Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5192–5204
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A studentized permutation test for the non-parametric Behrens–Fisher problem

Karin Neubert*, Edgar Brunner

Department of Medical Statistics, University of Goettingen, Humboldtallee 32, 37073 Goettingen, Germany

Available online 27 June 2006

Abstract

For the non-parametric Behrens–Fisher problem a permutation test based on the studentized rank statistic of Brunner and Munzel is proposed. This procedure is applicable to count or ordered categorical data. By applying the central limit theorem of Janssen, it is shown that the asymptotic permutational distribution of this test statistic is a standard normal distribution. For very small and very different sample sizes, frequently occurring in medical and biological applications, an extensive simulation study suggests that this permutation test works well for data from several underlying distributions. The proposed test is applied to data from a clinical trial.

Keywords: Studentized statistic; Asymptotic distribution; Brunner–Munzel test; Ordered categorical data; Rank test; Count data

1. Introduction

In many areas of applications two-sample comparisons of possibly heteroscedastic samples are frequently of interest. A particular challenge, e.g. in medical or biological applications, is the small number of available observations.

The classical parametric approach considers the hypothesis of equal means in the presence of potentially different variances. For underlying normal distributions various tests have been developed and were extensively studied in literature (e.g. Smith, 1936; Welch, 1937; Satterthwaite, 1946; Cochran, 1964; Moser and Stevens, 1992). For small samples permutation procedures have been suggested, which asymptotically keep the preassigned level (cf. Janssen, 1997; Pesarin, 2001).

In the case of non-normal data, the Wilcoxon–Mann–Whitney (WMW) test is used to compare two independent samples (Wilcoxon, 1945; Mann and Whitney, 1947). Here, the hypothesis $F_1 = F_2$ is considered, where F_1 and F_2 are the distributions of the two samples. If the two distributions, however, differ in dispersion, the WMW test does not maintain its level (Pratt, 1964). In semi-parametric models modifications of the WMW test have been described by Fligner and Policello (1981), extending the applicability to heteroscedastic situations, and by Babu and Padmanabhan (2002), allowing also for non-symmetric distributions. Fligner and Policello (1981) also note that their procedure is applicable for testing the hypothesis: $\int F_1 \, \mathrm{d}F_2 = \frac{1}{2}$. This hypothesis can equivalently be formulated using the relative effect $p = \int F_1 \, \mathrm{d}F_2$ introduced by Mann and Whitney (1947). This effect is defined as the probability that the observations in one sample tend to be larger (or smaller) than those in the other sample. A procedure based on this hypothesis also admitting discrete distributions is the rank test proposed by Brunner and Munzel (2000) and its t-approximation for

* Corresponding author.
E-mail address: karin.neubert@medizin.uni-goettingen.de (K. Neubert).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.05.024

small sample sizes. This rank test can be applied in rather general models, where arbitrary distributions of the data are allowed (only the trivial case of one-point distributions is excluded). This means that this procedure is also applicable for the analysis of ordered categorical or count data. A likelihood ratio test for this hypothesis was presented by Troendle (2002) and bootstrap procedures were discussed by Chen and Kianifard (2000) and Reiczigel et al. (2005).

When applying tests based on resampling methods to a Behrens–Fisher situation, it has to be noted that even under the hypothesis the random variables are not identically distributed. In order to overcome this problem, several approaches of data transformation are suggested when using bootstrap procedures (Efron and Tibshirani, 1993; Reiczigel et al., 2005). A suitable transformation has to be chosen based on the nature of deviation of the data from the case of identical distributions.

Here, we suggest a studentized permutation test based on the rank statistic of Brunner and Munzel (2000). Adopting Janssen's (1997) central limit theorem for studentized permutation tests, we will show asymptotic normality of this test statistic. For small sample sizes the properties of the test will be investigated by a simulation study.

The paper is organized as follows. In Section 2 we describe the model, formulate the hypothesis, and introduce the test statistic. The application of the central limit theorem of Janssen (1997) to the Behrens–Fisher rank statistic is derived in Section 3. Results of a comprehensive simulation study for small sample sizes are presented in Section 4 and a real data example is considered in Section 5. The derivations of the above results are shifted to the Appendix.

2. Model, hypothesis, and test statistic

We consider two independent random samples $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$, where $X_{ik} \overset{\text{iid}}{\sim} F_i$, $i = 1, 2$, $k = 1, \ldots, n_i$. Here, $F_i = \frac{1}{2}\left(F_i^+ + F_i^-\right)$ denotes the normalized version of the distribution function (Ruymgaart, 1980). The F_i's may be arbitrary distributions (only the trivial one-point distributions are excluded). Let $N = n_1 + n_2$ denote the total number of observations.

To allow for non-continuous distribution functions, we consider the extension of the relative effect (Mann and Whitney, 1947) to the discontinuous case, namely

$$p = P\left(X_{11} < X_{21}\right) + \frac{1}{2}P\left(X_{11} = X_{21}\right) = \int F_1 \, \mathrm{d}F_2.$$

This effect describes a tendency towards larger values of one random variable with respect to the other. The random variable X_{11} is then said *to tend to be smaller* (*larger*) *than* X_{21} *if* $p > \frac{1}{2}$ $(p < \frac{1}{2})$. Thus, the hypothesis of no treatment effect (no tendency) can be expressed as

$$\mathrm{H}_0 : p = \tfrac{1}{2}.$$

We note that this hypothesis includes the parametric hypothesis of equal means in the presence of potentially unequal variances as a special case, when the underlying distributions are normal.

A natural estimator $\widehat{p}$ of p is obtained by replacing the distribution functions $F_i(x)$ with their empirical counterparts $\widehat{F}_i(x)$. An unbiased and consistent estimator of p is then given by

$$\widehat{p} = \int \widehat{F}_1 \, \mathrm{d}\widehat{F}_2 = \frac{1}{N}\left(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}\right) + \frac{1}{2}, \quad \overline{R}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik},$$

where $R_{ik} = \text{rank}\,(X_{ik})$ denotes the (mid-)rank of X_{ik} among all N observations. To test the hypothesis, $\mathrm{H}_0 : p = \frac{1}{2}$, we consider the statistic $K_N = \sqrt{N}(\widehat{p} - \frac{1}{2})$. The asymptotic distribution of K_N is derived by considering an *asymptotically equivalent* statistic of independent random variables, which follows from the so-called asymptotic equivalence theorem (see, e.g. Brunner and Munzel, 2002). If $N \to \infty$ such that $N/n_i \leqslant N_0 < \infty$, $i = 1, 2$, then

$$\sqrt{N}\left(\widehat{p} - p\right) \doteqdot \sqrt{N}\left(\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} + 1 - 2p\right), \quad \overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik},$$

where the unobservable random variables $Y_{1k} = F_2\left(X_{1k}\right)$ and $Y_{2k} = F_1\left(X_{2k}\right)$ are the so-called *asymptotic rank transforms* (ARTs). The symbol $\doteqdot$ denotes asymptotic equivalence, i.e. the difference of the two sequences on the left- and right-hand sides of $\doteqdot$ converges to 0 in probability.

