Stat Comput (2014) 24:283–296
DOI 10.1007/s11222-012-9370-4

Bootstrapping and permuting paired t-test type statistics

Frank Konietschke · Markus Pauly

Received: 15 August 2012 / Accepted: 15 November 2012 / Published online: 8 January 2013

Abstract We study various bootstrap and permutation 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 different permutation approaches as valid procedures for finite sample sizes. It will be shown that permutation or bootstrap schemes, which neglect the dependency structure in the data, are asymptotically valid. Simulation studies show that these new tests improve the power of the t-test under non-normality.

Keywords Bootstrap · Heteroscedasticity · Matched pairs · Permutation tests

1 Introduction

In many psychological, biological and medical experiments, data are collected in terms of a matched pairs design, e.g. when a homogeneous group of subjects is repeatedly observed under two conditions called time points in the terminology of repeated measures designs. Hereby different variances of the observations occur in a natural way, e.g. when data are collected over time. The data of such trials can be modeled by independent and identically distributed random vectors

$$\mathbf{X}_i = (X_{i,1}, X_{i,2})', \quad i = 1, \ldots, n, \tag{1.1}$$

with expectation $E(\mathbf{X}_1) = \boldsymbol{\mu} = (\mu_1, \mu_2)'$ and an arbitrary positive definite covariance matrix $\mathrm{Var}(\mathbf{X}_1) = \boldsymbol{\Sigma}$. Our aim is to test the null hypothesis $H_0 : \mu_1 = \mu_2$, or $H_0^{(1)} : \mu_1 \leq \mu_2$, in this semi-parametric framework.

The paired t-test type statistic $|T_{n,stud}|$ with

$$T_{n,stud} = \sqrt{n}\, \overline{D}_n / V_n \tag{1.2}$$

is the commonly used statistic for testing H_0, where $D_i = X_{i,1} - X_{i,2}$ denote the differences of the pairs for $i = 1, \ldots, n$, $\overline{D}_n = n^{-1} \sum_{i=1}^n D_i = \overline{X}_1 - \overline{X}_2$ is the difference of the means, and $V_n^2 = (n-1)^{-1} \sum_{i=1}^n (D_i - \overline{D}_n)^2$ denotes the sample variance of the D_i's. As commonly known, $T_{n,stud}$ is exactly $T(n-1)$-distributed under H_0, if the differences are normal, even for arbitrary $\boldsymbol{\Sigma}$. Under non-normality, the distribution of $T_{n,stud}$ may be approximated by a $T(n-1)$-distribution, which follows from the central limit theorem. For large sample sizes, the null hypothesis $H_0 : \mu_1 = \mu_2$ will be rejected if $|T_{n,stud}| \geq t_{1-\alpha/2}$, where $t_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile from the $T(n-1)$-distribution. Thus, the t-test can be equivalently written as

$$\varphi_t = \mathbf{1}_{(t_{1-\alpha/2}, \infty)}\big(|T_{n,stud}|\big). \tag{1.3}$$

For testing $H_0^{(1)}$ the t-test φ_t can be redefined by using $T_{n,stud}$ as the test statistic in (1.3) and replacing the critical value $t_{1-\alpha/2}$ by $t_{1-\alpha}$. In a variety of papers and applications, however, it has already been shown that the rate of convergence from $T_{n,stud}$ to its asymptotic normality is rather slow, particularly for skewed distributions of the differences. For a detailed explanation we refer the reader to Munzel (1999).

F. Konietschke (✉)
Department of Medical Statistics, University of Goettingen, Humboldtallee 32, 37073 Goettingen, Germany
e-mail: fkoniet@gwdg.de

M. Pauly
Institute of Mathematics, University of Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany
e-mail: markus.pauly@uni-duesseldorf.de

It is the aim of the present paper to discuss the limit behaviour of various resampling versions of $T_{n,stud}$ to improve its small sample properties under non-normality. Specific examples are all kind of bootstrap and permutation resampling statistics. Although the data may not be exchangeable in model (1.1), an accurate and (asymptotically) valid level α resampling test for H_0 can be derived if (i) the resampling distribution of the statistic is asymptotically independent from the distribution of the data; (ii) the resampling distribution has a limit; and (iii) if the distribution of the test statistic and the conditional resampling distribution (asymptotically) coincide, see Janssen (1997, 1999a, 1999b, 2005), Janssen and Pauls (2003, 2005), Neubert and Brunner (2007), Pauly (2011) or Omelka and Pauly (2012). The items (i)–(iii) will be referred to *the permanence property* of resampling tests.

More details on theory and applications of bootstrap and permutation tests can be found in the monographs of Basso et al. (2009), Good (2005) as well as Pesarin and Salmaso (2010b). Moreover, when comparing more than one aspect of the data, Brombin et al. (2011) also discuss permutation tests for paired observations with an useful application. In particular, permutation approaches for multivariate data are intensively discussed by Pesarin and Salmaso (2012) and Brombin and Salmaso (2009). Both papers provide a detailed summary of existing procedures and some new developments. Regarding repeated measures designs, Pesarin and Salmaso (2010a) apply permutation tests by investigating finite-sample properties.

The intuitive resampling or permutation strategy is to draw the differences with replacement D_i from the data, or to permute the variables $X_{i,1}$ and $X_{i,2}$ within the pairs, respectively. The lack of both resampling schemes is that only a few permutations (2^n) are available, or that a small variety within the resamplings occurs when n is rather small. The counterintuitive resampling or permutation strategies are either drawing the variables $X_{i,s}$ with replacement from all $2n$ observations $X_{1,1},\ldots,X_{n,2}$, drawing the variables $X_{i,s}-\overline{X}_s$ with replacement from each marginal sample $X_{1,s},\ldots,X_{n,s}$, $s=1,2$, separately, or to permute all $2n$ observations in $\mathbf{X}=(X_{1,1},X_{1,2},\ldots,X_{n,2})'$, and then repeatedly compute (e.g. 10,000 times) the paired t-test statistic. On the one hand, these counterintuitive resampling methods increase the resampling variability, on the other hand, the dependency structure within the pairs is neglected. In this paper, it will be shown that both kinds of the intuitive and also the counterintuitive resampling strategies, which neglect the dependency structure in the data, fulfill the permanence property, and thus, the corresponding resampling tests are asymptotically valid. Extensive simulation studies show that especially permutation-based approaches improve the paired t-test, even for extremely small sample sizes. The paper is organized as follows: In Sect. 2 we explain how resampling and permutation tests work and explain in detail why the resulting tests are asymptotically valid. In Sect. 3 extensive simulations are conducted to compare the different resamplings with the paired t-test. The paper closes with a discussion of the results. All technical details and proofs are given in the Appendix.

