Test (2011) 20: 163–179
DOI 10.1007/s11749-010-0194-2

ORIGINAL PAPER

Discussion about the quality of F-ratio resampling tests for comparing variances

Markus Pauly

Received: 18 August 2009 / Accepted: 20 April 2010 / Published online: 14 May 2010

Abstract The goal of this paper is to describe the quality of different two-sample bootstrap and permutation tests for comparing variances. It is thereby inter alia shown that studentized resampling versions of the classical F-ratio test are asymptotically effective in a general nonparametric setting. This means that there is asymptotic no loss of power under contiguous alternatives. Moreover it is indicated that these tests are asymptotically consistent for fixed alternatives.

Keywords Heterogeneous null distributions · Permutation tests · Bootstrap tests · Power · Studentized statistics · Two-sample tests

Mathematics Subject Classification (2000) 62G09 · 62G10

1 Introduction and motivation

Unlike their bootstrap counterparts, two-sample permutation tests have the advantage that they are of exact level α for the null hypothesis of exchangeability, see, e.g., Lehmann and Romano (2005). However, some authors avoid their usage even for comparing means or variances, see, e.g., Hayes (1997, 2000). This is caused by the fact that the typical nonstudentized permutation tests are of asymptotic level α only for homogeneous or special heterogeneous null hypotheses, see Romano (1990),

Communicated by Domingo Morales.

This article has been developed while the author was a fellow of the 'Gründerstiftung for promotion of research of young academics' at the Heinrich-Heine Universität Düsseldorf.

M. Pauly (✉)
Department of Mathematics, Heinrich-Heine Universität Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany
e-mail: markus.pauly@uni-duesseldorf.de

Boos et al. (1989), Sakaori (2002), and Theorem 2 below. For a wide class of heterogeneous null distributions, Janssen (1997, 2005) has shown how to solve this problem in the case of comparing means by using a studentized version of the test statistic. In this paper we discuss the behavior of different bootstrap and permutation tests for comparing variances. Following the programme of Janssen and Pauly (2009), our main goal is to construct consistent and asymptotically effective tests with respect to a robust modification of Fisher's classical F-ratio test. Motivated by an article of Boos et al. (1989), it will turn out that this can be achieved by studentizing modified bootstrap and permutation versions of the F-ratio test.

1.1 The model and formulation of the problem

Consider the situation of Fisher's classical F-ratio test for comparing variances of two independent and normally distributed random samples. The model is given by the joint distribution

$$\mathcal{L}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = N\big(\mu_1, \sigma_1^2\big)^{n_1} \otimes N\big(\mu_2, \sigma_2^2\big)^{n_2}, \tag{1}$$

where μ_i and σ_i^2 are unknown for $i = 1, 2$, and $n_1, n_2 \geq 2$. Here $X_1, \ldots, X_{n_1}$ indicate the random variables of the first sample, and $Y_1, \ldots, Y_{n_2}$ the random variables of the second one. They are all defined on a probability space $(\Omega, \mathcal{A}, P)$ that runs in the background.

In the case that we are interested in testing one-sided hypotheses

$$H_0 : \big\{\sigma_1^2 \leq \sigma_2^2\big\} \quad \text{versus} \quad H_1 : \big\{\sigma_1^2 > \sigma_2^2\big\}, \tag{2}$$

the optimal level α test is well known, see, for example, Lehmann and Romano (2005). It is given by $\varphi_n = \mathbf{1}_{(F_{n_1-1,n_2-1;\alpha},\infty)}(F_n)$, where $F_{n_1-1,n_2-1;\alpha}$ is the $(1-\alpha)$-quantile of the F_{n_1-1,n_2-1} distribution, and the test statistic is just the ratio of the empirical variances of the two groups, i.e.,

$$F_n = F_n(X, Y) := \frac{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X}_{n_1})^2}{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y}_{n_2})^2} =: \frac{\hat{\sigma}_{n_1}^2(X)}{\hat{\sigma}_{n_2}^2(Y)}. \tag{3}$$

The problem of this test is that its optimality relies on the underlying normal distribution and does in general not hold in a semi- or nonparametric setting. This is caused by a high sensitivity against deviations from the kurtosis of the normal distribution. In nonparametric settings (like (4) below) it can scilicet be shown that the test φ_n is only of asymptotic level α if the kurtoses of both sample distributions are equal to 3 (the kurtosis of the normal distribution), see, e.g., Boos and Brownie (1989). Therefore it is convenient to carry out the test as a resampling (i.e., bootstrap or permutation) test. In this paper we investigate the quality of (studentized) resampling versions of the test φ_n in a general two-sample model. Thereby, the quality of the resampling tests is discussed in terms of the asymptotic effectiveness criterion introduced by Janssen and Pauly (2009), see (6) below. Our model is now given by the product distributions

$$\mathcal{L}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = G_1^{n_1} \otimes G_2^{n_2}, \tag{4}$$

where G_1 and G_2 are unknown distribution functions with finite fourth moments. Let $E(X_1) = \mu_1$, $E(Y_1) = \mu_2$, $\sigma_1^2 := \mathrm{Var}(X_1)$, $\sigma_2^2 := \mathrm{Var}(Y_1)$, and $\rho_1^4 := E((X_1 - \mu_1)^4)$, $\rho_2^4 := E((Y_1 - \mu_2)^4)$ denote the means, variances, and centered fourth moments of the two groups. To avoid trivialities it is throughout assumed that the variances are positive.

It is worth mentioning that the extended Behrens–Fisher model, which has been considered by Boos and Brownie (1989), Boos et al. (1989), and Janssen (1997), is also included. It is given by

$$\begin{aligned} X_i &= \mu_1 + \sigma_1 Z_i \quad \text{for } 1 \le i \le n_1, \\ Y_j &= \mu_2 + \sigma_2 Z_{n_1+j} \quad \text{for } 1 \le j \le n_2, \end{aligned} \tag{5}$$

where $Z_1, \ldots, Z_n$, $n := n_1 + n_2$, are i.i.d. random variables with $E(Z_1) = 0$, $\mathrm{Var}(Z_1) = 1$, and $E(Z_1^4) < \infty$.

As mentioned above, our aim is to find asymptotic effective bootstrap and permutation tests with respect to a robust modification of the F-ratio test. In our model (4) we call a sequence of tests ψ_n^* *asymptotically effective* (with respect to a benchmark test ψ_n) if, for $\sigma_1^2 = \sigma_2^2$,

$$E\big(|\psi_n - \psi_n^*|\big) \to 0 \tag{6}$$

as $n \to \infty$. What are the main advantages of this approach? To answer this question, suppose that the benchmark ψ_n is asymptotically exact for (4), i.e., for $\sigma_1^2 = \sigma_2^2$, $E(\psi_n) \to \alpha$ as $n \to \infty$. In this case (6) implicates that ψ_n^* is also asymptotically exact *and* reaches the power of the benchmark test asymptotically for contiguous alternatives. Hence we only have to compute this power for ψ_n, which is often easier. Especially under local asymptotic normality, the behavior of contiguous alternatives is well known, see, e.g., Van der Vaart (1998, Chaps. 6 and 7) for more details. For further investigations about asymptotic effectiveness, we refer to Janssen and Pauly (2009).

