w) \supseteq ... \supseteq \mathcal{E}_{XY_h^-}(w) \supseteq ...$ and $\mathcal{E}_{XY_1^+}(w) \subseteq \mathcal{E}_{XY_2^+}(w) \subseteq ... \subseteq \mathcal{E}_{XY_h^-}(w) \subseteq ...$. Furthermore

$$\bigcup_{h=1}^{\infty} \mathcal{E}_{XY_h^-}(w) = \mathcal{E}_{XY}(w) = \bigcap_{h=1}^{\infty} \mathcal{E}_{XY_h^+}(w) \qquad \forall \; w \in [0, mn) \;\; .$$

Observing that $\mathcal{W}_F(w;\Delta_h^+) = 1 - P\left[\mathcal{E}_{XY_h^+}(w)\right]$ and $\mathcal{W}_F(w;\Delta_h^-) = 1 - P\left[\mathcal{E}_{XY_h^-}(w)\right]$, the equality of the limits a_w and b_w follows from the continuity properties of a probability measure. Then, the corollary is proved.

Under the *lsm* it is also possible to define the power function. In particular, the function $\pi_{\mathbf{n},\alpha}(\Delta;F) = 1 - \mathcal{W}_F(w_{1-\alpha};\Delta)$, considered as a function of Δ over the range $-\infty < \Delta < \infty$, is the power function of the WRS test under the *lsm*. It is known (see Lehmann (1998), pp. 65-69) that the power function $\pi_{\mathbf{n},\alpha}(\Delta;F)$ is non-decreasing in Δ. Thanks to the Corollary 2 it is now possible to refine this result, as formalized in the following proposition.

Proposition 1. *Under the location shift model, the power function of the test (3) for testing hypotheses (13) is continuous and strictly increasing in* Δ *for all* $\Delta \in (-r, r)$.

P . Immediate consequence of the Corollary 2.

Following the same scheme of proof, similar results to those given in Corollary 2 and Proposition 1 can be stated also for the models described below.

1. **Lehmann's Alternatives**. Under the Lehmann's alternative hypotheses (see Lehmann (1953)) it is assumed that $G \in \mathcal{G}_F^q$ where
$$\mathcal{G}_F^q = \{G \in \mathcal{G}_F : G(x) = F(x)^q \; \forall x \in \mathbb{R}, \; q > 1\} \;\; .$$
The testing problem then becomes:
$$H_0 : \; G = F \qquad vs \qquad H_1 : \; G \in \mathcal{G}_F^q \;\; .$$
In this context we shall denote $\mathcal{W}_{FG}(w)$ by $\mathcal{W}_F(w;q)$. Following the scheme of proof of Corollary 2, it can be shown that $\mathcal{W}_F(w;q)$ is continuous and strictly decreasing in h over the range $(0,\infty)$. Then, the power function $\pi_{\mathbf{n},\alpha}(q;F) = 1 - \mathcal{W}_F(w_{1-\alpha};q)$ is continuous and strictly increasing in q over the range $(0,\infty)$.
2. **Rescaling model**. Assume that the random variables X and Y are positive. The alternative hypothesis of the *rescaling model* is $H_1 : \; G \in \mathcal{G}_F^k$ where $\mathcal{G}_F^k = \{G \in \mathcal{G}_F : G(x) = F(x/k) \;\; \forall \; x \in \mathbb{R}; \;\; k > 1\}$. Even for the rescaling model it can be shown that $\mathcal{W}_F(w;k)$ is continuous and strictly decreasing in k over the range (a,b), where $(a,b) \equiv (0,\infty)$ if $\inf\{\mathcal{S}_F\} = 0$ and/or $\sup\{\mathcal{S}_F\} = 0$ and $(a,b) \equiv (\inf\{\mathcal{S}_F\}/\sup\{\mathcal{S}_F\}, \sup\{\mathcal{S}_F\}/\inf\{\mathcal{S}_F\})$ otherwise. Consequently, the power function $\pi_{\mathbf{n},\alpha}(k;F) = 1 - \mathcal{W}_F(w_{1-\alpha};k)$ is continuous and strictly increasing in k over the range (a,b).