2 How do paired bootstrap and permutation tests work?

In this section we will study various resampling versions of the paired t-Test. Among others we like to point out why special bootstrap and permutation tests, which neglect the dependency structure of the data within their resampling scheme, are asymptotically valid level α tests for H_0. Let $\mathbf{X}^*=(\mathbf{X}_1^*,\ldots,\mathbf{X}_n^*)'$, with $\mathbf{X}_i^*=(X_{i,1}^*,X_{i,2}^*)$, denote n resampling vectors for $i=1,\ldots,n$, given the original data $\mathbf{X}$, where

(I) $\mathbf{X}^*$ is a random permutation of all data $\mathbf{X}=(X_{1,1},X_{1,2},\ldots,X_{n,2})'$, or
(II) $\mathbf{X}_i^*$ is a random permutation of the sample unit $\mathbf{X}_i'=(X_{i,1},X_{i,2})$, or
(III) $X_{i,s}^*$ is randomly drawn with replacement from all data $\mathbf{X}$, or
(IV) $X_{i,s}^*$ is randomly drawn with replacement from each centered marginal sample $\mathbf{X}_s=(X_{1,s}-\overline{X}_s,\ldots,X_{n,s}-\overline{X}_s)'$, $s=1,2$, respectively.

The conditional resampling statistic of $T_{n,stud}$ is then given by

$$T_{n,stud}^*=\sqrt{n}\,\overline{D}_n^*/V_n^*, \tag{2.1}$$

where $D_i^*=X_{i,1}^*-X_{i,2}^*$ denotes the differences of the resampling variables for $i=1,\ldots,n$, $\overline{D}_n^*=n^{-1}\sum_{i=1}^n D_i^*$ denotes their mean, and $V_n^{*2}=(n-1)^{-1}\sum_{i=1}^n(D_i^*-\overline{D}_n^*)^2$ denotes the sample variance of the differences D_i^*.

Here we like to point out that the denominator in (2.1) is part of the resampling procedures, which is in accordance with the guidelines for bootstrap testing, see Hall and Wilson (1991), Beran (1997), Bickel and Freedman (1981), and Janssen (2005). Delaigle et al. (2011) have further shown that studentized resampling t-statistics are more robust and accurate than non-studentized statistics. The following gives an explanation how the corresponding resampling tests can be computed.

The introduced conditional resampling tests rely on a reference distribution $\mathcal{L}(T_{n,stud}^*|\mathbf{X})$ given the data $\mathbf{X}$. This means that the data are treated as fixed values, and quantiles from the conditional resampling distribution of $T_{n,stud}^*$ are estimated to compute critical values. Denote by $c_n^*(1-\alpha)$ the $(1-\alpha)$-quantile of $\mathcal{L}(T_{n,stud}^*|\mathbf{X})$. Then, according to the

definition of the paired t-test in (1.3), conditional resampling tests can be written as

$$\varphi_n^* = \mathbf{1}_{(-\infty, c_n^*(\alpha/2))}(T_{n,stud}) + \mathbf{1}_{(c_n^*(1-\alpha/2),\infty)}(T_{n,stud}). \quad (2.2)$$

Next we will prove that $T_{n,stud}^*$ as given in (2.1) is asymptotically standard normal under all of the different resampling schemes described above. In particular, we will show that the permanence property is fulfilled, thus, φ_n^* is an asymptotically valid test for H_0. Its asymptotic normality is particularly derived under arbitrary alternatives, i.e. we do not assume that H_0 is true. To give an answer to the question "How do paired Bootstrap and Permutation tests work?" we will introduce the following criterion from Janssen and Pauly (2010), which uses the paired t-test as a benchmark for the resampling procedures.

Definition 2.1 The conditional tests φ_n^* defined in (2.2) are called

(i) *asymptotically effective under H_0 with respect to the paired t-test*, iff

$$E\left(\left|\varphi_n^* - \varphi_t\right|\right) \longrightarrow 0 \quad \text{as } n \to \infty, \quad \text{and} \quad (2.3)$$

(ii) *consistent* iff

$$E\left(\varphi_n^*\right) \longrightarrow \mathbf{1}\{\mu_1 > \mu_2\} \quad (2.4)$$

for $\mu_1 \neq \mu_2$ as $n \to \infty$.

Now we can formulate

Theorem 2.1 *The resampling tests φ_n^* defined in (2.1) are asymptotically effective with respect to φ_t and consistent under all resampling schemes* (I) *through* (IV).

From the proof it can be seen that a similar result also holds for one-sided versions of the tests. For further details see the Appendix. Specifically, Theorem 2.1 shows that the counterintuitive resampling procedures (I), (III) and (IV) are asymptotically valid, because studentized statistics are resampled. Roughly speaking, the studentization of the resampling variables "deletes" the dependency structure in the data when n is sufficiently large.

2.1 Resampling the differences $\mathbf{D}_i$

In this subsection we will also introduce resampling methods, particularly wild boostrap methods, which are based on the differences D_i. The wild bootstrap technique is motivated by the residual bootstrap commonly applied in regression analysis, see Wu (1986), Mammen (1992) and Beran (1997), and in time-series testing problems, see Kreiss and Paparoditis (2011). It is also proposed in the context of survival analysis, see Lin (1997) or Beyersmann et al. (2012). Here, we adapt the wild bootstrap to the simple matched pairs design and we will compare the accuracy of the resulting test procedures with the resampling tests in (2.1) in extensive simulation studies. Let D_i^* denote n resampling variables given the original differences $\mathbf{D} = (D_1, \ldots, D_n)'$, where D_i^* denotes the observed value from

(V) drawing with replacement from all differences $\mathbf{D}$, or

(VI) from a wild bootstrap method with $D_i^* = W_i D_i$, where $W_i, i = 1, \ldots, n$, denote independent and identically distributed random variables, which are independent from the D_i's, with $E(W_1) = 0$ and $\mathrm{Var}(W_1) = 1$.

The corresponding resampling tests are then defined as in (2.2) with the paired t-test type resampling statistic

$$T_{n,stud}^* = \sqrt{n}\, \overline{D}_n^* / V_n^*, \quad (2.5)$$

where now $\overline{D}_n^* = n^{-1} \sum_{i=1}^n D_i^*$ denotes the mean of the resampled differences, and $V_n^{*2} = (n-1)^{-1} \sum_{i=1}^n (D_i^* - \overline{D}_n^*)^2$ denotes the sample variance of the D_i^*'s. The effectiveness of these resampling procedures is given in the next theorem.

Theorem 2.2 *The resampling tests φ_n^* defined in (2.5) are asymptotically effective with respect to φ_t and consistent under both resampling schemes* (V) *and* (VI).

Example and Remark 2.1 In our simulation study in Sect. 3, we will focus on the following weight examples. However, there are of course others that may be of interest for particular situations.

