A note on permutation tests

W. Liu[a], F.Bretz[b], H. Dette[c]

[a]S3RI and School of Mathematics,

University of Southampton, UK

w.liu@maths.soton.ac.uk   Tel: 44 2380 598688

[b]Novartis Pharma AG, Basel, CH-4002, Switzerland

[c]Department of Mathematics,

Ruhr-Universität Bochum, Germany

Abstract

Permutation tests were first introduced in Eden and Yates (1933), Fisher (1935) and Pitman (1937a, 1937b, 1938), and are popular nowadays due to several nice properties they possess and the cheap availability of computation power of modern computers. In this paper, we demonstrate potential power loss of permutation tests using the prototype permutation test for comparing two populations that may possibly be different only in locations. Specifically, we show that the reference distribution used for this permutation test depends on the true value of the unknown parameter that is being tested and this may reduce the power in comparison with the standard parametric test especially for small sample sizes. The observations made for this particular permutation test in this paper is applicable to numerous other permutation tests and so users should be aware of this potential power loss of permutation tests for small sample sizes.

Keywords: Hypotheses testing; Permutation test; Reference distribution; Sampling without replacement; Statistical simulation.

Mathematics Subject Classification: 62G10, 62G20, 62E17, 62E20

1 Introduction

Permutation tests have a long history going back to Eden and Yates (1933), Fisher (1966, First edition 1935) and Pitman (1937a, 1937b, 1938); see David (2008) for a description of the early development. Advantages of permutation tests include being nonparametric, asymptotically as powerful as standard parametric tests (cf. Hoeffding, 1952), making full use of the observations rather than just the ranks and so often more powerful than rank-based tests (cf. Kempthorne and Doerfler, 1969, Good, 2005, p.47), and reliable for both small and large sample sizes. With the cheap availability of computation power, permutation tests are popular nowadays. Readers are referred to the books Edgington and Onghena (2007), Good (2005), Manly (2006) and Pesarin and Salmaso (2010) for excellent overviews.

The purpose of this paper is to illuminate potential power loss of permutation tests in comparison with standard parametric tests especially when sample sizes are small. Numerous permutation tests for different settings have been proposed in the statistical literature and this paper focuses on the permutation test for comparing two populations, which may possibly differ in locations only, both for ease of exposition of the main idea of this paper and for this test being the prototype of permutation tests considered first by Pitman (1937a) and used routinely in the statistical literature to illustrate the key ideas of permutation tests. Similar observations as described in this paper can be made for permutation tests in more general situations.

A brief description of how the permutation test for this particular problem is carried out is given in Section 2. In Section 3, we point out from the large sample viewpoint the reason why this permutation test may have lower power than the standard large sample two-sample test. This potential loss of power is demonstrated by statistical simulation in Section 4 when the sample sizes are small. Some concluding remarks are contained in Section 5. Finally, the Appendix provides some mathematical details for the statements in Section 3.

2 The permutation test

Assume $X_1, \ldots, X_m$ are i.i.d. observations from the first population with probability density function (pdf) f and, independently, $Y_1, \ldots, Y_n$ are i.i.d. observations from the second population with pdf $f(\cdot - \delta)$, where $\delta \in \mathbb{R}$ is an unknown parameter. We are interested in testing the hypothesis $H_0 : \delta = 0$ against the (two-sided) alternative $H_a : \delta \neq 0$, without assuming a specific form for $f(\cdot)$. Below is how the permutation (PM) test should be carried out; see, e.g., Canay *et al.* (2017), Chung and Romano (2013), Edgington and Onghena (2007), Good (2005), Manly (2006) and Ernst (2004).

To be precise let $N = m + n$ denote the total sample size, define

$$Z = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n), \tag{2.1}$$

as the vector of all observations, ket $\pi = (\pi(1), \ldots, \pi(N))$ be a permutation of the N indexes $\{1, \ldots, N\}$ and denote by $\mathbf{G}_N$ the set of all $N!$ permutations of $\{1, \ldots, N\}$. Define the test statistic

$$\begin{aligned} T_{m,n}(Z) &= T_{m,n}(Z_1, \ldots, Z_N) = \sqrt{N} \left(\frac{Z_1 + \ldots + Z_m}{m} - \frac{Z_{m+1} + \ldots + Z_N}{n} \right) \\ &= \sqrt{N} \left(\bar{X} - \bar{Y} \right), \end{aligned}$$

where $\bar{X} = \sum_{i=1}^m X_i/m$ and $\bar{Y} = \sum_{i=1}^m Y_i/m$ are the corresponding sample means. Under the null hypothesis $H_0 : \delta = 0$, the joint distribution of $Z_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ is the same as that of Z for any $\pi \in \mathbf{G}_N$, and so the $N!$ values $T_{m,n}(Z_\pi)$ corresponding to the$N!$ permutations $\pi \in \mathbf{G}_N$ are equally likely to occur conditioning on the observed vector Z.

We denote the order statistic of the sample $|T_{m,n}(Z_{\pi_1})|, \ldots |T_{m,n}(Z_{\pi_{N!}})|$ by

$$|T_{m,n}^{(1)}(Z)| \leq \cdots \leq |T_{m,n}^{(N!)}(Z)|$$

and define for a given nominal level $\alpha \in (0,1)$ the quantity $k = N! - \lceil \alpha N! \rceil$ where $\lceil \alpha N! \rceil$ denotes the largest integer less than or equal to $\alpha N!$. Let $M^+(Z)$ and $M^0(Z)$ be the number of values $|T_{m,n}^{(j)}(Z)|$ $(j = 1, \ldots, N!)$ that are greater than $|T_{m,n}^{(k)}(Z)|$ and equal to $|T_{m,n}^{(k)}(Z)|$, respectively, and set

$$a(Z) = \frac{\alpha N! - M^+(Z)}{M^0(Z)}.$$

The permutation test for the hypotheses $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ is finally defined as

$$\phi_{PM}(Z) = \begin{cases} 1 & \text{if } |T_{m,n}(Z)| > |T_{m,n}^{(k)}(Z)| \\ a(Z) & \text{if } |T_{m,n}(Z)| = |T_{m,n}^{(k)}(Z)| \\ 0 & \text{if } |T_{m,n}(Z)| < |T_{m,n}^{(k)}(Z)|. \end{cases} \tag{2.2}$$

Under the null hypothesis $H_0 : \delta = 0$, since the $N!$ values $|T_{m,n}(Z_\pi)|$ (as well as $T_{m,n}(Z_\pi)$) corresponding to the $N!$ permutations $\pi \in \mathbf{G}_N$ are equally likely to occur conditioning on the given Z, we have

$$\mathbb{P}_\Pi \left\{ H_0 \text{ is rejected by } \phi_{PM} \,|\, Z \right\} = \frac{M^+(Z)}{N!} + \frac{M^0(Z)}{N!} a(Z) = \alpha,$$

where $\Pi = (\Pi(1), \ldots, \Pi(N))$ denotes the random permutation that has a uniform distribution on $\mathbf{G}_N$. Since the test ϕ_{PM} in (2.2) is of size α conditioning on Z, it is also of size α unconditionally.

