TEST (2018) 27:811–825
https://doi.org/10.1007/s11749-017-0571-1

ORIGINAL PAPER

Exact testing with random permutations

Jesse Hemerik[1] · **Jelle Goeman**[1]

Received: 30 May 2017 / Accepted: 15 November 2017 / Published online: 30 November 2017

Abstract When permutation methods are used in practice, often a limited number of random permutations are used to decrease the computational burden. However, most theoretical literature assumes that the whole permutation group is used, and methods based on random permutations tend to be seen as approximate. There exists a very limited amount of literature on exact testing with random permutations, and only recently a thorough proof of exactness was given. In this paper, we provide an alternative proof, viewing the test as a "conditional Monte Carlo test" as it has been called in the literature. We also provide extensions of the result. Importantly, our results can be used to prove properties of various multiple testing procedures based on random permutations.

Keywords Permutation test · Nonparametric test · Resampling

Mathematics Subject Classification 62G09 · 62G10

1 Introduction

Permutation tests are nonparametric tests that are used in particular when the null hypothesis implies distributional invariance under certain transformations (Fisher 1936; Lehmann and Romano 2005; Ernst et al. 2004). Apart from permutations, other groups of transformations can be used, such as rotations (Langsrud 2005).

✉ Jesse Hemerik
j.b.a.hemerik@lumc.nl

[1] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Postzone S5-P, Postbus 9600, 2300 RC Leiden, The Netherlands

When the set of transformations used is not a group, a permutation test can be very conservative or anti-conservative. The first author who explicitly assumed a group structure is Hoeffding (1952). The role of the group structure has recently been emphasized (Southworth et al. 2009; Goeman and Solari 2010). Southworth et al. (2009) note that in particular the set of 'balanced permutations' cannot be used, since it is not a group.

Often it is computationally infeasible to use the whole group of permutations, due to its large cardinality. In that case, random permutations are used, as was first proposed by Dwass (1957). Often a permutation p value based on random permutations is simply seen as an estimate of the permutation p value.

It is known that naively using random permutations instead of all possible permutations can lead to extreme anti-conservativeness (Phipson and Smyth 2010), especially when combined with multiple testing procedures. Therefore, sometimes the identity permutation, which corresponds to the original observation, is included with the random permutations (Ge et al. 2003; Lehmann and Romano 2005). Lehmann and Romano (2005) (p. 636) state that when the identity is added, the estimated p value is stochastically larger than the uniform distribution on [0, 1] under the null. Phipson and Smyth (2010) note that adding the identity can make the permutation test exact, i.e. of level α exactly. They do not mention the role of the underlying group structure. Instead, they view the permutation test as a Monte Carlo test, which is known to be exact in some situations if the original observation is added.

Referring to Monte Carlo is not sufficient, because despite being related, a Monte Carlo test is very different from a permutation test. Monte Carlo samples are drawn from the null distribution. In the permutation context, the random permutations of the data are instead drawn from a conditional null distribution, i.e. the permutation distribution. Hence, the proof by Phipson and Smyth (2010) is incomplete and it remained unclear what assumptions (e.g. a group structure) are essential for the validity of random permutation tests. For example, it is unclear from Phipson and Smyth that random sampling from balanced permutations would lead to invalid tests.

In Hemerik and Goeman (2017), a test is given based on random transformations. In the present paper, we extend this work, investigating fundamental properties of random permutation tests. Our main focus is on the level of tests. Other properties, e.g. consistency, do not generally hold, but can be established for more specific scenarios (Lehmann and Romano 2005; Pesarin 2015; Pesarin and Salmaso 2013) by using results presented here. Our results are general and can be used to prove properties of various multiple testing methods based on random permutations, such as Westfall and Young (1993), Tusher et al. (2001), Meinshausen and Bühlmann (2005) and Meinshausen (2006). In the literature, there are two approaches to proving permutation tests with fixed permutations: a conditioning-based approach (Pesarin 2015) and a more direct approach (Hoeffding 1952; Lehmann and Romano 2005). We will give proofs with both approaches.

The structure of the paper is as follows. In Sect. 2, we review known results on the level of a permutation test based on a fixed group of transformations. The concepts and definitions from Sect. 2 are used throughout the paper. Testing with random permutations is covered in Sect. 3. In Sect. 3.1, permutation tests are contrasted with Monte Carlo tests. Estimation of p values is discussed in Sect. 3.2. Exact tests and

p values based on random transformation are given in Sects. 3.3 and 3.4. In Sect. 4, some additional applications of these results are mentioned.

2 Fixed transformations

Here we discuss tests that use the full group of transformations.

2.1 Basic permutation test

Let X be data taking values in a sample space $\mathcal{X}$. Let G be a finite set of transformations $g : \mathcal{X} \to \mathcal{X}$, such that G is a group with respect to the operation of composition of transformations. This means that G satisfies the following three properties: G contains an identity element (the map $x \mapsto x$); every element of G has an inverse in G; for all $a_1, a_2 \in G$, $a_1 \circ a_2 \in G$. This assumption of a group structure for G is fundamental throughout the paper, since it ensures that $Gg = G$ for all $g \in G$, i.e. that the set G is permutation invariant.

