Another look at the Lady Tasting Tea and permutation-based randomization tests

Jesse Hemerik∗† and Jelle Goeman‡

December 6, 2019

Summary

Fisher’s famous Lady Tasting Tea experiment is often referred to as the first permutation test or as an example of such a test. Permutation tests are special cases of the general group invariance test. Recently it has been emphasized that the set of permutations used within a permutation test should have a group structure, in the algebraic sense. If not, the test can be very anti-conservative. In this paper, however, we note that in the Lady Tasting Tea experiment, the type I error rate is controlled even if the set of permutations used does not correspond to a group. We explain the difference between permutation-based tests that fundamentally rely on a group structure, and permutation-based tests that do not. The latter are tests based on randomization of treatments. When using such tests, it can be useful to consider a randomization scheme that does correspond to a group. In particular, we can use randomization schemes where the number of possible treatment patterns is larger than in standard permutation-based randomization tests. This leads to exact p-values of improved resolution, providing increased power for very small significance levels. We discuss applications in clinical trials and elsewhere.

keywords: Permutation test; Lady Tasting Tea; Group invariance test; Randomization test

1 Introduction

The well-known “Lady Tasting Tea” experiment, described in Fisher (1935a, Ch. II), is commonly referred to as the first published permutation test (Wald and Wolfowitz, 1944; Hoeffding, 1952; Anderson and Robinson, 2001; Lehmann and Romano, 2005; Langsrud, 2005; Mielke and Berry, 2007; Phipson and Smyth, 2010; Winkler et al., 2014) or as a representative example of a permutation test (Freedman and Lane, 1983). Indeed, this test is based on permutations and, like other permutation-based tests, falls under the definition of group invariance tests. This is a general class of tests based on transformations of data, such as permutations or rotations (Langsrud, 2005). Details are in Section 2.

∗Oslo Centre for Biostatistics and Epidemiology, University of Oslo
†Address for correspondence: Jesse Hemerik, Oslo Centre for Biostatistics and Epidemiology, P.O. Box 1122 Blindern, 0317 Oslo, Norway. e-mail: jesse.hemerik@medisin.uio.no
‡Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands
In 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 will 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 judgment 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 had milk added first. According to most sources, Fisher actually performed the experiment (Box, 1978; Berry et al., 2014). The test is detailed in Section 3.1.

It has been emphasized by Southworth et al. (2009), among others, that for permutation tests to have proven properties, it is important that the set of permutations used has a group structure, in the algebraic sense, 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 anti-conservative test. Balanced permutations (not to be confused with stratified permutations) have been used in several publications (Fan et al., 2004; Jones-Rhoades et al., 2007) and Southworth et al. (2009) warn against their use.

Surprisingly, as we will show, the Lady Tasting Tea experiment still controls the type I error rate if the set of permutations used does not have a group structure. How is it possible that this test is a permutation test, but does not require a group structure?

As we will show, the reason is that the Lady Tasting Tea experiment can be viewed as a test based on randomization of treatments, i.e., as a randomization test in the sense of for example Kempthorne and Doerfler (1969), Edgington and Onghena (2007) and Rosenberger et al. (2019). Indeed, the experiment involves an experimenter, who randomizes the true pattern of cups. Unlike archetypical permutation tests, such randomization tests can generally still control the type I error rate, even if the set of permutations used is not a group. For instance, in the Lady Tasting Tea experiment, both the permutations that the experimenter randomly chooses from and the permutations that the lady is told to pick from, need not be groups. Randomization tests which are not based on groups, are uncommon but not new in the literature on randomization tests (Onghena and Edgington, 1994; Rosenberger et al., 2019).

In this paper, we explain the difference between permutation-based tests that require a group structure and permutation-based tests that do not, the latter being tests based on randomization of treatments. This explicit distinction has not been made before, to our knowledge. The further contributions of this paper are related to this distinction and are as follows.

First of all, since permutation-based randomization tests do not require a group structure, it can be useful to consider a randomization scheme that does not correspond to a group. We introduce the idea of using an alternative randomization 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 $\alpha$.