Let

$$T_{m,n}(Z_\Pi) = \sqrt{N}\left(\frac{Z_{\Pi(1)} + \ldots + Z_{\Pi(m)}}{m} - \frac{Z_{\Pi(m+1)} + \ldots + Z_{\Pi(N)}}{n}\right). \tag{2.3}$$

Then, after accounting for discreteness, the PM-test (2.2) rejects the null hypothesis if $T_{m,n}(Z)$ is either less than the $\alpha/2$ quantile or larger than the $1-\alpha/2$ quantile of the distribution of $T_{m,n}(Z_\Pi)$. Hence the conditional (on the given Z) distribution of $T_{m,n}(Z_\Pi)$ is the **reference distribution** used by the PM-test to judge whether or not the observed $T_{m,n}(Z)$ is too extreme and so H_0 should be rejected.

The (permutation) distribution of $T_{m,n}(Z_\Pi)$ involves $N!$ terms, which are reduced actually to $\binom{N}{m}$ terms for the test statistic $T_{m,n}(Z)$ used here and can be prohibitively large, even for moderate values of m and n. Only when the number $\binom{N}{m}$ is small, all the $\binom{N}{m}$ possible values of $T_{m,n}(Z_\Pi)$ can be easily enumerated and in this case the PM-test (2.2) be computed exactly.

In practice the permutation distribution is approximated routinely by statistical simulation in the following way. Generate a random permutation Π and compute $T_{m,n}(Z_\Pi)$. Repeat this for a large number of times, R, say $R = 10,000$ for example, and these R values of $T_{m,n}(Z_\Pi)$ allow the permutation distribution or its quantiles to be estimated accurately. For example, the $(1-\alpha/2)$-quantile of $|T_{m,n}(Z_\Pi)|$ can be estimated by the $1-\alpha$ sample quantile of the R values of $|T_{m,n}(Z_\Pi)|$, which is the $\lceil(1-\alpha)R\rceil$-th largest value of the $|T_{m,n}(Z_\Pi)|$ values. Hence the PM-test (2.2) rejects H_0 only if $|T_{m,n}(Z)|$ exceeds this sample quantile. Alternatively, the p-value is computed as the proportion of the R simulated $|T_{m,n}(Z_\Pi)|$ values that are equal to or larger than the observed $|T_{m,n}(Z)|$; H_0 is rejected only if the p-value is less than α.

3 The potential power loss of the PM-test

Note that the reference distribution $T_{m,n}(Z_\Pi)$ in (2.3) is conditional on, and built from, the two observed samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ by using the random permutation Π. But the unknown parameter δ of the second population pdf $f(x-\delta)$ from which $Y_1, \ldots, Y_n$ have been generated may well be **not equal to zero**. If the true status of nature that has generated the $Y_1, \ldots, Y_n$ is $\delta \neq 0$. this true status of nature does not change after the utterance of the words “let's assume H_0 is true”. Hence it is clear that the reference distribution of the PM-test works as intended only if the true value of δ is equal to zero. Although in statistical practice it not known whether δ is equal to zero or not (since we are only presented with the observed Z) the effect of δ on the reference distribution has not been investigated in the literature.

In the following discussion we will answer this question. Theorem 3.1 below gives the asymptotic distribution of $T_{m,n}(Z_\Pi)$, both conditionally (on the observed Z) and unconditionally, as the sample sizes m and n become large. The proof is given in the Appendix. It is immediately clear from the theorem how the asymptotic distribution of $T_{m,n}(Z_\Pi)$ depends through the variance on the true but unknown value δ inherited by the Y_i's. Throughout this paper $\mathcal{N}(\mu, \tau^2)$ denotes a normal distribution with mean μ and variance τ^2.

Theorem 3.1. *Assume $X_1, \ldots, X_m$ are i.i.d. with pdf $f(x)$ and, independently, $Y_1, \ldots, Y_n$ are i.i.d. with pdf $f(y - \delta)$ for some $\delta \in \mathbb{R}$, with $Var(X_1) = Var(Y_1) = \sigma^2 < \infty$. Let $m \to \infty$ and $n \to \infty$ with $m/N \to \lambda \in (0,1)$ and $n/N \to 1 - \lambda \in (0,1)$. Then, conditional on the sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$,*

$$T_{m,n}(Z_\Pi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^2 + \sigma^2/(\lambda(1-\lambda))) \quad a.s.$$

that is, given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, the conditional distribution of $T_{m,n}(Z_\Pi)$ is asymptotically normal with mean 0 and variance $\delta^2 + \sigma^2/(\lambda(1-\lambda))$. Furthermore, the unconditional distribution of $T_{m,n}(Z_\Pi)$ has the same asymptotic normal distribution, i.e.

$$T_{m,n}(Z_\Pi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^2 + \sigma^2/(\lambda(1-\lambda))) \; .$$

