Electronic Journal of Statistics
Vol. 6 (2012) 1358–1372
ISSN: 1935-7524
DOI: 10.1214/12-EJS714

A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data

Frank Konietschke

Department of Medical Statistics
University of Göttingen
Humboldtallee 32
37073 Göttingen, Germany
e-mail: fkoniet@gwdg.de

and

Markus Pauly

Institute of Mathematics
University of Düsseldorf
Universitätsstrasse 1
40225 Düsseldorf, Germany
e-mail: markus.pauly@uni-duesseldorf.de

Abstract: We consider nonparametric ranking methods for matched pairs, whose distributions can have different shapes even under the null hypothesis of no treatment effect. Although the data may not be exchangeable under the null, we investigate a permutation approach as a valid procedure for finite sample sizes. In particular, we derive the limit of the studentized permutation distribution under alternatives, which can be used for the construction of $(1-\alpha)$-confidence intervals. Simulation studies show that the new approach is more accurate than its competitors. The procedures are illustrated using a real data set.

Keywords and phrases: Confidence intervals, heteroscedasticity, matched pairs, nonparametric Behrens-Fisher problem, rank statistics, resampling.

Received February 2012.

Contents

1. Introduction

In many psychological, biological, or medical trials, data are collected in terms of a matched pairs design, e.g. when each subject is observed repeatedly under two different treatments or time points. When the normality assumption of the data (or of the differences of the paired observations) is violated, e.g. in case of skewed distributions or even ordered categorical data, nonparametric ranking procedures, which use ranks over all dependent and independent observations, are preferred for making statistical inferences (Munzel, 1999b). Most of these approaches, however, are restricted to testing problems and cannot be used for the computation of confidence intervals for the treatment effects. In randomized clinical trials, the construction of confidence intervals is consequently required by regulatory authorities: "Estimates of treatment effects should be accompanied by confidence intervals, whenever possible..." (ICH, 1998, E 9 Guideline, ch. 5.5, p. 25). Particularly, different variances of the paired observations occur in a natural way, e.g. when data are collected over time. The derivation of nonparametric procedures, which allow the data to have different variances or shapes even under the null hypothesis, is a challenge.

We consider n independent and identically distributed random vectors

$$X_k = (X_{k,1}, X_{k,2})^T, k = 1, \ldots, n, \tag{1.1}$$

with marginal distributions F_i, i.e. $X_{k,i} \sim F_i, i = 1, 2; k = 1, \ldots, n$. To allow for continuous and discontinuous distributions in a unified way, $F_i(x) = 1/2(F_i^+(x) + F_i^-(x))$ denotes the normalized version of the distribution function (Akritas, Arnold and Brunner, 1997; Akritas and Brunner, 1997; Munzel, 1999a; Ruymgaart, 1980). Hereby, $F_i^+(x) = P(X_{1,i} \leq x)$ denotes the right-continuous and $F_i^-(x) = P(X_{1,i} < x)$ denotes the left-continuous version of the distribution function, respectively. The general model (1.1) does not entail any parameters by which a difference between the distributions could be described. Therefore, the marginal distributions F_1 and F_2 are used to define a treatment effect by

$$p = \int F_1 dF_2 = P(X_{1,1} < X_{2,2}) + 1/2P(X_{1,1} = X_{2,2}), \tag{1.2}$$

which is known as relative marginal effect (Brunner, Domhof and Langer, 2002; Brunner and Puri, 1996; Konietschke et al., 2010, p. 38). Note that p is not equivalent to the sign effect $\widetilde{p} = P(X_{1,1} < X_{1,2}) + 1/2P(X_{1,1} = X_{1,2})$, i.e. to the probability of the first observation being smaller than the second observation. The relative marginal effect p uses more information from the data. In particular, compared to the Wilcoxon-signed-rank test, inference methods for p do not need the assumption of symmetric distributed differences (Munzel, 1999b). For independent ordered categorical data, p is also known as ordinal effect size measure (Ryu, 2009). If $p > 1/2$, the observations with distribution F_2 tend to be larger than those with distribution F_1. In case of $p = 1/2$, the observations in neither one of the two marginal samples tend to be smaller or larger than

in the other sample. Thus, we characterize the case of "no treatment effect" as $p = 1/2$. Clearly, if $F_1 = F_2$, then $p = 1/2$ is fulfilled. The other implication, however, is not true in general. This can be seen by the counterexample that $p = 1/2$ if $X_{1,1}$ and $X_{2,2}$ are both normally distributed with a common mean μ and possibly heteroscedastic variances σ_1^2 and σ_2^2, respectively. Therefore, testing the null hypothesis $H_0 : p = 1/2$ is known as the *nonparametric Behrens-Fisher problem* (Brunner and Munzel, 2000).

Munzel (1999b) proposed a test procedure for $H_0 : p = 1/2$ and a small sample approximation in matched pairs. His test is widely used by practitioners (Krone et al., 2008; Obenauer et al., 2002, among others). Simulation studies show that the procedure tends to maintain the type-I error level quite accurately when $n \geq 20$. For smaller $n < 20$, however, the test tends to be very liberal or conservative, depending on whether $X_{1,1}$ and $X_{1,2}$ are negatively or positively correlated.