In the following Sect. 2 we will first construct a studentized version $\varphi_{n,\mathrm{stud}}$ of the F-ratio test that is asymptotically exact for $\sigma_1^2 = \sigma_2^2$ in the model (4). This test will serve as our benchmark test. Since $\varphi_{n,\mathrm{stud}}$ performs poorly for small sample sizes (see Sect. 4), we will discuss the asymptotic effectiveness of different bootstrap and permutation versions with respect to $\varphi_{n,\mathrm{stud}}$ in a next step. In this context Boos et al. (1989, Example (I)) have pointed out that the classical bootstrap and permutation procedures can in general not be applied for general nonparametric models like (4). This is caused by a wrong limit distribution of the resampling version of the test statistic. However, the problem can be solved by modifying the resampling procedure in the following way:

1. On the one hand, we do not resample (with or without replacement) from the pooled sample $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ but rather from the *centered pooled sample* $X_1 - \overline{X}_{n_1}, \ldots, X_{n_1} - \overline{X}_{n_1}, Y_1 - \overline{Y}_{n_2}, \ldots, Y_{n_2} - \overline{Y}_{n_2}$.
2. On the other hand, we use a *studentized version* of the test statistic.

We will see that both modifications are needed to construct valid resampling tests. Thereby the first modification effects the conditional convergence in distribution of

the permuted test statistic to a normal distribution under the whole model (4). Since this limit distribution has in general a wrong variance, we correct it with the help of an appropriate studentization V_n, see Sect. 2 for the definition of V_n. The two resulting (asymptotic effective) resampling tests are then given by

$$\varphi^*_{n,\mathrm{stud}} = \begin{cases} 1 \\ 0 \end{cases} \quad \text{if } \tilde{T}_n := \left(\frac{n_1 n_2}{n}\right)^{1/2} \cdot \frac{\log(F_n)\mathbf{1}_{\{V_n^2>0\}}}{V_n} \begin{matrix} > \\ \leq \end{matrix} c^*_{n,\mathrm{stud}}(\alpha),$$

where we set $\tilde{T}_n = 0$ if the numerator is 0. Here $n = n_1 + n_2$ holds, and $c^*_{n,\mathrm{stud}}(\alpha)$ denotes the conditional $(1-\alpha)$-quantile of the modified bootstrap or permutation distribution of $\tilde{T}_n$, see Sect. 2 below. All proofs are presented in Sect. 3. Remark that the corresponding two-sided problem $\tilde{H}_0 = \{\sigma_1^2 = \sigma_2^2\}$ versus $\{\sigma_1^2 \neq \sigma_2^2\}$ can be treated in an analogous manner by analyzing the behavior of the sum of two one-sided tests of level $\alpha/2$.

The above-mentioned first modification (resampling the centered pooled sample instead of the pooled sample) goes back to Boos et al. (1989). The second modification (using a studentized version of the test statistic for variance correction) has first been used by Neuhaus (1993) for permutation tests in the context of random censoring problems. As mentioned above, it has also been applied by Janssen (1997) for comparing means in the extended Behrens–Fisher model.

The problem of constructing distribution-free and asymptotically correct tests for comparing variances has already been discussed by several authors. Especially the articles of Boos and Brownie (1989) and Boos et al. (1989) can be seen as a motivation for the current paper. Boos and Brownie (1989) discuss the construction of bootstrap tests for equality of variances in the extended Behrens–Fisher model (5) (even for k-sample problems, $k \geq 2$). Boos et al. (1989) have, amongst others, analyzed the asymptotic bootstrap and permutation distribution of the transformed F-ratio test statistic $\sqrt{n_1 n_2/n}\log(F_n)$ to construct asymptotically correct bootstrap and permutation tests φ_n^{bs} and φ_n^{per} for H_0 in the model (5). These tests will be analyzed in more detail in the following sections. The related problem of testing equality of covariance matrices via resampling methods has, for example, been studied by Zhang and Boos (1993) and Zhu et al. (2002). For more detailed discussions about procedures for comparing variances and covariances, we refer to the survey article of Boos and Brownie (2004) and the references therein.

In addition to the mathematical justification, the literature contains a lot of simulation studies that suggest the usage of bootstrap and permutation procedures. For example, the articles of Boos et al. (1989), Boos and Brownie (2004), Janssen (1997), Janssen and Pauls (2003b), the monographs of Manly (1997), Good (2005), and Edgington and Onghena (2007), and the references therein contain numerical justification for different situations. In addition, you can find a simulation study at the end of the paper that compares the performance of $\varphi_{n,\mathrm{stud}}$, the studentized bootstrap and permutation tests $\varphi^*_{n,\mathrm{stud}}$, and the tests φ_n^{bs} and φ_n^{per} proposed by Boos et al. (1989). It will turn out that the studentized resampling tests are more competitive than the others.

2 The studentized F-ratio test and its resampling versions

As mentioned above, the classical F-ratio test φ_n is in general not asymptotically exact for the model (4). To avoid this hitch (and to find an adequate benchmark test), we will first construct a studentized version of φ_n that is distribution free and asymptotically exact. In doing so, we first have to analyze the asymptotic distribution of F_n. Since the limit law of F_n is degenerate, we have to transform it first. This is done by a logarithmic transformation. The result is stated in the following proposition. There and throughout this paper, we will assume that the asymptotic ratios of the first and second groups $\lim_{n\to\infty} n_1/n =: p$ and $q := 1 - p$ exist with $p \in (0, 1)$.

Proposition 1 *Let* $a_n := (\frac{n_1 \cdot n_2}{n})^{1/2}$ *and suppose that* (4) *holds with* $\sigma_1^2 = \sigma_2^2$. *In this case we have the convergence in distribution, as* $n \to \infty$,

$$T_n := a_n \log(F_n) \xrightarrow{\mathcal{D}} Z \quad \text{with } P^Z = N\big(0, q\beta_2^{(1)} + p\beta_2^{(2)} - 1\big), \tag{7}$$

where $\beta_2^{(i)} := \rho_i^4/\sigma_i^4$, $i = 1, 2$, *denote the kurtoses of the two groups.*

Since the limit variance $\xi_{q,p}^2 := q\beta_2^{(1)} + p\beta_2^{(2)} - 1$ is in general not known, we need a consistent estimator for it. We will see that V_n given by

$$V_n^2 := \frac{\tilde{V}_n^2}{(a_n^2(\frac{1}{n_1}\hat{\sigma}_{n_1}^2 + \frac{1}{n_2}\hat{\sigma}_{n_2}^2))^2} \tag{8}$$

is a good choice. Here $\tilde{V}_n^2$ is defined as

$$\tilde{V}_n^2 := a_n^2\left(\frac{1}{n_1}\hat{\rho}_1^4 + \frac{1}{n_2}\hat{\rho}_2^4\right) - \left(a_n^2\left(\frac{1}{n_1}\hat{\sigma}_{n_1}^2 + \frac{1}{n_2}\hat{\sigma}_{n_2}^2\right)\right)^2, \tag{9}$$

where $\hat{\rho}_1^4 := \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \overline{X}_{n_1})^4$ and $\hat{\rho}_2^4 := \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \overline{Y}_{n_2})^4$ are consistent estimators for the centered fourth moments ρ_1^4 and ρ_2^4, respectively. Our desired benchmark test for H_0 is now given by

$$\varphi_{n,\text{stud}} := \mathbf{1}_{(u_{1-\alpha},\infty)}\left(\frac{T_n}{V_n}\mathbf{1}_{\{V_n^2>0\}}\right), \tag{10}$$

where $u_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution. Its properties are summarized in the following proposition.

