International Statistical Review

International Statistical Review (2020) doi:10.1111/insr.12431

Another Look at the Lady Tasting Tea and Differences Between Permutation Tests and Randomisation Tests

Jesse Hemerik[1] and Jelle J. Goeman[2]

[1] *Biometris, Wageningen University & Research, Droevendaalsesteeg 1, Wageningen, 6708 PB, The Netherlands*
E-mail: jesse.hemerik@wur.nl

[2] *Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, Leiden, 2333 ZC, The Netherlands*

Summary

The statistical literature is known to be inconsistent in the use of the terms ‘permutation test’ and ‘randomisation test’. Several authors successfully argue that these terms should be used to refer to two distinct classes of tests and that there are major conceptual differences between these classes. The present paper explains an important difference in mathematical reasoning between these classes: a permutation test fundamentally requires that the set of permutations has a group structure, in the algebraic sense; the reasoning behind a randomisation test is not based on such a group structure, and it is possible to use an experimental design that does not correspond to a group. In particular, we can use a randomisation scheme where the number of possible treatment patterns is larger than in standard experimental designs. This leads to exact *p* values of improved resolution, providing increased power for very small significance levels, at the cost of decreased power for larger significance levels. We discuss applications in randomised trials and elsewhere. Further, we explain that Fisher's famous Lady Tasting Tea experiment, which is commonly referred to as the first permutation test, is in fact a randomisation test. This distinction is important to avoid confusion and invalid tests.

Key words: group invariance test; lady tasting tea; permutation test; randomisation test.

1 Introduction

The statistical literature is very inconsistent in the use of the terms ‘permutation tests’ and ‘randomisation tests’ (Onghena, 2018; Rosenberger *et al.*, 2019). Both terms are often used to refer to tests that involve permutations. Sometimes, these two terms are considered to refer to strictly distinct classes, sometimes to the same, and sometimes to partly overlapping classes. The confusion surrounding differences between such tests is an important issue, because there are major differences between permutation tests and randomisation tests in the sense of, for example, Onghena (2018), Kempthorne and Doerfler (1969) and Rosenberger *et al.* (2019), whose definitions we will follow here. Those authors use these terms to refer to strictly distinct classes of tests and discuss the terms in detail. Permutation tests are based on random sampling from populations, and randomisation tests are based on experimental randomisation of treatments. With 'randomization of treatments', we refer to the (physical) randomisation that is part of the experimental design. With 'permutation-based tests', we will refer to all tests involving permutations, including randomisation tests involving permutations.

Another point of confusion has been the role of a *group structure*, in the algebraic sense, in permutation-based tests. Southworth *et al.* (2009), among others, explain that for permutation tests to have proven properties, it is important that the set of permutations used has such a group structure, as we discuss in Section 2. For example, the set of *balanced permutations*, which is a subset of a permutation group, does not have a group structure, and using it within a permutation test tends to lead to a very anticonservative test. A balanced permutation, roughly speaking, is a permutation map that moves exactly 50% of the cases to the control group and 50% of the controls to the case group. Balanced permutations (not to be confused with stratified permutations) have been used in several publications (Fan *et al.*, 2004; Jones-Rhoades *et al.*, 2007), but Southworth *et al.* (2009) warn against their use.

A common reference in the permutation literature is the 'Lady Tasting Tea' experiment, described in Fisher (1935a, Ch. II). This experiment is commonly referred to as the first published permutation test (Wald & Wolfowitz, 1944; Hoeffding, 1952; Anderson & Robinson, 2001; Lehmann & Romano, 2005; Langsrud, 2005; Mielke & Berry, 2007; Phipson & Smyth, 2010; Winkler *et al.*, 2014). Indeed, like permutation tests, this test is based on permutations. We will see, however, that it is not a permutation test, but a randomisation test, in the sense of Onghena (2018), Rosenberger *et al.* (2019) and Kempthorne and Doerfler (1969).

A main goal of the present paper is to explain the difference between two classes of tests involving permutations (or other transformations): tests that fundamentally rely on a group structure and tests that do not, in an appropriate sense. To our knowledge, this explicit distinction has not been made before. It is connected to the difference between permutation tests and randomisation tests: the former fundamentally rely on a group structure and the latter do not—with the caveat that a randomisation test should reflect the randomisation scheme of the (physical) experiment. If the randomisation scheme corresponds to a group, then the test also involves that group; otherwise, the test does not involve a group. The fundamental point of this paper is as follows: the mathematical reasoning underlying a randomisation test is not based on a group structure (even if a group happens to be used); the reasoning underlying a permutation test, on the other hand, is always based on a group structure and is completely different from the reasoning underlying randomisation tests. In the existing literature, many randomisation tests involve a group, but randomisation tests that do not involve a group are also often considered (Onghena & Edgington, 1994; 2005; Rosenberger *et al.*, 2019). The further contributions of this paper are as follows.

First of all, because permutation-based randomisation tests do not require a group structure, it can be useful to consider a randomisation scheme that does not correspond to a group. We introduce the idea of using an alternative randomisation scheme to increase the number of possible treatment patterns. This increases the resolution of the p value, thus improving power for very small significance levels α, at the price of power loss for larger α.

In addition, this paper provides the caveat that the lady tasting tea experiment is rather different from permutation tests (in the sense of, for example, Onghena, 2018). Referring to the lady tasting tea experiment as an example of a permutation test, as is often done, can put readers on the wrong foot, because the reasoning underlying this experiment is not based on a group structure. Referring to the lady tasting tea may have contributed to the confusion that has led researchers to design invalid permutation tests without a group structure (Southworth *et al.*, 2009). The purpose of this paper is not to identify the first permutation test, which would not be straightforward (Berry *et al.*, 2014).

This paper is built up as follows. In Section 2, we review existing results on permutation and group invariance tests, emphasising the key role of the group structure of the permutations. In Section 3.1, we discuss the lady tasting tea experiment, emphasising why it does *not* require a group structure to control the type I error rate. In Section 3.2, we generalise the test of Section 3.1, providing a general randomisation test and mentioning applications. In Section 3.3, we apply the general randomisation test in a randomised trial setting, discussing how we can obtain higher resolution p values than with a canonical permutation-based test. The performance of our alternative tests is illustrated with simulations in Section 4. We end with a discussion.