In this paper, we investigate the conditional studentized permutation distribution of Munzel's linear rank statistic to achieve a valid procedure for finite sample sizes. Although the data may not be exchangeable in model (1.1), an accurate and (asymptotically) valid level α permutation test for $H_0 : p = 1/2$ can be derived, if (i) the permutation distribution of the statistic is asymptotically independent from the distribution of the data; (ii) the permutation distribution has a limit; and (iii) if the distribution of the test statistic and the conditional permutation distribution (asymptotically) coincide (Janssen and Pauls, 2003). The conditions (i)-(iii) are also known as the invariance property of permutation tests (Neubert and Brunner, 2007). Moreover, our proposed permutation test has the additional advantage that it is even exact for finite sample sizes if the pairs are exchangeable.

For independent observations, Janssen (1997), Janssen and Pauls (2003) and Janssen (2005) investigate studentized permutation approaches for the parametric Behrens-Fisher problem, whereas Janssen (1999b) and Neubert and Brunner (2007) also consider ranking approaches for $H_0 : p = 1/2$. The theoretical results obtained in these papers, however, are not valid in our model and the permutation scheme carried out for independent observations (i.e. to permute all data $X_{1,1}, X_{1,2}, \ldots, X_{n,1}, X_{n,2}$) needs to be changed. Following Janssen (1999a) and Munzel and Brunner (2002), we permute the sample units $X_{k,1}$ and $X_{k,2}$ within each matched pair $X_k = (X_{k,1}, X_{k,2})^T$ to protect the dependency structure of the data. Based on the 2^n possible permutations within the sample units, conditional central limit theorems as well as test consistency results will be derived in this paper. It will be shown that the items (i)-(iii) mentioned above are fulfilled. In particular, the studentized permutation distribution under an arbitrary alternative $p \neq 1/2$ will be investigated, which can be used for the computation of approximate $(1 - \alpha)$-confidence intervals for p.

Munzel and Brunner (2002) suggest an exact paired rank test for the null hypothesis of exchangeability. However, this approach cannot be used for the computation of confidence intervals, and particularly, it is not valid under the assumption of heteroscedasticity.

2. Munzel's (1999b) unconditional procedure

To estimate the relative marginal effect size p defined in (1.2), let

$$\widehat{F}_i(x) = n^{-1} \sum_{k=1}^{n} 1/2 \left(\mathbf{1}\{X_{k,i} \leq x\} + \mathbf{1}\{X_{k,i} < x\} \right), \; i = 1, 2,$$

denote the marginal empirical distribution functions. Here, $\widehat{F}_i(x)$ denotes the normalized version of the empirical distribution function (Ruymgaart, 1980). An estimator $\widehat{p}$ of p is then obtained by replacing F_1, and F_2 with $\widehat{F}_1$, and $\widehat{F}_2$, respectively. The estimator

$$\widehat{p} = \int \widehat{F}_1 d\widehat{F}_2 = (2n)^{-1}(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}) + 1/2 \tag{2.1}$$

can easily be calculated by using the (mid-)ranks $R_{k,2}$ and $R_{k,1}$ from $X_{k,2}$ and $X_{k,1}$ among all $2n$ observations, respectively. Here, $\overline{R}_{i\cdot} = n^{-1} \sum_{k=1}^{n} R_{k,i}$ denotes their means within marginal sample i, $i = 1, 2$. Brunner, Puri and Sun (1995), Akritas and Brunner (1997), Brunner and Puri (1996) and Munzel (1999b) have shown that the linear rank statistic $T_n = n^{1/2}(\widehat{p} - p)$ follows, asymptotically, as $n \to \infty$, a normal distribution with expectation 0 and variance $\sigma^2 = var(F_2(X_{1,1}) - F_1(X_{1,2}))$. To estimate the unknown variance σ^2 consistently, let $R_{k,i}^{(i)}$ denote the rank of $X_{k,i}$ among all n observations in marginal sample i and define the normed placements (Orban and Wolfe, 1982) $Z_k = \widehat{F}_1(X_{k,2}) - \widehat{F}_2(X_{k,1}) = n^{-1}(R_{k,2} - R_{k,2}^{(2)} - R_{k,1} + R_{k,1}^{(1)})$. Finally, the unknown variance σ^2 can be consistently estimated by the empirical variance

$$\widehat{\sigma}^2 = (n-1)^{-1} \sum_{k=1}^{n} \left(Z_k - \overline{Z}_{\cdot} \right)^2. \tag{2.2}$$

Under the null hypothesis $H_0 : p = 1/2$, the linear rank statistic

$$M_n = n^{1/2}(\widehat{p} - 1/2)/\widehat{\sigma} \tag{2.3}$$

follows, asymptotically, as $n \to \infty$, a standard normal distribution if $\sigma^2 > 0$ holds. Thus, an asymptotic unconditional two-sided test is given by $\varphi_n = \mathbf{1}\{M_n \leq -z_{1-\alpha/2}\} + \mathbf{1}\{M_n \geq z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile from the standard normal distribution. A one-sided test for the null hypothesis $H_0 : p \leq 1/2$ is given by $\varphi_{n,1} = \mathbf{1}\{M_n \geq z_{1-\alpha}\}$. Without loss of generality we only consider two-sided testing problems throughout this paper. Here we like to point out that the theoretical results in this paper, however, are not restricted to one or two-sided testing problems. Asymptotic $(1-\alpha)$-confidence intervals for p are given by

$$C = \left[\widehat{p} \pm z_{1-\alpha/2}(\widehat{\sigma} n^{-1/2}) \right]. \tag{2.4}$$

Munzel (1999b) suggests to approximate the distribution of M_n by a central t-distribution with $n-1$ degrees of freedom. For inferences, the quantiles $z_{1-\alpha/2}$ used above are replaced by the $(1-\alpha/2)$-quantile from the t_{n-1}-distribution. Simulation studies show, however, that the quality of the approximation depends on the dependency structure in the data which is, of course, not a desirable property of a procedure for matched pairs. Therefore, as a finite correction, the studentized permutation distribution from Munzel's linear rank statistic M_n defined in (2.3) will be considered in the next section.