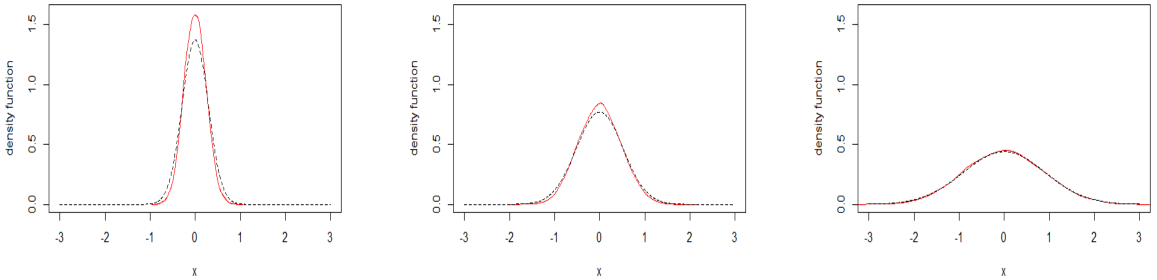

Figure 1: *The conditional pdf's (solid curves) and the corresponding asymptotic normal pdf's (dotted curves) of the statistic $T_{m,n}(Z_\Pi)$ in (2.3) for $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$, with $m = 20$ and $n = 29$. Left panel: $\delta = 0$; middle panel: $\delta = 3$; right panel: $\delta = 6$.*

In Figure 1 we presents three pdf's of the conditional distributions of $T_{m,n}(Z_\Pi)$ in (2.3) corresponding to $\delta = 0, 3, 6$ with $m = 20$, $n = 29$ and $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$, each based on one randomly observed $Z = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ and $R = 10,000$ simulations of Π. Each pdf (solid curve) is produced from these simulated values of $T_{m,n}(Z_\Pi)$ by using the kernel density estimate (cf. Wand and Jones, 1995, and the companion R package KernSmooth). The three conditional pdf's are given by the three solid curves, while the three dotted curves give the corresponding asymptotic normal pdf's from Theorem 3.1. We observe that the small sample conditional distribution of $T_{m,n}(Z_\Pi)$ is close to the large sample asymptotic normal distribution and that the distribution of $T_{m,n}(Z_\Pi)$ clearly depends on the true value of δ. The ideal reference distribution for the PM-test is given by the steepest pdf in the left panel of Figure 1 which happens only when the true value of δ is equal to zero. But the reference distribution actually used by the PM-test may well be a flatter pdf that corresponds to a non-zero δ value.

Now let us look at the reference distribution of the standard large sample (LS) test. Let $\tau_\lambda^2 = \sigma^2/(\lambda(1-\lambda))$. Since under the assumptions of Theorem 3.1 the quantity $T_{m,n}(Z)+\sqrt{N}\delta$ converges in distribution to $\mathcal{N}(0,\tau_\lambda^2)$, the asymptotic level α test is defined by

$$\phi_{LS}(Z) = \begin{cases} 1 & \text{if } |T_{m,n}(Z)| > z_{1-\alpha/2}\hat{\sigma}/\sqrt{\lambda(1-\lambda)} \\ 0 & \text{if } |T_{m,n}(Z)| \le z_{1-\alpha/2}\hat{\sigma}/\sqrt{\lambda(1-\lambda)} \end{cases} \tag{3.1}$$

where z_q denotes the q-quantile of standard normal distribution and

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{m}(X_i-\bar{X})^2 + \sum_{j=1}^{n}(Y_j-\bar{Y})^2\right)/(N-2)$$

is the pooled estimate of the common variance. Since $\hat{\sigma}^2 \overset{a.s.}{\to} \sigma^2$, the LS-test essentially uses $\mathcal{N}(0,\tau_\lambda^2)$ as the reference distribution in judging whether or not the observed $T_{m,n}(Z)$ is too extreme and so H_0 should be rejected.

The reference distribution $\mathcal{N}(0,\tau_\lambda^2)$ of the LS-test is therefore clearly different from the asymptotic reference distribution $\mathcal{N}(0,\delta^2+\tau_\lambda^2)$ of the PM-test unless the true value of δ in the density $f(y-\delta)$ of $Y_1,\ldots,Y_m$ is equal to zero. As a result of Theorem 3.1, asymptotically, the PM-test rejects H_0 if and only if

$$|T_{m,n}(Z)| > z_{1-\alpha/2}\sqrt{\delta^2+\tau_\lambda^2}$$

while the LS-test rejects H_0 if and only if

$$|T_{m,n}(Z)| > z_{1-\alpha/2}\sqrt{\tau_\lambda^2}\,.$$

The corresponding p-values of the PM-and LS-test for the observed $T_{m,n}(Z)$ are given asymptotically by

$$\mathbb{P}\{\ |\mathcal{N}(0,\delta^2+\tau_\lambda^2)| \ge |T_{m,n}(Z)|\ \}$$

and

$$\mathbb{P}\{\ |\mathcal{N}(0,\tau_\lambda^2)| \ge |T_{m,n}(Z)|\ \}$$

respectively. These expressions indicate that, asymptotically, the PM-test rejects H_0 less frequently, and so is less powerful, than the LS-test when $\delta \neq 0$.

In order to investigate the difference between the two tests in more detail we derive more accurate approximations to the powers of both tests. By noting that $T_{m,n}(Z) + \sqrt{N}\delta$ has an asymptotic distribution $\mathcal{N}(0, \tau_\lambda^2)$ for any $\delta \in \mathbb{R}$, a straightforward calculation shows that the power of the PM-test can be approximated by

$$\begin{aligned}\mathbb{P}\{H_0 \text{ is rejected by } \phi_{PM}\} &\approx \mathbb{P}\{\ |\mathcal{N}(-\sqrt{N}\delta, \tau_\lambda^2)| > z_{1-\alpha/2}\sqrt{\delta^2 + \tau_\lambda^2)}\ \} \\ &= \Phi\left(-z_{\alpha/2}\sqrt{1 + \tfrac{\delta^2}{\tau_\lambda^2}} - \tfrac{\sqrt{N}\delta}{\tau_\lambda}\right) + \Phi\left(-z_{\alpha/2}\sqrt{1 + \tfrac{\delta^2}{\tau_\lambda^2}} + \tfrac{\sqrt{N}\delta}{\tau_\lambda}\right)\end{aligned}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution. On the other hand the power of the LS-test can be approximated by

$$\begin{aligned}\mathbb{P}\{H_0 \text{ is rejected by } \phi_{LS}\} &\approx \mathbb{P}\{\ |\mathcal{N}(-\sqrt{N}\delta, \tau_\lambda^2)| > z_{1-\alpha/2}\tau_\lambda\ \} \\ &= \Phi\left(-z_{\alpha/2} - \tfrac{\sqrt{N}\delta}{\tau_\lambda}\right) + \Phi\left(-z_{\alpha/2} + \tfrac{\sqrt{N}\delta}{\tau_\lambda}\right)\end{aligned}$$