Considering a general group of transformations rather than only permutations is useful, since in many practical situations the group consists of, for example, rotations (Langsrud 2005; Solari et al. 2014) or maps that multiply part of the data by -1 (Pesarin and Salmaso 2010, pp. 54 and 168). We write $g(X)$ as gX. Consider any test statistic $T : \mathcal{X} \to \mathbb{R}$. Throughout this paper, we are concerned with testing the following null hypothesis of permutation invariance.

Definition 1 Let H_{p} be any null hypothesis which implies that the joint distribution of the test statistics $T(gX)$, $g \in G$, is invariant under all transformations in G of X. That is, writing $G = \{a_1, \ldots, a_{\#G}\}$, under H_{p}

$$\big(T(a_1X), \ldots, T(a_{\#G}X)\big) \stackrel{d}{=} \big(T(a_1gX), \ldots, T(a_{\#G}gX)\big) \tag{1}$$

for all $g \in G$.

Note that (1) holds in particular when for all $g \in G$

$$X \stackrel{d}{=} gX.$$

Composite null hypotheses are usually not of the form H_{p}, but for specific scenarios, properties of tests of such hypotheses can be established using results in this paper.

The most basic permutation test rejects H_{p} when $T(X) > T^{(k)}(X)$, where

$$T^{(1)}(X) \leq \cdots \leq T^{(\#G)}(X)$$

are the sorted test statistics $T(gX)$, $g \in G$, and $k = \lceil (1-\alpha)\#G \rceil$ with $\alpha \in [0, 1)$. As is known and stated in the following theorem, this test has level at most α.

Theorem 1 *Under H_p, $\mathbb{P}\{T(X) > T^{(k)}(X)\} \leq \alpha$.*

We now give two proofs: a conditioning-based approach and an approach without conditioning. Both approaches are more or less known. The conditioning-based proof is similar to that in Pesarin (2015), but the setting is more general. For each $x \in \mathcal{X}$, define O_x to be the orbit of x, which is the set $\{gx : g \in G\} \subseteq \mathcal{X}$.

Proof Let $A = \{x \in \mathcal{X} : T(x) > T^{(k)}(x)\}$ be the set of elements of the sample space that lead to rejection. Suppose H_p holds. By the group structure, $Gg = G$ for all $g \in G$. Consequently, $T^{(k)}(gX) = T^{(k)}(X)$ for all $g \in G$. Thus, $\#\{g \in G : gX \in A\} =$

$$\#\{g \in G : T(gX) > T^{(k)}(gX)\} = \#\{g \in G : T(gX) > T^{(k)}(X)\} \leq \alpha \#G.$$

Endow the space of orbits with the σ-algebra that it inherits from the σ-algebra on $\mathcal{X}$. Analogously to the proof of Theorem 15.2.2 in Lehmann and Romano (2005), we obtain

$$\mathbb{P}(X \in A \mid O_X) = \frac{1}{\#G}\#\{g \in G : gX \in A\}.$$

By the argument above, this is bounded by α. Hence,

$$\mathbb{P}(X \in A) = \mathbb{E}\{\mathbb{P}(X \in A \mid O_X)\} \leq \alpha$$

as was to be shown. □

We now state a different proof without conditioning. A similar proof can be found in Hoeffding (1952) and Lehmann and Romano (2005) (p. 634).

Proof By the group structure, $Gg = G$ for all $g \in G$. Hence, $T^{(k)}(gX) = T^{(k)}(X)$ for all $g \in G$. Let h have the uniform distribution on G. Then, under H_p, the rejection probability is

$$\begin{aligned}
&\mathbb{P}\{T(X) > T^{(k)}(X)\} \\
&\quad = \mathbb{P}\{T(hX) > T^{(k)}(hX)\} \\
&\quad = \mathbb{P}\{T(hX) > T^{(k)}(X)\}.
\end{aligned}$$

The first equality follows from the null hypothesis, and the second equality holds since $T^{(k)}(X) = T^{(k)}(hX)$. Since h is uniform on G, the above probability equals

$$\mathbb{E}\left[(\#G)^{-1} \cdot \#\{g \in G : T(gX) > T^{(k)}(X)\}\right] \leq \alpha,$$

as was to be shown. □

The test of Theorem 1 is not always exact. When the data are discrete, then the basic permutation test is often slightly conservative, due to a nonzero probability of tied values in X. Under the following condition, which is often satisfied for continuous data, but usually not for discrete data, the test is exact for certain values of α.

Condition 1 There is a partition $\{G_1, \ldots, G_m\}$ of G with $id \in G_1$ and $\#G_1 = \cdots = \#G_m$, such that under H_p with probability 1 for all $g, g' \in G$, $T(gX) = T(g'X)$ if and only if g and g' are in the same set G_i.

Proposition 1 *Under Condition 1, the test of Theorem 1 is exact if and only if* $\alpha \in \{0, 1/m, \ldots, (m-1)/m\}$.