In addition, this paper provides the caveat that the Lady Tasting Tea experiment is rather different from archetypical permutation tests (in the sense of Onghena, 2018). Using the Lady Tasting Tea experiment as an example of a permutation test, as is often done, can put readers on the wrong foot, since 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.
structure (Southworth et al., 2009). Instead of referring to Fisher (1935a, Ch. II) as an example of a permutation test, it may better to refer to the example in Fisher (1936, pp. 58-59), in which statures of Frenchmen and Englishmen are compared. This example is discussed in Section 2. This is a typical permutation test, which is not based on randomization of treatments, but on permuting random samples from populations. The argument underlying this test (which is implicit in that article) is based on the group structure of the set of permutations, unlike the argument underlying the Lady Tasting Tea experiment.

This paper is built up as follows. In Section 2 we review existing results on permutation and group invariance tests, emphasizing the key role of the group structure of the permutations. In Section 3.1 we discuss the Lady Tasting Tea experiment, emphasizing why this test does not require a group structure to control the type I error rate. In Section 3.2 we generalize the test of Section 3.1, obtaining a general randomization test and mentioning applications. In Section 3.3 we apply the general randomization test in a clinical trial setting, discussing how we can obtain higher-resolution \( p \)-values than with a canonical permutation-based test. The performance of our alternative test is illustrated with simulations in Section 4. We end with a discussion.

2 Permutation tests and group invariance tests

The terms “permutation test” and “randomization test” have been used somewhat inconsistently in the literature. Sometimes the class of permutation tests is understood to include randomization tests (Edgington and Onghena, 2007, p.1). Rosenberger et al. (2019) write that “Many statisticians use the terms permutation tests and randomization tests interchangeably. The first author has regrettably made this mistake himself.” Onghena (2018) and Kempthorne and Doerfler (1969) compare the two terms in detail.

A typical example of a permutation test in the sense of Onghena (2018) 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 Kempthorne and Doerfler (1969), Onghena (2018) and Rosenberger et al. (2019). Note that in this example, there is no randomization 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 below.

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 randomization tests also fall under it. The principle underlying the group invariance test can also be used to prove properties of various permutation-based multiple testing methods (Westfall and Young, 1993; Tusher et al., 2001; Meinshausen and Bühlmann, 2005; Hemerik and Goeman, 2018a; Hemerik et al., 2019).

A general definition of a group invariance test is as follows. Generalizations of this framework, such as two-sided tests, are possible. Let \( X \) be data taking values in a sample space
\[X\]. Consider a set \(G\) of permutation maps or other transformations \(g : X \to X\). We will assume that \(G\) is finite, although generalizations are possible. The set \(G\) is assumed to have a group structure with respect to the operation of composition of maps, which means that:
\(G\) contains the identity map \(x \mapsto x\); every element in \(G\) has an inverse; and for all \(g, h \in G\), \(g \circ h \in G\) (Hoeffding, 1952). Further, we consider some test statistic \(T : X \to X\). Consider a null hypothesis \(H_0\) which implies that the joint distribution of all test statistics \(T(g(X))\) with \(g \in G\) is invariant under all transformations in \(G\) of \(X\). This holds in particular if the data \(X\) are themselves transformation-invariant, i.e., if
\[g(X) \overset{d}{=} X\]
for every \(g \in G\).

A typical example of such a setting is the thought experiment from Fisher (1936, pp. 58-59), mentioned above. Let \(X_1, ..., X_{100}\) be the statures of the Englishmen and let \(X_{101}, ..., X_{200}\) be the statures of the Frenchmen. The test statistic considered in Fisher (1936, pp. 58-59) is
\[T(X) = \frac{1}{100} \sum_{i=1}^{100} X_i - \frac{1}{100} \sum_{i=101}^{200} X_i.\] 

The null hypothesis \(H_0\) is that \(X_1, ..., X_{200}\) are i.i.d.. The group \(G\) consists of all permutation maps \(g : \mathbb{R}^{200} \to \mathbb{R}^{200}\). Here, every \(g \in G\) is of the form
\[(x_1, ..., x_{200}) \mapsto (x_{\pi_1}, ..., x_{\pi_{200}}),\]
where \((\pi_1, ..., \pi_{200})\) is a permutation of \((1, ..., 200)\). Note that \(X\) is then \(G\)-invariant under \(H_0\), i.e., (1) holds for every \(g \in G\).