It is clear that both asymptotic power functions approach α as $\delta \to 0$ and 1 as $|\delta| \to \infty$ as expected. Moreover, if $N \to \infty$ and $\delta \to 0$ in such a way that $\sqrt{N}\delta$ approaches a non-zero constant, say $\xi\,\tau_\lambda$, we obtain the approximations (by using additionally that $\sqrt{1+x} = 1 + x/2 + o(x)$ as $x \to 0$)

$$\begin{aligned}\mathbb{P}\{H_0 \text{ is rejected by } \phi_{PM}\} &\approx \Phi\left(-z_{\alpha/2} - \xi - z_{\alpha/2}\,\tfrac{\xi^2}{2N}\right) + \Phi\left(-z_{\alpha/2} + \xi - z_{\alpha/2}\,\tfrac{\xi^2}{2N}\right), \\ \mathbb{P}\{H_0 \text{ is rejected by } \phi_{LS}\} &\approx \Phi\left(-z_{\alpha/2} - \xi\right) + \Phi\left(-z_{\alpha/2} + \xi\right).\end{aligned}$$

This implies (as $N \to \infty$, $\sqrt{N}\delta \to \xi\,\tau_\lambda > 0$)

$$\mathbb{P}\{H_0 \text{ is rejected by } \phi_{PM}\} - \mathbb{P}\{H_0 \text{ is rejected by } \phi_{LS}\} \to 0,$$

which agrees with Hoeffding (1952, pp.172) and indicates that for large sample sizes both tests behave very similar.

Nevertheless, for small sample sizes, there might be discernible differences between the powers of the two tests. We demonstrate by simulation in the next section that the power of the PM-test could be considerably smaller than the power of the LS-test for small sample sizes m and n.

4 A simulation study of power

We set $f(\cdot)$ to a standard normal pdf, $\alpha = 1\%$, $\lambda = 1/2$ (and so $m = n = N/2$ for $N = 12, 18$), and $\Delta = (\delta/\sigma)\sqrt{\lambda(1-\lambda)}$ over the grid points $seq(0, 1, 0.04)$ by noting that both the power functions of the PM-test and LS-test depend on δ and σ only through the ratio δ/σ. Hence, without loss of generality, σ is set to be 1 and so $\delta = \Delta/\sqrt{\lambda(1-\lambda)} = 2\Delta$. Note that $f(\cdot)$ is chosen as a normal pdf in order that the approximate size α two-sample LS-test will be replaced by the exact size α two-sample t-test for a fairer power comparison with the PM-test. Note that $\lambda = 1/2$ maximizes $\Delta = (\delta/\sigma)\sqrt{\lambda(1-\lambda)}$ over $\lambda \in (0,1)$. The cases of $\lambda \neq 1/2$ or $\alpha = 5\%$ are also considered in our simulation study.

To simulate the power of a test at a given δ value, one independent $Z = (X_1, \dots, X_m, Y_1, \dots, Y_n)$ is generated with $X_1, \cdots, X_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\delta, 1)$, and the outcome of the test for this Z, either rejection of H_0 or not, is recorded. Repeat this $R = 10,000$ times, and the proportion of times out of the R generated independent Z values that H_0 is rejected is taken as the power of the test at δ. For each independent Z generated, the LS-test is carried out by using the rejection rule in (3.1) but with $z_{1-\alpha/2}$ being replaced by the larger $(1-\alpha/2)$-quantile of the t distribution with $N-2$ degrees of freedom in order to keep the size equal to α for any $N > 2$ (in other words the two-sample t-test is actually used instead of the asymptotic test). Note that the PM-test requires another inner loop to compute the p-value of each Z generated, as described in the penultimate paragraph of Section 2 and involving $R = 10,000$ samples of the random permutation Π, from which the rejection of H_0 or not is determined for that Z.

In Figure 2 we display the difference between the power of the t-test and the power of the PM-test (nominal level $\alpha = 0.01$) over a set of grid points Δ for the sample sizes $n = m = 6$ $(N = 12)$ and $n = m = 9$ $(N = 18)$. The results show that in the case $N = 12$ the power of the t-test can be 5% larger than that of the PM-test. This occurs around $\Delta = 0.9$, with the power of the PM-test being about 0.57 and the power of the t-test about 0.62. For larger sample sizes there is little difference as predicted by the asymptotic results in Section 3.

Given the results in Figure 2 we have further compared the power of the t-test and PM-test over the grid points $\delta = seq(0, 4, 0.1)$ for even smaller total sample sizes $N = seq(9, 16)$. In this case, the PM-test is carried out by using the formula in (2.2) and so involves enumerating all the $\binom{N}{m}$ combinations. In particular, for each N in $seq(9, 16)$, the sample sizes (m, n) are chosen so that the equal probability $1/\binom{N}{m}$ assigned to each of the $\binom{N}{m}$ possible $T_{m,n}(Z_\pi)$-values is less than the exact size $\alpha = 0.01$ of the test. For example, when $N = 9$, the (m, n) that satisfy this constraint are $(5, 4)$ only. The smallest N is set to be 9 because $1/\binom{8}{m} > \alpha = 0.01$ for any natural number $m \leq 8$.

Among the sample sizes (m, n) considered, Figure 3 presents exemplarily the powers and the difference between the powers of the t-test and PM-test over the grid points $\delta = seq(0, 4, 0.1)$ for the cases of $(m, n) = (5, 4)$ and $(m, n) = (5, 5)$. For example, we observe for $(m, n) = (5, 4)$ the power of the t-test can be larger than that of the PM-test by 14% at $\delta = 3.0$, with the t-test having power 0.791 and the PM-test having power 0.655.

We have also done the simulation study for the level $\alpha = 5\%$. In this case the smallest sample sizes are given by $m = 4$ and $n = 3$, and $N = seq(7, 16)$ is used. Among the cases of (m, n) considered, the largest power difference observed between the t-test and PM-test is 6% when $m = 4$ and $n = 3$, and the next largest power difference between the t-test and PM-test is 4.5% when $m = 4$ and $n = 4$. Figure 4 plots the power and the difference between the powers of the t-test and PM-test over the grid points $\delta = seq(0, 4, 0.1)$ for the cases of $(m, n) = (4, 3)$ and $(m, n) = (4, 4)$. Compared with Figure 3 the power difference between the t-test and PM-test is smaller for the level $\alpha = 0.05$ than for $\alpha = 0.01$.