The proof of this result is analogous to the proof of Theorem 1. As an example where Condition 1 holds, consider a randomized trial where $X \in \mathbb{R}^{2n}$ and the test statistic is

$$T(X) = \sum_{i=1}^{n} X_i - \sum_{i=n+1}^{2n} X_i, \tag{2}$$

where $X_1, \ldots, X_n$ are cases and $X_{n+1}, \ldots, X_{2n}$ are controls and all X_i are independent and identically distributed under the null. Let

$$m = \binom{2n}{n}.$$

If the observations are continuous, then the set of α for which the test is exact is $\{0, 1/m, \ldots, (m-1)/m\}$, reflecting the fact that there are m equivalence classes of size $n!n!$ of permutations that always give the same test statistic.

The test of Theorem 1 is often conservative when the data are discrete, since then Condition 1 is usually not satisfied. Moreover, in many cases, the value 0.05 is not in the set mentioned in Proposition 1, and hence, the permutation test for $\alpha = 0.05$ is conservative, even if Condition 1 is satisfied. The test can be adapted to be exact by randomizing it, i.e. by rejecting H_p with a suitable probability a in the boundary case that $T(X) = T^{(k)}$ (Hoeffding 1952). Here

$$a = a(X) = \frac{\alpha \#G - M^+(X)}{M^0(X)}, \tag{3}$$

where

$$M^+(X) := \#\{g \in G : T(gX) > T^{(k)}(X)\},$$
$$M^0(X) := \#\{g \in G : T(gX) = T^{(k)}(X)\}.$$

This adaptation has the advantage that it is always exact. Even if Condition 1 is satisfied, the adaptation can be useful to guarantee that the level of the test is exactly the nominal level α. On the other hand, this test is less reproducible than the test of Theorem 1, since its outcome may depend on a random decision. Which test is to be preferred would depend on the context.

When the set G is not a group, the test can be highly anti-conservative or conservative. For example, the set of *balanced permutations* is a subset of the set of all permutations, but is not a subgroup. These permutations have been used in various papers since they can have an intuitive appeal. They are discussed in Southworth et al.

(2009), who warn against their use since they lead to anti-conservative tests. The fact that permutations have been used incorrectly illustrates that more emphasis should be put on the assumption of a group structure.

Intuitively, the reason why a group structure is needed for Theorem 1 is the following. Suppose for simplicity that H_p implies that $X \overset{d}{=} gX$ for all $g \in G$. The permutation test works since under H_p, for every permutation $g \in G$ the probability $\mathbb{P}\{T(gX) > T^{(k)}(X)\}$ is the same. The reason is that under H_p, for every $g \in G$, the joint distribution of gX and the set GX, i.e. of (gX, GX), is the same. Indeed, since $G = Gg$ (group structure), the set GX is a function of gX, namely $GX = f(gX)$, with f given by $f(x) = Gx$. Thus, for $g, g' \in G$, $(gX, GX) = (gX, f(gX)) \overset{d}{=} (g'X, f(g'X)) = (g'X, GX)$. When G is not a group, the joint distribution of gX and the set GX is not generally independent of g.

2.2 Permutation *p* values

Permutation p values are p value based on permutations of the data. Here we will discuss permutation p values based on the full permutation group. p values based on random permutations are considered in Sect. 3.4.

It is essential to note that there is often no unique null distribution of $T(X)$, since H_p often does not specify a unique null distribution of the data. Correspondingly, $T^{(k)}(X)$ should not be seen as the $(1-\alpha)$-quantile of *the* null distribution.

When a test statistic t is a function (which is not random) of the data and has a unique distribution under a hypothesis H, then a p value in the strict sense, $\mathbb{P}_H(t \geq t_{obs})$, is defined where t_{obs} is the observed value of t. Since under H_p $T(X)$ does not always have a unique null distribution, often there exists no p value in the strict sense based on this test statistic. However, under Condition 1 the statistic

$$D = \#\big\{g \in G : T(gX) \geq T(X)\big\}$$

does have a unique null distribution. Thus, a p value in the strict sense based on $-D$ is then defined. Denoting by d the observed value of D, we have under H_p

$$\mathbb{P}(-D \geq -d) = \mathbb{P}(D \leq d) = \mathbb{P}\big\{T(X) > T^{(\#G-d)}(X)\big\} = \frac{d}{\#G}.$$

This is indeed what is usually considered to be the permutation p value. This equality holds under Condition 1. In other cases, such as when the observations are discrete, the null hypothesis often does not specify a unique null distribution of D. Thus, there is not always a p value in the strict sense based on D.

When H_p does not specify a unique null distribution of any sensible test statistic, as a resolution a 'worst-case' p value could be defined. However, sometimes better solutions are possible, e.g. the randomized p value p' in Sect. 3.4. In general, a p value in the weak sense can be considered, i.e. any random variable p satisfying $\mathbb{P}(p \leq c) \leq c$ for all $c \in [0, 1]$ for every distribution under the null hypothesis. For H_p, $D/\#G$ is always a p value in the weak sense.