ACCEPTED MANUSCRIPT

3. **Probit shift alternative**. The *probit shift alternative* has been recently proposed by Rosner and Glynn (2009). The alternative hypotheses of this model is $H_1 : G \in \mathcal{G}_F^{\mu}$ where

$$\mathcal{G}_F^{\mu} = \left\{G \in \mathcal{G}_F : G(x) = \Phi(\Phi^{-1}(F(x)) + \mu) \ \forall\, x \in \mathbb{R};\ \mu > 0\right\}$$

and $\Phi(\cdot)$ denotes the standard normal cdf. It can be shown that $\mathcal{W}_F(w;\mu)$ (or $\pi_{\mathbf{n},\alpha}(\mu;F)$) is continuous and strictly decreasing (respectively, increasing) in μ over the range $(0,\infty)$.

In order to extend to a more general context the results proved for the location shift model and outlined for the three models described above, it is useful to highlight some common features of these models. For this purpose, consider first the *lsm* and suppose, for simplicity, that $\mathcal{S}_F = \mathbb{R}$ (in this case $r = \infty$). Further, let us introduce the following set of distributions

$$\tilde{\mathcal{G}}_F^{\Delta} = \{G \in \tilde{\mathcal{G}}_F \ : \ G(x) = F(x - \Delta) \ \ \forall\, x \in \mathbb{R};\ \ \Delta < 0\}$$

and let $\mathcal{D}_F^{\Delta} = \mathcal{G}_F^{\Delta} \cup \tilde{\mathcal{G}}_F^{\Delta} \cup F$. Note that the set $\mathcal{D}_F^{\Delta}$ is totally ordered with respect to the usual stochastic ordering. Moreover, note that for every couple of distributions $(L, M) \in \mathcal{G}_F^{\Delta} \times \mathcal{G}_F^{\Delta}$ with $L \neq M$ the condition (8) is satisfied. Analogously, for every couple of distributions $(L, M) \in \tilde{\mathcal{G}}_F^{\Delta} \times \tilde{\mathcal{G}}_F^{\Delta}$ with $L \neq M$ the condition (12) is satisfied. These last properties makes the application of Theorem 1 and Theorem 2 possible in order to show the strict monotonicity of $\mathcal{W}_F(\cdot;\Delta)$ in Δ. Similarly, the Lehmann's alternatives give rise to the set of distributions $\mathcal{D}_F^q = \mathcal{G}_F^q \cup \tilde{\mathcal{G}}_F^q \cup F$ where $\tilde{\mathcal{G}}_F^q = \left\{G \in \tilde{\mathcal{G}}_F : G(x) = F(x)^q \ \forall x \in \mathbb{R},\ 0 < q < 1\right\}$. Also $\mathcal{D}_F^q$ is totally ordered with respect to the usual stochastic ordering. In addition, for every couple of distributions $(L, M) \in (\mathcal{G}_F^q \times \mathcal{G}_F^q) \cup (\tilde{\mathcal{G}}_F^q \times \tilde{\mathcal{G}}_F^q)$ such that $L \neq M$, one of the conditions (8) and (12) is satisfied. Then, Theorem 1 and Theorem 2 can be applied in order to deduce the strict monotonicity of $\mathcal{W}_F(\cdot;q)$ in q. The same observation can also be made concerning the sets of distributions $\mathcal{D}_F^k$ and $\mathcal{D}_F^{\mu}$ associated to the rescaling model and to the Probit Shift Alternatives, respectively.