The only power comparison of the two-sample t-test and PM-test for small sample sizes that we have been able to find in the statistical literature is Tanizaki (1997) for the cases of $m = n = 5, 7, 9$ and $\delta = 0, 0.5, 1.0$. The simulation results in Tanizaki (1997, Table 1, $\sigma = 1$) show that the power differences between the t-test and PM-test are always less than 1% and so the two tests barely differ in power. Unfortunately, it is not clear how many replications are used to simulate the power of the tests in Tanizaki (1997). Our simulation results show that the tests could differ in power by as much as 14% when $(m, n) = (5, 4)$, and so differ from the observations made in Tanizaki (1997).

The conclusion from this simulation study is that, if one is comfortable about the assumption on normality of the populations, then the two-sample t-test is clearly preferable to the PM-test, especially when the sample sizes are small. Otherwise, the PM-test is clearly preferable to the t-test since the t-test no longer controls the type I error rate at the nominal level α.

5 Conclusions

We considered the classical two sample problem with a potential difference δ in the mean. In contrast to the large sample test or t-test, the reference distribution of the permutation test depends (implicitly) on the true but unknown value δ. This dependence is clearly reflected in the asymptotic distribution of statistic $T_{m,n}(Z_\Pi)$ of the permutation test given in Theorem 3.1. As a consequence the power of the PM-test could potentially be considerably smaller than that of the t-test especially when the sample sizes are small. This is demonstrated by our simulation results. So the users should be aware of this potential power loss of the PM-test.

As a prototype we have concentrated on the simple two-sample location problem, but similar observations can be made for other permutation tests too. While the asymptotic arguments in the general case are very similar to the ones given in Theorem 3.1 it will be useful for practical purposes to investigate the power loss for the numerous permutation tests proposed in the statistical literature, if the sample sizes are small.

It has been proposed in ter Braak (1992) to use the reference distribution based on permuting the residuals, i.e. the mean-centered observations $Z^* = (X_1 - \bar{X}, \ldots, X_m - \bar{X}, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ for the two-sample problem considered in this paper. It can be shown that, under the assumptions of Theorem 3.1, the asymptotic distribution of $T_{m,n}(Z^*_\Pi)$ is $\mathcal{N}(0, \sigma^2/(\lambda(1-\lambda)))$ and so does not depend on δ. But by using the distribution of $T_{m,n}(Z^*_\Pi)$ as the reference distribution, the resultant PM-test no longer has its size equal to α for given sample sizes (m, n) since the components of $Z^* = (X_1-\bar{X}, \ldots, X_m-\bar{X}, Y_1-\bar{Y}, \ldots, Y_n-\bar{Y})$ are not exchangeable. Hence one of the key properties of the PM-test (2.2) would be sacrificed.

6 Appendix

We prove Theorems 3.1 in this appendix by appealing to the lemma below on combinatorial central limit theorem which is taken from DasGupta (2008, Pages 67-68, Theorem 5.5) and proved in Hoeffding (1951). The proof is elementary. Other proofs are possible. For example, one may use the coupling argument of Chung and Romano (2013, Section 5.3) to prove the theorem, but it requires considerable more efforts to understand the coupling argument first.

Lemma 6.1. *Let $a_N(i), b_N(i)$ $(i = 1, \ldots, N)$ be two double arrays of constants with $\bar{a}_N = \sum_{i=1}^N a_N(i)/N$ and $\bar{b}_N = \sum_{i=1}^N b_N(i)/N$. If*

$$N^{r/2-1} \frac{\sum_{i=1}^N \left(a_N(i) - \bar{a}_N\right)^r}{\left(\sum_{i=1}^N \left(a_N(i) - \bar{a}_N\right)^2\right)^{r/2}} = O(1) \quad \text{for any} \quad r > 2 \tag{6.1}$$

and

$$\frac{\max_{1 \le i \le N} \left(b_N(i) - \bar{b}_N\right)^2}{\sum_{i=1}^N \left(b_N(i) - \bar{b}_N\right)^2} = o(1)\,, \tag{6.2}$$

then $\left(S_N - E(S_N)\right)/\sqrt{Var(S_N)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, where $S_N = \sum_{i=1}^N a_N(i) b_N(\Pi(i))$ and Π is the random permutation that has a uniform distribution on $\mathbf{G}_N$. Furthermore, with $c_N(i,j) = a_N(i) b_N(j)$, $1 \le i, j \le N$,

$$E(S_N) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N c_N(i,j)$$

and

$$Var(S_N) = \frac{1}{N-1} \sum_{i=1}^N \sum_{j=1}^N d_N^2(i,j) \tag{6.3}$$

where

$$d_N(i,j) = c_N(i,j) - \frac{1}{N}\sum_{k=1}^{N} c_N(k,j) - \frac{1}{N}\sum_{k=1}^{N} c_N(i,k) + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} c_N(k,l). \qquad \square \tag{6.4}$$

In order to use this lemma to prove Theorem 3.1, the statistic $T_{m,n}(Z_\Pi)$ defined in (2.3) can be written as

$$T_{m,n}(Z_\Pi) = S_N = \sum_{i=1}^{N} a_N(i) b_N(\Pi(i))$$

with

$$a_N(i) = \begin{cases} \sqrt{N}/m, & 1 \le i \le m \\ -\sqrt{N}/n, & m+1 \le i \le N \end{cases} \quad \text{and} \quad b_N(i) = Z_i = \begin{cases} X_i, & 1 \le i \le m \\ Y_{i-m}, & m+1 \le i \le N. \end{cases}$$