As another example of group invariance, suppose \(X \in \mathbb{R}^n\) has independent entries and under \(H_0\), the entries are symmetric around 0. Then the distribution of \(X\) is invariant under all transformations in \(G\) under \(H_0\) if we define \(G\) to be the group of all sign-flipping maps of the form
\[(x_1, ..., x_n) \mapsto (s_1x_1, ..., s_nx_n),\]
with \((s_1, ..., s_n) \in \{-1,1\}^n\). This test already appears in Fisher (1935a, §21), albeit without explicit proof.

In both examples above, we can apply the general group invariance test to test \(H_0\). This test already appears in the literature (Hoeffding, 1952; Lehmann and Romano, 2005; Hemerik and Goeman, 2018b), but for completeness we include the result and its proof. Exact testing with randomly sampled permutations will not be discussed here, but is also possible (Hemerik and Goeman, 2018b). We will write \(gX = g(X)\) for short. Let \(T^{(1)}(X) \leq ... \leq T^{(|G|)}(X)\) be the sorted values \(T(gX)\) with \(g \in G\). Let \(k = \lceil (1 - \alpha)|G| \rceil\), the smallest integer which is larger than or equal to \((1 - \alpha)|G|\).

**Theorem 1.** Under \(H_0\), \(\mathbb{P}\{T(X) > T^{(k)}(X)\} \leq \alpha\).

**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 \(H_0\), the rejection probability is
\[
\mathbb{P}\{T(X) > T^{(k)}(X)\} =
\mathbb{P}\{T(hX) > T^{(k)}(hX)\} =
\mathbb{P}\{T(hX) > T^{(k)}(X)\}.
\]
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
\[\mathbb{E}\left[|G|^{-1} \cdot \left| \{ g \in G : T(gX) > T^{(k)}(X) \} \right| \right] \leq \alpha,\]
as was to be shown.

Under additional assumptions, the test is exact, i.e., the rejection probability is exactly $\alpha$ under $H_0$. In the above proof we used the group structure, which guarantees the symmetry property $Gg = G$ for all $g \in G$. A different proof, based on conditioning on the pooled sample, is also possible and also requires using this symmetry (first proof of Theorem 1 in Hemerik and Goeman, 2018b). Write $GX = \{ gX : g \in G \}$ and assume for convenience that $gX$ and $g'X$ are distinct with probability 1 if $g, g' \in G$ are distinct. The permutation test is based on the fact that under $H_0$, for every permutation $g \in G$ the probability $\mathbb{P}\{T(gX) > T^{(k)}(X)\}$ is the same. The reason is that under $H_0$, for every $g \in G$, the joint distribution of $(gX, GX)$ is the same. This is because if $g, g' \in G$, under $H_0$ we have
\[
(gX, GX) = (gX, GgX) \overset{d}{=} (X, GX) \overset{d}{=} (g'X, Gg'X) = (g'X, GX).
\]

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

The group structure of $G$ implies that $Gg = G$ for all $g \in G$. The reverse implication also holds, under the mild condition that all $g \in G$ are surjective. For example, if $Gg = G$ for all $g \in G$, there are $h, g \in G$ with $hg = g$. It follows that $G$ 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.

3 The Lady Tasting Tea and randomization tests

Here, we first discuss the Lady Tasting Tea experiment, explaining that this test essentially does not rely on a group structure. This experiment is a special case of a general randomization test, which we discuss in Section 3.2. In Section 3.3 we apply this test to provide higher-resolution $p$-values in clinical trials.

3.1 The Lady Tasting Tea experiment

As discussed in the Introduction, in the Lady Tasting Tea experiment, the lady receives eight cups. There are two types of cups and she receives four of each kind. 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 randomizes the true pattern, i.e., he chooses each pattern with equal probability. Thus, if we reject $H_0$ when the lady guesses the order correctly, then the probability of a type I error is 1/70. The probability that she labels three of the “tea first” cups correctly is $\binom{3}{3} \frac{70}{4} = 16/70$ and the probability of two correct picks is 36/70. Thus, for example, when we reject $H_0$ if at least three picks are correct, the level is $16/70 + 1/70 = 17/70$. 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 $2 \times 2$ table.