Each of the models just described has its own natural parameter. In detail, Δ is the natural parameter for the location shift model while q, k, and μ are the natural parameters for the Lehamnn's Alternatives, the Rescaling Model, and the Probit Shift Alternatives, respectively. Nevertheless, all the models considered can be described using the parameter $p_1 = \int F dG = p_1^{\bullet}(F, G)$ which is particularly meaningful in the context of the WRS test. In more detail, in Lehmann (1998) (p. 70) and Newcombe (2006) it is argued that p_1 can be viewed as the effect size in the context of the WRS test. It should also be noted that W_{XY} is, in practice, the natural estimator of $p_1 : \hat{p}_1 = W_{XY}/mn$. Consequently, the statistics W_{XY} can even be used for testing hypotheses on p_1 (e.g. $H_0 : p_1 = 1/2$ vs $H_1 : p_1 > 1/2$) as shown in Zaremba (1962).

For these reasons, in the general context described below we study the characteristics of $\mathcal{W}_{FG}$ and $\pi_{\mathbf{n},\alpha}(F, G)$ in respect of p_1.

Let $\mathcal{D}_F^*$ be the set of distributions generated by a particular model starting from F. Suppose that $\mathcal{D}_F^* = \mathcal{G}_F^* \cup \tilde{\mathcal{G}}_F^* \cup F$ where $\mathcal{G}_F^* \subset \mathcal{G}_F$ and $\tilde{\mathcal{G}}_F^* \subset \tilde{\mathcal{G}}_F$. The testing problem of interest is:

$$H_0 : \ G = F \qquad vs \qquad H_1 : \ G \in \mathcal{G}_F^* \ . \tag{15}$$

Corollary 3. *Let $\mathcal{D}_F^* = \mathcal{G}_F^* \cup \mathcal{G}_F^* \cup F$ be totally ordered with respect to the usual stochastic ordering. Further, assume that for every $(L, M) \in (\mathcal{G}_F^* \times \mathcal{G}_F^*) \cup (\tilde{\mathcal{G}}_F^* \times \tilde{\mathcal{G}}_F^*)$ such that $L \neq M$, one of the following relations holds:*

1. $L(x) \leq M(x) \qquad \forall\, x \in \mathbb{R} \ \ \wedge \ \ \exists\, x \in \mathcal{S}_F \ : \ \ L(x) < M(x)$;
2. $L(x) \geq M(x) \qquad \forall\, x \in \mathbb{R} \ \ \wedge \ \ \exists\, x \in \mathcal{S}_F \ : \ \ L(x) > M(x)$.

Let $Im_{p_1^{\bullet}}(\mathcal{D}_F^*)$ be the image of $\mathcal{D}_F^*$ under $p_1^{\bullet}$, keeping fixed F: $Im_{p_1^{\bullet}}(\mathcal{D}_F^*) = p_1^{\bullet}(F; \mathcal{D}_F^*)$. Let $\mathcal{W}_F(w; p_1)$ denote the distribution $\mathcal{W}_{FG}(w)$ when $p_1^{\bullet}(F, G) = p_1$ and $G \in \mathcal{D}_F^*$.

If $Im_{p_1^{\bullet}}(\mathcal{D}_F^*) \equiv (c, d)$ then $\mathcal{W}_F(w; p_1)$ is continuous and strictly increasing in p_1 over the range (c, d) for all $w \in [0, mn)$.

P . Keep F fixed and observe that, thanks to conditions 1. and 2., the functional $p_1^{\bullet}(F, \cdot)$ is a bijection from $\mathcal{D}_F^*$ to $\mathrm{Im}_{p_1^{\bullet}}(\mathcal{D}_F^*)$. This assures that, for every $w \in [0, mn)$, $\mathcal{W}_F(w; p_1)$ is a function of p_1 over the range (c, d). Thanks to conditions 1. and 2., we can apply Theorems 1 and 2 obtaining that $\mathcal{W}_F(w; p_1)$ is strictly increasing in p_1 over (c, d). Thanks to the hypotheses that $\mathrm{Im}_{p_1^{\bullet}}(\mathcal{D}_F^*)$ is an interval, the continuity of $\mathcal{W}_F(w; p_1)$ can now be proved following the method described in the proof of Corollary 2.

ACCEPTED MANUSCRIPT

As a direct consequence of Corollary 3 we have the following result.