It is straightforward to show that (6.1) is true in this case. To prove (6.2), note first that

$$\frac{1}{N}\sum_{i=1}^{N}\left(b_N(i) - \bar{b}_N\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(b_N(i)\right)^2 - \left(\bar{b}_N\right)^2$$

with

$$\bar{b}_N = \bar{Z}_N = \frac{1}{N}\sum_{i=1}^{m} X_i + \frac{1}{N}\sum_{i=1}^{n} Y_i \overset{a.s.}{\to} \lambda E(X_1) + (1-\lambda)E(Y_1)$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\left(b_N(i)\right)^2 = \frac{1}{N}\sum_{i=1}^{N} Z_i^2 = \frac{1}{N}\sum_{i=1}^{m} X_i^2 + \frac{1}{N}\sum_{i=1}^{n} Y_i^2 \overset{a.s.}{\to} \lambda E(X_1^2) + (1-\lambda)E(Y_1^2)$$

by the law of large numbers. Hence

$$\frac{1}{N}\sum_{i=1}^{N}\left(b_N(i) - \bar{b}_N\right)^2 \overset{a.s.}{\to} \lambda E(X_1^2) + (1-\lambda)E(Y_1^2) - \left(\lambda E(X_1) + (1-\lambda)E(Y_1)\right)^2 . \tag{6.5}$$

Furthermore, note that

$$\frac{1}{N}\max_{1\le i\le N}\left(b_N(i) - \bar{b}_N\right)^2 \le \frac{2}{N}\max_{1\le i\le m} X_i^2 + \frac{2}{N}\max_{1\le i\le n} Y_i^2 + \frac{2}{N}\left(\bar{b}_N\right)^2,$$

$\left(\bar{b}_N\right)^2/N \overset{a.s.}{\to} 0$, and that $\max_{1\le i\le m} X_i^2/N \overset{a.s.}{\to} 0$ and $\max_{1\le i\le n} Y_i^2/N \overset{a.s.}{\to} 0$ which follow directly from $\sum_{i=1}^{m} X_i^2/m \overset{a.s.}{\to} E(X_1^2)$ and $\sum_{i=1}^{n} Y_i^2/n \overset{a.s.}{\to} E(Y_1^2)$, respectively. Hence

$$\frac{1}{N}\max_{1\le i\le N}\left(b_N(i) - \bar{b}_N\right)^2 \overset{a.s.}{\to} 0. \tag{6.6}$$

Combination of (6.5) and (6.6) shows that (6.2) holds for almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$. It follows therefore from Lemma 6.1 that, for almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, the conditional distribution of $T_{m,n}(Z_\Pi)$ is asymptotically normal with mean

$$E(S_N) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} a_N(i)b_N(j) = \frac{1}{N}\sum_{i=1}^{N} a_N(i)\sum_{j=1}^{N} b_N(j) = 0$$

and variance $Var(S_N)$. To find $Var(S_N)$ from (6.3), note from (6.4) that

$$d_N(i,j) = a_N(i)Z_j - a_N(i)\bar{Z}_N, \quad d_N^2(i,j) = (a_N(i))^2\left(Z_j^2 - 2Z_j\bar{Z}_N + \left(\bar{Z}_N\right)^2\right),$$

and so

$$\begin{aligned} Var(S_N) &= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N} d_N^2(i,j) \\ &= \frac{1}{N-1}\sum_{i=1}^{N}(a_N(i))^2\left(\sum_{j=1}^{N} Z_j^2 - N\left(\bar{Z}_N\right)^2\right) \\ &= \frac{N}{N-1}\left(\frac{1}{m}+\frac{1}{n}\right)\left(\sum_{j=1}^{N} Z_j^2 - N\left(\bar{Z}_N\right)^2\right) \\ &\overset{a.s.}{\to} \frac{1}{\lambda(1-\lambda)}\left(\lambda E(X_1^2) + (1-\lambda)E(Y_1^2) - (\lambda E(X_1) + (1-\lambda)E(Y_1))^2\right) \\ &= \delta^2 + \sigma^2/(\lambda(1-\lambda)). \end{aligned}$$

This completes the proof of that, for almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, the conditional distribution of $T_{m,n}(Z_\Pi)$ is asymptotically $\mathcal{N}(0, \delta^2 + \sigma^2/(\lambda(1-\lambda)))$.

For the unconditional distribution, note that

$$T_{m,n}(Z_\Pi) = T_{m,n}(Z_\Pi^0) + T_{m,n}(\Delta_\Pi) \tag{6.7}$$

where $Z^0 = (X_1, \cdots, X_m, Y_1 - \delta, \cdots, Y_n - \delta)$ and $\Delta = (0, \cdots, 0, \delta, \cdots, \delta) \in \mathbb{R}^N$ whose first m components are zero and the last n components are δ. By using Lemma 6.1 again in a similar way as above but with $b_N(i) = 0$ for $1 \le i \le m$ and $b_N(i) = \delta$ for $m+1 \le i \le N$, it is straightforward to show that $T_{m,n}(\Delta_\Pi)$ is asmptotically $\mathcal{N}(0, \delta^2)$. For $T_{m,n}(Z_\Pi^0)$, since the components of Z^0 are i.i.d. each with pdf $f(x)$, it is clear that $T_{m,n}(Z_\Pi^0)$ has the same distribution as $T_{m,n}(Z^0)$ and has the asymptotic normal distribution $\mathcal{N}(0, \sigma^2/(\lambda(1-\lambda)))$. Furthermore, since the distribution of $T_{m,n}(Z_\Pi^0)$ has nothing to do with Π and so $T_{m,n}(Z_\Pi^0)$ and $T_{m,n}(\Delta_\Pi)$ are independent. Hence, from (6.7), $T_{m,n}(Z_\Pi)$ has the asymptotic distribution $\mathcal{N}(0, \delta^2 + \sigma^2/(\lambda(1-\lambda)))$. This completes the proof of Theorem 3.1.

Acknowledgments. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt T1)

References

1. Canay, I.A., J.P. Romano, and A.M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85, 1013-1030.

2. Chung, E. and J.P. Romano (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41, 484507.

3. David, H.A. (2008). The beginnings of randomization tests. *The American Statisticians*, 62, 70-72.

4. DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* New York: Springer.

5. Eden, T. and F. Yates (1933). On the validity of Fisher's z test when applied to an actual example of non-normal data. *The Journal of Agricultural Science*, 23, 6-17.

6. Edgington, E. and P. Onghena (2007). *Randomization Tests, 4th ed.*. New York: Chapman & Hall/CRC.

7. Ernst, M.D. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, 19(4), 676-685.

8. Fisher, R.A. (1966). *The Design of Experiments, 8th ed.*. Edinburgh: Oliver & Boyd. (Original work published in 1935)

9. Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd ed.*. New York: Springer.