(a) $W_i, i = 1, \ldots, n$ is a sequence of symmetric independent and identically distributed random variables with

$$P\left(W_1 = \frac{1+\sqrt{5}}{2}\right) = \frac{\sqrt{5}-1}{2\sqrt{5}} \quad \text{and}$$
$$P\left(W_1 = \frac{1-\sqrt{5}}{2}\right) = \frac{\sqrt{5}+1}{2\sqrt{5}}.$$

In this case it even holds that $E(W_1^3) = 1$. These wild bootstrap weights are typically used for studentized test statistics, see e.g. Kreiss and Paparoditis (2011). We will call the corresponding test *Rademacher* wild bootstrap.

(b) $W_i, i = 1, \ldots, n$, is a sequence of independent and identically distributed Gaussian random variables, i.e. $W_i \sim N(0, 1)$. This corresponds to the resampling procedure proposed by Lin (1997).

We note that Arlot et al. (2010a, 2010b) investigate wild bootstrap methods for multiple comparisons and confidence intervals in high-dimensional data using random signs $W_i, i = 1, \ldots, n$, with distribution $P(W_1 = -1) = P(W_1 = 1) = 1/2$. This resampling method, however, is equivalent

to the resampling scheme (II). For further details we refer the reader to Janssen (1999b).

Theorems 2.1 and 2.2 state that all the considered procedures fulfill the permanence property, thus, the corresponding tests φ_n^* are asymptotically valid. The numerical algorithm for the computation of the p-value is as follows

(1) Given the data $\mathbf{X}$, compute the paired t-test statistic $T_{n,stud}$ as given in (1.2).
(2) Repeat the resampling steps N times (e.g. $N = 10{,}000$), compute the values $T_{n,stud}^*$ and save them in $A_1, \ldots, A_N$.
(3) Estimate the two-sided p-value by

$$\text{p-value} = \min\{2p_1, 2 - 2p_1\},$$

$$\text{where } p_1 = \frac{1}{N}\sum_{\ell=1}^{N} \mathbf{1}\{A_\ell \le T_{n,stud}\}.$$

In comparison to that the one-sided p-value is given by p_1.

3 Simulations

For testing the two-sided null hypothesis $H_0 : \mu_1 = \mu_2$ formulated above, we consider the unconditional t-test φ_t based on the $T(n-1)$-approximation of the statistic $T_{n,stud}$ in (1.2) and the various conditional resampling tests φ_n^* based on the resampling schemes (I) through (VI) as described in Sect. 2. The simulation studies are performed to investigate their behaviour with regard to maintaining the pre-assigned type-I error level under the hypothesis, and the power of the statistics under alternatives. The observations $\mathbf{X}_i = (X_{i,1}, X_{i,2})', i = 1, \ldots, n$, were generated using marginal distributions F_s and varying correlations $\rho \in (-1, 1)$. We hereby generate exchangeable matched pairs having a bivariate normal, exponential, log-normal or uniform distribution, each with correlation $\rho \in (-1, 1)$, as well as non-exchangeable data by simulating

(a) $F_1 = N(0, 1)$ and $F_2 = N(0, 2)$,
(b) $F_1 = N(0, 1)$ and $F_2 = N(0, 4)$,
(c) $F_1 = N(3, 4)$ and $F_2 = \chi_3^2$, and
(d) $F_1 = N(\exp(0.5), 3)$ and $F_2 = LN(0, 1)$,

each with correlation ρ, respectively. Routine calculations show that $\mu_1 = \mu_2$ is fulfilled in all of these considerations. 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 (www.r-project.org), each with $nsim = 10{,}000$ and $N = 10,000$ bootstrap runs. The simulation results for exchangeable normally, exponentially, log-normally, and uniformly distributed matched pairs with the very small sample size of $n = 7$ and different correlations ρ are displayed in Table 1.

It follows from Table 1 that the paired t-test is an accurate procedure for symmetric distributions (normal and uniform), even for the very small sample size of $n = 7$. When the data are skewed (exponential and log-normal), the t-test tends to be conservative. It is apparent that both the wild bootstrap methods using the Rademacher weights as defined in Remark 2.1(b) and the Gaussian weights given in Remark 2.1(c) are inappropriate tests for such small sample sizes. The resampling test with Rademacher weights is very liberal. This can be explained by the fact that these weights are very skewed distributed. Roughly speaking, both wild bootstrap resampling distributions are too far away from the distribution of $T_{n,stud}$, when n is rather small and the original data are not resampled. Simply drawing the differences from the data with replacement can not be recommended either. The corresponding test tends to be quite liberal when the data are skewed. This occurs, because the resampling variability (i.e. the variability within the resampling variables D_i^*) is rather small when $n = 7$. However, drawing with replacement from either all $2n$ observations or from each marginal separately, results in more accurate test decisions. Comparing these results with the permutation based approaches, it is easily seen that both kind of permutation tests (i.e. to permute all data, or to permute within the sample unit) control the type-I error level for all distributions and all dependencies ρ in the data. Next we investigate the behaviour of the different resampling tests for larger $n = 10$. The simulation results are displayed in Table 2.

From Table 2 an interesting phenomenon of replacement procedures with resampling scheme (IV) and (V) can be observed: The rejection rates do not converge linearly in n to α. The tests are more liberal with $n = 10$ than with $n = 7$. Their liberality increases with an increasing n up to the breakpoint $n \approx 15$. With larger n (e.g. $n \ge 30$), all resampling test based on drawing with replacements are accurate. The liberality of the wild bootstrap tests using Rademacher or Gaussian weights decrease. Both kind of permutation approaches, however, are still the most accurate procedures.

Now we investigate how accurate the tests control the type-I error level when both marginal distributions are different. The simulation results for different non-exchangeable distributions (a) through (d) with $n = 7$ and varying correlations are displayed in Table 3.

It follows from Table 3 that both the permutation approaches are accurate, even for non-exchangeable distributions, $n = 7$, and permutations of all data $\mathbf{X}$. When two distributions with extremely different shapes and negative correlations (normal versus log-normal) are compared, they tend to be slightly liberal. The same conclusions, however, can be drawn for the t-test. In Table 4 the simulation results for $n = 10$ and the same non-exchangeable distributions are given.

Table 1 Type-I error level ($\alpha = 5\ \%$) simulations for very small sample sizes ($n = 7$) with exchangeable distributions