Proposition 2 *Suppose that the conditions of Proposition* 1 *are fulfilled and that we have* $\xi_{q,p}^2 > 0$. *Then we have the convergences*

$$T_{n,\text{stud}} := a_n \frac{\log(F_n)}{V_n}\mathbf{1}_{\{V_n^2>0\}} \xrightarrow{\mathcal{D}} \tilde{Z} \quad \text{with } P^{\tilde{Z}} = N(0, 1) \tag{11}$$

and $E(\varphi_{n,\text{stud}}) \to \alpha$ *as* $n \to \infty$, *i.e.*, $\varphi_{n,\text{stud}}$ *is asymptotically exact.*

In addition, we have that $\varphi_{n,\mathrm{stud}}$ *is consistent, i.e.,* $E(\varphi_{n,\mathrm{stud}}) \to \mathbf{1}_{\{\sigma_1^2 > \sigma_2^2\}}$ *for unequal variances* $\sigma_1^2 \neq \sigma_2^2$.

Remark that the condition $\xi_{q,p}^2 > 0$ is needed to get a nondegenerate limit distribution. It is fulfilled iff at least one of the kurtoses $\beta_2^{(1)}$ and $\beta_2^{(2)}$ is greater than zero.

In the next step we will analyze bootstrap and permutation versions of φ_n. Since Boos et al. (1989) have already found asymptotically exact resampling tests for the extended Behrens–Fisher model (5), we start with the behavior of their tests in our model (4). Therefore we denote the centered pooled sample by $Z := (Z_{n,1}, \ldots, Z_{n,n}) := (X_1 - \overline{X}_{n_1}, \ldots, X_{n_1} - \overline{X}_{n_1}, Y_1 - \overline{Y}_{n_2}, \ldots, Y_{n_2} - \overline{Y}_{n_2})$ and assume that $\tau : (\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}) \to \mathcal{S}_n$ is a uniformly distributed random variable on the symmetric group $\mathcal{S}_n$ (the set of all permutations of $(1, \ldots, n)$). For constructing a permutation test, we have to assume that τ and $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ are independent random variables on the joint probability space $(\Omega \times \tilde{\Omega}, \mathcal{A} \otimes \tilde{\mathcal{A}}, P \otimes \tilde{P})$. Boos et al.'s permutation version of the F-ratio test φ_n is now given by

$$\varphi_n^{\mathrm{per}} = \begin{cases} 1 & > \\ \gamma_n & \text{for } T_n = c_n^{\mathrm{per}}(\alpha), \\ 0 & < \end{cases} \tag{12}$$

where $c_n^{\mathrm{per}}(\alpha) = c_n^{\mathrm{per}}(\alpha, \omega)$ is the $(1-\alpha)$-quantile of the conditional permutation distribution $\mathcal{L}(T_n((Z_{n,\tau(i)})_{i \le n})|Z)(\omega, \cdot)$. Remark that permuting the centered pooled sample has the disadvantage that the permutation test looses the exactness for the smaller null hypothesis of exchangeability. It will turn out that the randomization γ_n does not play an important role in our (asymptotic) investigations. Hence we can set $\gamma_n = 0$. In contrast to φ_n^{per}, their bootstrap version of φ_n is given by $\varphi_n^{\mathrm{bs}} = \mathbf{1}_{(c_n^{\mathrm{bs}}(\alpha), \infty)}(T_n)$, where now $c_n^{\mathrm{bs}}(\alpha)$ is the $(1-\alpha)$-quantile of the conditional bootstrap distribution $\mathcal{L}(T_n((Z_{n,i}^*)_{i \le n})|Z)(\omega, \cdot)$. Here the bootstrap array $Z_{n,1}^*, \ldots, Z_{n,n}^*$ is rowwise i.i.d. (given Z) with conditional distribution function $H(x|Z) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(-\infty, x]}(Z_{n,i})$. The limit behavior of the conditional bootstrap and permutation distributions given the data is discussed in the following theorem. Here d denotes a distance for probability measures on the real line $\mathcal{M}_1(\mathbb{R}, \mathcal{B})$ that metrizes weak convergence, e.g., the Levy distance. As usual, we denote by $\mathcal{L}(T|X)$ the conditional distribution of T given X.

Theorem 1 *Under the conditions of Proposition 1, we have the following conditional weak convergences in* P*-probability as* $n \to \infty$:

$$d\big(\mathcal{L}\big(T_n\big((Z_{n,i}^*)_{i \le n}\big)|Z\big), N\big(0, \xi_{p,q}^2\big)\big) \xrightarrow{P} 0, \tag{13}$$

$$d\big(\mathcal{L}\big(T_n\big((Z_{n,\tau(i)})_{i \le n}\big)|Z\big), N\big(0, \xi_{p,q}^2\big)\big) \xrightarrow{P} 0, \tag{14}$$

where the limit variance is given by $\xi_{p,q}^2 := p\beta_2^{(1)} + q\beta_2^{(2)} - 1$.

Remark 1

1. In the classical case (resampling from the pooled sample) the convergences (13) and (14) hold for $\mu_1 = \mu_2$, see the proof of Theorem 1 below and the example for (5) in Boos et al. (1989).
2. Suppose that (4) holds. Then the above theorem shows that the resampling tests φ_n^{per} and φ_n^{bs} are in general not even of asymptotic level α and therefore neither applicable nor asymptotically effective. The only exceptions are the cases of (asymptotically) equal sample sizes $p = q = 1/2$ or equal kurtoses $\beta_2^{(1)} = \beta_2^{(2)}$ (which is fulfilled under (5)), see Boos et al. (1989) for further discussions.
3. The proof of (13) and (14) (under more restrictive moment conditions) has already been shown in Boos et al. (1989).

The crux of the above problem is that the resampling procedure interchanges p and q in the limit variance. To overcome this hitch we have chosen V_n^2, see (8), in good foresight in such a way that its resampling version also interchanges p and q in the limit. The corresponding studentized bootstrap and permutation tests are now given by

$$\varphi_{n,\text{stud}}^{\text{bs}} = \begin{cases} 1 \\ 0 \end{cases} \quad \text{if } T_{n,\text{stud}} = \frac{T_n}{V_n} \mathbf{1}_{\{V_n^2 > 0\}} \begin{matrix} > \\ \leq \end{matrix} c_{n,\text{stud}}^{\text{bs}}(\alpha) \tag{15}$$

and $\varphi_{n,\text{stud}}^{\text{per}} := \mathbf{1}_{(c_{n,\text{stud}}^{\text{per}}(\alpha),\infty)}(T_{n,\text{stud}})$. Here $c_{n,\text{stud}}^{\text{bs}}(\alpha)$ and $c_{n,\text{stud}}^{\text{per}}(\alpha)$ denote the $(1-\alpha)$-quantiles of the conditional bootstrap and permutation distributions of $T_{n,\text{stud}}$ given the data.