10. Hoeffding, W. (1951). A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, 22, 558-566.

11. Hoeffding, W. (1952). The Large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23, 169-192.

12. Kempthorne, O. and Doerfler, T.E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika*, 56, 231-248.

13. Manly, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology, 3rd ed.*. New York: Chapman & Hall/CRC.

14. Pesarin, F. and L. Salmaso (2010). *Permutation Tests for Complex Data: Theory, Applications and Software.* New York: Wiley

15. Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, 4, 119-130.

16. Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any population. II. The correlation coefficient test. *Royal Statistical Society Supplement*, 4, 225-232.

17. Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population. Part III. The analysis of variance test. *Biometrika*, 29, 322335.

18. Tanizaki, H. (1997). Power comparison of non-parametric tests: small sample properties from Monte Carlo experiments. *Journal of Applied Statistics*, 24, 603-632.

19. ter Braak, C.J.F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA, in *Bootstrapping and Related Techniques* (ed. K.H. Jöckel), Springer-Verlag, pp. 79-86.

20. Wand, M. and M.C. Jones (1995). *Kernel Smoothing.* New York: Springer.

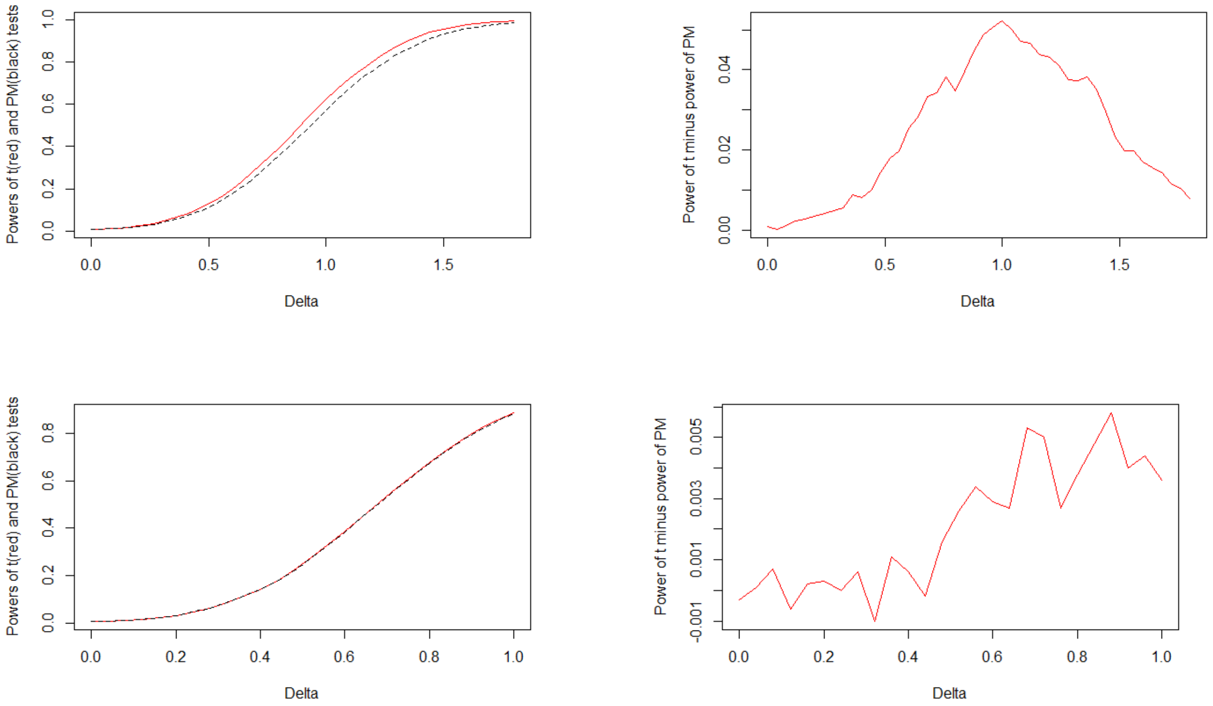

Figure 2: *The simulated power (left panel) and the difference between the powers (right panel) of the t-test and PM-test (with $\alpha = 1\%$) over a grid points of Δ. The samples size is $n = m = 6$ (upper row) and $n = m = 9$ (lower row) .*