3. A studentized permutation test

Let $\tau_1, \ldots, \tau_n$ denote n independent and identically distributed permutations on the symmetric group $\mathcal{S}_2$,, i.e. on the set of permutations $\{(1,2),(2,1)\}$, and let $X_k^\tau = (X_{k,\tau_{k(1)}}, X_{k,\tau_{k(2)}})^T$ denote the permuted pairs for $k = 1, \ldots, n$. The data X_k and the permuted data X_k^τ are collected in $X = (X_1, \ldots, X_n)^T$ and in $X^\tau = (X_1^\tau, \ldots, X_n^\tau)^T$, respectively. Further let $\widehat{p} = \widehat{p}(X)$ and $\widehat{\sigma}^2 = \widehat{\sigma}^2(X)$ denote the estimators of p and σ^2 as defined in (2.1) and in (2.2). For convenience, let $\widehat{p}^\tau = \widehat{p}(X^\tau)$ and $\widehat{\sigma}_\tau^2 = \widehat{\sigma}^2(X^\tau)$ denote the quantities $\widehat{p}$ and $\widehat{\sigma}^2$ being computed with the permuted variables.

It turns out, that the distribution of $T_n(X) = n^{1/2}(\widehat{p} - 1/2)$ and the conditional permutation distribution of $T_n^\tau(X^\tau) = n^{1/2}(\widehat{p}^\tau - 1/2)$ differ under heteroscedasticity, and a valid level α test can not be achieved in this setup. Therefore, we consider the distribution of the test statistic $M_n(X) = n^{1/2}(\widehat{p} - 1/2)/\widehat{\sigma}$ defined in (2.3) and the conditional studentized permutation distribution of the statistic

$$M_n^\tau = n^{1/2}(\widehat{p}^\tau - 1/2)/\widehat{\sigma}_\tau, \tag{3.1}$$

i.e. of the studentized quantity $T_n^\tau(X^\tau)$. In the next steps, we will investigate the invariance property of the conditional distribution of M_n^τ. The limiting distribution of M_n^τ will be derived in the next theorem.

Theorem 1. *Let M_n^τ as given in (3.1) and denote by Φ the standard normal distribution function. If $\sigma^2 > 0$, then we have convergence under the null as well as under the alternative*

$$\sup_{x \in \mathbb{R}} \mid P(M_n^\tau \leq x \mid X) - \Phi(x) \mid \to 0 \quad \textit{in probability, as } n \to \infty. \tag{3.2}$$

Theorem 1 states that the limiting standard normal distribution of M_n^τ does not depend on the distribution of the data, particularly, it is achieved for arbitrary p, i.e. it even holds under the alternative $p \neq 1/2$. Let $\varphi_n^\tau = \mathbf{1}\{M_n \leq z_{\alpha/2}^\tau\} + \mathbf{1}\{M_n \geq z_{1-\alpha/2}^\tau\}$, where $z_{1-\alpha/2}^\tau$ denotes the $(1-\alpha/2)$-quantile from the conditional studentized permutation distribution. Note that for notational convenience we only focus on this non-randomized version. However, the following results also hold for a randomized version of the permutation test. In the next theorem we will show that Munzel's unconditional test φ_n and the conditional permutation test φ_n^τ are asymptotically equivalent.

Theorem 2.

(i) Under the null hypothesis $H_0 : p = 1/2$, the studentized permutation test φ_n^τ is asymptotically exact at α-level of significance, i.e. $E(\varphi_n^\tau) \longrightarrow \alpha$, and asymptotically equivalent to φ_n, i.e.

$$E(|\varphi_n^\tau - \varphi_n|) \longrightarrow 0 \quad \text{as } n \to \infty. \tag{3.3}$$

(ii) The permutation test φ_n^τ is consistent, i.e. for any choice of p we have convergence

$$E(\varphi_n^\tau) \longrightarrow \alpha \mathbf{1}\{p = 1/2\} + \mathbf{1}\{p \neq 1/2\} \quad \text{as } n \to \infty.$$

Theorem 1 and Theorem 2 show that the studentized permutation test φ_n^τ fulfills the invariance property, thus, it is an appropriate level α test procedure for $H_0 : p = 1/2$. The numerical algorithm for the computation of the p-value is as follows

1. Given the data X, compute Munzel's statistic M_n as given in (2.3).
2. For each of the 2^n possible permutations, compute the values M_n^τ and save them in $A_1, \ldots, A_{2^n}$.
3. Estimate the two-sided p-value by

$$\text{p-value} = \min\{2p_1, 2 - 2p_1\}, \text{ where } p_1 = 2^{-n} \sum_{\ell=1}^{2^n} \mathbf{1}\{M_n \leq A_\ell\}.$$

In particular, Theorem 1 states that the distributions of the pivotal quantity $M_n^p = n^{1/2}(\widehat{p} - p)/\widehat{\sigma}$ and of the studentized permutation statistic M_n^τ asymptotically coincide. This means, approximate $(1-\alpha)$- two-sided confidence intervals for p can be obtained from

$$C^\tau = \left[\widehat{p} - z_{1-\alpha/2}^\tau(\widehat{\sigma} n^{-1/2}); \widehat{p} - z_{\alpha/2}^\tau(\widehat{\sigma} n^{-1/2})\right]. \tag{3.4}$$

Remark 1. For larger $n > 10$ the number of permutations increases rapidly and the numerical computation of the p-value can be cumbersome. We recommend to use random permutations of the data in those cases. Simulation studies show that $10,000$ random permutations are sufficient for an adequate p-value estimation.