Theorem 2 *Under the conditions of Proposition 2, we have the following conditional weak convergences in P-probability as $n \to \infty$:*

$$d\big(\mathcal{L}\big(T_{n,\text{stud}}\big((Z_{n,i}^*)_{i\leq n}\big)|Z\big), N(0,1)\big) \xrightarrow{P} 0, \tag{16}$$

$$d\big(\mathcal{L}(T_{n,\text{stud}}((Z_{n,\tau(i)})_{i\leq n})|Z), N(0,1)\big) \xrightarrow{P} 0. \tag{17}$$

Moreover the studentized resampling tests $\varphi_{n,\text{stud}}^{\text{bs}}$ and $\varphi_{n,\text{stud}}^{\text{per}}$ are asymptotically effective with respect to $\varphi_{n,\text{stud}}$.

The next theorem describes the behavior of the power functions under fixed and local alternatives.

Theorem 3 *Suppose that our model (4) holds. Then the studentized resampling tests $\varphi_{n,\text{stud}}^{\text{bs}}$ and $\varphi_{n,\text{stud}}^{\text{per}}$ are consistent for unequal variances $\sigma_1^2 \neq \sigma_2^2$.*

Furthermore, for $c_n > -a_n, c_n \to c \in \mathbb{R}$, and local alternatives $\sigma_1^2 = \sqrt{1 + c_n/a_n}\sigma_2^2$, the limit of the power functions of $\varphi_{n,\text{stud}}$ and $\varphi_{n,\text{stud}}^{\text{per}}$ is given by $\lim_{n\to\infty} E(\varphi_{n,\text{stud}}) = \lim_{n\to\infty} E(\varphi_{n,\text{stud}}^{\text{per}}) = \Phi(c/\xi_{q,p} - u_{1-\alpha})$.

3 The proofs

It will turn out that the following lemma is quite helpful for the proofs of the above theorems. As in the above sections, we will again assume that $n_1/n \to p = 1 - q \in (0,1)$ as $n \to \infty$ and that τ is a uniformly distributed random variable on the symmetric group $\mathcal{S}_n$.

Lemma 1 *Let $X_{n,1}, \ldots, X_{n,n}$ be a triangular array of real-valued random variables that is independent from τ. Denote its rowwise mean by $\overline{X}_n$. If the convergences in probability*

$$\max_{1\le i\le n} \frac{|X_{n,i}|}{\sqrt{n}} \xrightarrow{P} 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_n)^2 \xrightarrow{P} \hat{\sigma}^2 \tag{18}$$

hold as $n \to \infty$, then we have the conditional convergence

$$d\left(\mathcal{L}\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}(X_{n,\tau(i)}-\overline{X}_n)\,|\,X_{n,1},\ldots,X_{n,n}\right), N\left(0,q\hat{\sigma}^2\right)\right)\xrightarrow{P}0. \tag{19}$$

Proof We want to apply Theorem 2.1 of Janssen (2005) for the weights $W_{n,i} := c_{n,\tau(i)}, 1 \le i \le n$, where $c_{n,i} := \sqrt{n/(n_1 n_2)}\mathbf{1}_{\{1,\ldots,n_1\}}(i), 1 \le i \le n$. Therefore we have to check whether our weights fulfill the conditions of his theorem. Because of $\overline{W}_n = \overline{c}_n = \sqrt{\frac{n_1}{nn_2}}$, we obtain

$$\max_{i\le n}\left|W_{n,i}-\overline{W}_n\right| = \max\left\{\overline{c}_n, \sqrt{\frac{n_1 n}{n_2}}\left(\frac{1}{n_1}-\frac{1}{n}\right)\right\} \to 0$$

as $n \to \infty$. In addition, we have the convergence $\sum_{i=1}^{n}(W_{n,i}-\overline{W}_n)^2 = n_2\overline{c}_n^2 + n_1\frac{n_2}{nn_1} \to p+q = 1$ as $n \to \infty$. Hence it remains to show that

$$\sqrt{n}(W_{n,1}-\overline{W}_n)\xrightarrow{\mathcal{D}}\frac{1}{\sqrt{pq}}\mathbf{1}_A - \sqrt{\frac{p}{q}},$$

where $A \in \tilde{\mathcal{A}}$ is a set with $\tilde{P}(A) = p$. This can be obtained by Slutzky's Lemma and the convergences $\tilde{P}(\sqrt{n}W_{n,1} = 0) = \frac{n_2}{n} \to q$ and $\tilde{P}(\sqrt{n}W_{n,1} = \frac{n}{\sqrt{n_1 n_2}}) = \frac{n_1}{n} \to p$. Applying Theorem 2.1 of Janssen (2005) yields

$$d\left(\mathcal{L}\left(\sqrt{\frac{n}{n_2 n_1}}\sum_{i=1}^{n_1}\left(X_{n,\tau(i)}-\overline{X}_n\right)|X_{n,1},\ldots,X_{n,n}\right), N\left(0,\hat{\sigma}^2\right)\right)\xrightarrow{P}0 \tag{20}$$

as $n \to \infty$. This implies (19). □

Remark, that in the case $p \in \{0,1\}$, a similar version of Lemma 1 can be proven by applying the methods of Del Barrio et al. (2009).

Proposition 1 follows from results in Boos et al. (1989, Proposition on p. 331). However, since we want to apply some of the following proof steps in other situations (see Remark 2 and the proof of Theorem 1 below), we prove the result in a different way.

Proof of Proposition 1. By the shift and scale invariance of F_n we can assume without loss of generality that $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$.

The Taylor expansion of the logarithm shows that

$$\begin{aligned} T_n &= a_n\big(\log\big(\hat\sigma_{n_1}^2(X)\big) - \log\big(\hat\sigma_{n_2}^2(Y)\big)\big) \\ &= a_n\big(\hat\sigma_{n_1}^2(X) - \hat\sigma_{n_2}^2(Y) + R_2\big(\hat\sigma_{n_1}^2(X)\big) - R_2\big(\hat\sigma_{n_2}^2(Y)\big)\big) \\ &= a_n\big(\hat\sigma_{n_1}^2(X) - \hat\sigma_{n_2}^2(Y)\big) + o_P(1), \end{aligned} \tag{21}$$

where the remainder R_2 fulfills $\frac{R_2(x)}{x-1} \to 0$ as $x \to 1$. To accept the last equality, remark that Example 3.2 of Van der Vaart (1998), Slutzky's Lemma, and the SLLN together imply the convergence in probability

$$a_n \cdot R_2\big(\hat\sigma_{n_1}^2(X)\big) = \sqrt{\frac{n_2}{n}} \cdot \sqrt{n_1}\big(\hat\sigma_{n_1}^2(X) - \sigma_1^2\big) \cdot \frac{R_2(\hat\sigma_{n_1}^2(X))}{(\hat\sigma_{n_1}^2(X) - \sigma_1^2)} \xrightarrow{P} 0. \tag{22}$$

The same holds for the other remainder $a_n R_2(\hat\sigma_{n_2}^2(Y))$. Thus the convergence (7) can be obtained by applying Example 3.2 of Van der Vaart (1998) to (21). □

Proof of Proposition 2. Since V_n^2 is a consistent estimator of $\xi_{q,p}^2$, the convergence (11) follows from Proposition 1 and Slutzky's Lemma. Moreover this implicates the asymptotic exactness of $\varphi_{n,\mathrm{stud}}$. Hence it remains to investigate the behavior of $T_{n,\mathrm{stud}}$ for fixed variances $\sigma_1^2 \neq \sigma_2^2$. For the denominator, we get the almost sure convergence

$$V_n^2 \longrightarrow \frac{q\rho_1^4 + p\rho_2^4 - (q\sigma_1^2 + p\sigma_2^2)^2}{(q\sigma_1^2 + p\sigma_2^2)^2} =: \varsigma^2$$

as $n \to \infty$. Since $F_n \to \sigma_1^2/\sigma_2^2$ a.s. and $\varsigma > 0$, we also have the almost sure convergence

$$a_n \frac{\log(F_n)}{V_n} \longrightarrow +\infty \cdot \mathbf{1}_{\{\sigma_1^2 > \sigma_2^2\}} - \infty \cdot \mathbf{1}_{\{\sigma_1^2 < \sigma_2^2\}} \tag{23}$$

as $n \to \infty$. This shows the consistency of $\varphi_{n,\mathrm{stud}}$. □

Proof of Theorem 1. Convergence (13) follows from Theorem 1 and Remark 3 in Boos et al. (1989). We now prove (14).