Proposition 2. *Under the conditions of Corollary 3, the power function of the test (3) for testing hypotheses (15) is continuous and strictly increasing in p_1 for all $p_1 \in (c, d)$.*

4. Reproducibility probability estimation for the WRS test.

In this section we show that the WRS test (3) can be performed using a properly defined RP estimator as test statistic. In order to demonstrate the equivalence between the RPE-based test and the classical one, we follow the methodological scheme outlined in De Martini (2008): a point/conservative RP estimator is defined starting from a point/conservative estimator of the parameter of interest and then by applying the plug in principle to the power function. Finally, it is shown that the RP estimator so defined can substitute the test statistics in order to perform the test.

As observed earlier, in the context of the WRS test, it is natural to choose p_1 as the parameter of interest. Then, we first provide a point/conservative estimator of p_1.

Lemma 1. *Let $\mathcal{D}_F^*$ satisfy the conditions of Corollary 3 and let $\gamma \in (0,1)$. If $Im_{p_1^\bullet}(\mathcal{D}_F^*) \equiv (0,1)$, then the solution $\hat{p}_1^\gamma$ of the equation $\mathcal{W}_F(W_{XY}; \hat{p}_1^\gamma) = 1-\gamma$ is a lower bound at level $(1-\gamma)$ for p_1. In particular:*

$$P(\hat{p}_1^\gamma \leq p_1) \cong 1-\gamma \tag{16}$$

P . First, we must ensure that the random variable $\hat{p}_1^\gamma$ is well defined, i.e. we had to assure that the equation $\mathcal{W}_F(W_{XY}; \hat{p}_1^\gamma) = 1-\gamma$ has a solution whatever the observed value of W_{XY} is. For that purpose note that $lim_{p_1 \to 0} \mathcal{W}_F(w; p_1) = 1$ and $lim_{p_1 \to 1} \mathcal{W}_F(w; p_1) = 0$ for every $w \in [0, mn)$. Consequently, thanks to the continuity and to the strict monotonicity of $\mathcal{W}_F(\cdot; p_1)$, whatever the value of $\gamma \in (0,1)$ is and assuming that the observed value of W_{XY} is different from mn, we have that the solution of equation $\mathcal{W}_F(W_{XY}; \hat{p}_1^\gamma) = 1-\gamma$ exists and is unique. If $W_{XY} = mn$, we assume that $\hat{p}_1^\gamma = 1$. In this way $\hat{p}_1^\gamma$ is well defined. The fact that $\hat{p}_1^\gamma$ is a lower bound for p_1 at level $(1-\gamma)$ can now be proved following the proof scheme of Lemma 1 in De Martini (2008). Otherwise, $\hat{p}_1^\gamma$ is a lower bound for p_1 at level $(1-\gamma)$ thanks to the inversion method described in Casella and Berger (2002), Theorem 9.2.14 (p. 434).

Remark 4. As highlighted in (16), the random variable $\hat{p}_1^\gamma$ is a lower bound for p_1 at level *just approximatively* equal to $(1-\gamma)$. This is due to the fact that W_{XY} is a discrete random variable.

Thanks to Proposition 2, the random variable $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) = \pi_{\mathbf{n},\alpha}(\hat{p}_1^\gamma, F)$ is a conservative RP-estimator at level $(1-\gamma)$. That is:

$$P(\hat{\pi}_{\mathbf{n},\alpha}(\hat{p}_1^\gamma, F) \leq \pi_{\mathbf{n},\alpha}(p_1, F)) \cong 1-\gamma \quad .$$

As formalized in the following proposition, the conservative RP-Estimator just introduced defines a test equivalent to the WRS test.

Proposition 3. *Under the conditions of Corollary 3, if $Im_{p_1^\bullet}(\mathcal{D}_F^*) \equiv (0,1)$ the test*

$$\Phi_\alpha(\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F)) = \begin{cases} 1 & \text{iff} \quad \hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) > \gamma \\ 0 & \text{iff} \quad \hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) \leq \gamma \end{cases} \tag{17}$$

for testing hypotheses (15) is equivalent to the test (3).