Remark 2. It may occur that $\widehat{\sigma}$ or $\widehat{\sigma}^\tau$ are 0. We recommend to replace them by $1/n$ in those cases (Neubert and Brunner, 2007).

4. Simulations and data analysis

4.1. Monte-Carlo simulations

For testing the null hypothesis $H_0 : p = 1/2$ formulated above, we consider the unconditional Munzel test φ_n based on the t_{n-1}-approximation of the statistic

M_n in (2.3) and the conditional permutation test φ_n^τ as described above, respectively. The simulation studies are performed to investigate their behaviour with regard to (i) maintaining the pre-assigned type-I error level under the hypothesis, (ii) the power of the statistics under alternatives, and (iii) maintaining the pre-assigned coverage probability of the corresponding confidence intervals. The observations $X_k = (X_{k,1}, X_{k,2})^T, k = 1, \ldots, n$, were generated using marginal distributions F_i and varying correlations $\rho \in (-1, 1)$. We hereby generate exchangeable matched pairs having a bivariate normal, exponential, log-normal, or a contaminated normal distribution (where we have rounded to one decimal) each with correlation $\rho \in (-1, 1)$, as well as non-exchangeable data by simulating $F_1 = 0.7N(4, 1) + 0.3N(8, 1)$ and $F_2 = 0.3N(2.07, 2) + 0.7N(6.21, 2)$; $F_1 = N(2.5745, 2)$ and $F_2 = \chi^2_3$; $F_1 = N(0, 1)$ and $F_2 = N(0, 2)$; and $F_1 = N(0, 1)$ and $F_2 = N(0, 4)$, each with correlation ρ, respectively. Routine calculations show that $H_0 : p = 1/2$ is fulfilled in all of these considerations (Neubert and Brunner, 2007). We only consider the small sample sizes $n = 7$ and $n = 10$ throughout this paper. All simulations were conducted with the help of R-computing environment, version 2.13.2 (R Development Core Team, 2010), each with $nsim = 10,000$ runs. The simulation results for normally, exponentially, log-normally, and contaminated normally distributed matched pairs are displayed in Figure 1.

It can be readily seen from Figure 1, that Munzel's test does not control the type-I error level constantly over the range of correlations ρ in the data in case of small sample sizes. It is very liberal when the data are negatively correlated, and vice versa quite conservative in case of positive correlated data. For very small sample sizes ($n = 7$), the test never rejects the null hypothesis in case of strongly positive correlations. This behaviour of the test does not depend

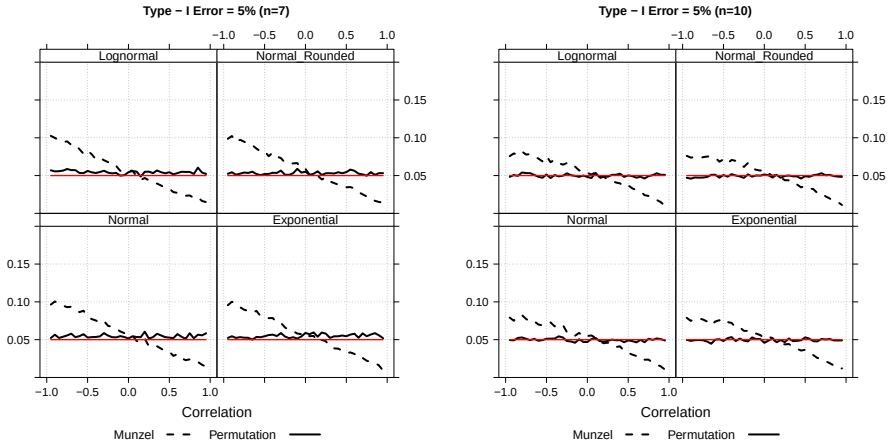

FIG 1. *Type-I error level (y-axis) simulation results ($\alpha = 5\%$) for Munzel's test (dashed line) and the studentized permutation test (solid line) for different exchangeable distributions with varying correlations ρ (x-axis), each with $n = 7$ (left) and $n = 10$ (right).*

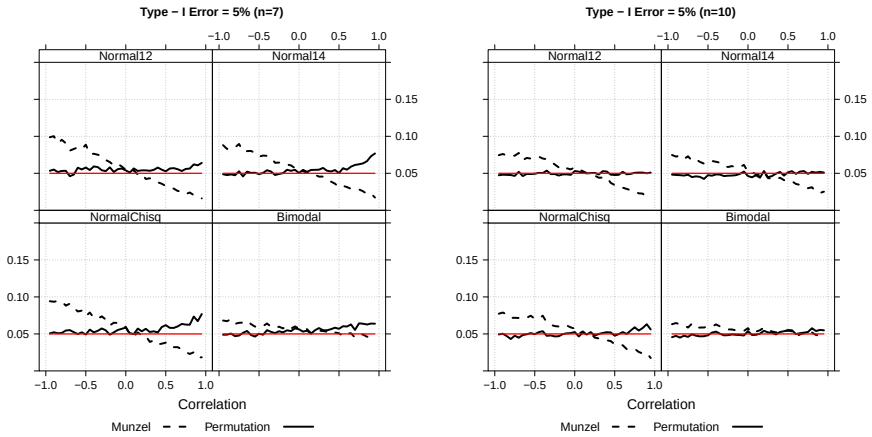

FIG 2. *Type-I error level (y-axis) simulation results ($\alpha = 5\%$) for Munzel's test (dashed line) and the studentized permutation test (solid line) for different non-exchangeable distributions with varying correlations ρ (x-axis), each with $n = 7$ (left) and $n = 10$ (right).*

on the underlying distributions and is identical for all considered setups. The studentized permutation test, however, is a (nearly) exact level α test in these cases and it is not affected by the dependency structure. For $n = 7$, however, only $2^7 = 128$ permutations are possible and the simulation results indicate an maximal estimated type-I error level of 5.6%. For $n = 10$, 1024 permutations can be computed and the exactness of the procedure is apparent. We note that a randomized version of the permutation test would be an exact level α test, because the underlying distributions are exchangeable.