Remark first that we can again assume without loss of generality that $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$. This is caused by the definition of $Z_{n,i}$, $1 \le i \le n$, and the shift and scale invariance of F_n. We now set $Z_n^{(\tau)} := (Z_{n,\tau(i)})_{i \le n}$, $Z_{n_1}^{(\tau)} := (Z_{n,\tau(i)})_{i \le n_1}$ and

$Z_{n2}^{(\tau)} := (Z_{n,\tau(i)})_{n_1+1\le i\le n}$ and show foremost that the permuted test statistic $T_n(Z_n^{(\tau)})$ fulfills a decomposition as in (21). Therefore consider the equation

$$\begin{aligned}
&\sqrt{n_1}\left(\frac{n_1-1}{n_1}\hat{\sigma}_{n_1}^2\big(Z_{n_1}^{(\tau)}\big)-1\right)\\
&\quad=\sqrt{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}Z_{n,\tau(i)}^2-\frac{1}{n}\sum_{j=1}^{n}Z_{n,j}^2\right)+\sqrt{n_1}\left(\frac{1}{n}\sum_{j=1}^{n}Z_{n,j}^2-1\right)\\
&\qquad-\sqrt{n_1}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}Z_{n,\tau(j)}\right)^2\\
&\quad=:A_n+B_n-C_n.
\end{aligned}$$

We start with the treatment of the first remainder A_n. Since the fourth moments exist in both groups, the pooled sample fulfills the Lindeberg condition. This, together with the WLLN, implies that the array $(Z_{n,i}^2)_{i\le n}$ fulfills conditions (18) of Lemma 1 with $\hat{\sigma}^2 := p\rho_1^4+q\rho_2^4-1$.

Thus Lemma 1 shows that

$$d\left(\mathcal{L}\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(Z_{n,\tau(i)}^2-\frac{1}{n}\sum_{j=1}^{n}Z_{n,j}^2\right)\middle|Z_{n,1}^2,\dots,Z_{n,n}^2\right),N\big(0,q\hat{\sigma}^2\big)\right)\xrightarrow{P}0. \quad (24)$$

Since A_n only depends on the squared random variables, the convergence in (24) will also hold if we condition with respect to $Z_{n,1},\dots,Z_{n,n}$. We can now investigate the behavior of the second remainder B_n. Straightforward calculations and the WLLN show that it fulfills

$$B_n=\frac{n_1}{n}\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\big(X_i^2-1\big)\right)+\frac{n_2}{n}\left(\frac{1}{\sqrt{n_2}}\sum_{Y=1}^{n_2}\big(Y_j^2-1\big)\right)+o_P(1).$$

Here the first two summands each converge in distribution to a normally distributed random variable. Hence it remains to investigate the asymptotic behavior of the last remainder C_n. Here Lemma 1 can be applied to implicate that the conditional distribution of $\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}Z_{n,\tau(j)}$ converges weakly to a normal distribution (in probability). Thus Slutzky's Lemma yields $C_n=o_{P\otimes\tilde{P}}(1)$. Altogether, this shows that

$$a_nR_2\big(\hat{\sigma}_{n_1}^2\big)\big(Z_{n_1}^{(\tau)}\big)=\sqrt{\frac{n_2}{n}}\sqrt{n_1}\big(\hat{\sigma}_{n_1}^2\big(Z_{n_1}^{(\tau)}\big)-1\big)\frac{R_2(\hat{\sigma}_{n_1}^2)(Z_{n_1}^{(\tau)})}{(\hat{\sigma}_{n_1}^2(Z_{n_1}^{(\tau)})-1)}\xrightarrow{P\otimes\tilde{P}}0.$$

Thus by Slutzky's Lemma we get a decomposition of the permutation version of the test statistic as in (21):

$$\begin{aligned}
T_n\big(Z_n^{(\tau)}\big)&=a_n\big(\hat{\sigma}_{n_1}^2\big(Z_{n_1}^{(\tau)}\big)-\hat{\sigma}_{n_2}^2\big(Z_{n_2}^{(\tau)}\big)\big)+o_{P\otimes\tilde{P}}(1)\\
&=a_n\left(\frac{1}{n_1}\sum_{i=1}^{n_1}Z_{n,\tau(i)}^2-\frac{1}{n_2}\sum_{i=n_1+1}^{n}Z_{n,\tau(i)}^2\right)+o_{P\otimes\tilde{P}}(1).
\end{aligned}$$

We can now rewrite the last line with the help of the regression coefficients of the two-sample problem $e_{n,i} := a_n(n_1^{-1}\mathbf{1}_{\{1,\ldots,n_1\}}(i) - n_2^{-1}\mathbf{1}_{\{n_1+1,\ldots,n\}}(i))$. More precisely, we have $a_n \log(F_n(Z_{n,\tau(i)})_{i\le n}) = \sum_{i=1}^n e_{n,i} Z_{n,\tau(i)}^2 + o_{P\otimes\tilde{P}}(1)$. As in the proof of Lemma 1, we can now apply Theorem 2.1 of Janssen (2005) for the weights $W_{n,i} := e_{n,\tau(i)}$, $1 \le i \le n$. Together with Slutzky's Lemma, this deduces the conclusion (14). □

Remark 2 In the case $\mathcal{L}(X_1) = \mathcal{L}(Y_1)$ the proof can be simplified by applying Lemma 2 of Janssen and Pauly (2009) directly to (21).