P . The equivalence of tests (17) and (3) is proved by showing that $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) \leq \gamma$ if and only if $W_{XY} \leq w_{1-\alpha}$. We begin by proving the first implication: if $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) \leq \gamma$ then $W_{XY} \leq w_{1-\alpha}$. By definition, if $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) = \gamma$ then p_1^γ satisfies the equation $\mathcal{W}_F(w_{1-\alpha}; \hat{p}_1^\gamma) = 1-\gamma$ and Lemma 1 assures that $W_{XY} = w_{1-\alpha}$. Analogously, if $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) < \gamma$ then $\mathcal{W}_F(w_{1-\alpha}; \hat{p}_1^\gamma) > 1-\gamma$. Further, Lemma 1 asserts that p_1^γ is the solution of $\mathcal{W}_F(W_{XY}; \hat{p}_1^\gamma) = 1-\gamma$. Consequently, the inequality $W_{XY} < w_{1-\alpha}$ holds because $\mathcal{W}_F(w; p_1)$ is non-decreasing in w for all $p_1 \in (0,1)$. The first implication is then proved. We now focus on the second implication: if $W_{XY} \leq w_{1-\alpha}$ then $\hat{\pi}_{\mathbf{n},\alpha}^\gamma(F) \leq \gamma$. Let $W_{XY} = w_{1-\alpha}$.

ACCEPTED MANUSCRIPT

In this case, Lemma 1 assures that $\hat{p}_1^{\gamma}$ is the solution of the equation $\mathcal{W}_F(w_{1-\alpha}; \hat{p}_1^{\gamma}) = 1 - \gamma$ and, consequently $\hat{\pi}_{\mathbf{n},\alpha}^{\gamma}(F) = \gamma$. Analogously, let $W_{XY} < w_{1-\alpha}$. In this case, Lemma 1 asserts that $\hat{p}_1^{\gamma}$ is the solution of the equation $\mathcal{W}_F(W_{XY}; \hat{p}_1^{\gamma}) = 1 - \gamma$ and the inequality $\hat{\pi}_{\mathbf{n},\alpha}(F) < \gamma$ holds because $\mathcal{W}_F(w; p_1)$ is non-decreasing in w for all $p_1 \in (0, 1)$. Then, also the second implication is proved.

The proof of Proposition 3 is similar to the proofs of Corollaries 1 and 2 in De Martini (2008). However, these last two corollaries concern the special cases $\gamma = \alpha$ and $\gamma = 1/2$, respectively, while Proposition 3 concerns the general case $\gamma \in (0, 1)$.

5. Example of application

The result given in Proposition 3 is merely theoretical because the exact distribution of W_{XY} under H_1 and, consequently, the exact power of the WRS test are, in practice, unknown. However, for practical purposes, it is possible to use the asymptotic distribution of W_{XY} given in (5). This latter large sample distribution can be used in order to define a lower bound for p_1 by the inversion method showed in Lemma 1 and it can be used to approximate the power of the WRS test (as shown in (6)). Clearly, the use of the asymptotic approximation implies that Proposition 3 holds only approximately, i.e. the classical WRS test (3) and the RP-based test (17) do not exactly correspond.

In order to obtain the lower bounds for p_1, we adopt method 5 in Newcombe (2006), since it provides good coverage accuracies. In detail, the conservative estimator $\hat{p}_1^{\gamma}$ is derived inverting the asymptotic distribution of $\hat{p}_1$ (stemming from (5)) obtained under the assumption that X is exponentially distributed and $Y \stackrel{d}{=} kX$ with $k > 1$ (i.e. under the rescaling model with exponential distribution). See Newcombe (2006) for details. Here, it is interesting to note that the conservative estimator $\hat{p}_1^{1/2}$ coincide with the point estimator $\hat{p}_1$.