In the next step, type-I error simulations for the case of non-exchangeable data under the null hypothesis $H_0 : p = 1/2$ will be considered. The simulation results are displayed in Figure 2. For non-exchangeable data, the studentized permutation test exhibits a much better control of the pre-assigned type-I error level than Munzel's test. In case of very small sample sizes ($n = 7$) the procedure gets slightly liberal. This can be explained by the fact that only $2^7 = 128$ permutations are possible. With an increasing sample size ($n = 10$), the procedure is accurate and a valid testing procedure for the null hypothesis $H_0 : p = 1/2$. The powers of the two competing procedures were investigated for homoscedastic bivariate normal distributions with correlation $\rho \in (-1/2, 0, 1/2)$, varying expectations $\mu = (0, \delta)^T$ and moderate sample sizes $n = 20$ and $n = 30$. We have chosen these sample sizes for a fair power comparison. The simulation results are displayed in Figure 3.

It can be seen that both procedures have a comparable power, which is in line with Theorem 2. The power simulations for non-exchangeable data did not show any significant difference and are therefore not shown here. Finally, we investigate the maintaining of the pre-assigned coverage probability of the unconditional confidence interval C for p as given in (2.4) as well as of the conditional confidence interval C^τ given in (3.4) for small sample sizes. Hereby,

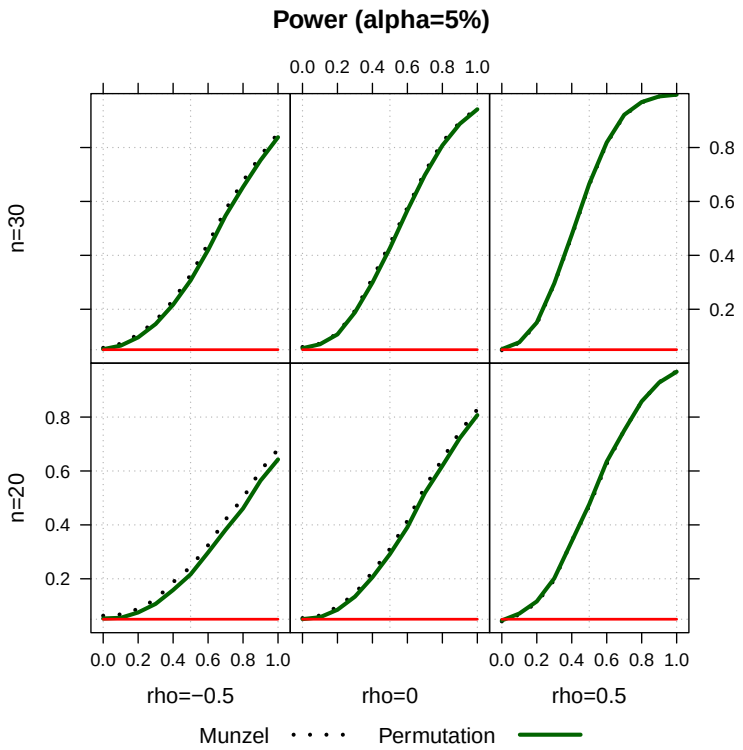

FIG 3. *Power simulation results ($\alpha = 5\%$) for Munzel's test (dashed line) and of the studentized permutation test (solid line).*

TABLE 1
95-% coverage probabilites (%)

	$n = 7$		$n = 10$	
p	C	C^τ	C	C^τ
0.50	94.44	94.61	94.16	94.80
0.55	94.17	94.93	94.11	94.74
0.60	93.70	94.84	94.34	94.90
0.65	93.51	94.84	93.69	94.57
0.70	93.07	93.84	93.55	94.30
0.75	92.58	93.29	92.43	93.63
0.80	92.03	93.41	91.09	93.50

we generate a bivariate normal distribution with correlation $\rho = 0$ and estimate the coverage probabilities for varying true underlying treatment effects p. The simulation results are displayed in Table 1.

It can be seen from Table 1 that the confidence intervals based on the studentized permutation distribution C^τ tend to maintain the pre-assigned coverage probability better than the unconditional version C. For large effects $p \geq 0.8$ both types of confidence intervals tend to be liberal. This occurs, because the distribution of M_n is very skewed in terms of high effects and small sample sizes.