2 Permutation Tests and Group Invariance Tests

In the present section, we explain the mathematical reasoning behind permutation tests, focusing on type I error control. As mentioned, the terms ‘permutation test’ and ‘randomisation test’ have been used inconsistently in the literature. A typical example of a permutation test in the sense of Onghena (2018), Kempthorne and Doerfler (1969) and Rosenberger *et al.* (2019) is discussed in Fisher (1936, pp. 58–59). In this thought experiment, measurements of the statures of 100 Englishmen and 100 Frenchmen are considered. These observations are assumed to be randomly sampled from their respective populations. Such a model, where observations are randomly sampled from their populations, is typical for permutation tests in the sense of, for example, Onghena (2018), Kempthorne and Doerfler (1969) and Rosenberger *et al.* (2019). Note that in this example, there is no randomisation of treatments as in, for example, clinical trials. In the example in Fisher (1936, pp. 58–59), to test whether ‘the two populations are homogeneous’, the difference between the two sample means is computed, and this is repeated for each permutation of the 200 observations. The null hypothesis is rejected if the original difference is larger than most of the differences obtained after permutation. We will return to this example shortly.

Permutation tests are special cases of the general group invariance test. The definition of the group invariance test in, for example, Hoeffding (1952), Lehmann and Romano (2005) and Hemerik and Goeman (2018b) is rather general, so that many randomisation tests also fall under it. The relationships between permutation tests, randomisation tests and group invariance tests are illustrated in Figure 1. The principle underlying the group invariance test can also be used to prove properties of various permutation-based multiple testing methods (Hemerik & Goeman, 2018; Hemerik *et al.*, 2019; Meinshausen & Bühlmann, 2005; Tusher *et al.*, 2001; Westfall & Young, 1993).

A general definition of a group invariance test is as follows. Generalisations of this framework, such as two-sided tests, are possible. Let X model random data with support in a space $\mathcal{X}$. For example, X could be a random vector or matrix. Consider a set $\mathcal{G}$ of permutation maps or other transformations $g : \mathcal{X} \to \mathcal{X}$. We will assume that $\mathcal{G}$ is finite, although generalisations are possible. The set $\mathcal{G}$ is assumed to have a group structure with respect to the operation of composition of maps, which means that $\mathcal{G}$ contains the identity map $x \mapsto x$; every element in $\mathcal{G}$ has an inverse; and for all $g, h \in \mathcal{G}$, $g \circ h \in \mathcal{G}$ (Hoeffding, 1952). Further, we consider some test statistic $T : \mathcal{X} \to \mathbb{R}$. Consider a null hypothesis H_0 that implies that the joint distribution of all test statistics $T(g(X))$ with $g \in \mathcal{G}$ is invariant under all transformations of X in $\mathcal{G}$ (Hemerik & Goeman, 2018b). This holds in particular if

$$g(X) \overset{d}{=} X, \tag{1}$$

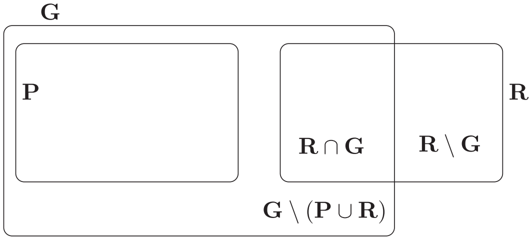

Figure 1. *A Venn diagram showing relationships between permutation tests (**P**), randomisation tests (**R**) and group invariance tests (**G**). Permutation tests (**P**) are a subclass of group invariance tests (**G**) and are based on random sampling from populations. Randomisation tests (**R**) are based on randomisation of treatments. The randomisation scheme sometimes corresponds to a group (**R** ∩ **G**) and sometimes does not (**R** \ **G**). If the scheme corresponds to a group, then this is often a set of permutations, but not always. There are also group invariance tests that fall outside all these categories (**G** \ (**P** ∪ **R**)). An example of a test from class **P** is Fisher's thought experiment with Englishmen and Frenchmen described in Section 2. An example from the class **R** ∩ **G** is Fisher's lady tasting tea experiment. A different example, based on sign flipping instead of permutation, is the test in Fisher (1935a, section 21). An example from the class **R** \ **G** is in Section 3.3. Examples from the class **G** \ (**P** ∪ **R**) are certain tests based on rotations (Solari et al., 2014) or sign flipping (Winkler et al., 2014)*

for every $g \in \mathcal{G}$.

A typical example of such a setting is the thought experiment from Fisher (1936, pp. 58–59), mentioned earlier. Let $X_1, \ldots, X_{100}$ be the statures of the Englishmen, and let $X_{101}, \ldots, X_{200}$ be the statures of the Frenchmen. The test statistic considered in Fisher (1936, pp. 58–59) is

$$T(X) = \left| \frac{1}{100} \sum_{i=1}^{100} X_i - \frac{1}{100} \sum_{i=101}^{200} X_i \right|. \tag{2}$$

The null hypothesis H_0 is that $X_1, \ldots, X_{200}$ are i.i.d. The null hypothesis is rejected if the original test statistic is larger than most of the statistics obtained after permutation. The group $\mathcal{G}$ that Fisher considers consists of all permutation maps $g : \mathbb{R}^{200} \to \mathbb{R}^{200}$. Here, every $g \in \mathcal{G}$ is of the form

$$(x_1, \ldots, x_{200}) \mapsto (x_{\pi_1}, \ldots, x_{\pi_{200}}),$$

where $(\pi_1, \ldots, \pi_{200})$ is a permutation of $(1, \ldots, 200)$. Note that under H_0, (1) holds for every $g \in \mathcal{G}$.

As another example, consider random data X with support in $\mathbb{R}^n$, with independent entries that are symmetric about their means. Suppose that under H_0, the entries have mean 0. This may for example be the case if each entry of X is the difference of two paired observations. Then the distribution of X is invariant under all transformations in $\mathcal{G}$ under H_0 if we define $\mathcal{G}$ to be the group of all sign-flipping maps of the form

$$(x_1, \ldots, x_n) \mapsto (s_1 x_1, \ldots, s_n x_n), \tag{3}$$

with $(s_1, \ldots, s_n) \in \{-1, 1\}^n$. We may take $T((x_1, \ldots, x_n)) = \sum_{i=1}^n x_i$. If the data-generating mechanism involves randomisation of treatments (rather than sampling from populations), such a test can be considered a randomisation test. The test already appears in Fisher (1935a, section 21), albeit without explicit proof. Refer also to Basu (1980).