As regard the RP estimators, the following alternatives can be considered:

1. Let $\hat{p}_2$ and $\hat{p}_3$ be consistent estimators of p_2 and p_3, respectively. The following RP estimator can be defined by applying the plug in principle to the power approximation (6):

$$\hat{\pi}_{\mathbf{n},\alpha}(\hat{p}_1^{\gamma}, \hat{p}_1, \hat{p}_2, \hat{p}_3) = \Phi\left[\frac{mn\left(\hat{p}_1^{\gamma} - \frac{1}{2}\right) - z_{1-\alpha}\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V(\hat{p}_1, \hat{p}_2, \hat{p}_3)}}\right] \tag{18}$$

2. Following Noether (1987), we can assume that the difference between F and G is quite small. Consequently, the variance of W_{XY} is well approximated by the value it takes under H_0, i.e. $V_0 = \sqrt{mn(m+n+1)/12}$. Replacing $V(p_1, p_2, p_3)$ with V_0 in expression (5) and plugging $\hat{p}_1^{\gamma}$ into the resulting formula, the following RP estimator is obtained:

$$\hat{\pi}_{\mathbf{n},\alpha}(\hat{p}_1^{\gamma}) = \Phi\left[\sqrt{\frac{12mn}{n+m+1}}\left(\hat{p}_1^{\gamma} - \frac{1}{2}\right) - z_{1-\alpha}\right] \tag{19}$$

3. Following the method 5 of Newcombe (2006), we hypothesize the rescaling model with exponential distribution. In this case $p_2 = p_1/(2 - p_1)$ and $p_3 = 2p_1^2/(1 + p_1)$ and the variance $V(p_1, p_2, p_3)$ becomes

$$V^E(p_1) = mnp_1(1 - p_1)\left[1 + (n-1)\frac{1 - p_1}{2 - p_1} + (m-1)\frac{p_1}{1 + p_1}\right] .$$

Replacing $V(p_1, p_2, p_3)$ with $V^E(p_1)$ in expression (5) and plugging $\hat{p}_1^{\gamma}$ into the resulting formula, we obtain the following RP estimator:

$$\hat{\pi}_{\mathbf{n},\alpha}(\hat{p}_1^{\gamma}) = \Phi\left[\frac{mn\left(\hat{p}_1^{\gamma} - \frac{1}{2}\right) - z_{1-\alpha}\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V^E\left(\hat{p}_1^{\gamma}\right)}}\right] \tag{20}$$

Note that, for the three estimators proposed above, we have that $\hat{\pi}_{\mathbf{n},\alpha}^{1/2} > 1/2$ if and only if $(w_{xy} - mn/2)/\sqrt{mn(m+n+1)/12} > z_{1-\alpha}$. We can, therefore, conclude that the test (17), defined assuming $\gamma = 1/2$ and using one of the estimators just proposed, is equivalent to the WRS test when the critical value is determined through

ACCEPTED MANUSCRIPT

the asymptotic approximation (4). Compared with the exact WRS test, there is, however, some slight disagreements. For example, when $m = n = 20$ and $\alpha = 0.05$, the exact WRS test leads to the rejection of H_0 if $w_{xy} \geq 262$. Nevertheless, when $w_{xy} = 261$ the pointwise estimates of the power are equal to 0.5020752 and 0.5022004, by applying the (19) and the (20), respectively. Consequently, H_0 is accepted by the test (3) but it is rejected by the test (17) when it is defined starting from the RP-estimators (19) and (20). Indeed, all the differences between the exact WRS test and an approximated RPE-based test are explained by their joint distribution

$\Phi_\alpha(\hat{\pi}_{\mathbf{n},\alpha})$ / $\Phi_\alpha(W_{XY})$	1	0	
1	$\pi - \epsilon_1$	ϵ_1	π
0	ϵ_2	$1 - \pi - \epsilon_2$	$1 - \pi$
	π'	$1 - \pi'$	1