By assumption, the ARTs, Y_{ik}, $k=1,\ldots,n_i$, $i=1,2$, are uniformly bounded, independent and identically distributed random variables with variances $\sigma_1^2=Var\,(F_2\,(X_{11}))$ and $\sigma_2^2=Var\,(F_1\,(X_{21}))$, respectively. If $\sigma_1^2,\sigma_2^2>0$, then by applying the central limit theorem, it follows under H_0 that

$$\frac{K_N}{\sigma_N}=\frac{\sqrt{N}}{\sigma_N}\left(\widehat{p}-\frac{1}{2}\right) \tag{1}$$

has, asymptotically, a standard normal distribution where

$$\sigma_N^2=\frac{N}{n_1n_2}\left(n_1\sigma_2^2+n_2\sigma_1^2\right)$$

is unknown and has to be estimated consistently from the data. Brunner and Munzel (2002) derive a consistent estimator V_N^2 of σ_N^2, namely

$$V_N^2=N\left(\frac{1}{n_2}\widehat{\sigma}_1^2+\frac{1}{n_1}\widehat{\sigma}_2^2\right),$$

where

$$\widehat{\sigma}_i^2=\frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(R_{ik}-R_{ik}^{(i)}-\overline{R}_{i\cdot}+\frac{n_i+1}{2}\right)^2.$$

Here, R_{ik} is the overall rank of X_{ik} among all N observations and $R_{ik}^{(i)}$ is the internal rank of X_{ik} among the n_i observations $X_{i1},\ldots,X_{in_i}$ within group i, $i=1,2$. Replacing σ_N in (1) by V_N, we finally obtain the test statistic

$$T_N=\frac{\overline{R}_{2\cdot}-\overline{R}_{1\cdot}}{V_N}\sqrt{\frac{n_1n_2}{N}}, \tag{2}$$

which has, asymptotically, a standard normal distribution under H_0.

As noted above, the normality of T_N only holds asymptotically. Simulation studies showed that quite large sample sizes are necessary to get a satisfactory approximation. In many medical and biological applications, however, the number of available observations may be rather small. Moreover, it is often not appropriate to make any assumptions of continuity for the distribution functions. Brunner and Munzel (2000), for example, suggested to approximate the distribution of the test statistic T_N by a $t_{\widehat{f}}$-distribution. The degrees of freedom $\widehat{f}$ is derived as in the parametric case by a Satterthwaite–Smith–Welch approximation and estimated by

$$\widehat{f}=\frac{\left(\sum_{i=1}^2\widehat{\sigma}_i^2/(N-n_i)\right)^2}{\sum_{i=1}^2\left(\widehat{\sigma}_i^2/(N-n_i)\right)^2/(n_i-1)}.$$

Two other methods recommended for this rather general setting are the likelihood ratio test described by Troendle (2002) and a bootstrap test by Reiczigel et al. (2005). The likelihood ratio approach of Troendle is a recursive method reducing the numerical problem to one dimension and thus making a numerical solution actually feasible. A detailed description of the testing procedure can be found in Troendle (2002). The bootstrap test of Reiczigel et al. (2005) is based on the rank Welch test. Their simulation study suggests that this bootstrap test behaves similar to the *t*-approximation of Brunner and Munzel.

Simulation studies show that the above procedures have shortcomings in attaining the preassigned type I error if the underlying distributions show deviations from symmetry or if sample size is small. Troendle's likelihood ratio procedure, e.g. behaves somewhat liberal for bimodal distributions and the bootstrap test of Reiczigel et al. (2005) tends to become slightly conservative, especially for small sample sizes. The *t*-approximation, however, gets a little liberal when using two-sided tests at the 5% level.

Because of these problems we suggest the application of a permutation test in this situation. In the parametric setting Janssen (1997) proposed to base a permutation test for the Behrens–Fisher problem on a studentized test statistic. This idea dates back to Neuhaus (1993), who applied it to survival problems. Here, we want to adopt this idea to the

standardized rank statistic T_N by calculating a permutation test based on this statistic. That means we base the decision, to accept or reject the hypothesis, on the permutational distribution of the statistic T_N.

Consider all permutations of N elements $\pi \in \mathscr{S}_N$ which are uniformly distributed according to some probability measure P^* independent of $X_{ik},\ i = 1, 2,\ k = 1, \ldots, n_i$. The vectors of permuted random variables are denoted by $\mathbf{X}_\pi,\ \mathbf{R}_\pi$, and $\mathbf{R}_\pi^{(1)}, \mathbf{R}_\pi^{(2)}$, where

$$\mathbf{X}_\pi = \left(X_{\pi(11)}, \ldots, X_{\pi(1n_1)}, X_{\pi(21)}, \ldots, X_{\pi(2n_2)}\right),$$

$$\mathbf{R}_\pi = \left(R_{\pi(11)}, \ldots, R_{\pi(1n_1)}, R_{\pi(21)}, \ldots, R_{\pi(2n_2)}\right),$$

$$\mathbf{R}_\pi^{(i)} = \left(R^{(i)}_{\pi(i1)}, \ldots, R^{(i)}_{\pi(in_i)}\right), \quad i = 1, 2.$$

When replacing the original observations by the permuted observations, the recalculated variance and test statistic will be referred to as T_N^* and V_N^{*2}, where

$$T_N^* = T_N\left(\mathbf{R}_\pi, \mathbf{R}_\pi^{(1)}, \mathbf{R}_\pi^{(2)}\right), \quad V_N^{*2} = V_N^2\left(\mathbf{R}_\pi, \mathbf{R}_\pi^{(1)}, \mathbf{R}_\pi^{(2)}\right).$$

By applying a permutation to the observations, their assignment to the groups is changed and with this the values of the internal ranks. Thus, for each permutation the variance V_N^{*2} has to be recalculated in order to obtain the new value of the test statistic T_N^*. The p-value of the corresponding test is then given by the proportion of permutations where the value of the test statistic of the original observations T_N is smaller or equal to the value of the test statistic calculated of the permuted observations T_N^*. If we conduct $n_{\text{sim}} = \#S$ permutations with $\pi \in S \subset \mathscr{S}_N$ we get for the p-value $p = \#\left\{\pi \in S | T_N \leqslant T_N^*\right\}/n_{\text{sim}}$.