3 Random transformations

In Sect. 3, we extend the results of the previous section to tests based on random transformations. Since permutation testing with random permutations is often confused with Monte Carlo testing, in Sect. 3.1 the differences between the two are made explicit. Since random permutations are often used for estimation (rather than exact computation) of p values, estimation of permutation p values is discussed in Sect. 3.2. Exact tests and p values are given in Sects. 3.3 and 3.4, respectively. These two sections contain most of the novel results of the paper.

3.1 Comparison of Monte Carlo and permutation tests

In a basic Monte Carlo experiment, the null hypothesis H_0 is that X follows a specific distribution. A Monte Carlo test is used when there is no analytical expression for the $(1-\alpha)$-quantile of the null distribution of $T(X)$, such that the observed value of $T(X)$ cannot simply be compared to this quantile. To test H_0, independent realizations $X_2, \ldots, X_w$ are drawn from the null distribution of X. Assume that $T(X), T(X_2), \ldots, T(X_w)$ are continuous. Writing $X_1 = X$, let

$$B' = \#\{1 \leq j \leq w : T(X_j) \geq T(X)\}$$

and let b' denote its observed value. It is easily verified that under H_0, B' has the uniform distribution on $\{1, \ldots, w\}$.

The Monte Carlo test rejects H_0 when $T(X) > T^{(k')}$, where $k' = \lceil (1-\alpha)w \rceil$ and $T^{(1)} \leq \cdots \leq T^{(w)}$ are the sorted test statistics $T(X_j)$, $1 \leq j \leq w$. Note that $T^{(k')}$ is not the exact $(1-\alpha)$-quantile of the null distribution of $T(X)$, but nevertheless the test is exact. The reason is that the null distribution of B' is known. The test rejects H_0 if and only if $-B'$ exceeds the $(1-\alpha)$-quantile of its null distribution. Equivalently, it rejects when the Monte Carlo p value

$$\mathbb{P}_{H_0}(B' \leq b') = b'/w,$$

where b' is the observed value of B', is at most α.

The validity of a random permutation test is not as obvious. Let $g_2, \ldots, g_w$ be random permutations from G. (There are various ways of sampling them, which we discuss later.) One permutation, g_1, is fixed to be $id \in G$, reflecting the original observation. Then, similarly to a Monte Carlo test, the permutation test rejects H_p if and only if $T(X) > T^{(k')}(X)$, where now $T^{(1)} \leq \cdots \leq T^{(w)}$ are the sorted test statistics $T(g_j X)$, $1 \leq j \leq w$.

Note that contrary to the Monte Carlo sample $X_1, \ldots, X_w$, the permutations $g_1 X, \ldots, g_w X$ are not independent under the null. Thus, the random permutation test is not analogous to the Monte Carlo test. To prove the validity of the test based on random permutations, we must use that $g_1 X, \ldots, g_w X$ are independent and identically distributed conditionally on the orbit O_X. It is, however, not obvious what properties G should have in order that $g_1 X = X$ can be seen as a random draw from O_X condi-

tionally on O_X. It will be seen that it suffices that G is a group. In that case, the test can be said to be a 'conditional Monte Carlo test'.

3.2 Estimated p values

In practice, it is often computationally infeasible to calculate the permutation p value based on the whole permutation group, $D/\#G$. To work around this problem, there are two approaches in the literature. In both approaches, random permutations are used. The first approach is *calculating* (rather than estimating) a p value based on the random permutations. This is discussed in Sect. 3.4. The second approach is *estimating* the p value $D/\#G$, which we discuss now.

In practice, the p value $p = D/\#G$ is often estimated using random permutations. The random permutations are typically all taken to be uniform on G and can be drawn with or without replacement. The estimate of p is often taken to be $\hat{p} = B/w$, with B as defined above. This is an unbiased estimate of p, i.e. $\mathbb{E}\hat{p} = p$, and usually $\lim_{w\to\infty} \hat{p} = p$.

A more conservative estimate $\tilde{p} = (B + 1)/(w + 1)$ is sometimes also used. This formula is discussed in Sect. 3.4.

Using the unbiased estimate $\hat{p} = B/w$ can be very dangerous, as Phipson and Smyth (2010) thoroughly explain. The reason is that $\hat{p}$ is almost never stochastically larger than the uniform distribution on [0, 1] under H_p. This is immediately clear from the fact that $\hat{p}$ usually has a strictly positive probability of being zero. Consequently, if H_p is rejected if $\hat{p} \leq c$ for some cut-off c, then the type I error rate can be larger than c. Often this difference will be small for large w. However, when c is itself small due to, for example, Bonferroni's multiple testing correction, then $\mathbb{P}(\hat{p} \leq c)$ can become many times larger than c under H_p. This is because this probability does not converge to zero as $c \downarrow 0$ for fixed w. Thus, as Phipson and Smyth (2010) note, using $\hat{p}$ in combination with, for example, Bonferroni can lead to completely faulty inference. Appreciable anti-conservativeness also occurs if very few (e.g. 25–100) random permutations are used [as in, for example, Byrne et al. (2013) and Schimanski et al. (2013)].