Proof of Theorem 2 By construction of the random array $Z_{n,i}$, without loss of generality, we can again assume that $\mu_1 = \mu_2 = 0$. We start with the proof of (17). By Theorem 1 we only need to show that, for $\sigma_1^2 = \sigma_2^2$,

$$V_n^2\big(Z_n^{(\tau)}\big) \xrightarrow{P\otimes\tilde{P}} p\beta_2^{(1)} + q\beta_2^{(2)} - 1 \tag{25}$$

as $n \to \infty$. Therefore consider the decomposition

$$\tilde{V}_n^2\big(Z_n^{(\tau)}\big) = \frac{n_2}{n}\hat{\rho}_1^4\big(Z_{n_1}^{(\tau)}\big) + \frac{n_1}{n}\hat{\rho}_2^4\big(Z_{n_2}^{(\tau)}\big) - \left(\frac{n_2}{n}\hat{\sigma}_{n_1}^2\big(Z_{n_1}^{(\tau)}\big) + \frac{n_1}{n}\hat{\sigma}_{n_2}^2\big(Z_{n_2}^{(\tau)}\big)\right)^2. \tag{26}$$

We start with the treatment of the first two summands. In the proof of Theorem 1 we have already seen that $\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{n,\tau(j)} \to 0$ in $P\otimes\tilde{P}$-probability as $n \to \infty$. Thus one can suspect that

$$\hat{\rho}_1^4\big(Z_{n_1}^{(\tau)}\big) = \frac{1}{n_1}\sum_{i=1}^{n_1} Z_{n,\tau(i)}^4 + o_{P\otimes\tilde{P}}(1).$$

This can indeed be proven by the same arguments that are used in the following to analyze the limit behavior of $\frac{1}{n_1}\sum_{i=1}^{n_1} Z_{n,\tau(i)}^4$. Hence we can rewrite the first two summands of (26) as $\sum_{i=1}^n d_{n,i} Z_{n,\tau(i)}^4 + o_{P\otimes\tilde{P}}(1)$. Here we have used the same regression coefficients as Janssen (1997), $d_{n,i} := n_2/(n_1 n)\mathbf{1}_{\{1,\ldots,n_1\}}(i) + n_1/(n_2 n)\mathbf{1}_{\{n_1+1,\ldots,n\}}(i))$. Since $n\overline{d}_n := \sum_{i=1}^n d_{n,i} = 1$, we obtain the almost sure convergence

$$\begin{aligned} E\left(\sum_{i=1}^n d_{n,i} Z_{n,\tau(i)}^4 | Z\right) &= \sum_{i=1}^n d_{n,i}\left(\frac{n_1}{n}\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{n,j}^4 + \frac{n_2}{n}\frac{1}{n_2}\sum_{k=n_1+1}^{n} Z_{n,k}^4\right) \\ &\longrightarrow p\rho_1^4 + q\rho_2^4 \end{aligned} \tag{27}$$

as $n \to \infty$. Suppose for the moment that the eighth moments $E(X_1^8) + E(Y_1^8) < \infty$ are finite in both groups. Then the convergence $\sum_{i=1}^n (d_{n,i} - \overline{d}_n)^2 \to 0$ as $n \to \infty$, the WLLN, and Theorem 3 in the monograph of Hájek et al. (1999, p. 61f) together imply the almost sure convergence

$$\mathrm{Var}\left(\sum_{i=1}^n d_{n,i} Z_{n,\tau(i)}^4 | Z\right) = \sum_{k=1}^n (d_{n,k} - \overline{d}_n)^2 \frac{1}{n-1}\sum_{i=1}^n \left(Z_{n,i}^4 - \frac{1}{n}\sum_{j=1}^n Z_{n,j}^4\right)^2 \to 0$$

as $n \to \infty$. This yields the convergence in $P \otimes \tilde{P}$-probability

$$\sum_{i=1}^{n} d_{n,i} Z_{n,\tau(i)}^4 \xrightarrow{P\otimes\tilde{P}} p\rho_1^4 + q\rho_2^4 \tag{28}$$

as $n \to \infty$. Since our model (4) does not postulate the finiteness of eight moments, we must consider the trimmed random variables $X_{i,k}^{(1)} := Z_{n,i} \mathbf{1}_{[-k,k]}(Z_{n,i})$ and $X_{i,k}^{(2)} := Z_{n,i} - X_{i,k}^{(1)}$ for $1 \le i \le n$ and $k \in \mathbb{N}$. Set $X_{n_1+i} := Y_i$ for $1 \le i \le n_2$. Since $|\overline{X}_{n_1}| + |\overline{Y}_{n_2}| \to 0$ a.s., the inequalities $\mathbf{1}_{[-k+\epsilon,k-\epsilon]}(X_i) \le \mathbf{1}_{[-k,k]}(Z_{n,i}) \le \mathbf{1}_{[-k-\epsilon,k+\epsilon]}(X_i)$ hold for every $\epsilon > 0$ on sets A_n with $P(A_n) \to 1$. Thus arithmetics similar to the one used for getting (28) imply, for fixed $k \in \mathbb{N}$ and $n \to \infty$ (since $\mu_1 = \mu_2 = 0$),

$$\sum_{i=1}^{n} d_{n,i} \big(X_{\tau(i),k}^{(1)}\big)^4 \xrightarrow{P\otimes\tilde{P}} \rho_k^4 := pE\big(X_1^4 \mathbf{1}_{[-k,k]}(X_1)\big) + qE\big(Y_1^4 \mathbf{1}_{[-k,k]}(Y_1)\big).$$

Moreover we have, as $k \to \infty$,

$$\frac{1}{\epsilon} \limsup_{n\to\infty} E\left(\sum_{i=1}^{n} d_{n,i} \big(X_{\tau(i),k}^{(2)}\big)^4 | Z\right) = \frac{1}{\epsilon}\big(p\rho_1^4 + q\rho_2^4 - \rho_k^4\big) \to 0.$$

Thus an application of the Markov inequality, together with Theorem 4.2. of Billingsley (1968), shows that

$$\frac{n_2}{n} \hat{\rho}_1^4\big(Z_{n_1}^{(\tau)}\big) + \frac{n_1}{n} \hat{\rho}_2^4\big(Z_{n_2}^{(\tau)}\big) \xrightarrow{P\otimes\tilde{P}} p\rho_1^4 + q\rho_2^4 - 1.$$

Because of $\hat{\sigma}_{n_1}^2((Z_{n,i})_i) = \frac{n_1}{n_1-1}(\frac{1}{n_1}\sum_{i=1}^{n_1} Z_{n,i}^2)$, the $P \otimes \tilde{P}$-convergence of $\frac{n_2}{n}\hat{\sigma}_{n_1}^2((Z_{n,\tau(i)})_{i\le n}) + \frac{n_1}{n}\hat{\sigma}_{n_2}^2((Z_{n,\tau(i)})_{i\le n})$ to $p\sigma_1^2 + q\sigma_2^2 = \sigma_1^2 = \sigma_2^2$ can be obtained in the same manner, see Janssen (1997) for the special case of the extended Behrens–Fisher model. This shows (25) and thus (17) by Slutzky's Lemma. The asymptotic effectiveness of $\varphi_{n,\mathrm{stud}}^{\mathrm{per}}$ now follows by applying Lemma 1 of Janssen and Pauls (2003a), see also Janssen and Pauly (2009, Lemma 1).