In the next section we will show that the permutational distribution of our studentized statistic is asymptotically normal by applying the central limit theorem of Janssen (1997). Small sample properties of the permutation test and the other three procedures mentioned are analysed in Section 4 by a simulation study.

3. The studentized permutational procedure

In this section we apply the central limit theorem of Janssen (1997) for the conditional permutation distribution of linear studentized test statistics to show that our permutation test asymptotically maintains its preassigned level. We note that this theorem does not assume independence of the random variables and, thus, can be applied to the rank statistic T_N given in (2) (cf. Janssen, 1997). To this end, we rewrite T_N as

$$\begin{aligned} T_N &= \frac{\overline{Z}_{2\cdot} - \overline{Z}_{1\cdot}}{\sqrt{\frac{1}{N}\sum_i \sum_k (N - n_i)/(n_i - 1)\,(N/(N - n_i))^2 \left(Z_{ik} - \overline{Z}_{i\cdot}\right)^2}} \sqrt{\frac{n_1 n_2}{N}} \\ &= \frac{\sum_i \sum_k c_{ik} Z_{ik}}{\sqrt{\sum_i \sum_k (N/(N - n_i))\,(1/(n_i - 1))\left(Z_{ik} - \overline{Z}_{i\cdot}\right)^2}} \quad \text{with } c_{ik} = \begin{cases} -\sqrt{\frac{n_1 n_2}{N}}\frac{1}{n_1}, & i = 1, \\ \sqrt{\frac{n_1 n_2}{N}}\frac{1}{n_2}, & i = 2, \end{cases} \end{aligned} \tag{3}$$

where

$$Z_{ik} = \frac{1}{N}\left(R_{ik} - R_{ik}^{(i)} - \frac{N - n_i}{2}\right) \tag{4}$$

and

$$\overline{Z}_{i\cdot} = \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{n_i + 1}{2} - \frac{N - n_i}{2}\right) = \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{N + 1}{2}\right).$$

We note that the random variables Z_{ik} are uniformly bounded by $\frac{1}{2}$, i.e. $P(|Z_{ik}| > \frac{1}{2}) = 0$.

For the central limit theorem of Janssen (1997) to hold, a couple of conditions on the coefficients and the variance estimator of the test statistic have to be fulfilled. If these conditions are met, Janssen proves that there exists a constant $\tau^2 > 0$ such that

$$\sup_{t\in\mathbb{R}}(|P^*(T_N(\mathbf{R}_\pi)\leqslant t|\mathbf{R}) - \Phi(t/\tau)|) \xrightarrow{P} 0,$$

where Φ denotes the standard normal distribution function. One possible choice of such conditions is

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik}^2 = 1 \quad \forall N \in \mathbb{N}, \tag{5}$$

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik} = 0 \quad \forall N \in \mathbb{N}, \tag{6}$$

$$\max_{i,k} |c_{ik}| \to 0 \quad \text{if } N \to \infty, \tag{7}$$

$$\lim_{N\to\infty} \inf \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(Z_{ik} - \overline{Z}_{..}\right)^2 > 0 \quad P\text{-a.s.} \tag{8}$$

$$\exists\tau > 0 : \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(Z_{ik} - \overline{Z}_{..}\right)^2 V_N^{*-1} \xrightarrow{P\times P^*} \tau^2 \quad \text{if } N \to \infty, \tag{9}$$

$$\lim_{N\to\infty} \sup \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(Z_{ik} - \overline{Z}_{..}\right)^2 1_{[d,\infty)}\left(\left|Z_{ik} - \overline{Z}_{..}\right|\right) \longrightarrow 0 \quad P\text{-a.s. } d \to \infty. \tag{10}$$

Conditions (5)–(7) follow immediately from the linear representation of T_N as given in (3). The validity of the remaining conditions, which involve some detailed mathematical derivations, are presented in the Appendix.

For the above conditions to hold, the variances $\widetilde{\sigma}_1^2 = Var\left(\widehat{F}_2\left(X_{11}\right)\right)$ and $\widetilde{\sigma}_2^2 = Var\left(\widehat{F}_1\left(X_{21}\right)\right)$ must be strictly positive. In addition we assume that there exists some $\kappa \in (0, 1)$ such that

$$\frac{n_1}{N} \xrightarrow{N\to\infty} \kappa \quad \Rightarrow \quad \frac{n_2}{N} \xrightarrow{N\to\infty} 1 - \kappa.$$

4. Simulation results

A comprehensive simulation study was conducted to evaluate the small sample properties of the studentized permutation test derived in the previous sections. Level and power were simulated using data from two normal distributions, two bimodal distributions and, to see the influence of skewed distributions, a normal against a χ_3^2-distribution. To study the influence of very small and very different sample sizes, simulations were performed for sample sizes

$$(n_1, n_2) \in \{(15, 15), (15, 7), (7, 15), (7, 7)\}.$$

To analyse the power of the tests, the mean, μ, of one sample in each distributional setting was shifted by 0.5 and 1, respectively. For simulating the level μ was set to 0.

- Two normal distributions $\mathcal{N}_{n_1}\left(0, \sigma_1^2\right)$, $\mathcal{N}_{n_2}\left(\mu, \sigma_2^2\right)$: Different variances were used for the two samples, where the larger variance was applied to the larger as well as to the smaller sample.

$$\left(\sigma_1^2, \sigma_2^2\right) \in \{(1, 1), (1, 2), (1, 4)\}.$$

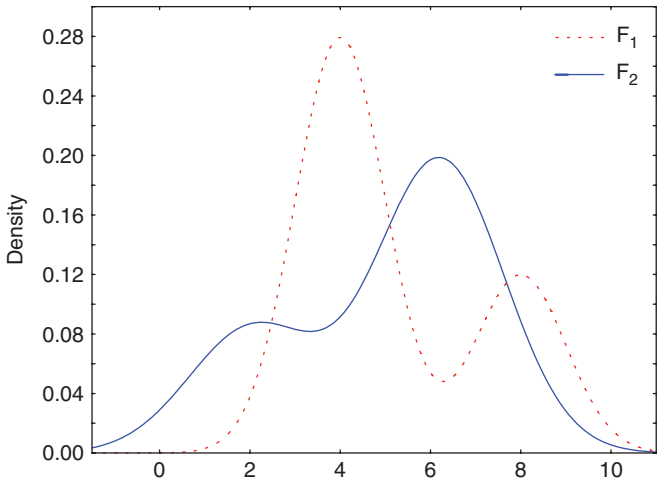


Fig. 1. Densities of two bimodal distributions under hypothesis $p = \frac{1}{2}$.