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
Normal	−0.90	4.64	4.58	4.98	4.70	4.45	4.95	14.61	7.14
	−0.50	4.82	4.69	5.04	4.90	4.93	5.07	15.18	7.24
	−0.30	4.82	4.74	5.05	4.80	5.14	5.04	14.84	7.46
	0.00	4.97	5.01	4.80	5.13	5.02	5.23	14.55	7.53
	0.30	5.12	5.17	5.16	5.20	4.67	5.26	15.03	7.52
	0.50	5.04	5.02	5.35	5.02	4.89	5.16	15.22	7.46
	0.90	4.82	4.77	5.10	4.87	5.08	5.32	15.13	7.40
Exp	−0.90	4.24	5.68	5.06	5.10	6.80	5.65	15.45	7.25
	−0.50	4.22	5.39	5.41	5.01	6.80	6.34	16.26	7.22
	−0.30	4.21	5.55	5.40	4.99	5.77	6.42	16.21	7.64
	0.00	3.99	5.25	5.42	4.69	5.82	6.64	16.84	7.47
	0.30	3.68	5.00	5.28	4.53	5.99	7.05	17.53	7.60
	0.50	3.50	4.38	5.34	4.21	5.61	7.12	17.36	7.73
	0.90	3.42	4.12	5.44	4.13	5.00	7.09	17.89	7.32
Uniform	−0.90	5.79	5.27	5.09	5.69	4.12	4.02	13.82	7.31
	−0.50	4.95	4.68	4.52	4.71	3.21	3.73	13.54	7.03
	−0.30	5.44	5.29	5.32	5.27	3.60	4.76	14.34	7.37
	0.00	5.17	4.96	5.12	5.07	3.04	4.86	14.69	7.65
	0.30	4.80	4.48	5.39	4.77	3.62	5.04	15.70	7.58
	0.50	4.54	4.34	5.02	4.60	3.35	5.64	15.06	7.50
	0.90	4.49	4.15	5.17	4.38	3.73	6.25	16.27	7.54
LNorm	−0.90	3.65	5.18	5.01	5.04	7.72	6.66	16.95	6.83
	−0.50	3.02	5.24	5.23	4.37	7.07	7.16	17.60	6.91
	−0.30	3.11	5.37	5.41	4.07	7.07	6.82	18.28	6.53
	0.00	2.67	4.78	5.02	3.80	5.90	6.92	18.13	6.49
	0.30	2.51	4.23	5.01	3.63	5.79	6.43	18.02	6.18
	0.50	2.64	4.78	5.33	3.67	5.27	7.41	18.60	7.00
	0.90	2.40	4.91	5.21	2.92	4.11	6.82	18.71	6.46

Table 2 Type-I error level ($\alpha = 5\ \%$) simulations for moderate sample sizes ($n = 10$) with exchangeable distributions

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
Normal	−0.90	4.58	4.56	4.89	4.70	5.14	4.44	12.13	6.08
	−0.50	4.76	4.73	4.77	4.83	4.81	4.76	12.67	6.46
	−0.30	4.98	5.03	5.32	5.03	4.76	5.13	12.95	6.88
	0.00	4.33	4.30	4.15	4.35	4.90	4.63	12.21	5.77
	0.30	5.08	5.02	4.85	5.13	4.91	4.99	12.54	6.71
	0.50	4.63	4.65	4.50	4.64	5.25	4.76	12.03	6.28
	0.90	4.83	4.86	4.98	4.80	5.35	5.00	12.08	6.53
Exp	−0.90	4.58	5.49	5.11	4.95	7.27	6.20	13.91	6.65
	−0.50	4.67	5.47	5.27	5.23	7.76	6.57	13.59	6.61
	−0.30	4.49	5.27	5.17	5.08	7.10	7.00	14.49	6.67
	0.00	4.07	5.06	4.85	4.62	7.08	7.39	14.42	6.61
	0.30	3.93	4.60	4.80	4.45	6.94	7.36	14.17	6.44
	0.50	3.86	4.44	4.64	4.24	6.24	7.74	14.79	6.55
	0.90	3.99	4.18	5.12	4.25	5.10	8.20	15.02	7.00
Uniform	−0.90	5.39	5.38	5.08	5.27	3.67	3.13	11.54	6.46
	−0.50	5.61	5.38	5.25	5.51	3.53	3.65	11.84	6.57
	−0.30	5.51	5.26	5.25	5.31	3.44	3.97	11.99	6.79
	0.00	5.09	4.83	4.78	5.01	3.70	4.67	12.10	6.99
	0.30	5.12	5.00	4.99	4.96	3.69	5.16	12.93	6.60
	0.50	4.79	4.69	4.69	4.72	4.13	5.34	12.53	6.35
	0.90	4.68	4.57	5.01	4.58	3.66	6.48	13.77	6.86
LNorm	−0.90	3.51	5.50	4.81	4.53	8.92	7.61	14.61	6.23
	−0.50	3.18	4.93	4.58	4.15	8.93	7.72	14.42	5.88
	−0.30	3.33	5.39	5.07	4.37	8.60	8.48	15.35	6.40
	0.00	3.02	5.03	5.02	3.99	8.24	8.87	15.92	6.28
	0.30	2.89	4.21	4.86	3.67	7.97	8.71	16.46	6.18
	0.50	2.87	4.42	5.00	3.61	7.31	8.43	15.90	5.96
	0.90	2.82	4.40	4.97	3.43	5.27	8.46	16.54	6.06

Table 3 Type-I error level ($\alpha = 5\ \%$) simulations for very small sample sizes ($n = 7$) with non-exchangeable distributions (a) through (d) as described in the text

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
(a)	−0.90	5.21	5.21	5.30	5.26	5.12	5.76	15.31	7.90
	−0.50	4.84	4.96	5.47	5.00	5.20	5.01	15.23	7.56
	−0.30	4.71	4.75	5.04	4.87	5.35	5.06	14.95	6.92
	0.00	4.72	4.65	4.87	4.83	4.98	4.86	14.60	7.11
	0.30	5.15	5.19	5.03	5.14	4.92	5.47	14.82	7.51
	0.50	4.97	5.00	5.28	4.95	5.15	5.58	15.64	7.68
	0.90	4.90	4.88	5.10	4.94	5.00	5.62	15.24	7.62
(b)	−0.90	4.93	5.01	5.10	5.12	5.06	4.83	15.58	7.19
	−0.50	4.82	4.71	4.94	5.00	4.89	5.08	14.63	7.19
	−0.30	5.13	5.16	5.07	5.39	4.97	5.18	14.78	7.65
	0.00	5.12	5.41	5.42	5.51	4.83	5.24	15.65	7.75
	0.30	5.18	5.25	5.32	5.20	5.02	5.24	14.77	7.54
	0.50	4.93	4.97	5.28	5.12	5.06	5.03	15.92	7.76
	0.90	5.09	5.35	5.50	5.43	5.00	5.42	14.96	7.75
(c)	−0.90	6.14	6.63	6.22	6.49	6.51	5.86	16.93	8.76
	−0.50	5.85	5.99	5.76	6.11	5.92	5.73	15.66	8.29
	−0.30	5.44	5.80	5.66	5.86	5.51	5.85	15.58	7.86
	0.00	5.10	5.31	5.50	5.45	5.52	5.80	15.45	7.73
	0.30	5.36	5.67	5.19	5.54	5.44	6.05	16.08	8.03
	0.50	5.11	5.40	5.29	5.37	5.64	5.63	15.42	7.66
	0.90	5.50	5.53	5.73	5.56	5.16	6.15	16.02	8.17
(d)	−0.90	6.94	7.51	6.94	7.48	6.91	6.84	17.18	9.32
	−0.50	6.06	6.51	5.95	6.56	6.44	6.55	16.59	8.87
	−0.30	5.53	5.94	5.35	6.16	6.26	6.11	15.85	7.96
	0.00	5.42	6.01	5.43	6.07	6.22	6.06	16.09	7.93
	0.30	5.04	5.71	5.36	5.36	5.50	5.82	15.79	7.51
	0.50	5.00	5.45	4.95	5.38	5.51	5.39	15.32	7.50
	0.90	5.91	6.13	5.17	6.15	6.19	5.98	16.56	8.48