When possible, computing exact p values is always to be preferred over estimating p values. Exact p values based on random permutations are given in Sect. 3.4.

3.3 Random permutation tests

Here we discuss exact tests based on random transformations. Apart from Theorem 2 (Hemerik and Goeman 2017), the results in this section are novel.

Phipson and Smyth (2010) also consider exact p values based on random permutations. The proofs in Phipson and Smyth (2010) are incomplete, since they do not show the role of the group structure of the set of all permutations. Lehmann and Romano (2005) (p. 636) remark without proof that if G is a group, then under H_p the p value $(B + 1)/(w + 1)$ is always stochastically larger than uniform on [0, 1], but they state no other properties. In Hemerik and Goeman (2017) for the first time, a theoretical

foundation is given for the random permutation test, using the group structure of the set G. Here this work is extended with additional results.

Theorem 2 states that the permutation test with random permutations has level at most α if the identity map is added. This was remarked several times in the literature and proved in Hemerik and Goeman (2017). We first define the vector of random transformations.

Definition 2 Let G' be the vector $(id, g_2, \ldots, g_w)$, where id is the identity in G and $g_2, \ldots, g_w$ are random elements from G. Write $g_1 = id$. The transformations can be drawn either with or without replacement: the statements in this paper hold for both cases. If we draw $g_2, \ldots, g_w$ *without* replacement, then we take them to be uniformly distributed on $G \backslash \{id\}$, otherwise uniform on G. In the former case, $w \leq \#G$.

Theorem 2 *Let G' be as in Definition 2. Let $T^{(1)}(X, G') \leq \cdots \leq T^{(w)}(X, G')$ be the ordered test statistics $T(g_j X)$, $1 \leq j \leq w$. Let $\alpha \in [0, 1)$ and recall that $k' = \lceil (1-\alpha)w \rceil$.*

Reject H_p when $T(X, G') > T^{(k')}(X, G')$. Then, the rejection probability under H_p is at most α.

A proof of Theorem 2 is in Hemerik and Goeman (2017), and we recall it here.

Proof From the group structure of G, it follows that for all $1 \leq j \leq w$, $G' g_j^{-1}$ and G' have the same distribution, if we disregard the order of the elements. Let j have the uniform distribution on $\{1, \ldots, w\}$ and write $h = g_j$. Under H_{p},

$$\begin{aligned} &\mathbb{P}\big\{T(X) > T^{(k')}(X, G')\big\} \\ &\quad = \mathbb{P}\big\{T(X) > T^{(k')}(X, G'h^{-1})\big\} \\ &\quad = \mathbb{P}\big\{T(hX) > T^{(k')}(hX, G'h^{-1})\big\}. \end{aligned}$$

Since $(G'h^{-1})(hX) = G'(h^{-1}hX)$, the above equals

$$\begin{aligned} &\mathbb{P}\big\{T(hX) > T^{(k')}(h^{-1}hX, G')\big\} \\ &\quad = \mathbb{P}\big\{T(hX) > T^{(k')}(X, G')\big\}. \end{aligned}$$

Since $h = g_j$ with j uniform, this equals

$$\mathbb{E}\Big[w^{-1}\#\big\{1 \leq j \leq w : T^{(j)}(X, G') > T^{(k')}(X, G')\big\}\Big] \leq \alpha,$$

as was to be shown. □

We now prove Theorem 2 with a conditioning-based approach, viewing the test as a "conditional Monte Carlo" test as it has been called in the literature.

Proof We prove the result for the case of drawing with replacement. The proof for drawing without replacement is analogous. Note that (X, G') takes values in $\mathcal{X} \times$

$\{id\} \times G^{w-1}$. Let $A \subset \mathcal{X} \times \{id\} \times G^{w-1}$ be such that the test rejects if and only if $(X, G') \in A$.

Endow the space of orbits with the σ-algebra that it inherits from the σ-algebra on $\mathcal{X}$. Suppose H_p holds. Assume that almost surely O_X contains $\#G$ distinct elements. In case not, the proof is analogous. Analogously to the proof of Theorem 15.2.2 in Lehmann and Romano (2005), we obtain

$$\mathbb{P}\big\{(X, G') \in A \mid O_X\big\} = \frac{\#\big(O_X \times \{id\} \times G^{w-1}\big) \cap A}{\#O_X \times \{id\} \times G^{w-1}}. \tag{4}$$

We now argue that this is at most α. Fix X. Let $\tilde{X}$ have the uniform distribution on O_X. It follows from the group structure of G that the entries of $G'\tilde{X}$ are just independent uniform draws from O_X. Thus, from the Monte Carlo testing principle it follows that $\mathbb{P}\{(\tilde{X}, G') \in A\big\} \leq \alpha$. Since $(\tilde{X}, G')$ was uniformly distributed on $O_X \times \{id\} \times G^{w-1}$, it follows that (4) is at most α. We conclude that

$$\mathbb{P}\big\{(X, G') \in A\big\} = \mathbb{E}\Big[\mathbb{P}\big\{(X, G') \in A \mid O_X\big\}\Big] \leq \alpha,$$

as was to be shown. □

Theorem 2 implies that $(B + 1)/(w + 1)$ is always a p value in the weak sense if all random permutations (including g_1) are uniform draws with replacement from G or without replacement from $G\backslash\{g_1\}$. Under more specific assumptions, Theorem 2 can be extended to certain composite null hypotheses. Proposition 2 states that under Condition 1 and suitable sampling, the test with random permutations is exact. The formula in Sect. 3.4 for the p value under sampling without replacement is equivalent to the last part of this result.