Mathematically, we can describe the experiment as follows. Let $\mathcal{D} \subset \{0,1\}^8$ be the set of vectors containing four 0’s and four 1’s, so that the cardinality of $\mathcal{D}$ is $m := |\mathcal{D}| = 70$. Let the decision of the lady be encoded as $D' \in \mathcal{D}$ and let $D \in \mathcal{D}$ be the true order, i.e., the random decision by the experimenter. The experimenter’s order $D$ is assumed to be uniformly distributed on $\mathcal{D}$. The null hypothesis is

$$H_0 : D' \text{ is independent of } D.$$ 

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 $\alpha$ is called attainable in the Lady Tasting Tea experiment, meaning that we obtain a test of exactly level $\alpha$ (Pesarin, 2015). If $\alpha$ is not attainable, then we obtain a test with level strictly less than $\alpha$.

Let $T : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ be a test statistic such that high values of $T(D, D')$ indicate that the patterns $D$ and $D'$ are similar, i.e., that there is evidence against $H_0$. Let

$$T^{(1)}(D') \leq \ldots \leq T^{(70)}(D')$$

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

$$T(D, D') = \sum_{i=1}^{8} \{D_i = 1\} \cap \{D'_i = 1\}$$

and it can be seen the sorted statistics do not depend on $D'$. Let $\lceil (1-\alpha)m \rceil$ be the smallest integer which is at least $(1-\alpha)m$. We have the following result (Fisher, 1935a).

**Theorem 2.** The test that rejects $H_0$ if and only if $T(D, D') > T^{(\lceil (1-\alpha)m \rceil)}$ has size as most $\alpha$.

**Proof.** Assume $H_0$ holds. Conditional on $D'$, $D$ is uniformly distributed on $\mathcal{D}$ and $T^{(\lceil (1-\alpha)m \rceil)}(D')$ is known. Hence, conditional on $D'$, the rejection probability is

$$\mathbb{P}(D \in \{B \in \mathcal{D} : T(B, D') > T^{(\lceil (1-\alpha)m \rceil)}(D')\}) = \frac{1}{m} |\{B \in \mathcal{D} : T(B, D') > T^{(\lceil (1-\alpha)m \rceil)}(D')\}| \leq \alpha.$$ 

Thus, marginal over $D'$, the rejection probability is also at most $\alpha$. \qed

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

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

by the argument at the beginning of this section. If $\alpha \in (0,1) \setminus A$, the level is strictly smaller than $\alpha$. 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, since the sorted test statistics may depend on \( D' \). 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 randomization \( D \) of the researcher is by design independent of the reference set \( \{(B, D') : B \in \mathcal{D}\} \). Further considerations follow below.

### 3.2 A general randomization 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 picks randomly from some set \( \mathcal{D} \) of 69 patterns, with or without the lady’s knowledge. Denote the set that the lady chooses from by \( \mathcal{D}' \). Then \( D \) and \( D' \) will still be independent and Theorem 2 still applies if we let \( m = |\mathcal{D}| = 69 \) and let

\[
T^{(1)}(D) \leq \ldots \leq T^{(69)}(D)
\]

be the sorted test statistics \( T(B, D') \), \( B \in \mathcal{D} \). Indeed, conditional on \( D' \), \( D \) will have a uniform distribution on \( \mathcal{D} \). In fact, we have the following very general randomization test, of which the Lady Tasting Tea experiment is a special case. We refer to this result as a randomization test since in most applications of the theorem, the variable \( D \) will encode experimental randomization of treatments (Kempthorne and Doerfler, 1969; Onghena, 2018).

The idea of the theorem is certainly not new.

**Theorem 3.** Let \( \mathcal{D} \) and \( \mathcal{D}' \) be nonempty sets, where \( \mathcal{D} \) is assumed to be finite. Write \( m = |\mathcal{D}| \). Let \( \mathcal{D}' \) be a variable taking values in \( \mathcal{D}' \) and assume \( D \) is uniformly distributed on \( \mathcal{D} \). Let \( T : \mathcal{D} \times \mathcal{D}' \to \mathbb{R} \) be some test statistic. Consider a null hypothesis \( H_0 \) which implies that \( D' \) is independent of \( D \). Let \( T^{(1)}(D') \leq \ldots \leq T^{(m)}(D') \) be the sorted values \( T(B, D') \) with \( B \in \mathcal{D} \). Then the result of Theorem 2 still applies.