- Two bimodal distributions: Distributions were derived from the combination of two normal distributions each. The choice of the expectations, the variances, and the proportion of the mixing parts ensure $p = \frac{1}{2}$ if $\mu = 0$ (cf. Fig. 1).

$$F_1 = \tfrac{7}{10}\mathcal{N}_{n_1}(4, 1) + \tfrac{3}{10}\mathcal{N}_{n_1}(8, 1),$$

$$F_2 = \tfrac{3}{10}\mathcal{N}_{n_2}(2.07 + \mu, 2) + \tfrac{7}{10}\mathcal{N}_{n_2}(3(2.07 + \mu), 2).$$

- A normal distribution $F_1 = \mathcal{N}\left(\mu_0 + \mu, \sigma^2\right)$ against one χ^2-distribution $F_2 = \chi^2_3(0)$: Simulations showed that $p = \frac{1}{2}$ is fulfilled when a normal distribution with expectation $\mu_0 = 2.5745$ and variance $\sigma^2 = 2$ is used.

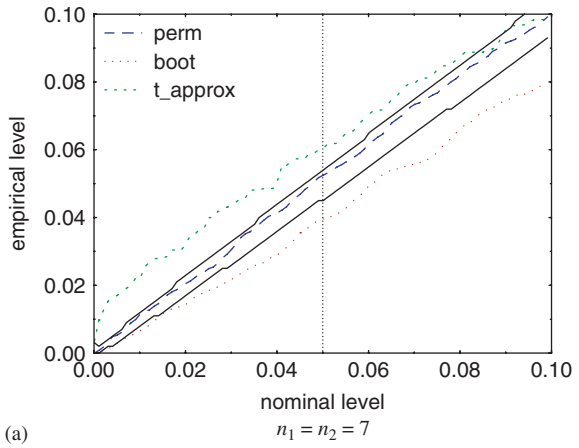
For comparison, the level and power of the t-approximation by Brunner and Munzel (2000), the bootstrap test of Reiczigel et al. (2005) as well as the likelihood ratio test of Troendle (2002) were simulated. Furthermore, "reliability bounds" $\left(\alpha_{\text{low}}, \alpha_{\text{up}}\right)$ were calculated for the empirical level $\widetilde{\alpha}$ of the simulated tests such that

$$\mathbb{P}(\alpha_{\text{low}} \leqslant \widetilde{\alpha} \leqslant \alpha_{\text{up}}) = 0.95.$$

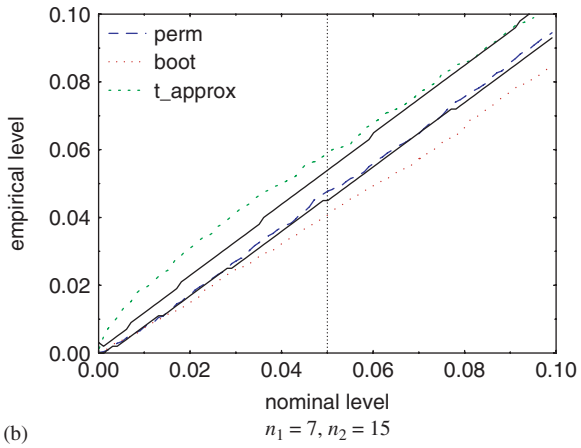
Since the number of all possible permutations (bootstrap samples) and, thus, the simulation time increases rapidly with sample size, only 10,000 random permutations (bootstrap samples) were performed for each test. The level and power calculations are based on 10,000 simulation runs. Due to long computing times the likelihood ratio test was only evaluated at nominal level of $\alpha = 0.05$.

When the variance estimator V_N^2 was 0, i.e. if all the observations in one group had lower ranks than the observations in the other group, it was replaced by some reasonable small value larger than 0, thus leading to a conservative test decision for the permutation test and the t-approximation in this case. This lower bound of V_N^2 is attained when exactly two observations, one of each sample, take the same value, whereas the other observations of the two samples are completely separated and have no ties. Hence, the lowest value of the variance estimator V_N^2 larger than 0 is $N/(2n_1n_2)$. The simulations for the permutation test, the bootstrap test, and the t-approximation were conducted in SAS IML 9.1. For the likelihood ratio test Dr. Troendle kindly provided his FORTRAN 77 simulation program. All tests are performed as two-sided test.

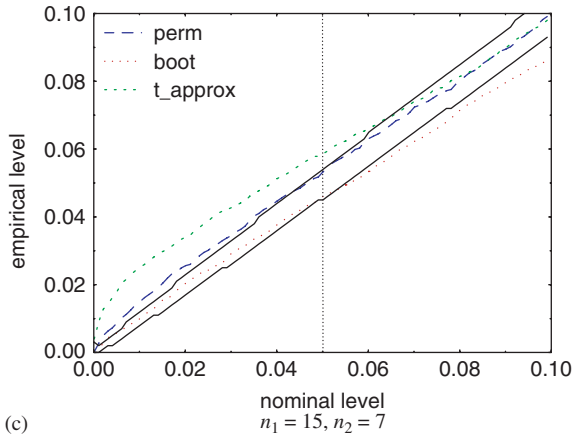
When analysing two normal distributions with equal variances for sample size 15 in both groups all three tests keep the level, as their level curves follow the lines of the upper (t-approximation) or lower (bootstrap test) reliability bound or run within these bounds (permutation test). If the sample size is reduced in one sample only, the t-approximation becomes a little anti-conservative, whereas the curves of the other two tests almost stay the same. When there are only seven observations in both groups the level curve of the permutation test follows the upper reliability bound, the curve of the bootstrap test runs somewhat below the lower bound and the t-approximation is liberal. If the variance in the second sample is 2 and samples size is 7 in both groups (cf. Fig. 2(a)) the empirical level of the permutations



(a) $n_1 = n_2 = 7$



(b) $n_1 = 7, n_2 = 15$



(c) $n_1 = 15, n_2 = 7$

test	$\mu = 0$	$\mu = 0.5$	$\mu = 1$
Φ_{Perm}	**0.053**	0.125	0.343
Φ_{Boot}	**0.046**	0.108	0.305
$\Phi_{\text{t-approx}}$	0.059	0.132	0.349
Φ_{LR}	0.056	0.126	0.327