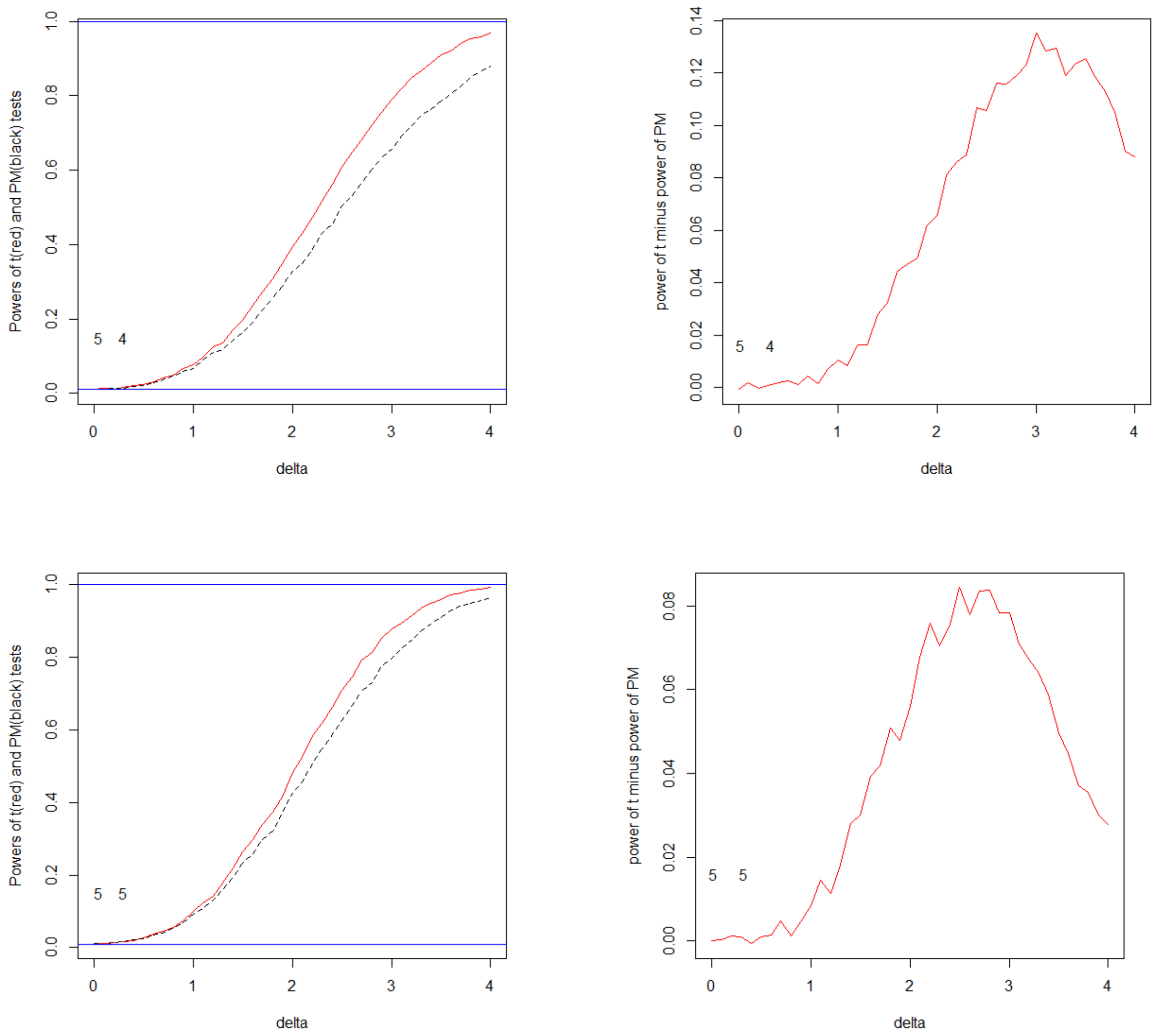

Figure 3: *The power (left panel) and the difference between the powers (right panel) of the t-test and PM-test (nominal level* $\alpha = 1\%$*) over a grid points of* δ*. Upper row:* $(m, n) = (5, 4)$ *as indicated by the numbers 5 and 4 in the figures; lower row* $(m, n) = (5, 5)$ *as indicated by the numbers 5 and 5 in the figures.*

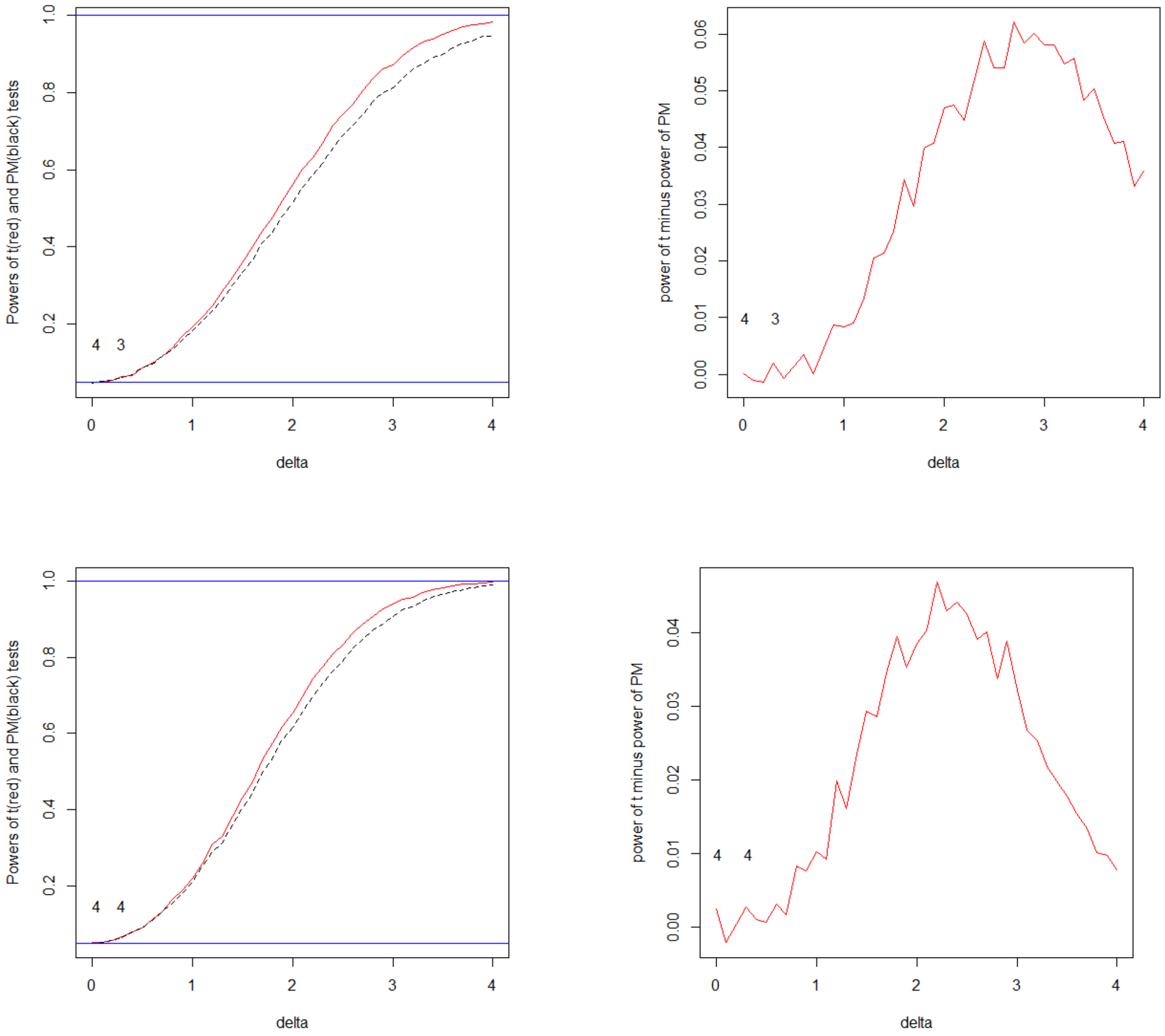

Figure 4: *The power (left panel) and the difference between the powers (right panel) of the t-test and PM-test (nominal level* $\alpha = 5\%$*) over a grid points of* δ*. Upper row:* $(m, n) = (4, 3)$ *as indicated by the numbers 4 and 3; lower row* $(m, n) = (4, 4)$ *as indicated by the numbers 4 and 4.*