Table 4 Type-I error level ($\alpha = 5\ \%$) simulations for moderate sample sizes ($n = 10$) with non-exchangeable distributions (a) through (d) as described in the text

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
(a)	−0.90	4.99	4.91	4.90	5.06	4.96	4.90	12.88	6.57
	−0.50	4.94	4.91	4.92	4.92	4.70	4.98	12.42	6.31
	−0.30	4.82	4.77	4.75	4.70	5.11	4.75	11.74	6.28
	0.00	4.61	4.71	4.63	4.70	4.95	4.66	12.72	5.90
	0.30	4.97	5.09	4.92	4.94	4.66	5.08	12.93	6.42
	0.50	4.96	4.95	4.74	5.06	5.03	5.17	12.35	6.50
	0.90	4.86	4.91	5.02	4.97	5.35	4.90	12.38	6.48
(b)	−0.90	4.89	4.98	4.86	4.88	4.16	5.18	12.61	6.53
	−0.50	5.11	5.20	4.99	5.12	5.46	5.17	13.00	6.62
	−0.30	4.80	4.97	4.56	5.05	4.56	4.66	12.37	6.18
	0.00	5.13	5.31	5.12	5.26	4.99	5.00	13.09	6.81
	0.30	4.76	4.78	4.49	4.86	5.15	4.82	12.60	6.22
	0.50	5.05	5.13	4.82	5.18	4.66	5.09	13.06	6.66
	0.90	4.98	5.02	4.77	5.15	5.20	4.71	12.47	6.45
(c)	−0.90	5.83	5.85	5.81	5.75	6.53	5.22	13.51	7.23
	−0.50	6.22	6.47	6.18	6.14	5.81	6.06	14.49	8.02
	−0.30	5.82	6.09	5.80	6.07	5.70	5.63	13.74	7.36
	0.00	5.42	5.75	5.39	5.68	5.25	5.88	13.71	7.61
	0.30	5.09	5.27	4.96	5.14	5.37	5.42	13.26	6.64
	0.50	5.11	5.24	5.15	5.09	5.17	5.52	12.99	6.82
	0.90	5.38	5.57	5.41	5.50	5.41	5.80	13.89	7.30
(d)	−0.90	6.38	6.45	6.29	6.40	7.07	6.27	14.62	7.89
	−0.50	6.02	6.41	6.27	6.12	6.07	6.09	14.02	7.63
	−0.30	5.66	6.06	5.89	5.87	6.19	6.25	14.57	7.79
	0.00	5.11	5.43	5.50	5.33	6.23	5.59	13.43	6.93
	0.30	4.90	5.12	5.05	5.16	5.94	5.16	13.21	6.43
	0.50	4.82	5.03	4.95	4.97	5.40	5.60	13.24	6.48
	0.90	6.17	6.32	6.17	6.26	6.22	5.84	14.48	7.83

Table 5 Power ($\alpha = 5\,\%$) simulations for moderate sample sizes ($n = 10$) and $\rho = 1/2$

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
Normal	0.00	4.91	4.85	5.00	5.00	4.79	5.01	12.99	6.61
	0.10	5.75	5.74	5.56	5.54	5.88	5.84	14.48	7.65
	0.20	8.85	8.82	8.74	8.76	8.73	8.33	18.67	10.95
	0.30	14.41	14.30	14.03	14.23	13.99	13.18	27.46	17.55
	0.40	20.44	20.47	20.11	20.39	20.28	19.50	37.07	24.21
	0.50	28.78	28.71	28.34	28.81	28.45	27.25	47.83	33.98
	0.60	39.80	39.60	39.04	39.87	38.97	36.96	60.28	45.34
	0.70	51.24	51.20	50.61	51.17	50.29	47.47	70.98	56.96
	0.80	60.12	60.08	58.79	59.82	59.42	55.71	78.51	65.81
	0.90	72.13	71.89	71.07	72.34	70.78	66.49	86.86	77.03
	1.00	80.64	80.60	80.18	80.85	79.69	74.49	92.27	84.28
LNorm	0.00	2.90	4.15	4.66	3.69	7.29	8.43	15.69	6.07
	0.10	3.50	4.66	5.75	4.19	7.99	9.35	17.21	6.90
	0.20	5.18	7.18	8.26	6.51	10.84	12.06	19.95	9.43
	0.30	7.97	10.68	11.70	9.70	14.44	15.92	24.75	13.05
	0.40	11.11	14.21	15.74	12.93	18.59	19.55	29.56	16.62
	0.50	16.26	20.09	21.38	18.40	23.71	24.70	34.85	22.44
	0.60	21.46	26.15	27.36	24.19	29.47	30.20	40.48	28.50
	0.70	27.60	32.56	33.50	30.38	35.45	35.85	47.83	34.97
	0.80	33.03	37.92	38.63	35.54	40.01	39.69	52.45	39.62
	0.90	37.24	42.23	43.08	40.05	43.5	43.53	55.89	43.71
	1.00	43.94	49.24	49.36	46.90	49.47	48.32	61.85	51.21

Table 6 Power ($\alpha = 5\,\%$) simulations for moderate sample sizes ($n = 20$) and $\rho = 1/2$

Distribution	ρ	$T_{n,stud}$	Permutation tests		Bootstrap tests			Wild bootstrap (VI)	
			Overall (I)	Per unit (II)	Overall (III)	Marginal (IV)	Differences (V)	Rademacher	Normal
Normal	0.00	5.04	4.98	4.95	5.00	4.83	4.84	9.49	5.68
	0.10	6.84	6.81	6.99	6.78	6.65	6.72	12.23	7.76
	0.20	13.58	13.45	13.47	13.57	13.27	13.24	21.25	14.68
	0.30	24.03	24.07	23.97	23.90	24.03	23.52	35.03	25.98
	0.40	39.84	39.51	39.99	39.87	39.40	38.37	52.23	42.28
	0.50	56.02	56.21	55.70	55.92	55.88	54.35	67.88	58.51
	0.60	72.30	72.15	72.13	72.12	71.77	71.50	82.33	74.73
	0.70	84.72	84.75	84.57	84.74	84.67	83.78	91.16	86.24
	0.80	92.21	92.12	92.21	92.26	91.92	91.34	96.21	93.25
	0.90	96.96	96.93	96.94	96.95	96.73	96.34	98.70	97.38
	1.00	98.97	99.01	98.98	98.94	98.90	98.57	99.64	99.09
LNorm	0.00	3.47	4.70	5.20	4.28	8.34	9.38	13.72	5.73
	0.10	4.32	5.59	6.01	5.18	9.64	10.78	14.78	6.90
	0.20	7.33	9.10	9.90	8.24	13.64	14.69	18.91	10.69
	0.30	11.83	13.80	14.35	13.00	18.09	19.11	24.72	15.44
	0.40	17.60	20.15	21.26	19.22	24.50	25.59	31.92	22.38
	0.50	24.23	27.32	28.19	26.17	30.81	31.30	38.34	29.21
	0.60	31.55	34.60	35.31	33.37	37.69	37.97	46.48	36.60
	0.70	39.55	42.54	43.07	41.39	44.37	44.55	52.78	44.57
	0.80	47.33	50.23	50.61	49.26	50.33	49.97	59.55	51.91
	0.90	55.02	57.81	58.39	56.86	57.19	56.51	66.15	59.26
	1.00	60.97	63.91	64.33	62.96	61.60	60.62	71.13	65.38