In these scenarios, we can apply the general group invariance test to test H_0. This test already appears in the literature (Hoeffding, 1952; Hemerik & Goeman, 2018b; Lehmann & Romano, 2005), but for completeness, we include the result and its proof.

We will write $gX = g(X)$ for short. Let $k = \lceil (1-\alpha)|\mathcal{G}| \rceil$, the smallest integer that is larger than or equal to $(1-\alpha)|\mathcal{G}|$. Let $T^{(1)}(X) \leq T^{(2)}(X) \leq \ldots \leq T^{(k)}(X) \leq \ldots \leq T^{(|\mathcal{G}|)}(X)$ be the sorted values $T(gX)$ with $g \in \mathcal{G}$.

Theorem 1. *The size of the group invariance test is at most* α, *that is, under* H_0, $\mathbb{P}\left\{T(X) > T^{(k)}(X)\right\} \leq \alpha$.

Proof. By the group structure, $\mathcal{G}g = \mathcal{G}$ for all $g \in \mathcal{G}$. Hence $T^{(k)}(gX) = T^{(k)}(X)$ for all $g \in \mathcal{G}$. Let G have the uniform distribution on $\mathcal{G}$. Then under H_0, the rejection probability is

$$\begin{aligned}
&\mathbb{P}\left\{T(X) > T^{(k)}(X)\right\} = \\
&\mathbb{P}\left\{T(GX) > T^{(k)}(GX)\right\} = \\
&\mathbb{P}\left\{T(GX) > T^{(k)}(X)\right\}.
\end{aligned}$$

The first equality follows from the null hypothesis, and the second equality holds because $T^{(k)}(X) = T^{(k)}(GX)$. Because G is uniform on $\mathcal{G}$, the above probability equals

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

as was to be shown.

Under additional assumptions, the test is exact, that is, the rejection probability is exactly α under H_0 (Hemerik & Goeman, 2018b). In the above proof, we used the group structure, which guarantees the symmetry property $\mathcal{G}g = \mathcal{G}$ for all $g \in \mathcal{G}$. A different proof, based on conditioning on the pooled sample, is also possible and also requires using this symmetry property (first proof of Theorem 1 in Hemerik & Goeman, 2018b).

Write $\mathcal{G}X = \{gX : g \in \mathcal{G}\}$ and assume for convenience that gX and $g'X$ are distinct with probability 1 if $g, g' \in \mathcal{G}$ are distinct. This is usually the case if X is continuous. The permutation test is based on the fact that under H_0, for every permutation $g \in \mathcal{G}$, the probability $\mathbb{P}\{T(gX) > T^{(k)}(X)\}$ is the same. The reason is that under H_0, for every $g \in \mathcal{G}$, the joint distribution of $(gX, \mathcal{G}X)$ is the same. This is because if $g, g' \in \mathcal{G}$, under H_0, we have

$$(gX, \mathcal{G}X) = (gX, \mathcal{G}gX) \overset{d}{=} (X, \mathcal{G}X) \overset{d}{=} (g'X, \mathcal{G}g'X) = (g'X, \mathcal{G}X).$$

When $\mathcal{G}g = \mathcal{G}$ does not hold for all $g \in \mathcal{G}$, then the above does not generally hold under H_0.

The group structure of $\mathcal{G}$ implies that $\mathcal{G}g = \mathcal{G}$ for all $g \in \mathcal{G}$. Under the mild condition that all $g \in \mathcal{G}$ are surjective, the reverse implication also holds, that is, if $\mathcal{G}g = G$ for all $g \in \mathcal{G}$, then $\mathcal{G}$ is a group. For example, if $\mathcal{G}g = \mathcal{G}$ for all $g \in \mathcal{G}$, there are $h, g \in \mathcal{G}$ with $hg = g$. It follows that $\mathcal{G}$ contains an identity element and the other group properties also easily follow. We conclude that in the argument underlying the permutation test, the group structure is key.

In practice, it is often computationally infeasible to use a test based on the full group $\mathcal{G}$ of transformations. Researchers then usually resort to using a limited number of random transformations, uniformly sampled from the group $\mathcal{G}$. It is then still possible to obtain an exact test (refer to Hemerik & Goeman, 2018b, and Phipson & Smyth, 2010, for a detailed treatment).

3 The Lady Tasting Tea and Randomisation Tests

Unlike a permutation test, a randomisation test is based on data collected in an experiment involving randomisation of treatments. The randomisation scheme of the physical experiment does not necessarily correspond to a group, and if it does not, the statistical test does not involve a group either. Even if a group is used, the reasoning underlying a randomisation test is not based on the group structure and is very different from the reasoning underlying permutation tests.

In this section, we first discuss the lady tasting tea experiment, explaining that the reasoning underlying the test is not based on a group structure, because it is based on randomisation. This experiment is a special case of a general randomisation test that we discuss in Section 3.2. In Section 3.3, we apply this test to provide higher resolution p values in randomised trials.

3.1 The Lady Tasting Tea Experiment

In the lady tasting tea experiment (Fisher, 1935a, Ch. II), the null hypothesis is that a particular lady cannot distinguish between two types of cups of tea with milk: cups in which the tea was added first and cups in which the milk was added first. To test the null hypothesis, which we denote by H_0, the experimenter 'mixes eight cups of tea, four in one way and four in the other', and presents them 'to the subject for judgement in a random order'. The experimental setup is made known to the lady. The lady then tastes from the cups and has to determine which four cups in the sequence of eight had milk added first. Fisher actually performed the experiment (Box, 1978; Berry *et al.*, 2014).