4.2. Analysis of the panic disorder longitudinal trial

We reconsider the panic disorder longitudinal trial which was performed to observe the effect of a specific physical exercise therapy for $n = 15$ patients (Munzel and Brunner, 2002). The response variable is the patient rated global impression (PGI) score being observed at baseline and after 4 weeks of treatment. The lower the score, the better the clinical impression. The original data can be found in the appendix from (Munzel and Brunner, 2002), who already evaluated this trial with an exact paired rank test for the null hypothesis of exchangeability. This approach, however, cannot be used for the computation of confidence intervals for the treatment effect and is not robust against variance heterogeneity. Here, we will analyze the data with the studentized permutation approach. Since the data are observed on an ordinal scale, mean based approaches are inappropriate for this study. Using Spearman's rank correlation coefficient we estimate the correlation and obtain $\widehat{r} = 0.61$, thus, a positive statistical dependence. We obtain an estimated treatment effect of $\widehat{p} = 0.29$. This means, the PGI scores tend to be smaller after 4 weeks of treatment. The null hypothesis $H_0 : p = 1/2$ is significantly rejected at 5% level of significance (p-value=0.006). The 95% confidence intervals for p is given by $[0.16; 0.43]$. Applying Munzel's approach yields the confidence interval $[0.15; 0.43]$ and corresponding p-value = 0.007, thus, comparable results for this study.

Acknowledgement

The authors are grateful to an Associate Editor for helpful comments which considerably improved the paper. This work was supported by the German Research Foundation projects DFG-Br 655/16-1 and HO 1687/9-1

Appendix A: Asymptotic results

For simplicity, we first rewrite the relative marginal effect $p = \int F_1 dF_2$ by $p = \int H dF_2 - \int H dF_1 + 1/2$, where $H(x) = 1/2(F_1(x) + F_2(x))$ denotes the mean distribution function. It is easily seen that the null hypotheses $H_0 : p = 1/2$ and $H_0 : p_1 = p_2$ are equivalent, where $p_i = \int H dF_i$. Further let $\widehat{H}(x) = 1/2(\widehat{F}_1(x) + \widehat{F}_2(x))$ denote the mean empirical distribution function. Note that $\widehat{p}$ as given in (2.1) can be equivalently written as $\widehat{p} = \int \widehat{H} d\widehat{F}_2 - \int \widehat{H} d\widehat{F}_1 + 1/2$.

A.1. Proof of Theorem 1

We start by showing conditional convergence of the numerator

$$\sup_{x \in \mathbb{R}} \mid P\Big(T_n(X^\tau) \leq x \mid X\Big) - \Phi(x/\sigma_\tau) \mid \to 0 \quad \text{in probability,} \tag{A.1}$$

where $\sigma_\tau^2 = E([H(X_{1,1}) - H(X_{1,2})]^2)$ holds.

Note that we can write

$$\widehat{p}_2 - \widehat{p}_1 = \frac{1}{n}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big).$$

Since

$$\begin{aligned}
&\widehat{H}(t) = \widehat{H}(X,t)\\
= \;& \frac{1}{n}\sum_{k=1}^{n}\frac{1}{2}\left(\mathbf{1}\{X_{k,1}\le t\} + \mathbf{1}\{X_{k,1}< t\} + \mathbf{1}\{X_{k,2}\le t\} + \mathbf{1}\{X_{k,2}< t\}\right)\\
= \;& \frac{1}{2n}\sum_{k=1}^{n}\big(\mathbf{1}\{X_{k,\tau_k(1)}\le t\} + \mathbf{1}\{X_{k,\tau_k(1)}< t\} + \mathbf{1}\{X_{k,\tau_k(2)}\le t\} + \mathbf{1}\{X_{k,\tau_k(2)}< t\}\big)\\
= \;& \widehat{H}(X^\tau,t)
\end{aligned}$$

is not affected by the permutation we can rewrite

$$\begin{aligned}
T_n(X^\tau) &= n^{-1/2}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,\tau_k(2)}) - \widehat{H}(X_{k,\tau_k(1)})\big)\\
&\overset{\mathcal{D}}{=} n^{-1/2}\sum_{k=1}^{n} W_k\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big),
\end{aligned}$$

where $(W_k)_k$ is a sequence of independent and identically distributed Rademacher variables with distribution $\frac{1}{2}(\delta_1+\delta_{-1})$ that are independent from the data. Here $\overset{\mathcal{D}}{=}$ means that both expression are equal in distribution. Now, given the data X,

$$E_{k,n} = n^{-1/2}W_k(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})), \quad k = 1,\dots,n,$$

defines an array of row-wise independent random variables. It fulfills

$$E(E_{k,n}|X) = 0 \quad \text{and} \quad var(E_{k,n}|X) = n^{-1}(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1}))^2.$$

Hence we can expand the conditional variance of T_n^W as

$$\begin{aligned}
var(T_n^W \mid X) &= \frac{1}{n}\sum_{k=1}^{n}(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1}))^2\\
&= \frac{1}{n}\sum_{k=1}^{n}\widehat{H}^2(X_{k,2}) + \frac{1}{n}\sum_{k=1}^{n}\widehat{H}^2(X_{k,1}) - \frac{2}{n}\sum_{k=1}^{n}\widehat{H}(X_{k,2})\widehat{H}(X_{k,1})\\
&= V_{n,1} + V_{n,2} - 2V_{n,3}.
\end{aligned}$$

In the following the limit behavior of $V_{n,1}, V_{n,2}$ and $V_{n,3}$ will be analyzed separately. For $V_{n,1}$ we have

$$\begin{aligned}
V_{n,1} &= \frac{1}{n}\sum_{k=1}^{n}H^2(X_{k,2}) + \frac{1}{n}\sum_{k=1}^{n}(\widehat{H}(X_{k,2}) - H(X_{k,2}))^2\\
&- \frac{2}{n}\sum_{k=1}^{n}(\widehat{H}(X_{k,2}) - H(X_{k,2}))H(X_{k,2}).
\end{aligned}$$