For larger $n = 10$, both permutation approaches are accurate and demonstrate a similar behaviour to the t-test.

To compare the power of the tests, we generate bivariate normally and log-normally distributed matched pairs with $n = 10$ and $n = 20$, respectively, each with correlation $\rho = 1/2$. Hereby, we shifted the data under time-point 2 with $\delta \in (0, 1)$. The simulation results for $n = 10$ are displayed in Table 5. Although both the wild bootstrap methods using Rademacher and Gaussian weights, as well as the resampling tests based on scheme (III)–(V) were quite liberal in these situations, we included them in the power simulation study. However, to give a fair comparison between the procedures, we will not grade them in detail and concentrate on the t-test and the permutation based approaches.

It follows from Table 5 that both the permutation approaches have a comparable power to the t-test under normality. Under non-normality, the power of the permutation based approaches is remarkably higher. The same conclusions can be drawn for $n = 20$, as can be seen from Table 6.

4 Discussion

We analyzed two different permutation approaches for testing $H_0 : \mu_1 = \mu_2$ with paired data under non-normality. Particularly, we demonstrated that the usual assumption of exchangeability is not necessary for the construction of permutation tests. We have analytically shown that permutation approaches, which are based on permutations of all observed data (i.e. neglecting the dependency structure), are asymptotically valid procedures. The results are obtained by investigating the conditional permutation distribution of studentized statistics. All results in this paper would not hold without the studentization. The investigation of permutation techniques in heteroscedastic repeated measures designs will be part of future research.

In this paper, only mean based approaches were considered. Rank-based studentized permutation tests are proposed by Konietschke and Pauly (2012).

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

Appendix

The next Lemma explains that it suffices to analyze the limit of the conditional distribution for proving all theorems. In the sequel '$\xrightarrow{P}$' will denote convergence in probability as $n \to \infty$.

Lemma 5.1 *Let φ_n^* be one of the resampling tests* (I)–(VI) *defined as in* (2.2). *If we have convergence*

$$c_n^*(1-\alpha) \xrightarrow{P} \Phi^{-1}(1-\alpha) \tag{5.1}$$

for all $\alpha \in (0, 1)$ and general $\boldsymbol{\mu} \in \mathbb{R}^2$, the test φ_n^* is asymptotically effective with respect to φ_t and consistent.

Proof For completeness we start by giving a short proof for the asymptotic exactness of φ_t: By the multivariate central limit theorem we have for $E(\boldsymbol{X}_1) = \boldsymbol{\mu}$ convergence in distribution

$$\boldsymbol{S}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \boldsymbol{Y} = (Y_1, Y_2)' \sim N(0, \boldsymbol{\Sigma}).$$

Since $T_n = (1, -1)\boldsymbol{S}_n$ is a linear transformation of $\boldsymbol{S}_n$ we get from the continuous mapping theorem and Polya's Theorem that under $H_0 : \mu_1 = \mu_2$ $\sup_{x \in \mathbb{R}} |P(T_n \le x) - \Phi(x/\sigma)| \to 0$, where $\sigma^2 = (1, -1)\boldsymbol{\Sigma}(1, -1)^T = \sigma_1^2 - 2\sigma_{12} + \sigma_2^2 = \mathrm{Var}(D_1)$ with $\sigma_j^2 := \mathrm{Var}(X_{1,j})$, $j = 1, 2$ and $\sigma_{12} = Cov(X_{1,1}, X_{1,2})$. Since V_n^2 is a consistent estimator of σ^2 the result follows from Slutzky's Theorem.

Note that by (5.1) Lemma 1 in Janssen and Pauls (2003) implies (2.3). Moreover, since the convergence (5.1) also holds under alternatives $\mu_1 \neq \mu_2$, the result follows from the convergence

$$T_{n,stud}(\mathbf{X}) = T_{n,stud}\big((X_i - \mu)_{1 \le i \le n}\big) + \sqrt{n}\frac{(\mu_1 - \mu_2)}{V_n}$$
$$\xrightarrow{P} \mathrm{sign}(\mu_1 - \mu_2)\infty. \qquad \square$$

In the following we will apply Lemma 5.1. Note, that in order to prove (5.1) it suffices to show that the conditional resampling distribution converges weakly to a standard normal distribution in probability, i.e.

$$\sup_{x \in \mathbb{R}} \big| P\big(T_{n,stud}^* \le x\big) - \Phi(x) \big| \xrightarrow{P} 0. \tag{5.2}$$

Proof of Theorem 2.1 We start by analyzing the resampling scheme (I) which is based on permuting the pooled sample:

Let $Z_1, \ldots, Z_{2n}$ with $Z_i = X_{i,1}$ for $1 \le i \le n$ and $Z_i = X_{i,2}$ for $n+1 \le i \le 2n$ denote the pooled sample. For studying the permutation test based on the resampling scheme (I) let π be a random permutation of $(1, \ldots, 2n)$, i.e. a random variable that is uniformly distributed on the symmetric group $\mathcal{S}_{2n}$, that is independent from $\mathbf{X}$. Consider the modified studentized version of $T_{n,stud}$

$$\tilde{T}_{n,stud} := \sqrt{n}\,\overline{D}_n \Bigg((n-1)^{-1} \Bigg[\sum_{i=1}^{n} (X_{i,1} - \overline{X}_1)^2$$

$$+\sum_{i=1}^{n}(X_{i,2}-\overline{X}_2)^2\Bigg]\Bigg)^{-1/2}$$
$$=:\frac{T_n}{\tilde{V}_n},$$

see Eqs. (4.1)–(4.2) in Janssen (2005). We will complete the proof for (I) as follows: First we prove that the permutation version of $\frac{T_n}{\tilde{V}_n}$ derived from (I) fulfills (5.2). After that we argue that $T^*_{n,stud}$ has the same asymptotic limit behaviour by discussing the different permutation versions of the standardizations. For the first part we apply Theorem 4.1. in Janssen (2005). Note, that we have convergences in probability