There are $\binom{8}{4} = 70$ possible orders, with respect to the two types of cups. Suppose H_0 is true. If the lady guesses every pattern with probability 1/70, then the probability that she chooses the correct order is 1/70. Even if she has an a priori preference for a certain order, the probability of guessing correct is 1/70. Indeed, it is assumed that the researcher randomises the true pattern, that is, he chooses each pattern with equal probability. Thus, if we reject H_0 when the lady identifies all four 'milk first' cups correctly, then the probability of a type I error is 1/70 (1.4%). The probability that she labels three of the 'milk first" cups correctly is $\binom{4}{3}\binom{4}{1}/70 = 16/70$ (22.9%), and the probability of two correct picks is 36/70 (51.4%). Thus, for example, when we reject H_0 if at least three picks are correct, the level is $16/70 + 1/70 = 17/70$ (24.3%). The test is equivalent to an instance of 'Fisher's exact test' (Yates, 1934; Fisher, 1935b; Berry *et al.*, 2014) with pre-fixed marginal frequencies in the 2×2 table. Fisher's exact test, however, was not originally motivated from a permutation or randomisation testing perspective.

Mathematically, we can describe the experiment as follows. Let $\mathcal{W} \subset \{0,1\}^8$ be the set of vectors containing four 0s and four 1s, so that the cardinality of $\mathcal{W}$ is $R := |\mathcal{W}| = 70$. Let the decision of the lady be denoted by Y, and let W denote the true order, that is, the random decision by the experimenter. Here, Y and W are random variables taking values in $\mathcal{W}$. Note that W represents the 'treatments' given by the experimenter and Y represents the lady's 'responses'. The experimenter's order W is assumed to be uniformly distributed on $\mathcal{W}$. The null hypothesis is

$$H_0 : Y \text{ is independent of } W.$$

Let $\alpha \in (0,1)$ be the desired type I error rate. If $\alpha \in A = \{1/70, 17/70, 53/70, 69/70\}$, then α is called *attainable* in the lady tasting tea experiment, meaning that we obtain a test of exactly level α (Pesarin, 2015). If α is not attainable, then we obtain a test with level strictly less than α.

Let $T : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a test statistic such that high values of $T(w,y)$ indicate that the patterns w and y are similar, that is, that there is evidence against H_0. Let

$$T^{(1)}(Y) \leq \ldots \leq T^{(70)}(Y)$$

be the sorted statistics $T(w, Y)$ with $w \in \mathcal{W}$. Whether the vector of sorted statistics $(T^{(1)}, \ldots, T^{(70)})$ actually depends on Y or not, depends on the definition of T. In Fisher (1935a), the test statistic is

$$T(W,Y) = \sum_{i=1}^{8} \mathbb{1}_{\{W_i=1\}\cap\{Y_i=1\}} = \sum_{i=1}^{8} W_i Y_i, \tag{4}$$

and it can be seen the sorted statistics do not depend on Y, because Y always has four entries that are 1. Let $\lceil(1-\alpha)R\rceil$ be the smallest integer that is at least $(1-\alpha)R$. We have the following result (Fisher, 1935a).

Theorem 2. *The test that rejects H_0 if and only if $T(W, Y) > T^{(\lceil(1-\alpha)R\rceil)}$ has size as most α.*

Proof. Assume that H_0 holds. Conditional on Y, W is uniformly distributed on $\mathcal{W}$, and $T^{(\lceil(1-\alpha)R\rceil)}(Y)$ is known. Hence, conditional on Y, the rejection probability is

$$\mathbb{P}\left(W \in \{w \in \mathcal{W} : T(w,Y) > T^{(\lceil(1-\alpha)R\rceil)}(Y)\}\right) =$$

$$\frac{1}{R}|\{w \in \mathcal{W} : T(w,Y) > T^{(\lceil(1-\alpha)R\rceil)}(Y)\}| \leq \alpha.$$

Thus, marginal over Y, the rejection probability is also at most α.

Observe that when we use the test statistic (4), then taking $\alpha \in A$ indeed results in an exact test. This follows from the fact that

$$\begin{aligned} T^{(1)} < T^{(2)} = T^{(3)} = \ldots = T^{(17)} < T^{(18)} = T^{(19)} \\ = \ldots = T^{(53)} < T^{(54)} = \ldots = T^{(69)} < T^{(70)}, \end{aligned}$$

by the argument at the beginning of this section. If $\alpha \in (0, 1) \setminus A$, the level is strictly smaller than α. If the experimenter does not choose randomly from all 70 possible patterns, but uses some smaller set of patterns for him and the lady to choose from, then there may not be any $\alpha \in (0, 1)$ for which the test is exact, because the sorted test statistics may depend on Y. This is one of the reasons why using the full set of patterns, in combination with a suitable test statistic T, is useful. However, to prove Theorem 2, we did not need to use the group structure of the permutations. The reason is that in the lady tasting tea experiment, under H_0, the randomisation W of the researcher is by design independent of the reference set $\{(w, Y) : w \in \mathcal{W}\}$. Further considerations follow below.

3.2 A general randomisation test

Theorem 2 still applies if the researcher uses a set of permutations that does not correspond to a group. Suppose for example that the researcher omits one of the patterns, thus picking randomly from some set $\mathcal{W}$ of 69 patterns, with or without the lady's knowledge. Denote the set that the lady chooses from by $\mathcal{Y}$. Then W and Y will still be independent, and Theorem 2 still applies if we let $R = |\mathcal{W}| = 69$ and let

$$T^{(1)}(W) \leq \ldots \leq T^{(69)}(W)$$

be the sorted test statistics $T(w, Y)$, $w \in \mathcal{W}$. Indeed, conditional on Y, W will have a uniform distribution on $\mathcal{W}$. In fact, we have the following very general randomisation test, of which the lady tasting tea experiment is a special case. We refer to this result as a randomisation test because in most applications of the theorem, the variable W will represent experimental randomisation of treatments (Kempthorne & Doerfler, 1969; Onghena, 2018). The idea of the theorem is certainly not new; it is at least implicitly present in earlier works (Morgan *et al.*, 2012).