Here the first term converges in probability to $E(H^2(X_{1,2}))$ by the law of large numbers and we will now show that the two others are negligible. To see this, note that we have for the second term

$$\begin{aligned} & \frac{1}{n}\sum_{k=1}^{n}(\widehat{H}(X_{k,2}) - H(X_{k,2}))^2 \leq \sup_{x\in\mathbb{R}}(\widehat{H}(x) - H(x))^2 \\ \leq \quad & \frac{1}{4}\Big[\sup_{x\in\mathbb{R}} \mid \widehat{H}^+(x) - H^+(x) \mid + \sup_{x\in\mathbb{R}} \mid \widehat{H}^-(x) - H^-(x) \mid \Big]^2 \\ \to \quad & 0 \end{aligned}$$

in probability by the Glivenko-Cantelli-Theorem. In the same way we get convergence in probability

$$\begin{aligned} & \frac{2}{n}\sum_{k=1}^{n}(\widehat{H}(X_{k,2}) - H(X_{k,2}))H(X_{k,2}) \\ \leq \quad & \sup_{x\in\mathbb{R}} \mid \widehat{H}(x) - H(x) \mid \ \frac{2}{n}\sum_{k=1}^{n} H(X_{k,2}) \to 0. \end{aligned} \tag{A.2}$$

Altogether this shows that $V_{n,1}$ converges in probability to $E(H^2(X_{1,2}))$. With similar arguments the convergences $V_{n,2} \to E(H^2(X_{1,1}))$ and $V_{n,3} \to E(H(X_{1,2})H(X_{1,1}))$ can be proved. Hence the convergence in probability

$$var(T_n^W \mid X) \to \sigma_\tau^2 \tag{A.3}$$

holds as $n \to \infty$. Since we also have

$$\sum_{k=1}^{n} E(E_{k,n}^2 \mathbf{1}\{\mid E_{k,n} \mid \geq \epsilon\} \mid X) \leq E(\mathbf{1}\{1 \geq \epsilon n^{1/2}\}) \longrightarrow 0$$

for all $\epsilon > 0$ the convergence (A.1) follows from Lindeberg's central limit theorem.

Hence by Slutzky's Lemma it remains to check that $\widehat{\sigma}^2(X^\tau)$ converges in probability to σ_τ^2 to complete the proof. Note first that for $i = 1, 2$,

$$\begin{aligned} \widehat{F}_{i,\tau}(t) \quad &= \quad \widehat{F}_i(X^\tau, t) = \frac{1}{2n}\sum_{k=1}^{n}(\mathbf{1}\{X_{k,\tau_k(i)} \leq t\} + \mathbf{1}\{X_{k,\tau_k(i)} < t\}) \\ &= \quad \frac{1}{2}\big(\widehat{F}_{i,\tau}^+(t) + \widehat{F}_{i,\tau}^-(t)\big). \end{aligned}$$

Now fix $i \in \{1, 2\}$. Then, given the data X, the r.v.s $X_{k,\tau_k(i)}, 1 \leq k \leq n$, are independent with distribution function $\widehat{G}_{i,k}^+(t) = \frac{1}{2}(\mathbf{1}\{X_{k,1} \leq t\} + \mathbf{1}\{X_{k,2} \leq t\})$. Thus the Extended Glivenko-Cantelli-Theorem, see e.g. (Shorack and Wellner, 1986, Theorem 1, p.106), shows (conditioned on X) convergence in probability

$$\sup_{t\in\mathbb{R}} \mid \widehat{F}_{i,\tau}^+(t) - \frac{1}{n}\sum_{k=1}^{n}\widehat{G}_{i,k}^+(t) \mid \to 0$$

for $i = 1, 2$. In the same way it follows that (again conditioned on X) $\sup_{t\in\mathbb{R}} | \widehat{F}^{-}_{i,\tau}(t) - \frac{1}{n}\sum_{k=1}^{n} \widehat{G}^{-}_{i,k}(t) | \to 0$ in probability, so that an application of the triangular inequality gives convergence in probability (conditioned on X) for $i = 1, 2$

$$\sup_{t\in\mathbb{R}} | \widehat{F}_{i,\tau}(t) - \widehat{H}(t) | \to 0. \tag{A.4}$$

Applying (A.4) on $\widehat{\sigma}^2(X^\tau)$ we obtain using similar arguments as used for the convergence (A.2) above (again conditioned on X) that

$$\begin{aligned} \frac{n-1}{n}\widehat{\sigma}^2(X^\tau) &= \frac{1}{n}\sum_{k=1}^{n}\big(\widehat{F}_{1,\tau}(X_{k,\tau_k(2)}) - \widehat{F}_{2,\tau}(X_{k,\tau_k(1)})\big)^2 \\ &- \Big(\frac{1}{n}\sum_{k=1}^{n}\big(\widehat{F}_{1,\tau}(X_{k,\tau_k(2)}) - \widehat{F}_{2,\tau}(X_{k,\tau_k(1)})\big)\Big)^2 \end{aligned}$$

is asymptotically equivalent to

$$\begin{aligned} Z_n &= \frac{1}{n}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,\tau_k(2)}) - \widehat{H}(X_{k,\tau_k(1)})\big)^2 \\ &- \Big(\frac{1}{n}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,\tau_k(2)}) - \widehat{H}(X_{k,\tau_k(1)})\big)\Big)^2 \\ &\overset{\mathcal{D}}{=} \frac{1}{n}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,i}) - \widehat{H}(X_{k,i})\big)^2 \\ &- \Big(\frac{1}{n}\sum_{k=1}^{n} W_k\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big)\Big)^2 + o_P(1), \end{aligned}$$

where again W_k are independent and identically distributed Rademacher variables. By (A.3) the first summand of the last expression converges in probability to σ^2_τ. Hence the proof is completed if we show that the second summand converges in probability to zero. But this follows from

$$E\Big(\frac{1}{n}\sum_{k=1}^{n} W_k\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big) \mid X\Big) = 0$$

and the convergence

$$\begin{aligned} var\Big(\frac{1}{n}\sum_{k=1}^{n} W_k\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big) \mid X\Big) &= \frac{1}{n^2}\sum_{k=1}^{n}\big(\widehat{H}(X_{k,2}) - \widehat{H}(X_{k,1})\big)^2 \\ &\leq \frac{1}{n} \to 0. \end{aligned}$$