$$\frac{1}{2n}\sum_{i=1}^{2n}\left(Z_i-\frac{1}{2n}\sum_{j=1}^{2n}Z_j\right)^2$$
$$=\frac{1}{2n}\sum_{i=1}^{n}X_{i,1}^2+\frac{1}{2n}\sum_{i=1}^{n}X_{i,2}^2$$
$$-\left(\frac{1}{2n}\sum_{i=1}^{n}X_{i,1}+\frac{1}{2n}\sum_{i=1}^{n}X_{i,2}\right)^2$$
$$\xrightarrow{P}\frac{1}{2}\left(\sigma_1^2+\sigma_2^2\right)+\frac{1}{4}(\mu_1+\mu_2)^2>0,$$

by the law of large numbers, and

$$\frac{1}{\sqrt{2n}}\max_{1\le i\le 2n}\left|Z_i-\frac{1}{2n}\sum_{j=1}^{2n}Z_j\right|$$
$$\le\frac{4}{\sqrt{n}}\left(\max_{1\le i\le n}|X_{i,1}|+\max_{1\le i\le n}|X_{i,2}|\right)\xrightarrow{P}0, \tag{5.3}$$

by the fact that $(X_{i,j}/\sqrt{n})_{1\le i\le n}$ fulfill the Lindeberg condition for each $j=1,2$, which is more restrictive. Hence Condition (1.12) in his paper is fulfilled and Theorem 4.1 in Janssen (2005) implies a conditional central limit theorem for the permutation version of $\tilde{T}_{n,stud}$

$$\sup_{x\in\mathbb{R}}\left|P\left(\tilde{T}_{n,stud}\left((Z_{\pi(i)})_{1\le i\le 2n}\right)\le x\right)-\Phi(x)\right|\xrightarrow{P}0.$$

Note that $T_{n,stud}$ and $\tilde{T}_{n,stud}$ only differ in their standardizations. Hence, to complete the proof, we have to show that the difference $\tilde{V}_n^2((Z_{\pi(i)})_{1\le i\le 2n})-V_n^2((Z_{\pi(i)})_{1\le i\le 2n})$ converges in probability to zero (note that both are positive on a set with probability tending to 1). Straightforward calculations show that

$$\tilde{V}_n^2\left((Z_{\pi(i)})_{1\le i\le 2n}\right)-V_n^2\left((Z_{\pi(i)})_{1\le i\le 2n}\right)$$
$$=2\Bigg[\frac{1}{n-1}\sum_{i=1}^{n}Z_{\pi(i)}Z_{\pi(n+i)}$$
$$-\left(\frac{1}{n}\sum_{j=1}^{n}Z_{\pi(j)}\right)\left(\frac{1}{n}\sum_{j=1}^{n}Z_{\pi(n+j)}\right)\Bigg]$$
$$=:2\left(R^{\pi}_{n,1}-R^{\pi}_{n,2}\right).$$

We will first study $R^{\pi}_{n,1}$. The conditional expectation fulfills

$$\frac{n-1}{n}E\left(R^{\pi}_{n,1}|\mathbf{X}\right)$$
$$=\frac{1}{2n(2n-1)}\sum_{1\le i\ne j\le 2n}Z_iZ_j$$
$$=\frac{2n}{2n-1}\left(\frac{1}{2n}\sum_{j=1}^{2n}Z_i\right)^2+\frac{1}{2n(2n-1)}\sum_{i=1}^{2n}Z_i^2$$
$$=\frac{1}{4}(\mu_1+\mu_2)^2+o_P(1).$$

Here $o_P(1)$ stands for a sequence that converges in probability to zero as $n\to\infty$. Moreover, for the conditional second moment we have by the law of large numbers

$$\frac{n-1}{1}E\left(\left(R^{\pi}_{n,1}\right)^2|\mathbf{X}\right)$$
$$=\frac{1}{n^2}\sum_{1\le i\ne j\le n}E(Z_{\pi(j)}Z_{\pi(n+j)}Z_{\pi(i)}Z_{\pi(n+i)}|\mathbf{X})$$
$$+\frac{1}{n^2}\sum_{k=1}^{n}E\left(Z^2_{\pi(k)}Z^2_{\pi(n+k)}|\mathbf{X}\right)$$
$$=\frac{n-1}{n}\frac{1}{2n(2n-1)(2n-2)(2n-3)}$$
$$\times\sum_{\substack{1\le i_1,i_2,i_3,i_4\le 2n\\ \text{all }\ne}}Z_{i_1}\cdots Z_{i_4}$$
$$+\frac{1}{n}\frac{1}{2n(2n-1)}\sum_{1\le i\ne j\le 2n}Z_i^2Z_j^2$$
$$=\frac{1}{2n}\left(\sum_{i=1}^{2n}Z_i\right)^4+o_P(1)$$
$$=\left(\frac{1}{2}(\mu_1+\mu_2)\right)^4+o_P(1)$$
$$=E\left(R^{\pi}_{n,1}|\mathbf{X}\right)^2+o_P(1).$$

Note that the third step comprised iterated applications of the law of large numbers together with inequalities that involve the convergence in probability $\max_{1\le i\le 2n}Z_i^2/n\xrightarrow{P}0$. Altogether this shows $\mathrm{Var}(R^{\pi}_{n,1})\xrightarrow{P}0$ so that $R^{\pi}_{n,1}$ converges in probability to $\frac{1}{4}(\mu_1+\mu_2)^2$. For $R^{\pi}_{n,2}$ similar cal-

culations as above show that

$$E\left(\frac{1}{n}\sum_{j=1}^{n} Z_{\pi(j)}\middle|\mathbf{X}\right) \xrightarrow{P} \frac{1}{2}(\mu_1+\mu_2) \quad \text{and}$$

$$\mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} Z_{\pi(j)}\middle|\mathbf{X}\right) \xrightarrow{P} 0.$$

Thus $\frac{1}{n}\sum_{j=1}^{n} Z_{\pi(j)}$ converges in probability to $\frac{1}{2}(\mu_1+\mu_2)$. Since the same holds true for $(\frac{1}{n}\sum_{j=1}^{n} Z_{\pi(n+j)})$ it follows that $R_{n,2}$ converges in probability to $\frac{1}{4}(\mu_1+\mu_2)^2$ which completes the proof for the resampling scheme (I).

The proof for scheme (III), where we draw the resample with replacement from the pooled sample, can be obtained with similar methods.