(d) n_1=15, n_2=7

Fig. 2. Two normal distributions. Figures show nominal against empirical levels where $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, solid lines indicating the reliability bounds. Simulated levels and power are displayed in (d) for $\alpha = 5\%$, where bold letters indicate levels within the reliability bounds: (a) $n_1 = n_2 = 7$, (b) $n_1 = 7, n_2 = 15$, (c) $n_1 = 15, n_2 = 7$ and (d) $n_1 = 15, n_2 = 7$.

test is for all levels within the reliability bounds, whereas the curve of the bootstrap test is a little conservative. For different sample sizes in the two groups the level curve of the permutation test gets close to the lower (Fig. 2(b)) or upper (Fig. 2(c)) reliability bound, respectively. For the larger variance occurring in the larger group, the bootstrap test is slightly conservative (Fig. 2(b)), whereas in the opposite situation its level curve runs within the reliability bounds for nominal levels up to 5%, crossing the lower bound there and staying just below for all higher levels (Fig. 2(c)). The *t*-approximation is slightly liberal in all three settings for nominal levels up to 6% and has level within or very close to the reliability region above that. For variance 4, the observed effects fortify, where the bootstrap test shows the smallest changes. Fig. 2(d) compares all four tests at $\alpha = 5\%$ regarding level and power, where bold letters indicate that the empirical level is within the reliability bounds. We present the case of sample sizes $n_1 = 15$ and $n_2 = 7$, where the smaller sample has variance 2. In this case the level of the permutation and bootstrap test lies within the reliability bounds. All four tests have comparable power under the two alternative settings.

The influence of changing variances is illustrated in Fig. 3. As expected, the power decreases as the variance increases. Furthermore, if the larger variance is observed in the smaller sample, the power is smaller than in the opposite situation. Though the power of the permutation test is lower than the power of the *t*-approximation and the likelihood ratio test when the larger variance occurs in the larger sample, its power is comparable to the liberal *t*-approximation in the opposite situation.

If the underlying distributions are very differently shaped, e.g. bimodal, it can be seen from Fig. 4 that the *t*-approximation and the permutation test behave very similar as with underlying normal distributions. For samples

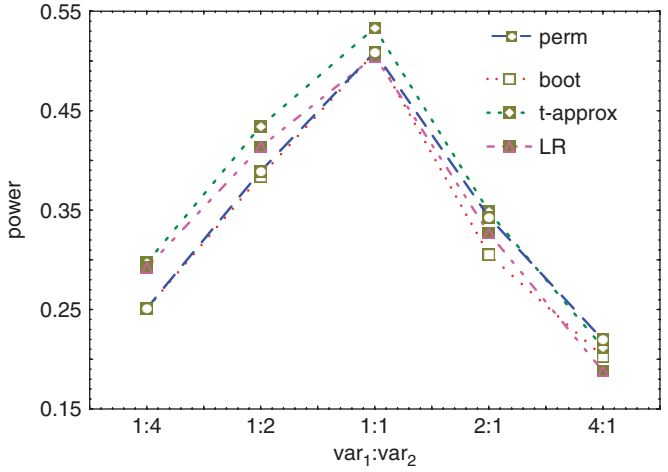
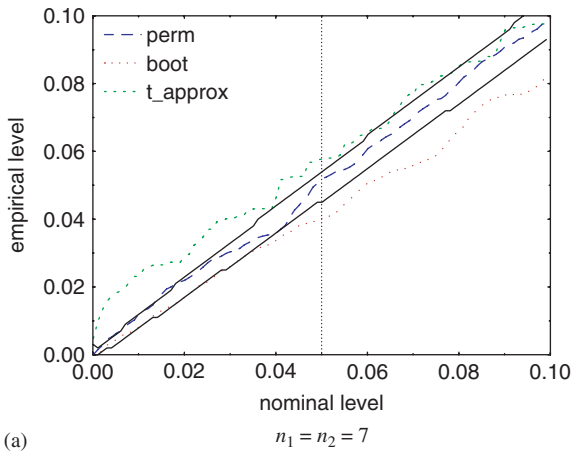
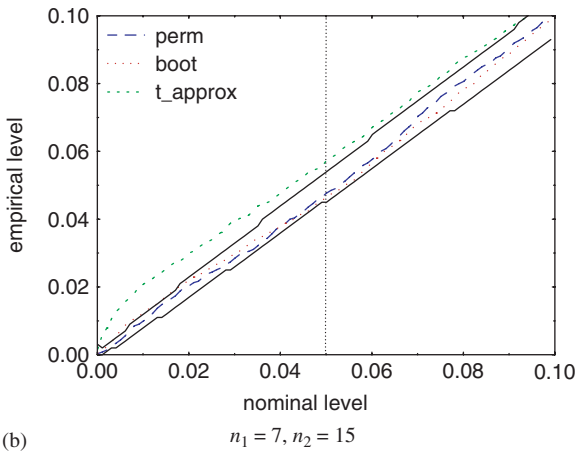


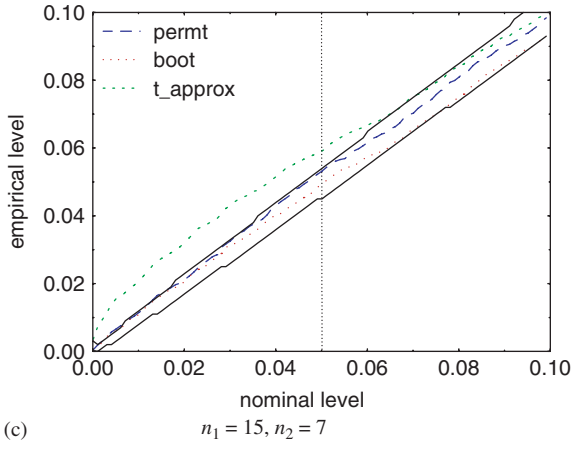
Fig. 3. Powercurves for varying variances for two normal distributions with $n_1 = 7$, $n_2 = 15$, $\alpha = 5\%$, $\mu = 1$.