Proposition 2 *Suppose Condition 1 holds. Let $h_1 \in G_1, \ldots, h_m \in G_m$. Then, the result of Theorem 2 still holds if $g_2, \ldots, g_w$ are drawn with replacement from $\{h_1, \ldots, h_m\}$ or without replacement from $\{h_2, \ldots, h_m\}$. Moreover, in the latter case, the test of Theorem 2 is exact for all $\alpha \in \{0/w, 1/w, \ldots, (w-1)/w\}$.*

Proof We consider the case that $g_2, \ldots, g_w$ are drawn without replacement from $\{h_2, \ldots, h_m\}$ and show that the test is exact for $\alpha \in \{0/w, \ldots, (w-1)/w\}$. Write $G' = (g_1, \ldots, g_w)$. Let h have the uniform distribution on $\{g_1, \ldots, g_w\}$. For each $g \in G$ let $i(g) \in \{1, \ldots, m\}$ be such that $g \in G_{i(g)}$. Suppose H_p holds. From the group structure of G, it follows that the sets $\big\{i(g_1), \ldots, i(g_w)\big\}$ and $\big\{i(g_1 h^{-1}), \ldots, i(g_w h^{-1})\big\}$ have the same distribution. Consequently,

$$\begin{aligned}&\mathbb{P}\big\{T(X) > T^{(k')}(X, G')\big\}\\ &\quad = \mathbb{P}\big\{T(X) > T^{(k')}(X, G'h^{-1})\big\}.\end{aligned}$$

As in the above proof of Theorem 2, we find that this equals $\mathbb{P}\big\{T(hX) > T^{(k')}(X, G')\big\}$.

Since $\alpha \in \{0/w, \ldots, (w-1)/w\}$ and $T(g_1 X), \ldots, T(g_w X)$ are distinct, it holds with probability one that

$$\#\left\{1 \le j \le w : T(g_j X) > T^{(k')}\right\} = \alpha w.$$

Since h is uniform, it follows that $\mathbb{P}\left\{T(hX) > T^{(k')}(X, G')\right\} = \alpha$. □

Using this result, it can be shown that specific tests with random permutations are unbiased. The test of Theorem 2 can be slightly conservative if α is not chosen suitably or due to the possibility of ties. Recall that the same holds for the basic permutation test that uses all transformations in G. The adaptation by Hoeffding at (3) then guarantees exactness. The following is a generalization of Hoeffding's result to random transformations.

Proposition 3 *Consider the setting of Theorem 2. Let*

$$a = a(X, G') = \frac{w\alpha - M^+(X, G')}{M^0(X, G')}, \tag{5}$$

where

$$M^+(X, G') := \#\{1 \le j \le w : T(g_j X) > T^{(k')}(X, G')\},$$
$$M^0(X, G') := \#\{1 \le j \le w : T(g_j X) = T^{(k')}(X, G')\}.$$

Reject if $T(X) > T^{(k')}(X, G')$ and reject with probability a if $T(X) = T^{(k')}(X, G')$. Then, the rejection probability is exactly α under H_p.

Proof Assume H_p holds. Note that

$$\mathbb{P}(\text{reject}) = \mathbb{E}\left\{\mathbb{1}_{\{T(X) > T^{(k')}(X,G')\}} + a(X, G')\mathbb{1}_{\{T(X) = T^{(k')}(X,G')\}}\right\}.$$

Write $M^+ = M^+(X, G')$ and $M^0 = M^0(X, G')$. Analogously to the first four steps of the second proof of Theorem 2, it follows for h as defined there that the above equals

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{1}_{\{T(hX) > T^{(k')}(X,G')\}} + a(X, G')\mathbb{1}_{\{T(hX) = T^{(k')}(X,G')\}}\right\} \\
&= \mathbb{E}\left\{\mathbb{1}_{\{T(hX) > T^{(k')}(X,G')\}}\right\} + \mathbb{E}\left\{a(X, G')\mathbb{1}_{\{T(hX) = T^{(k')}(X,G')\}}\right\} \\
&= \mathbb{E}\left\{M^+ w^{-1}\right\} + \mathbb{E}\left[E\left\{\frac{w\alpha - M^+}{M^0}\mathbb{1}_{\{T(hX) = T^{(k')}(X,G')\}} \mid M^+, M^0\right\}\right] \\
&= \mathbb{E}\left\{M^+ w^{-1}\right\} + \mathbb{E}\left[\frac{w\alpha - M^+}{M^0} E\left\{\mathbb{1}_{\{T(hX) = T^{(k')}(X,G')\}} \mid M^+, M^0\right\}\right] \\
&= \mathbb{E}\left\{M^+ w^{-1}\right\} + \mathbb{E}\left[\frac{w\alpha - M^+}{M^0} M^0 w^{-1}\right] = \alpha,
\end{aligned}$$

as was to be shown. □

The test of Proposition 3 entails a randomized decision: in case $T(X) = T^{(k')}$, the test randomly rejects with probability a. This is in itself not objectionable, since the test is randomized anyway due to the random transformations. Note that in the situation of Proposition 2 under drawing without replacement the test is already exact, such that Proposition 3 is not needed to obtain an exact test.