(21)

where π and π' denotes the power/level of the WRS test and the power/level of the RP-based test, respectively. The probability of disagreement between the two tests is given by $d = \epsilon_1 + \epsilon_2$. Here, we present a brief simulation study (with 10.000 replications) in order to evaluate the magnitude of the possible differences between the RPE-based tests and the classical one. In the simulation we analyze the features of the tests based on (18), (19), and (20) in the following scenario:

1. **model**: uniform distribution under location shift model;
2. **sample sizes**: $m = n = 20$ and $m = n = 60$;
3. **significance level**: $\alpha = 0.05$;
4. **shifts**: $\Delta = 0$; 0.1; 0.2; 0.3 when $m = n = 20$ and $\Delta = 0$; 0.055; 0.11; 0.165 when $n = m = 60$. The Δ values for the two sample sizes are chosen in order to give approximately the same power levels (ranging from α to 0.9) for both the sample sizes (see the values in bold in the third line of Table 2). Moreover we avoid the situation in which both the WRS test and the RP-based test have power 1 (or, very close to 1) since, in that situation, there is no disagreement (or, there is a very low disagreement) even if the two tests are not equivalent.

Then, the simulation study analyses $2 \times 4 \times 3 \times 3 = 72$ different combinations of sample sizes (2 values), Δ values (4 values), γ values (3 values), and RP-based tests (3 different tests). For each combination we simulate the joint distribution (21) and, then, we calculate the simulated disagreement rate (i.e. $100 \cdot d\%$) and the simulated powers (π and π'). The results concerning the disagreement rate and the power of the two tests are given in Table 1 and Table 2, respectively. As it can be noted from the Table 1, the disagreement rates are very low, and in 11 cases out of 72 they are lower than 0.1%. Moreover, as the sample size increases the magnitude of disagreement diminishes, whatever RP-Estimator is used. It can be observed that for "intermediate" values of Δ (and then for intermediate levels of the power) the disagreement is a bit higher. When $\gamma = 0.5$, all the RPE-based tests have the same disagreement. When γ decreases some differences appear. In particular, for a given value of Δ, the tests defined on the basis of the estimators (19) and (20) have approximately the same disagreement for all the γ's considered. To the contrary, as regards the test defined on the basis of the RP-estimator (18), for a given Δ the disagreement is quite variable in γ. In general, the magnitude of the disagreement can be considered negligible, except for the estimator (18) when $n = m = 20$ and $\gamma = \alpha$. In Table 2, we give the simulated powers of the WRS test (in bold in the third line of Table 2) and the simulated powers of the RPE-based tests. From that table it turns out that the simulated powers and levels of the RP-based tests are very close to those of the classical test and that they are generally a bit higher. Then, in the scenario considered, the test (17) defined through the estimators (19) or (20) approximates very well the classical test (3).

To conclude, we remark that the power of the WRS test can be approximated in many different ways, and that (17), (18) and (19) represent just a few of them. In a further study, we aim comparing some other RP estimators on some different scenarios, in order to provide a reliable tool for RP estimation and testing in the context of the WRS test.

6. Acknowledgement

Partial support was provided by the Italian Ministry of Research (MIUR) (protocol 2007 AYHZWC Statistical methods for learning in clinical research). Thanks are also due to an anonymous referee whose comments helped the improvement of the manuscript.

ACCEPTED MANUSCRIPT

γ	RP-Estimator	$m = n = 20$				$m = n = 60$			
	Δ	0.0	0.1	0.2	0.3	0.000	0.055	0.110	0.165
0.5	(18)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%
	(19)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%
	(20)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%
0.3	(18)	0.12%	0.11%	0.14%	0.05%	0.02%	0.11%	0.07%	0.04%
	(19)	0.20%	0.79%	1.10%	0.45%	0.04%	0.40%	0.39%	0.27%
	(20)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%
α	(18)	0.48%	2.36%	3.07%	1.39%	0.14%	0.64%	0.82%	0.33%
	(19)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%
	(20)	0.20%	0.79%	1.10%	0.45%	0.02%	0.24%	0.22%	0.14%

Table 1: Simulated rate of disagreement of the RPE based tests defined according to (18), (19), and (20).