(a) $n_1 = n_2 = 7$



(b) $n_1 = 7, n_2 = 15$

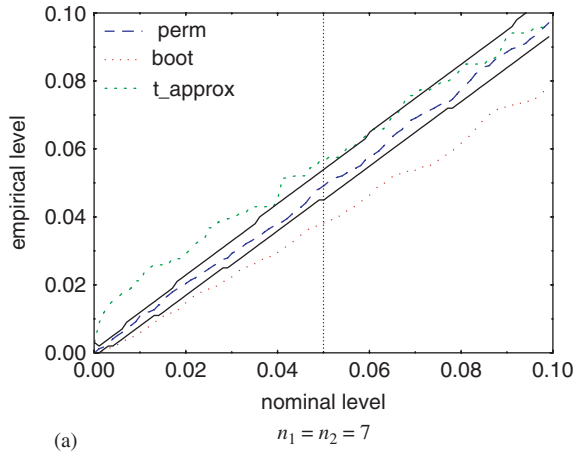


(c) $n_1 = 15, n_2 = 7$

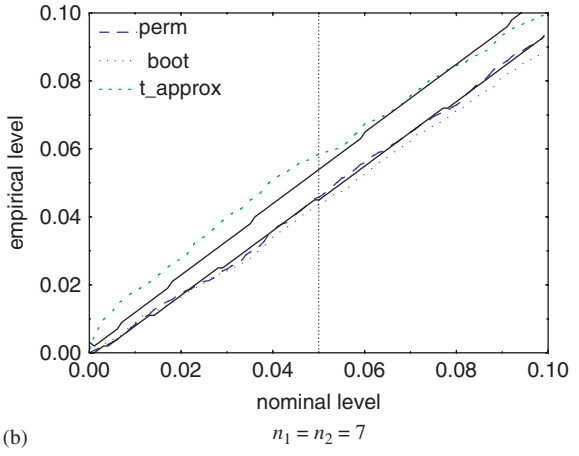
test	$\mu = 0$	$\mu = 0.5$	$\mu = 1$
Φ_{Perm}	**0.053**	0.129	0.288
Φ_{Boot}	**0.049**	0.122	0.269
$\Phi_{t\text{-approx}}$	0.059	0.134	0.274
Φ_{LR}	0.062	0.111	0.221

(d) n_1=15, n_2=7

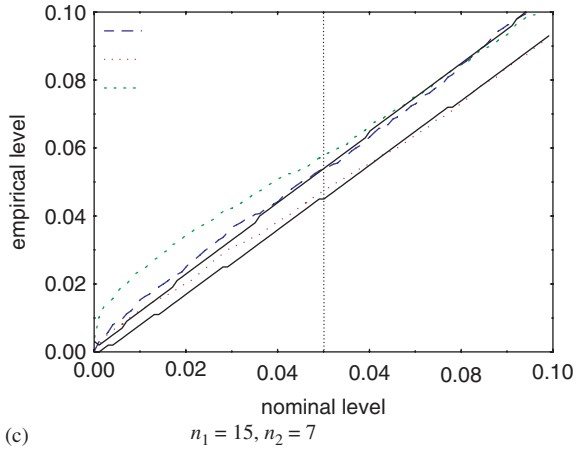
Fig. 4. Two bimodal distributions. Figures show nominal against empirical levels, solid lines indicating the reliability bounds. Simulated levels and power are displayed in (d) for $\alpha = 5\%$, where bold letters indicate levels within the reliability bounds: (a) $n_1 = n_2 = 7$, (b) $n_1 = 7$, $n_2 = 15$, (c) $n_1 = 15$, $n_2 = 7$ and (d) $n_1 = 15$, $n_2 = 7$.



(a) $n_1 = n_2 = 7$



(b) $n_1 = n_2 = 7$



(c) $n_1 = 15, n_2 = 7$

test	μ=0	μ=0.5	μ=1
Φ_{Perm}	**0.054**	0.090	0.175
Φ_{Boot}	**0.047**	0.083	0.163
$\Phi_{\text{t-approx}}$	0.058	0.094	0.175
Φ_{LR}	0.052	0.085	0.159

(d) n_1=15, n_2=7

Fig. 5. One normal against one χ^2_3-distribution. Figures show nominal against empirical levels, solid lines indicating the reliability bounds. Simulated levels and power are displayed in (d) for $\alpha = 5\%$, where bold letters indicate levels within the reliability bounds: (a) $n_1 = n_2 = 7$, (b) $n_1 = 7$, $n_2 = 15$, (c) $n_1 = 15$, $n_2 = 7$ and (d) $n_1 = 15$, $n_2 = 7$.

size 7 in both groups the level curve of the bootstrap test also looks very similar as in the normal case (Fig. 4(a)). For sample sizes 7 in one group and 15 in the other group its curve runs within the reliability bounds (cf. Fig. 4(b) and (c)). The powers given in Fig. 4(d) are calculated with $n_1 = 15$ and $n_2 = 7$. Under H_0 the likelihood ratio test and the *t*-approximation are somewhat liberal. Power values are very close for the permutation and bootstrap test. The likelihood ratio test has the lowest power despite its slight liberality.

Considering one symmetric against one skewed distribution, i.e. one normal against one χ^2-distribution, *t*-approximation and permutation test again have the same properties as with two normal distributions (Fig. 5). The level curve of the bootstrap test is also very similar as in the normal case for sample size 7 in both groups (Fig. 5(a)). For seven observations derived from a normal distribution and 15 observations derived from a χ^2_3-distribution its level curve runs along the lower reliability bound. In the opposite case the level attains values within the reliabilty bounds up to a nominal level of 6% and just below the lower bound afterwards. Fig. 5(d) shows that for $n_1 = 15$ and $n_2 = 7$ only the *t*-approximation is somewhat liberal. When comparing the power all tests achieve similar values.

5. Example

To illustrate the behaviour of the studentized permutation test the pain score data from the shoulder tip pain trial reported by Lumley (1996) were reanalysed. The score takes values from 1 (low pain) up to 5 (high pain) and was

Table 1
Data of the shoulder tip pain trial

Treatment	Pain scores
Y	1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 4, 1, 1
N	3, 3, 4, 3, 1, 2, 3, 1, 1, 5, 4

Table 2
p-values for the shoulder tip pain trial

Φ_{Perm}	Φ_{Boot}	Φ_t-approx	Φ_{LR}
0.0075	0.0019	0.0058	0.0078

measured on the third day after a surgery in 25 female patients. The patients were randomly assigned to two treatments: 14 patients receiving the active treatment (*Y*) and 11 patients receiving a control treatment (*N*). The data are listed in Table 1.

To facilitate comparisons with the likelihood ratio test suggested by Troendle (2002), two-sided tests were used, i.e. the *t*-approximation was recalculated as a two-sided test. As reported in Brunner and Munzel (2000), the rank means are $\overline{R}_{1\cdot} = 9.82$ (treatment) and $\overline{R}_{2\cdot} = 17.05$ (control) and the estimator of the relative effect becomes $\widehat{p} = 0.789$.

All four tests reject the hypothesis $\mathrm{H}_0 : \ p = \frac{1}{2}$ at $\alpha = 5\%$ (Table 2).