In Theorem 2, the requirement of using the whole group is replaced by suitable random sampling from the group. Interestingly, the following sampling scheme is also possible. Let $G^* \subseteq G$ be any finite subset of G, where we now allow G to be an infinite group as well. Write $k^* = \lceil (1-\alpha)\#G^* \rceil$. Let h be uniformly distributed on G^* and independent. Reject H_p if and only if

$$T(X) > T^{(k^*)}(X, G^* h^{-1}),$$

i.e. if $T(X)$ exceeds the $(1-\alpha)$-quantile of the values $T(gh^{-1})$, $g \in G^*$. This is a randomized rejection rule, since it depends on h, which is randomly drawn each time the test is executed. The rejection probability is at most α, which follows from an argument analogous to the last five steps of the first proof of Theorem 2. Note that if G^* is a group itself, then $G^* h^{-1} = G^*$ and this test becomes nonrandom, coinciding with the basic permutation test. Thus, it is a generalization thereof. This result allows using a permutation test when G is an infinite group of transformations, from which it may not be obvious how to sample uniformly. One simply uses any finite subset G^* of the infinite group.

3.4 *p* values based on random transformations

Phipson and Smyth (2010) give formulas for p values, when permutations are randomly drawn. Here we provide the required assumptions and proofs, which follow from Sect. 3.3. We then provide some additional results.

Write

$$B = \#\big\{1 \leq j \leq w : T(g_j X) \geq T(X)\big\}, \tag{6}$$

where $g_1, \ldots, g_w$ are random permutations with distribution to be specified. Let b be the observed value of B. Under Condition 1, Phipson and Smyth's p values are exactly equal to $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$. Under Condition 1, if $g_1, \ldots, g_w$ are drawn such that they are from distinct elements G_i of the partition and not from G_1, the p value $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$ is exactly

$$\frac{b+1}{w+1}.$$

The validity of this formula follows from Proposition 2. For the case that permutations are drawn with replacement, where $g_1, \ldots, g_w$ are independent and uniform on G, Phipson and Smyth also provide a formula for $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$, under Condition 1.

The formula $(B+1)/(w+1)$ simplifies to the formula B/w if the identity map is added to the random permutations. It follows that the permutation test based on random permutations becomes exact for certain α if the identity is added. Note that this only

holds if Condition 1 is satisfied and all permutations are from distinct equivalence classes G_i.

We now state some additional results that follow from Sect. 3.3. Corresponding to the randomized test of Proposition 3, a randomized p value can be defined as follows. The advantage of this p value is that it is always uniform on [0, 1] under H_p without requirement of additional assumptions, and it is easy to compute. Consider the randomized test of Proposition 3 (hence with G' as in Definition 2). Suppose without loss of generality that when $T(X) = T^{(k')}$, the test rejects if and only if $a > u$, where u is uniform on [0, 1] and independent. Define the randomized p value by

$$p' = \frac{\#\{1 \le j \le w : T(g_j X) > T(X)\}}{w} + u\frac{\#\{1 \le j \le w : T(g_j X) = T(X)\}}{w}.$$

This p value has the property that $p' \le \alpha$ if and only if the randomized test rejects. This implies in particular that p' is exactly uniform on [0, 1] under H_p. The fact that p' is randomized is in itself not objectionable, since it is randomized anyway due to the random transformations.

A simple upper bound to p' is

$$\frac{\#\{1 \le j \le w : T(g_j X) \ge T(X)\}}{w},$$

a p value in the weak sense, which translates to $(B+1)/(w+1)$ when $g_1, \ldots, g_w$ are for example all independent uniform draws from G. It is not exactly uniform on [0, 1] under H_p. However, when w is large and there are few ties among the test statistics, it tends to closely approximate p', so that it may be used for simplicity.

4 Applications

We briefly mention some applications where our results are particularly useful. We have considered data X that lie in an arbitrary space $\mathcal{X}$ and an arbitrary group of transformations G. For example, we allow X to be a vector of functions, which is the type of data investigated by functional data analysis (FDA) (Cuevas 2014; Goia and Vieu 2016). Cox and Lee (2008) consider permutation testing with such functional data. To formulate an exact random permutation test in such a setting, the present paper is useful.