Since case (II) is a special example of (VI), see Remark 2.1 above, the result follows from Theorem 2.2. Hence it remains to prove (IV). Therefore we can again proceed as in the proof of (I). First it follows from Theorem 4.2. in Janssen (2005) that $\tilde{T}^*_{n,stud} = \tilde{T}_{n,stud}(\mathbf{X}^*)$ is asymptotically standard normal, i.e. (5.2) holds with $T^*_{n,stud}$ replaced by $\tilde{T}^*_{n,stud}$. Again the asymptotic equivalence of the different studentizations follows as in case (I). □

Proof of Theorem 2.2 We start by verifying the result for (V). Therefore we will apply Theorem 3.1. in Janssen (2005) with the array $X_{n,i} := D_i/\sqrt{n}$. Note that by Eq. (3.4) in his paper the result follows from the convergences $\sum_{i=1}^{n}(X_{n,i}-\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2 - \overline{D}_n^2 \xrightarrow{P} \mathrm{Var}(D_1)$ and $\max_{1\le i\le n}|X_{n,i}| \xrightarrow{P} 0$. Here the last convergence is a consequence of (5.3). This finishes the proof for (V).

For the last case (VI) we analyze foremost the conditional distribution of the enumerator of the wild bootstrap t-type statistic $\sqrt{n}\,\overline{D}^*_n$. Note, that given the data $\mathbf{X}$

$$W_{n,i} := \frac{1}{\sqrt{n}} W_i D_i, \quad 1 \le i \le n$$

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

$$E(W_{n,i}|\mathbf{X}) = 0 \quad \text{and} \quad \mathrm{Var}(W_{n,i}|\mathbf{X}) = \frac{1}{n} D_i^2.$$

Hence the conditional variance of $\sqrt{n}\,\overline{D}^*_n$ fulfills

$$\mathrm{Var}\big(\sqrt{n}\,\overline{D}^*_n|\mathbf{X}\big) = \frac{1}{n}\sum_{i=1}^{n} D_i^2 \xrightarrow{P} E\big(D_1^2\big) =: \sigma_W^2.$$

Since we also have

$$\sum_{i=1}^{n} E\big(W_{n,i}^2 \mathbf{1}\{|W_{n,i}| \ge \epsilon\}\big|\mathbf{X}\big)$$

$$\le E\Big(W_1^2 \mathbf{1}\Big\{\max_{1\le i\le n}|W_i| \ge \epsilon\sqrt{n}\Big\}\Big) \longrightarrow 0$$

for all $\epsilon > 0$ by the dominated convergence theorem, Lindeberg's central limit theorem implies

$$\sup_{x\in\mathbb{R}}\big|P\big(\sqrt{n}\,\overline{D}^*_n \le x\big|\mathbf{X}\big) - \Phi(x/\sigma_W)\big| \xrightarrow{P} 0.$$

By Slutzky's Lemma it remains to prove that V_n^{*2} converges in probability to σ_W^2. But this follows from the law of large numbers since

$$\frac{n-1}{n} V_n^{*2} = \frac{1}{n}\sum_{i=1}^{n}(W_i D_i)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} W_j D_j\right)^2$$

converges in probability to $E((W_1 D_1)^2) - E(W_1 D_1)^2 = E(W_1^2)E(D_1^2) = \sigma_W^2$. □

References

Arlot, S., Blanchard, G., Roquain, E.: Some nonasymptotic results on resampling in high dimension, I: confidence regions. Ann. Stat. **38**, 51–82 (2010a)

Arlot, S., Blanchard, G., Roquain, E.: Some nonasymptotic results on resampling in high dimension, II: multiple tests. Ann. Stat. **38**, 83–99 (2010b)

Basso, D., Pesarin, F., Salmaso, L., Solari, A.: Permutation Tests for Stochastic Ordering and ANOVA. Springer, New York (2009)

Beran, R.: Diagnosing bootstrap success. Ann. Inst. Stat. Math. **49**, 1–24 (1997)

Beyersmann, J., Di Termini, S., Pauly, M.: Weak convergence of the wild bootstrap for the Aalen-Johansen estimator of the cumulative incidence function of a competing risk. Scand. J. Stat. (2012). doi:10.1111/j.1467-9469.2012.00817.x

Bickel, P.J., Freedman, D.A.: Some asymptotic theory for the bootstrap. Ann. Stat. **9**, 1196–1217 (1981)

Brombin, C., Salmaso, L.: Multi-aspect permutation tests in shape analysis with small sample size. Comput. Stat. Data Anal. **53**, 3921–3931 (2009)

Brombin, C., Salmaso, L., Ferronato, G., Galzignato, P.-F.: Multi-aspect procedures for paired data with application to biometric morphing. Commun. Stat., Simul. Comput. **40**, 3921–3931 (2011)

Delaigle, A., Hall, P., Jin, J.: Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. J. R. Stat. Soc. B **73**, 283–301 (2011)

Good, P.: Permutation, Parametric and Bootstrap Tests of Hypotheses, 3rd edn. Springer Series in Statistics. Springer, New York (2005)

Hall, P., Wilson, S.R.: Two guidelines for bootstrap hypothesis testing. Biometrics **47**, 757–762 (1991)

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

Janssen, A.: Testing nonparametric statistical functionals with application to rank tests. J. Stat. Plan. Inference **81**, 71–93 (1999a). Erratum: J. Stat. Plan. Inference **92**, 297 (2001)

Janssen, A.: Nonparametric symmetry tests for statistical functionals. Math. Methods Stat. **8**, 320–343 (1999b)

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

Janssen, A., Pauls, T.: How do bootstrap and permutation tests work? Ann. Stat. **31**, 768–806 (2003)

Janssen, A., Pauls, T.: A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. Comput. Stat. **20**, 369–383 (2005)

Janssen, A., Pauly, M.: Asymptotics and effectiveness of conditional tests with applications to randomization tests. Tech. Report, University of Duesseldorf (2010)

Konietschke, F., Pauly, M.: A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. Electron. J. Stat. **6**, 1358–1372 (2012)

Kreiss, J.-P., Paparoditis, E.: Bootstrap for dependent data: a review, with discussion, and a rejoinder. J. Korean Stat. Soc. **40**, 357–378, 393–395 (2011)

Lin, D.: Non-parametric inference for cumulative incidence functions in competing risks studies. Stat. Med. **16**, 901–910 (1997)

Mammen, E.: When Does Bootstrap Work? Asymptotic Results and Simulations. Springer, New York (1992)

Munzel, U.: Nonparametric methods for paired samples. Stat. Neerl. **53**, 277–286 (1999)

Neubert, K., Brunner, E.: A studentized permutation test for the nonparametric Behrens-Fisher problem. Comput. Stat. Data Anal. **51**, 5192–5204 (2007)

Omelka, M., Pauly, M.: Testing equality of correlation coefficients in an potentially unbalanced two-sample problem via permutation methods. J. Stat. Plan. Inference **142**, 1396–1406 (2012)

Pauly, M.: Discussion about the quality of F-ratio resampling tests for comparing variances. Test **20**, 163–179 (2011)

Pesarin, F., Salmaso, L.: Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. J. Nonparametr. Stat. **22**, 669–684 (2010a)

Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. Wiley, Chichester (2010b)

Pesarin, F., Salmaso, L.: A review and some new results on permutation testing for multivariate problems. Stat. Comput. **22**, 639–646 (2012)

Wu, C.: Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Stat. **14**, 1261–1295 (1986)