Theorem 3 (General randomisation test). *Let $\mathcal{W}$ and $\mathcal{Y}$ be nonempty sets, where $\mathcal{W}$ is assumed to be finite. Write $R = |\mathcal{W}|$. Let Y be a variable taking values in $\mathcal{Y}$ and assume W is uniformly distributed on $\mathcal{W}$. (Here, Y and W are variables in a general sense. For example, they may be random vectors.) Let $T : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$ be some test statistic. Consider a null hypothesis H_0 that implies that Y is independent of W. Let $T^{(1)}(Y) \leq \ldots \leq T^{(R)}(Y)$ be the sorted values $T(w, Y)$ with $w \in \mathcal{W}$. Consider the test that rejects H_0 if and only if $T(W, Y) > T^{(\lceil (1-\alpha)R \rceil)}(Y)$. Then the result of Theorem 2 still applies, that is, the test has size at most α.*

The proof is analogous to that of Theorem 2: under H_0, conditional on Y, W is uniformly distributed on $\mathcal{W}$ and $T^{(\lceil (1-\alpha)R \rceil)}(Y)$ is known. Hence, conditional on Y, the rejection probability is

$$\mathbb{P}\Big(W \in \{w \in \mathcal{W} : T(w, Y) > T^{(\lceil (1-\alpha)R \rceil)}(Y)\}\Big) \leq \alpha$$

as before.

Under additional assumptions, the test is exact, that is, the rejection probability is exactly α under H_0. We assumed that $\mathcal{W}$ is finite, but generalisations to infinite $\mathcal{W}$ are possible, as well as generalisations to non-uniform W. We can also define a two-sided test.

Note that in Theorem 3, Y might be a constant, conditional on W. In fact, the randomisation testing literature often views the outcomes as non-random, conditional on the treatments. This corresponds to the fact that randomisation tests can be used without an assumption that the responses are randomly sampled from populations (Cox, 2009; Onghena, 2018; Rosenberger *et al.*, 2019). We discuss this further in the context of randomised trials in Section 3.3.

The general randomisation test of Theorem 3 has many applications. Examples are agricultural experiments and randomised clinical trials. Randomised trials will be discussed in Section 3.3. We mention a few other interesting applications here.

First of all, Theorem 3 has implications for the lady tasting tea experiment. In Section 3.1, it is assumed that the lady knows beforehand that there are m cups of each type, where $2m$ is the total number of cups she receives. If for some reason she does not know that, then she might label, for example, $m + 1$ of the $2m$ items with the same label. In other words, she might pick a pattern from a set containing more patterns than the experimenter chooses from. Theorem 3 then says that the type I error probability will nevertheless be at most α under H_0. Indeed, in Theorem 3, $\mathcal{Y}$ is allowed to be any non-empty set, so in particular, it can be larger than $\mathcal{W}$.

A further application of Theorem 3 are general sensory tests, of which the lady tasting tea experiment is an example. It is interesting to note that in the literature on sensory tests, Fisher's experiment has been regarded a ‘forerunner of modern sensory analysis’ (Bi & Kuesten, 2015). For example, Harris and Kalmus (1949) perform a sensory experiment that is analogous to the lady tasting tea experiment, as follows: ‘The glasses are arranged at random. The subject is told that four of them contain the substance and four contain water, and he is asked to taste them all and to separate them into the two groups of four.'

Other applications of Theorem 3 are existing permutation-based randomisation tests that are used to evaluate whether some classification algorithm has any predictive ability (both in-sample and out-of-sample). Such tests can be used to evaluate algorithms for, for example, text categorisation, fraud detection, optical character recognition, and medical diagnosis. Tests of this type are discussed in, for instance, Golland *et al.* (2005), Airola *et al.* (2010), Ojala and Garriga (2010), Schreiber and Krekelberg (2013) and Rosenblatt *et al.* (2019).

3.3 Randomisation Testing Without a Group Structure: Higher Resolution p Values

In randomised trials, often we are interested in comparing two different treatments, for example, a drug and a placebo. In such a setting, there is a treatment assignment randomised by the experimenter. In that case, we can use the randomisation test of Theorem 3, as explained. As discussed, we then do not require a group structure to control the type I error rate. We now discuss such a setting in detail. The tests considered here will also be studied with simulations in Section 4.

Let $n \geq 2$ be an integer, assumed even for convenience, and suppose we have n subjects, $n/2$ of which receive one treatment and $n/2$ of which receive the other treatment. Let $W = (W_1, .., W_n)$ denote the treatments and $Y = (Y_1, \dots, Y_n)$ the responses, taking values in $\mathbb{R}^n$. The treatment pattern W is uniformly sampled from a set $\mathcal{W} \subseteq \{0, 1\}^n$. The most common type of randomised trial is the *forced balance procedure*, where

$$\mathcal{W} = \{w \in \{0, 1\}^n : w \text{ contains } n/2 \; 1's\} \tag{5}$$

(Rosenberger & Lachin, 2015; Lachin, 1988b; Braun & Feng, 2001). For each $1 \leq i \leq n$, the response $Y_i \in \mathbb{R}$ is independent of all the other subjects' treatments and responses. We consider the null hypothesis H_0 that Y is independent of W.

These assumptions are still rather general. It can be useful to consider a more specific randomisation model as in Pitman (1937, section 7) who assumes an additive treatment effect. An important property of randomisation models is that to test whether the treatment has an effect on our particular individuals, we do not need to assume that they are random draws from populations. We could consider the individuals as fixed and Y as constant, conditional on W (Pitman, 1937, section 3). Indeed, 'Any assumption that the units are, say, a random sample from a population of units [. . .] is additional to the specification' of the model (Cox, 2009). This property is discussed in detail in Onghena (2018) and Rosenberger *et al.* (2019).

We can invoke Theorem 3 to obtain a test that controls the type I error rate. We can also obtain an exact test, that is, a test that rejects with probability exactly α under H_0. Consider the test statistic $T : \mathcal{W} \times \mathbb{R}^n \to \mathbb{R}$ defined as

$$T(W, Y) = \sum_{\{i : W_i = 1\}} Y_i - \sum_{\{i : W_i = 0\}} Y_i. \tag{6}$$

Recall that Y may be viewed as random or constant, conditional on W. In either case, assume that Y is such that (with probability 1), for all distinct $w_1, w_2 \in \mathcal{W}$, $T(w_1, Y) \neq T(w_2, Y)$. This is satisfied in particular if $Y_1, \dots, Y_n$ have continuous distributions. Write $N = \binom{n}{n/2}$. The test is exact if $\alpha \in (0, 1)$ is a multiple of $1/|\mathcal{W}|$, where $|\mathcal{W}|$ equals N. An exact p value is

$$p(W, Y) = \frac{|\{w \in \mathcal{W} : T(w, Y) \geq T(W, Y)\}|}{|\mathcal{W}|}, \tag{7}$$

that is, if $\alpha \in (0, 1)$ is a multiple of $1/|\mathcal{W}|$, then $\mathbb{P}(p \leq \alpha) = \alpha$ under H_0. A two-sided exact test can be obtained analogously.