For proving the result (16), remark that (26) and (27) can be obtained for the bootstrap versions in the same way. Only the calculation of the conditional variance differs a little bit from the permutation case. Here we have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} d_{n,i}(Z_{n,i}^*)^4 | Z\right) &= \sum_{k=1}^{n} d_{n,k}^2 \mathrm{Var}\big((Z_{n,1}^*)^4 | Z\big) \\
&= \sum_{k=1}^{n} d_{n,k}^2 \left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}^8 - \left(\frac{1}{n}\sum_{j=1}^{n} Z_{n,j}^4\right)^2\right) \\
&= \sum_{k=1}^{n} d_{n,k}^2 \frac{1}{n}\sum_{i=1}^{n}\left(Z_{n,i}^4 - \frac{1}{n}\sum_{j=1}^{n} Z_{n,j}^4\right)^2.
\end{aligned}$$

Since we have again $\sum_{k=1}^{n} d_{n,k}^2 \to 0$ as $n \to \infty$, the remaining proof steps can be established as above. □

Proof of Theorem 3. Suppose first that the variances $\sigma_1^2 \neq \sigma_2^2$ are fixed. In this case the proof of Theorem 2 above shows the following convergences in $P \otimes \tilde{P}$-probability:

$$V_n^2\big((Z_{n,\tau(i)})_{i\le n}\big) \xrightarrow{P\otimes\tilde{P}} \frac{p\rho_1^4 + q\rho_2^4 - (p\sigma_1^2 + q\sigma_2^2)^2}{(p\sigma_1^2 + q\sigma_2^2)^2} =: \tilde{\zeta}^2$$

and

$$\log\big(F_n\big((Z_{n,\tau(i)})_{i\le n}\big)\big) \xrightarrow{P\otimes\tilde{P}} \log\left(\frac{p\sigma_1^2 + q\sigma_2^2}{p\sigma_1^2 + q\sigma_2^2}\right) = \log(1) = 0$$

as $n \to \infty$. Since the same holds for the bootstrap versions, the consistency of $\varphi_{n,\mathrm{stud}}^{\mathrm{bs}}$ and $\varphi_{n,\mathrm{stud}}^{\mathrm{per}}$ follows from (23).

We now consider the local alternatives $\sigma_1^2 = \sigma_{1,n}^2 = \sqrt{1 + c/a_n} \cdot \sigma_2^2$. By similar arithmetics (involving the Lindeberg–Feller Theorem) to the one used for the proof of Proposition 1, we get that $T_n - a_n \log(\sigma_1^2/\sigma_2^2)$ possesses the representation (21). Together with the fact that $a_n \log(\sigma_1^2/\sigma_2^2) \to c$, we get the convergence in distribution $T_n \xrightarrow{\mathcal{D}} Z + c \sim N(c, \xi_{q,p}^2)$ as $n \to \infty$. Since $1 + c/a_n \to 1$, the limit of V_n^2 remains unchanged. Hence we have the convergence in distribution $T_{n,\mathrm{stud}} \xrightarrow{\mathcal{D}} \tilde{Z} + c \sim N(c/\xi_{q,p}, 1)$, which shows that $E(\varphi_{n,\mathrm{stud}}) \to \Phi(c/\xi_{q,p} - u_{1-\alpha})$. Since Lemma 1 and Theorem 2.1 in Janssen (2005) can be applied for triangular arrays, we can go through the proof of Theorems 1 and 2 in the current setting to see that the convergence (17) remains unchanged. This completes the proof. □

4 Simulation study

In this section we carried out a simulation study in order to compare the five different approaches: the unconditional benchmark test $\varphi_{n,\mathrm{stud}}$, the permutation and bootstrap F-ratio tests φ_n^{per} and φ_n^{bs} of Boos et al. (1989), and their studentized versions $\varphi_{n,\mathrm{stud}}^{\mathrm{per}}$ and $\varphi_{n,\mathrm{stud}}^{\mathrm{bs}}$. Our hypotheses of interest were $H_0 : \{\sigma_1^2 \le \sigma_2^2\}$ versus $H_1 : \{\sigma_1^2 > \sigma_2^2\}$. Comparisons were made with respect to type I error probabilities both for homogeneous and heterogeneous situations under various sample size assumptions. We also show a small power comparison, see Table 3. For the type I error control under the boundary $\{\sigma_1^2 = \sigma_2^2\}$, we utilized some common standardized distributions, namely, the normal ($\mu_i = 0, \sigma_i^2 = 1$), logistic ($\mu_i = 0, \sigma_i^2 = 1$), double-exponential with density $f(x) = \frac{1}{2}\exp(-|x|)$ ($\mu_i = 0, \sigma_i^2 = 1$), and exponential ($\sigma_i^2 = 1$) distributions. As nominal level, we have taken $\alpha = 0.05$. All entries of the tables are based on $M = 10^4$ Monte Carlo trials with $B = 10^3$ resampling replications. The results are presented below.

Table 1 Type I error control comparison between the benchmark and resampling tests in case of equal distributions

	n_1	n_2	$\varphi_{n,\text{stud}}$	φ_n^{per}	φ_n^{bs}	$\varphi_{n,\text{stud}}^{\text{per}}$	$\varphi_{n,\text{stud}}^{\text{bs}}$
Normal	4	8	0.2190	0.0481	0.0481	**0.0504**	0.0538
Logistic			0.2361	0.0559	0.0604	**0.0521**	0.0574
Double-exp.			0.2546	0.0698	0.0769	**0.0576**	0.0646
Exponential			0.2509	0.1044	0.0963	0.0735	**0.0696**
Normal	8	16	0.1345	0.0533	0.0547	**0.0521**	0.0437
Logistic			0.1371	0.0528	0.0574	**0.0489**	0.0410
Double-exp.			0.1481	0.0589	0.0655	**0.0539**	0.0454
Exponential			0.1726	0.0887	0.0843	0.0662	**0.0619**
Normal	16	16	0.0901	0.0475	**0.0486**	0.0474	0.0437
Logistic			0.1041	0.0550	0.0558	0.0538	**0.0471**
Double-exp.			0.1052	0.0529	0.0544	**0.0517**	0.0438
Exponential			0.1240	0.0764	0.0715	0.0612	**0.0525**

Table 1 illustrates the Monte Carlo estimates of the true level in case of equal distributions, where the rowwise best value occurs in bold numbers. In this situation the unstudentized resampling tests φ_n^{per} and φ_n^{bs} are of asymptotic level α, see Theorem 1 and Remark 1 above. As mentioned in the introduction, one of the first eye-catching observations is that our benchmark test $\varphi_{n,\text{stud}}$ performs poorly for small sample sizes. Although the error probability converges to the correct direction with increasing sample sizes, the test is not even applicable for $n_1 = n_2 = 16$.

As expected, all resampling tests behave better than the benchmark. Moreover, in almost all situations the studentized tests $\varphi_{n,\text{stud}}^{\text{bs}}$ and $\varphi_{n,\text{stud}}^{\text{per}}$ have a much better control of the type I error probability than their unstudentized counterparts. The only exception is the normal case with equal sample sizes $n_1 = n_2 = 16$, where the unstudentized tests seem to be slightly better. Especially in all exponential cases and the nonnormal cases with unequal sample sizes, the studentized resampling tests perform much better than their unstudentized counterparts. The studentized resampling tests differ slightly in their behavior. Except for the exponential case, the permutation test φ_n^{per} seems to be better for small sample sizes ($n_1 = 4, n_2 = 8$ or $n_1 = 8, n_2 = 16$). In all other cases the studentized bootstrap test has a better control of the type I error probability.

In contrast to Table 1, Table 2 provides a situation where the unstudentized resampling tests do not have the correct asymptotic level α. This is caused by unequal sample sizes and the different kurtoses of the normal ($\beta_2 = 3$), the logistic ($\beta_2 = 4.2$), the double-exponential ($\beta_2 = 6$), and the exponential distributions ($\beta_2 = 9$), see Johnson and Kotz (1970). The simulations fit to our theoretical investigations since the studentized bootstrap and permutation tests are again better than their unstudentized counterparts.