In Hemerik and Goeman (2017), properties are proven of the popular method SAM ("Significance Analysis of Microarrays", Tusher et al. 2001). This is a permutation-based multiple testing method which provides an estimate of the false discovery proportion, the fraction of false positives among the rejected hypotheses. Using Theorem 2, Hemerik and Goeman (2017) showed for the first time how a confidence interval can be constructed around this estimate.

In a basic permutation test, the observed statistic $T(X)$ is compared to $T^{(k)} \in \mathbb{R}$, a quantile of the permutation distribution. The permutation-based multiple testing method by Meinshausen (2006), which provides simultaneous confidence bounds for

the false discovery proportion, also constructs a quantile based on the permutation distribution. There, however, $l \in \mathbb{N}$ hypotheses and hence l statistics $T_1(X), \ldots, T_l(X)$, are considered. (They consider p values as test statistics.) Correspondingly, the quantile which Meinshausen constructs is l-dimensional. It turns out that the crucial step of the proof (the second last line of the proof, p. 231) relies on the principle behind the basic permutation test. The present article can be used to make this method exact. (For example, in Meinshausen (2006), id should be added to the random permutations.)

In Goeman and Solari (2011), it is suggested to combine the method by Meinshausen (2006) with closed testing, which leads to a very computationally intensive method. Hence, preferably only a limited number of permutations (e.g. 100) would be used. The present paper allows using such a limited number of transformations, while still obtaining an exact method.

5 Discussion

This paper proves properties of tests with random permutations in a very general setting. Properties such as unbiasedness of tests of composite null hypotheses and consistency do not hold in general, but may be proved for more specific scenarios. For fixed permutations, there are many results regarding such properties (Hoeffding 1952; Lehmann and Romano 2005; Pesarin and Salmaso 2010, 2013) which may be extended to random permutations.

Aside from the permutation test, there are many multiple testing methods which employ permutations, some of which are mentioned in Sect. 4. Another example is the max$T$ method by Westfall and Young (1993). These methods are precisely based on the principle behind the permutation test. This paper can provide better insight into these procedures, when random permutations are used.

Acknowledgements We thank Aldo Solari and Vincent van der Noort for their valuable suggestions.

References

Byrne E, Carrillo-Roa T, Henders A, Bowdler L, McRae A, Heath A, Martin N, Montgomery G, Krause L, Wray N (2013) Monozygotic twins affected with major depressive disorder have greater variance in methylation than their unaffected co-twin. Transl Psychiatry 3(6):e269

Cox DD, Lee JS (2008) Pointwise testing with functional data using the westfall-young randomization method. Biometrika 95(3):621–634

Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23

Dwass M (1957) Modified randomization tests for nonparametric hypotheses. Ann Math Stat 28:181–187

Ernst MD et al (2004) Permutation methods: a basis for exact inference. Stat Sci 19(4):676–685

Fisher RA (1936) "The coefficient of racial likeness" and the future of craniometry. J Anthropol Inst G B Irel 66:57–63

Ge Y, Dudoit S, Speed TP (2003) Resampling-based multiple testing for microarray data analysis. Test 12(1):1–77
Goeman JJ, Solari A (2010) The sequential rejection principle of familywise error control. Ann Stat 38:3782–3810
Goeman JJ, Solari A (2011) Multiple testing for exploratory research. Stat Sci 26(4):584–597
Goia A, Vieu P (2016) An introduction to recent advances in high/infinite dimensional statistics. J Multivariate Anal 146:1–6
Hemerik J, Goeman JJ (2017) False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. J R Stat Soc Ser B (Stat Methodol). https://doi.org/10.1111/rssb.12238
Hoeffding W (1952) The large-sample power of tests based on permutations of observations. Ann Math Stat 23:169–192
Langsrud Ø (2005) Rotation tests. Stat Comput 15(1):53–60
Lehmann EL, Romano JP (2005) Testing statistical hypotheses. Springer, New York
Meinshausen N (2006) False discovery control for multiple tests of association under general dependence. Scand J Stat 33(2):227–237
Meinshausen N, Bühlmann P (2005) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. Biometrika 92(4):893–907
Pesarin F (2015) Some elementary theory of permutation tests. Commun Stat Theory Methods 44(22):4880–4892
Pesarin F, Salmaso L (2013) On the weak consistency of permutation tests. Commun Stat Simul Comput 42(6):1368–1379
Pesarin F, Salmaso L (2010) Permutation tests for complex data: theory, applications and software. Wiley, New York
Phipson B, Smyth GK (2010) Permutation p values should never be zero: calculating exact p values when permutations are randomly drawn. Stat Appl Genet Mol Biol 9(1):39
Schimanski LA, Lipa P, Barnes CA (2013) Tracking the course of hippocampal representations during learning: When is the map required? J Neurosci 33(7):3094–3106
Solari A, Finos L, Goeman JJ (2014) Rotation-based multiple testing in the multivariate linear model. Biometrics 70(4):954–961
Southworth LK, Kim SK, Owen AB (2009) Properties of balanced permutations. J Comput Biol 16(4):625–638
Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98(9):5116–5121
Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p value adjustment. Wiley, New York