Because Theorem 3 applies, the test essentially does not rely on a group structure. Hence, we may consider taking $\mathcal{W}$ to be a set that does not correspond to a group. In a different context, this is also done in Onghena and Edgington (1994,2005), where a set of permutations is used that is strictly smaller than the full set of permutations. This is done to avoid too repetitive treatment patterns such as ABBBBAAA. In our setting, if $n = 8$, instead of taking $\mathcal{W}$ to be the set of all permutations of $(0, 0, 0, 0, 1, 1, 1, 1)$ we could take $\mathcal{W}$ to be a subset that does not correspond to a group, and still obtain an exact test (for certain α). As Onghena and Edgington (1994) illustrates, this may be useful in some settings. However, in a typical randomised trial, there is no evident reason to only use a subset of the permutations, except to limit the number of permutations for computational reasons.

A more interesting alternative, when running the experiment, is to draw W from a set that is strictly larger than the set in (5), for example, from the set of all possible labellings, $\{0, 1\}^n$ (a *Bernoulli trial*, Imbens & Rubin, 2015). Indeed, if the standard randomisation scheme is used, that is, the forced balance trial, then the smallest possible p value that can be obtained is $1/N$, owing to the discreteness of the p value. If $n = 8$, for example, then $1/N = 1/70$. This means that if the significance level is $\alpha = 0.01$ for instance, we have a power of 0 to reject H_0. Such small α are often used nowadays, for example because of multiple testing. The discreteness of the permutation p value is a well-known downside of permutation-based tests (Berger, 2000). If we take $\mathcal{W} = \{0, 1\}^n$, however, then $|\mathcal{W}| = 2^8$, so that the smallest possible p value is $1/2^8 = 1/256$. If $1/256 \leq \alpha < 1/70$, this means a uniform improvement in power over the standard randomisation test. Note that there only is a power improvement for very small α; for larger α, the Bernoulli trial has less power than the forced balance trial.

Under H_0, if α is a multiple of $1/2^n$, the test with $\mathcal{W} = \{0, 1\}^n$ rejects with probability exactly α. Otherwise, the test rejects with probability less than α under H_0. For $\mathcal{W} = \{0, 1\}^n$, to our knowledge, it is not known what the optimal choice of T is for testing an additive treatment effect. In Section 4, we will take

$$T(W, Y) = \sum_{\{i:W_i=1\}} (Y_i - \overline{Y}) - \sum_{\{i:W_i=0\}} (Y_i - \overline{Y}), \tag{8}$$

where $\overline{Y} = n^{-1}(Y_1 + \ldots + Y_n)$. Using this, test statistic ensures that under H_0, the expected value of $T(W, Y)$ does not depend on the random labelling W. In Appendix A1, we show how the Bernoulli trial can be modified to enforce covariate balancing.

That it is possible to take $\mathcal{W} = \{0, 1\}^n$ has been noted by several authors (Pocock, 1979; Kalish & Begg, 1985; Lachin, 1988a; Wei & Lachin, 1988; Suresh, 2011; Imbens & Rubin, 2015; Rosenberger *et al.*, 2019). They do not recommend this approach, but merely mention it as a possibility, while focusing on more common randomisation schemes. Their main argument against taking $\mathcal{W} = \{0, 1\}^n$ is that it leads to less power than the usual approach of restricted randomisation (Pocock, 1979, p.188). This is true when α is large enough and it is then better to use the forced balance approach. When α is rather small, however, that test has 0 power, while the Bernoulli trial may have substantial power. The idea that using $\mathcal{W} = \{0, 1\}^n$ leads to higher resolution p values has not been mentioned before to our knowledge. Nowadays, the use of large multiple testing corrections is more common than in the past, so higher resolution, exact p values can clearly be of interest.

Suppose we use $\mathcal{W} = \{0, 1\}^n$. Then, if we happen to draw $W = (0, \ldots, 0)$ or $W = (1, \ldots, 1)$, the value of the statistic (8) is 0, and we can have no hope of rejecting H_0 (if $\alpha = 0.05$). Hence, we might exclude $(0, \ldots, 0)$ and $(1, \ldots, 1)$, and perhaps more elements, from $\mathcal{W}$. In any case, if $\alpha < 1/N$, it can be useful to consider a design with $|\mathcal{W}|$ larger than N. Note that in practice, we must choose $\mathcal{W}$ before administering the treatments. Once the treatments have been given, we cannot change our minds about $\mathcal{W}$. The randomisation test that uses all patterns from $\{0, 1\}^n$ except $(0, \dots, 0)$ and $(1, \dots, 1)$ is further studied with simulations in Section 4.

3.4 Randomisation Testing Under a Random Sampling Model

For completeness, we note the following, but it can be skipped at a first read. Existing permutation tests based on random sampling from populations rely on a group structure. However, in some cases, we can use an alternative approach to avoid the requirement of a group structure also in this setting. This is a novel idea, to our knowledge (and arguably falls outside the categories of Figure 1). The approach is analogous to the test in Section 3.3. Suppose that we are comparing two populations, for example a population of cases and a population of controls, or Englishmen and Frenchmen. Let W have the uniform distribution on $\mathcal{W} = \{0, 1\}^n$ or some subset thereof, as before. Then we could draw from the two populations as indicated by W, that is, for every $1 \le i \le n$, the i-th individual is drawn from the first population if $W_i = 0$ and from the second population if $W_i = 1$. For $1 \le i \le n$, let Y_i be the observation for the i-th individual, for example his or her stature. We can then perform a test exactly as in Section 3.3, using the test statistic (8) and the p value (7).

If we take $\mathcal{W}$ as in (5), then the test will be equivalent to a standard permutation test. For many other choices of $\mathcal{W}$, we obtain a novel type of test. If we take for instance $\mathcal{W} = \{0, 1\}^n$, then the number of observations drawn from each population will be random, with only the total number of observations being fixed at n. In many situations, this would be impractical, for example because there is only a limited, fixed number of cases. We will not pursue such tests further here.