Altogether this proves $\widehat{\sigma}^2(X^\tau) \to \sigma^2_\tau$ in probability and therefore the desired result.

A.2. Proof of Theorem 2

We remark that part (i) follows directly from Theorem 3 in Munzel (1999) together with the results from Theorem 1 above by applying Lemma 1 in Janssen and Pauls (2003). Suppose now that $p_1 \neq p_2$. Note first that in this situation (3.2) implies that the data dependent critical values $z^{\tau}_{\alpha/2}$ and $z^{\tau}_{1-\alpha/2}$ still converge in probability to the corresponding quantiles from the standard normal distribution, i.e. $z^{\tau}_{\alpha/2} \to u_{\alpha/2}$ and $z^{\tau}_{1-\alpha/2} \to u_{1-\alpha/2}$. We can now expand

$$T_n \quad = \quad n^{1/2}(\widehat{p}_1 - p_1) - n^{1/2}(\widehat{p}_2 - p_2) + n^{1/2}(p_1 - p_2).$$

Since $T_n - n^{1/2}(p_1 - p_2)$ is stochastically bounded, see e.g. (Brunner and Munzel, 2000) it follows that

$$T_n \to \text{sign}(\text{p}_1 - \text{p}_2) \cdot \infty$$

as $n \to \infty$. Moreover, $\widehat{\sigma}^2$ converges in probability to $var(F_1(X_{1,2}) - F_2(X_{1,1}))$ which is positive by assumption. This proves (ii).

References

Akritas, M. G., Arnold, S. F. and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association* **92** 258–265. MR1436114

Akritas, S. F. and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference* **61** 249–277. MR1457720

Brunner, E., Domhof, S. and Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. Wiley, New York. MR1865401

Brunner, E. and Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* **1** 17–21. MR1744561

Brunner, E., Puri, M. L. and Sun, S. (1995). Nonparametric methods for stratified two-sample designs with application to multi clinic trials. *Journal of the American Statistical Association* **90** 1004–1014. MR1354017

Brunner, E. and Puri, M. L. (1996). Nonparametric methods in design and analysis of experiments. *Handbook of Statistics* **13** 631–703. MR1492581

ICH, (1998). *Statistical Principles for Clinical Trials* ICH, Guideline.

Janssen, A. (1997). Studentized permutation test for non-i.i.d. hypotheses and the generalized Behrens – Fisher problem. *Statist. Probab. Lett.* **36** 9–21. MR1491070

Janssen, A. (1999a). Nonparametric symmetry tests for statistical functionals. *Math. Meth. Stat.* **8** 320–343. MR1735469

Janssen, A. (1999b). Testing nonparametric statistical functionals with application to rank tests. *J. Stat. Plan. Inference* **81** 71–93. MR1718393

Janssen, A. (2005). Resampling Student's t-type statistics. *Ann. Inst. Statist. Math.* **57** 507–529. MR2206536

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *The Annals of Statistics* **3** 768–806. MR1994730

Konietschke, F., Bathke, A. C., Hothorn, L. A. and Brunner, E. (2010). Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics and Data Analysis* **54** 1895–1905. MR2640294

Krone, B., Pohl, D., Rostasy, K., Kahler, E., Brunner, E., Oeffner, F., Grange, J. M., Gaertner, J. and Hanefeld, F. (2008). Common infectious agents in multiple sclerosis: a case-control study in children. *Multiple Sclerosis Journal* **14** 136–139.

Munzel, U. (1999a). Linear rank score statistics when ties are present. *Statistics and Probability Letters* **41** 389–395. MR1666092

Munzel, U. (1999b). Nonparametric methods for paired samples. *Statistica Neerlandica* **53** 277–286. MR1730628

Munzel, U. and Brunner, E. (2002). An Exact Paired Rank Test. *Biometrical Journal* **44** 584–593. MR1917920

Neubert, K. and Brunner, E. (2007). A studentized permutation test for the non-parametric Behrens – Fisher problem. *Computational Statistics and Data Analysis* **51** 5192–5204. MR2370717

Obenauer, S., Luftner-Nagel, S., von Heyden, D., Munzel, U., Baum, F. and Grabbe, E. (2002). Screen film vs full-field digital mammography: image quality, detectability and characterization of lesions. *European Radiology* **12** 1697–1702.

Orban, J. and Wolfe, D. A. (1982). A class of distribution-free two-sample tests based on placements. *Journal of the American Statistical Association* **77** 666–672. MR0675896

R Development Core Team, (2010). *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

Ruymgaart, F. H. (1980). A unified approach to the asymptotic distribution theory of certain midrank statistics. *Statistique non Parametrique Asymptotique, 118, J.P. Raoult. Lecture Notes on Mathematics* **821**. MR0604018

Ryu, E. (2009). Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. *Statistics In Medicine* **28** 3179–3188. MR2750413

Shorack, G. S. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York. MR0838963