γ	RP-Estimator	$m = n = 20$				$m = n = 60$			
	Δ	0.0	0.1	0.2	0.3	0.000	0.055	0.110	0.165
	Power	**0.0505**	**0.2634**	**0.6290**	**0.9039**	**0.0500**	**0.2610**	**0.6244**	**0.8965**
0.5	(18)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979
	(19)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979
	(20)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979
0.3	(18)	0.0517	0.2645	0.6304	0.9044	0.0502	0.2621	0.6251	0.8969
	(19)	0.0525	0.2713	0.6400	0.9084	0.0504	0.2650	0.6283	0.8992
	(20)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979
α	(18)	0.0459	0.2404	0.5985	0.8900	0.0486	0.2550	0.6162	0.8934
	(19)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979
	(20)	0.0525	0.2713	0.6400	0.9084	0.0502	0.2634	0.6266	0.8979

Table 2: Simulated level and power of the WRS test (in bold in the third line of the table) and of the RPE based tests defined according to (18), (19), and (20).

ACCEPTED MANUSCRIPT

References

Bean, R., Froda, S., van Eeden, C., 2004. The normal, Edgeworth, saddlepoint and uniform approximations to the Wilcoxon-Mann-Whitney null-distribution: a numerical comparison, Journal of Nonparametric Statistics. 16, 279-288.

Berger, J.O., Sellke, T., 1987. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence, JASA. 82, 112-122.

Casella, G., Berger, R.L., 2002. Statistical Inference, second ed. Duxbury

De Martini, D., 2008. Reproducibility Probability Estimation for Testing Statistical Hypotheses. Statistics & Probability Letters. 78, 1056-1061.

Di Bucchianico A., 1999. Combinatorics, computer algebra and the Wilcoxon-Mann-Whitney test. Journal of Statistical Planning and Inference. 79, 349-364.

Gibbons, J.D., Chakraborti, S., 2003. Nonparametric Statistical Inference. Dekker, New York.

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

Haynam G.E., Govindarajulu Z., 1966. Exact Power of Mann-Whitney Test for Exponential and Rectangular Alternatives. The Annals of Mathematical Statistics. 37, 945-953.

Hubbard, R., Bayarri, M.J., 2003. Confusion Over Measures of Evidence (p's) Versus Error (α's) in Classical Statistical Testing. The American Statistician. 57, 171-178.

Lehmann, E.L., 1951. Consistency and Unbiasedness of Certain Nonparametric Tests. The Annals of Mathematical Statistics. 22, 165-179.

Lehmann, E.L., 1953. The Power of Rank Tests. The Annals of Mathematical Statistics. 24, 23-43.

Lehmann, E.L., 1998. Nonparametrics: Statistical Methods Based on Ranks. Chapman & Hall, New York.

Newcombe, R.G., 2006. Confidence intervals for an effect size measure based on the Mann-Whitney statistics. Part 2: Asymptotic methods and evaluation. Statistics in Medicine. 25, 559-573.

Noether, G.E., 1987. Sample size determination for some common non-parametric tests. JASA. 82, 645-647.

Rosner, B., Glynn, R. J., 2009. Power and Sample Size Estimation for the Wilcoxon Rank Sum Test with Application to Comparisons of C Statistics from Alternative Prediction Models. Biometrics. 65, 188-197.

Shao, J., Chow, S.C., 2002. Reproducibility Probability in Clinical Trials. Statistics in Medicine. 21, 1727-1742.

Wilcoxon, F., 1945. Individual compariosons by ranking methods. Biometrics. 1, 80-83.

Witting,H., 1960. A generalized Pitmann efficiency for nonparametric tests. The Annals of Mathematical Statistics. 31, 405-414.

Zaremba, S.K., 1962. A generalization of Wilcoxon's test. Monatshefte fur Mathematik. 66, 359-370

ACCEPTED MANUSCRIPT