4 Empirical Example

Here, we illustrate the idea in Section 3.3 with a simple simulation study. We considered two randomisation tests: a standard randomisation test and an alternative test that provides higher resolution p values, as discussed in Section 3.3. The setting was as in the example in Section 3.3, with $n = 8$. Every Y_i was distributed as the absolute value of a $N(0, 1)$ variable if $W_i = 0$; if $W_i = 1$ it had the same distribution, but with an increase in mean of 2 in the power simulations. Under the null hypothesis, the distribution of Y_i does not depend on W_i. The first test considered was the standard randomisation test. This test uses $N = (n!)/((n/2)!(n/2)!) = 70$ permutations. The second test was based on all relabellings in $\{0, 1\}^n$ excluding $(0, \dots, 0)$ and $(1, \dots, 1)$. Thus, this test used $2^n - 2 = 254$ relabellings. We used the test statistic (8). By Theorem 3, both tests control the type I error rate. Moreover, the first test is exact if $\alpha \in (0, 1)$ is a multiple of 1/70. The second test is exact if α is multiple of 1/254. It is important to note that the two tests are based on different randomisation schemes, that is, on different data-gathering mechanisms. In practice, the type of test should already be decided upon before running the physical experiment.

In Table 1, for different values of the significance level α, the estimated level and power of the two tests are shown. Every estimate in the table is based on 10^4 repeated simulations. The regular randomisation test had no power for $\alpha < 1/70$, because of the fact that only 70 relabellings are available with this approach. Test 2, which is based on 254 relabellings, however, did have substantial power for $1/254 \le \alpha < 1/70$, as explained in Section 3.3. In the table, the

Table 1. *Performance of an alternative to the standard randomisation test. Test 1 is the standard test, based on a forced balance procedure. Test 2 is the alternative test, based on more relabellings.*

		α				
	Test	1/254 ($\approx$0.0039)	0.005	0.01	0.02	0.05
Size	Test 1	0	0	0	0.0122	0.0418
	Test 2	0.0034	0.0034	0.0076	0.0190	0.0464
Power	Test 1	0	0	0	0.9011	0.9725
	Test 2	0.5443	0.5443	0.7027	0.8436	0.9316

estimated size for $\alpha = 1/254$ is 0.0034, which is approximately the true size 1/254 $\approx$ 0.0039. Note that for $\alpha = 0.005$, the size and power are the same as for $\alpha = 1/254$. The reason is the discreteness of the p value: 0.005 lies between 1/254 and 2/254.

5 Discussion

In this paper, we have distinguished between two types of permutation-based tests: tests that fundamentally rely on a group structure and tests based on treatment randomisation, which do not necessarily require a group structure. We have discussed that in settings where treatments are randomly assigned, it can be useful to consider a randomisation scheme that does not correspond to a group. In particular, this allows obtaining higher resolution exact p values than are possible with standard randomisation tests. This paper also provides the caveat that referring to the lady tasting tea experiment as an example of a permutation test can be misleading, because the reasoning underlying this experiment is not based on a group structure.

The two types of tests between which we distinguish roughly correspond to respectively ‘permutation tests’ and ‘randomisation tests’ in the sense of Onghena (2018) and Rosenberger *et al.* (2019). As we mentioned, the use of these terms has been rather inconsistent throughout the literature. For example, Edgington and Onghena (2007, p.1) write that ‘*randomisation tests* are a subclass of statistical tests called *permutation tests*’, while Onghena (2018) proposes to use the terms for strictly distinct classes of tests. In any case, we propose to use the term ‘randomisation test’ only when there is some form of treatment randomisation. This is in line with Kempthorne and Doerfler (1969), Edgington and Onghena (2007), Onghena (2018) and Rosenberger *et al.* (2019).

As mentioned in the Section 1, it would not be straightforward to identify the first permutation test (Berry *et al.*, 2014). In any case, it is clear that, once the concepts of randomisation of treatments and random sampling from populations had been established in the 1920s (Rubin, 1990; Fisher, 1925; Neyman & Pearson, 1928), the way was paved for the theoretical development of permutation-based tests. However, until the 1980s, there was limited interest in permutation-based procedures, due to lack of access to fast computers. Nowadays, the opposite is true (Albajes-Eizagirre *et al.*, 2019; Hemerik *et al.*, 2019; Rao *et al.*, 2019), and this article discusses important differences between two types of tests involving permutations or other reassignments.

References

Airola, A., Pahikkala, T., Boberg, J. & Salakoski, T. (2010). Applying permutation tests for assessing the statistical significance of wrapper based feature selection. In *2010 Ninth international conference on machine learning and applications*, pp. 989–994, IEEE.

Albajes-Eizagirre, A., Solanes, A., Vieta, E. & Radua, J. (2019). Voxel-based meta-analysis via permutation of subject images (PSI): theory and implementation for SDM. *NeuroImage*, **186**, 174–184.

Anderson, M. J. & Robinson, J. (2001). Permutation tests for linear models. *Australian New Zealand J. Stat.*, **43**(1), 75–88.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. *J. American Stat. Assoc.*, **75**.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Stat. Med.*, **19**(10), 1319–1328.

Berry, K.J., Johnston, J. E. & Mielke Jr, P. W. (2014). *A Chronicle of permutation statistical methods*. Springer: Cham.

Bi, J. & Kuesten, C. (2015). Revisiting Fisher's 'Lady Tasting Tea' from a perspective of sensory discrimination testing. *Food Qual. Pref.*, **43**, 47–52.

Box, J. R. A. (1978). Fisher: the life of a scientist.

Braun, T. M. & Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. American Stat. Assoc.*, **96**(456), 1424–1432.

Cox, D. R. (2009). Randomization in the design of experiments. *Int. Stat. Rev.*, **77**(3), 415–429.

Edgington, E. & Onghena, P. (2007). *Randomization tests*. Chapman and Hall/CRC.

Fan, J., Tam, P., Woude, G. V. & Ren, Y. (2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Nat. Acad. Sci.*, **101**(5), 1135–1140.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Fisher, R. A. (1935a). *The Design of Experiments*. Oliver and Boyd.

Fisher, R. A. (1935b). The logic of inductive inference. *J. Royal Stat. Soc.*, **98**(1), 39–82.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. *J. Anthropological Inst. Great Britain Ireland*, **66**, 57–63.