6. Conclusion

The studentized permutation test based on the two-sample rank statistic of Brunner and Munzel (2000) presented here seems to be a robust and widely applicable method. Our simulations showed that this permutation test is robust against deviations from symmetry of the underlying distributions, e.g. when the distributions are skewed or bimodal. The level is maintained quite well under the hypothesis and a high power is achieved in alternative settings even if the sample sizes are small (down to 7 per group) or very different (e.g. $n_1 = 7, \ n_2 = 15$). Other methods proposed for small samples have comparable properties but show tendencies to somewhat liberal (*t*-approximation) or conservative (bootstrap test) behaviour, especially for very small sample sizes. These effects may be due to the variance estimators used. For the *t*-approximation, e.g. the variance estimator used is adopted from the parametric case. The bootstrap test of Reiczigel et al. (2005), on the other hand, is based on Welch's statistic using the corresponding variance estimator. Level and power of the likelihood ratio test is comparable to that of the permutation test, however, for underlying bimodal distributions it becomes slightly liberal and for sample size 7 in both samples the maximum likelihood estimator was not computable for up to 173 out of the 10,000 simulation runs. Asymptotically, the permutational distribution of the studentized test statistic is a normal distribution.

Since only trivial assumptions are necessary regarding the underlying distribution of the data, this permutation test is also applicable to count or ordered categorical data. A macro in SAS-IML will be provided on the website http://www.ams.med.uni-goettingen.de/de/sof/.

Appendix A

Following the lines of the derivations in Janssen (1997), it will be shown that the conditions (8)–(10) given in Section 3 are fulfilled by the random variables $Z_{ik} = (1/N)\left(R_{ik} - R_{ik}^{(i)} - (N - n_i)/2\right)$ as defined in (4).

Verification of (8): For the mean we obtain

$$\begin{aligned}\overline{Z}_{\cdot\cdot} &= \frac{1}{N}\sum_i\sum_k Z_{ik} = \frac{1}{N}\left(n_1\overline{Z}_{1\cdot} + n_2\overline{Z}_{2\cdot}\right)\\ &= \frac{1}{N}\left(\frac{1}{N}\sum_i\sum_k R_{ik} - \frac{(n_1+n_2)(N+1)}{2N}\right) = 0. \end{aligned} \tag{A.1}$$

Note that

$$E(Z_{ik}) = \frac{N-n_i}{N}\left(\int F_j \,\mathrm{d}F_i - \frac{1}{2}\right) = \frac{N-n_i}{N}\left(\widetilde{p}_i - \frac{1}{2}\right),$$

$$Var(Z_{ik}) = \left(\frac{N-n_i}{N}\right)^2 Var\left(\widehat{F}_j\left(X_{ik}\right)\right) = \left(\frac{N-n_i}{N}\right)^2 \widetilde{\sigma}_i^2 > 0,$$

where $i \neq j = 1, 2$. Thus, we can write $Z_{ik} = ((N-n_i)/N)\left(\widetilde{p}_i - \frac{1}{2} + \widetilde{\sigma}_i \xi_{ik}\right)$, where ξ_{ik} are iid random variables with expectation 0 and variance 1. Then we obtain

$$\begin{aligned}\frac{1}{N}\sum_i\sum_k Z_{ik}^2 &= \frac{1}{N}\left[\sum_{k=1}^{n_1}\left(\frac{n_2}{N}\right)^2\left(\widetilde{p}_1 - \frac{1}{2} + \widetilde{\sigma}_1\overline{\xi}_{1\cdot}\right)^2 + \sum_{k=1}^{n_2}\left(\frac{n_1}{N}\right)^2\left(\widetilde{p}_2 - \frac{1}{2} + \widetilde{\sigma}_2\overline{\xi}_{2\cdot}\right)^2\right]\\ &\xrightarrow{P\text{-a.s.}} \kappa(1-\kappa)\left[(1-\kappa)\left(\widetilde{p}_1^2 - \widetilde{p}_1 + \frac{1}{4} + \widetilde{\sigma}_1^2\right) + \kappa\left(\widetilde{p}_2^2 - \widetilde{p}_2 + \frac{1}{4} + \widetilde{\sigma}_2^2\right)\right]\\ &=: \widetilde{\rho} \quad \text{as } N \to \infty. \end{aligned} \tag{A.2}$$

To verify (8) it remains to show that the limit $\widetilde{\rho}$ is strictly positive. Note that

$$\widetilde{p}_i^2 - \widetilde{p}_i = \int F_j \,\mathrm{d}F_i\left(\int F_j \,\mathrm{d}F_i - 1\right) = -p(1-p).$$

Thus, it follows that

$$\widetilde{\rho} = \kappa(1-\kappa)\left[(-p)(1-p) + \tfrac{1}{4} + (1-\kappa)\widetilde{\sigma}_1^2 + \kappa\widetilde{\sigma}_2^2\right].$$

Since $p \in [0, \frac{1}{2}] \Rightarrow -p(1-p) + \frac{1}{4} \in [0, \frac{1}{4}]$. Furthermore, $\widetilde{\sigma}_i^2 > 0$ and $\kappa \in (0, 1)$ by assumption such that finally $\widetilde{\rho} > 0$, which proves (8).

Verification of (9): To calculate the conditional permutation distribution of V_N^2, consider the following decomposition $V_N^2 = W_1 - W_2^2 - W_3^2$, where

$$W_1 = \frac{N}{n_2}\frac{1}{n_1-1}\sum_{k=1}^{n_1} Z_{1k}^2 + \frac{N}{n_1}\frac{1}{n_2-1}\sum_{k=1}^{n_2} Z_{2k}^2,$$

$$W_2 = \sqrt{\frac{N}{n_2}\frac{n_1}{n_1-1}}\,\overline{Z}_{1\cdot}, \quad W_3 = \sqrt{\frac{N}{n_1}\frac{n_2}{n_2-1}}\,\overline{Z}_{2\cdot}.$$

Thus, we can derive the distribution for each part separately by calculating the conditional expectations and variances. Let W_i^* denote the above expression calculated from the permuted observations. Starting with W_2^* and W_3^* we obtain by (A.1)

$$E(W_2^*|\mathbf{R}) = \sqrt{\frac{N}{n_2}\frac{n_1}{n_1-1}}\,\overline{Z}_{\cdot\cdot} = 0, \quad E\left(W_3^*|\mathbf{R}\right) = \sqrt{\frac{N}{n_1}\frac{n_2}{n_2-1}}\,\overline{Z}_{\cdot\cdot} = 0.$$