Moreover the benchmark test performs even inferior to the above case. It seems to be that the performance gets worse with increasing differences of the group kurtoses. This could be explained by a reduced convergence speed of the centered fourth

Table 2 Type I error control comparison between the benchmark and resampling tests in case of unequal distributions and sample sizes

	n_1	n_2	$\varphi_{n,\text{stud}}$	φ_n^{per}	φ_n^{bs}	$\varphi_{n,\text{stud}}^{\text{per}}$	$\varphi_{n,\text{stud}}^{\text{bs}}$
Norm. vs. Logis.	4	8	0.2406	0.0595	0.0617	**0.0574**	0.0629
Norm. vs. Dexp.			0.2719	0.0806	0.0845	**0.0722**	0.0781
Norm vs. Exp.			0.3086	0.1125	0.1200	**0.0942**	0.1023
Logis. vs. Dexp.			0.2639	0.0747	0.0804	**0.0663**	0.0725
Norm. vs. Logis.	8	16	0.1567	0.0676	0.0716	0.0640	0.0565
Norm. vs. Dexp.			0.1883	0.0851	0.0891	0.0855	**0.0756**
Norm vs. Exp.			0.2226	0.1169	0.1224	0.1123	**0.0997**
Logis. vs. Dexp.			0.1655	0.0729	0.0785	0.0701	**0.0591**

Table 3 Power comparison between the studentized bootstrap and permutation tests $\varphi_{n,\text{stud}}^{\text{bs}}$ and $\varphi_{n,\text{stud}}^{\text{per}}$

	σ_1	σ_2	Norm. Norm.	Norm. Logis.	Norm. Dexp.	Norm. Exp.	Logis. Dexp.
	$n_1 = 4, n_2 = 8$						
Bootstrap	1.2	1	0.0870	0.1025	0.1236	0.1548	0.1121
Permutation			0.0792	0.0935	0.1101	0.1400	0.1020
Bootstrap	1.5		0.1541	0.1750	0.1863	0.2251	0.1804
Permutation			0.1329	0.1519	0.1683	0.2039	0.1612
Bootstrap	2		0.2725	0.2966	0.3045	0.3448	0.4866
Permutation			0.2303	0.2527	0.2672	0.3055	0.4308
	$n_1 = 8, n_2 = 16$						
Bootstrap	1.2	1	0.1122	0.1359	0.1537	0.1924	0.1234
Permutation			0.1304	0.1545	0.1703	0.2132	0.1424
Bootstrap	1.5		0.2788	0.2894	0.3153	0.3411	0.2639
Permutation			0.3118	0.3177	0.3450	0.3694	0.2924
Bootstrap	2		0.5677	0.5596	0.5632	0.5761	0.5004
Permutation			0.6085	0.5993	0.5980	0.6064	0.5402

moment estimators in heterogeneous situations. The same can be observed for the studentized resampling tests. Again $\varphi_{n,\text{stud}}^{\text{per}}$ has a better control for small sample sizes than $\varphi_{n,\text{stud}}^{\text{bs}}$ and $\varphi_{n,\text{stud}}^{\text{bs}}$ performs better for slightly larger sample sizes $n_1 = 8$, $n_2 = 16$. However, the values are not acceptable for situations with larger kurtoses differences (Norm vs. Exp. and Norm vs. Dexp.). Here larger sample sizes are needed for a better type I error control.

The above observations also explain the power behavior of the two studentized resampling tests presented in Table 3. Since $\varphi_{n,\text{stud}}^{\text{bs}}$ seems to be more liberal than $\varphi_{n,\text{stud}}^{\text{per}}$, for small sample sizes $n_1 = 4$, $n_2 = 8$, it has higher power in this case. For slightly larger sample sizes $n_1 = 8$, $n_2 = 16$, the situation is again the other way around.

5 Conclusions

Based on our theoretical and simulation results, it is clear that the studentized bootstrap and permutation tests both fit the bill of a completely nonparametric test. On the one hand, both tests are consistent and of asymptotic level α for general heterogeneous models like (4). This remains unchanged even for unbalanced designs. On the other hand, both tests possess a good type I error control for small sample sizes (especially in contrast to the competing tests $\varphi_{n,\mathrm{stud}}$, φ_n^{per}, and φ_n^{bs}). In spite of these excellent properties, it should be clear that other tests (like the classical F-ratio test) perform better if we were in a parametric (e.g., normal) setting.

Acknowledgements We are especially grateful to the Associate Editor and two referees in terms of their careful reading of our manuscript, which led to a much improved presentation. Special thanks go to our phd supervisor Arnold Janssen.

References

Billingsley P (1968) Convergence of probability measures. Wiley, New York
Boos DD, Brownie C (1989) Bootstrap methods for testing homogeneity of variances. Technometrics 31(1):69–82
Boos DD, Brownie C (2004) Comparing variances and other measures of dispersion. Stat Sci 19(4):571–578
Boos DD, Janssen P, Veraverbeke N (1989) Resampling from centered data in the two-sample problem. J Stat Plan Inference 21(3):327–345
Del Barrio E, Janssen A, Matrán C (2009) Resampling schemes with low resampling intensity and their applications in testing hypotheses. J Stat Plan Inference 139:184–202
Edgington ES, Onghena P (2007) Randomization tests. Statistics: textbooks and monographs. Chapman & Hall, Boca Raton
Good P (2005) Permutation, parametric and bootstrap tests of hypotheses, 3rd edn. Springer series in statistics. Springer, New York
Hájek J, Šidak Z, Sen PK (1999) Theory of rank tests. Academic Press, San Diego
Hayes AF (1997) Cautions in testing variance equality with randomization tests. J Stat Comput Simul 59:25–31
Hayes AF (2000) Randomization tests and the homoscedasticity assumption when comparing group means. Anim Behav 59:653–656
Janssen A (1997) Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. Stat Probab Lett 36:9–21
Janssen A (2005) Resampling Student's t-type statistics. Ann Inst Stat Math 57:507–529
Janssen A, Pauls T (2003a) How do bootstrap and permutation tests work? Ann Stat 31:768–806
Janssen A, Pauls T (2003b) A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. Comput Stat 20(3):369–383
Janssen A, Pauly M (2009). Asymptotics and effectiveness of conditional tests with applications to randomization tests. Technical report, University of Duesseldorf
Johnson NL, Kotz S (1970) Distributions in statistics: continuous univariate distributions, vols 1, 2. Houghton Mifflin, New York
Lehmann EL, Romano JP (2005) Testing statistical hypotheses. Springer, New York
Manly BFJ (1997) Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall, London
Neuhaus G (1993) Conditional rank tests for the two-sample problem under random censorship. Ann Stat 21:1760–1779
Romano JP (1990) On the behavior of randomization tests without a group invariance assumption. J Am Stat A 85:686–692
Sakaori F (2002) Permutation test for equality of correlation coefficients in two populations. Commun Stat, Simul Comput 31:641–651

Van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Zhang J, Boos DD (1993) Testing hypotheses about covariance matrices using bootstrap methods. Commun Stat, Theory Methods 22(3):723–739
Zhu L-X, Ng KW, Jing P (2002) Resampling methods for homogeneity tests of covariance matrices. Stat Sin 12(3):769–783