Golland, P., Liang, F., Mukherjee, S. & Panchenko, D. (2005). Permutation tests for classification. In *International conference on computational learning theory*, pp. 501–515, Springer.

Harris, H. & Kalmus, H. (1949). The measurement of taste sensitivity to phenylthiourea (PTC). *Annal. Eugenics*, **15**(1), 24–31.

Hemerik, J. & Goeman, J. J. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *J. Royal Stat. Soc. Ser. B (Stat. Method.)*, **80**(1), 137–155.

Hemerik, J. & Goeman, J. J. (2018b). Exact testing with random permutations. *TEST*, **27**(4), 811–825.

Hemerik, J., Solari, A. & Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, **106**(3), 635–649.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annal. Math. Stat.*, **23**, 169–192.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jones-Rhoades, M. W., Borevitz, J. O. & Preuss, D. (2007). Genome-wide expression profiling of the arabidopsis female gametophyte identifies families of small, secreted proteins. *PLoS genetics*, **3**(10), e171.

Kalish, L. A. & Begg, C. B. (1985). Treatment allocation methods in clinical trials: a review. *Stat. Med.*, **4**(2), 129–144.

Kempthorne, O. & Doerfler, T. E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika*, **56**(2), 231–248.

Lachin, J. M. (1988a). Properties of simple randomization in clinical trials. *Cont. Clinic. Trial.*, **9**(4), 312–326.

Lachin, J. M. (1988b). Statistical properties of randomization in clinical trials. *Cont. Clinic. Trial.*, **9**(4), 289–311.

Langsrud, O. (2005). Rotation tests. *Stat. Comput.*, **15**(1), 53–60.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Science & Business Media.

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.

Mielke, P. W. & Berry, K. J. (2007). *Permutation Methods: A Distance Function Approach*. Springer Science & Business Media.

Morgan, K. L., Rubin, D. B. et al. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, **40**(2), 1263–1282.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika*, **20**(1/2), 175–240.

Ojala, M. & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *J. Machine Learn. Res.*, **11**(Jun), 1833–1863.

Onghena, P. (2018). Randomization tests or permutation tests? A historical and terminological clarification. *Random. Mask. Alloc. Concealment*, 209–227.

Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs. *Behav. Res. Therapy*, **32**(7), 783–786.

Onghena, P. & Edgington, E. S. (2005). Customization of pain treatments: single-case design and analysis. *The Clinic. J. Pain*, **21**(1), 56–68.

Pesarin, F. (2015). Some elementary theory of permutation tests. *Commu. Stat. Theory Method.*, **44**(22), 4880–4892.

Phipson, B. & Smyth, G. K. (2010). Permutation *P*-values should never be zero: calculating exact *P*-values when permutations are randomly drawn. *Stat. Appl. Gen. Mole. Bio*, **9**(1), 39.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Supp. J. Royal Stat. Soc.*, **4**(1), 119–130.

Pocock, S. J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics*, 183–197.

Rao, K., Drikvandi, R. & Saville, B. (2019). Permutation and Bayesian tests for testing random effects in linear mixed-effects models. *Stat. Med.*

Rosenberger, W. F. & Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rosenberger, W. F., Uschner, D. & Wang, Y. (2019). Randomization: the forgotten component of the randomized clinical trial. *Stat. Med.*, **38**(1), 1–12.

Rosenblatt, J. D., Benjamini, Y., Gilron, R., Mukamel, R. & Goeman, J. J. (2019). Better-than-chance classification for signal detection. *Biostatistics*, **kxz035**.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat. Sci.*, **5**(4), 472–480.

Schreiber, K. & Krekelberg, B. (2013). The statistical analysis of multi-voxel patterns in functional imaging. *PLoS One*, **8**(7), e69328.

Solari, A., Finos, L. & Goeman, J. J. (2014). Rotation-based multiple testing in the multivariate linear model. *Biometrics*, **70**(4), 954–961.

Southworth, L. K., Kim, S. K. & Owen, A. B. (2009). Properties of balanced permutations. *J. Comput. Bio.*, **16**(4), 625–638.

Suresh, K. P. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *J. Human Rep. Sci.*, **4**(1), 8.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci.*, **98**(9), 5116–5121.

Wald, A. & Wolfowitz, J. (1944). Statistical tests based on permutations of the observations. *Annal. Math. Stat.*, **15**(4), 358–372.

Wei, L. J. & Lachin, J. M. (1988). Properties of the urn randomization in clinical trials. *Cont. Clinic. Trial.*, **9**(4), 345–364.

Westfall, P. H. & Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Vol. **279**. John Wiley & Sons.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, **92**, 381–397.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supp. J. Royal Stat. Soc.*, **1**(2), 217–235.

[*Received October 2019, Revised October 2020, Accepted October 2020*]

Appendix A: Enforced Covariate Balance in a Bernoulli Trial

Consider the Bernoulli trial discussed in Section 3.3. Here, we illustrate how we can use this type of approach while also enforcing covariate balance. We provide a simple example where there is one (binary) covariate, say sex, and we want the percentage of women to be the same in the two treatment groups. Generalisations can also be formulated.

Suppose that $n > 0$ is divisible by 4 and that there are $n/2$ men and $n/2$ women. The goal is to achieve covariate balance, which means that we want the same fraction of women in both treatment groups. One way to proceed is as follows. Randomly and independently allocate treatments to the $n/2$ women. There are $2^{n/2}$ ways to do this. Let l be the number of times that the women received treatment I. Then, randomly allocate treatment I to l of the men. In this way, covariate balance is achieved: l women and l men have received treatment I, and $n/2-l$ men and $n/2-l$ women have received treatment II. In each treatment group, the percentage of women is the same (although there is a small probability that one treatment group is empty).

We now compute the total number of possible treatment allocation patterns. Given l, there are

$$\binom{n/2}{l}\binom{n/2}{l}$$

possible patterns. Hence, the total number of possible patterns is

$$\sum_{l=0}^{n/2}\binom{n/2}{l}\binom{n/2}{l}. \tag{A1}$$

Correspondingly, the smallest possible p value is the inverse of this number. Note that with the common design that enforces equally large treatment groups, the total number of possible treatment allocation patterns is

$$\binom{n/2}{n/4}\binom{n/2}{n/4}.$$

This is smaller than (A1), which means that the smallest possible p value is larger than for the Bernoulli trial.