To calculate the variances, we apply the variance formula of Hájek and Ŝidák (1967, p. 61):

$$W_2 = \sum_i\sum_k m_{2ik} Z_{ik} \quad \text{where } m_{2ik} = \begin{cases}\sqrt{\dfrac{N}{n_2}\dfrac{n_1}{n_1-1}}\dfrac{1}{n_1}, & i = 1,\\ 0, & i = 2,\end{cases}$$

$$\begin{aligned} Var(W_2^*|\mathbf{R}) &= \sum_i\sum_k (m_{2ik} - \overline{m}_{2\cdot\cdot})^2 \frac{1}{N-1}\sum_i\sum_k \left(Z_{ik} - \overline{Z}_{\cdot\cdot}\right)^2\\ &= \frac{1}{n_1-1}\frac{N}{N-1}\left(\frac{1}{N}\sum_i\sum_k Z_{ik}^2 - \overline{Z}_{\cdot\cdot}^2\right) \xrightarrow{P\text{-a.s.}} 0 \quad \text{as } N \to \infty. \end{aligned}$$

Thus $W_2^* \to 0$ in $P \otimes P^*$ probability. Analogously we obtain the same result for W_3^*. To derive the conditional distribution for W_1^* we use the following representation:

$$W_1 = \sum_i \sum_k d_{ik} Z_{ik}^2 \quad \text{where } d_{ik} = \begin{cases} \dfrac{N}{n_2}\dfrac{1}{n_1 - 1}, & i = 1, \\ \dfrac{N}{n_1}\dfrac{1}{n_2 - 1}, & i = 2. \end{cases}$$

It is easily seen that, $N\overline{d}_{..} \to 1/(\kappa(1-\kappa))$ as $N \to \infty$. Using (A.2) the conditional expectation of W_1^* is given by

$$E(W_1^*|\mathbf{R}) = \sum_i \sum_k d_{ik} \frac{1}{N} \sum_i \sum_k Z_{ik}^2 \overset{P\text{-a.s.}}{\longrightarrow} \frac{1}{\kappa(1-\kappa)}\widetilde{\rho} \quad \text{as } N \to \infty.$$

Again using (II.3.1.c) in Hájek and Ŝidák (1967), we obtain for the conditional variance

$$Var(W_1^*|\mathbf{R}) = \sum_i \sum_k \left(d_{ik} - \overline{d}_{..}\right)^2 \frac{1}{N-1} \sum_i \sum_k \left(Z_{ik}^2 - \frac{1}{N}\sum_i \sum_k Z_{ik}^2\right)^2.$$

Further, Z_{ik}^4 is bounded by $C = 1/2^4$:

$$Z_{ik}^4 = \left(\frac{N - n_i}{N}\left(\widehat{Y}_{ik} - \frac{1}{2}\right)\right)^4 = \left(\frac{N - n_i}{N}\right)^4 C \leqslant C.$$

Noting that $\sum_i \sum_k \left(d_{ik} - \overline{d}_{..}\right)^2 \to 0$ as $N \to \infty$, we obtain

$$Var(W_1^*|\mathbf{R}) \overset{P\text{-a.s.}}{\longrightarrow} 0 \quad \text{as } N \to \infty.$$

Thus, $W_1^* \to \widetilde{\rho}/(\kappa(1-\kappa))$ in $P \otimes P^*$ probability. Combining these results, it follows that

$$V_N^{*2} = W_1^* - W_2^{*2} - W_3^{*2} \overset{P\otimes P^*}{\longrightarrow} \frac{1}{\kappa(1-\kappa)}\widetilde{\rho} \quad \text{as } N \to \infty.$$

Finally, we obtain the following limit for the ratio of sequences in condition (9):

$$\frac{(1/N)\sum_i \sum_k \left(Z_{ik} - \overline{Z}_{..}\right)^2}{V_N^{*2}} \overset{P\otimes P^*}{\longrightarrow} \kappa(1-\kappa) = \tau^2 \quad \text{as } N \to \infty.$$

Verification of (10): Consider

$$\left|Z_{ik} - \overline{Z}_{..}\right| = |Z_{ik}| \in \left[0, \tfrac{1}{2}\right].$$

Thus, the above expression is bounded by $\frac{1}{2}$ yielding that the limit in (10) is 0 P-a.s. as $d \to \infty$.

References

Babu, G.J., Padmanabhan, A.R., 2002. Resampling methods for the nonparametric Behrens–Fisher problem. Sankhyā Ser. A 64, 678–692.

Brunner, E., Munzel, U., 2000. The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation. Biometrical J. 42, 17–25.

Brunner, E., Munzel, U., 2002. Nichtparametrische Datenanalyse—Unverbundene Stichproben. Springer, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Tokyo.

Chen, M., Kianifard, F., 2000. A nonparametric procedure associated with a clinically meaningful efficacy measure. Biostatistics 1, 293–298.

Cochran, W.B., 1964. Approximate significance levels of the Behrens–Fisher test. Biometrics 20, 191–195.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Fligner, M.A., Policello, G.E., 1981. Robust rank procedures for the Behrens–Fisher problem. J. Amer. Statist. Assoc. 76, 162–178.

Hájek, J.A., Ŝidák, Z., 1967. Theory of Rank Tests. Academic Press, New York.

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

Lumley, T., 1996. Generalized estimating equations for ordinal data: a note on working correlation structures. Biometrics 52, 354–361.
Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18, 50–60.
Moser, B.K., Stevens, G.R., 1992. Homogeneity of variance in the two-sample means test. Amer. Statist. 46, 19–21.
Neuhaus, G., 1993. Conditional rank test for the two-sample problem under random censorship. Ann. Statist. 21, 1760–1779.
Pesarin, F., 2001. Multivariate Permutation Tests. Wiley, Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.
Pratt, J.W., 1964. Robustness of some procedures for the two-sample location problem. J. Amer. Statist. Assoc. 59, 665–680.
Reiczigel, J., Zakariás, I., Rózsa, L., 2005. A bootstrap test of stochastic equality of two populations. Amer. Statist. 59, 156–161.
Ruymgaart, F.H., 1980. A unified approach to the asymptotic distribution theory of certain midrank statistics. In: Raoult, J.P. (Ed.), Statistique non Parametrique Asymptotique. Lecture Notes in Mathematics, vol. 821. Springer, Berlin, pp. 1–18.
Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometrical Bull. 2, 110–114.
Smith, H.F., 1936. The problem of comparing the results of two experiments with unequal error. J. Council Sci. Indust. Res. 9, 211–212.
Troendle, J.F., 2002. A likelihood ratio test for the non-parametric Behrens–Fisher problem. Biometrical J. 44, 813–824.
Welch, B.L., 1937. The significance of the difference between two means when the population variances are unequal. Biometrika 29, 350–362.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1, 80–83.