**Proof.** The proof is analogous to that of Theorem 2. Assume \( H_0 \) holds. Conditional on \( D' \), \( D \) is uniformly distributed on \( \mathcal{D} \) and \( T^{(\lceil(1-\alpha)m\rceil)}(D') \) is known. Hence, conditional on \( D' \), the rejection probability is

\[
P(D \in \{B \in \mathcal{D} : T(B, D') > T^{(\lceil(1-\alpha)m\rceil)}(D')\}) = \frac{1}{m} \frac{|\{B \in \mathcal{D} : T(B, D') > T^{(\lceil(1-\alpha)m\rceil)}(D')\}|}{m} \leq \alpha.
\]

Thus, marginally over \( D' \), the rejection probability is also at most \( \alpha \).

We assumed that \( \mathcal{D} \) is finite, but generalizations to infinite \( \mathcal{D} \) are possible, as well as generalizations to non-uniform \( D \). We can also define a two-sided test. Moreover, under straightforward additional assumptions, we can prove that the test of Theorem 3 is exact for certain \( \alpha \), i.e., that the rejection probability is exactly \( \alpha \) under \( H_0 \).

Note that in Theorem 3, \( D' \) might be a constant, conditional on \( D \). In principle, randomization tests can be used without an assumption that the responses are randomly sampled from populations (Cox, 2009; Onghena, 2018; Rosenberger et al., 2019). This is a known property of randomization tests, which we discuss further in the context of clinical trials in Section 3.3.
The general randomization test of Theorem 3 has many applications. Examples are agricultural experiments and randomized clinical trials. The latter example 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 e.g. \( m + 1 \) of the \( 2m \) items with the same label. Theorem 3 then says that the type I error probability will nevertheless be at most \( \alpha \) under \( H_0 \). Indeed, in Theorem 3, \( D' \) is allowed to be any set, so in particular it can be larger than \( D \).

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 and Kuesten, 2015). For example, Harris and Kalnus (1949) perform a sensory experiment as follows: “The subject is now presented with eight tumblers, four of which contain a few c.c. of water and four containing a few c.c. of the solution [...] . 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.”

Another application of Theorem 3 are existing permutation-based randomization tests which are used to evaluate whether some classification algorithm has any predictive ability. Such tests can be used to evaluate algorithms for, for example, text categorization, 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. (2016).

### 3.3 Randomization testing without a group structure: higher-resolution \( p \)-values

In randomized clinical trials, often we are interested in comparing two different treatments, for example a drug and a placebo. In such a setting, there is obviously a treatment assignment, which we randomize. In that case, we can use the randomization test of Theorem 3. 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 \( Z = (Z_1,...,Z_n) \) encode the treatments and \( Y = (Y_1,...,Y_n) \) the responses. The treatment pattern \( Z \) is uniformly sampled from a set \( Z \subseteq \{0,1\}^n \). In a standard randomized trial, \( Z = \{z \in \{0,1\}^n : z \text{ contains } n/2 \text{ 1's}\} \) (Lachin, 1988b; Braun and Feng, 2001). For each \( 1 \leq i \leq n \), the response \( Y_i \in \mathbb{R} \) is independent of all other variables except (possibly) \( Z_i \). We consider the null hypothesis \( H_0 \) that \( Y \) is independent of \( Z \).

These assumptions are still rather general. It can be useful to consider a more specific randomization model as in Pitman (1937, §7), who assumes an additive treatment effect. An important property of randomization models, is that to test whether the treatment has an effect on our particular patients, we do not need to assume that they are random draws from
populations. We could consider the patients as fixed and $Y$ as constant, conditional on $Z$ (Pitman, 1937, §7). 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 which controls the type I error rate. Indeed, one can take $D = Z$, $D' = Y$ and $D = Z$ and note that $D$ is uniformly distributed on $D$. We can also obtain an exact test, i.e., a test that rejects with probability exactly $\alpha$ under $H_0$. Consider the test statistic $T : \mathcal{Z} \times \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies

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

Recall that $Y$ may be viewed as random or constant, conditional on $Z$. In either case, assume that $Y$ is such that (with probability 1), for all distinct $z_1, z_2 \in \mathcal{Z}$, $T(z_1, Y) \neq T(z_2, Y)$. This is satisfied in particular if $Y_1, ..., Y_n$ have continuous distributions. The test is exact if $\alpha \in (0, 1)$ is a multiple of $1/|\mathcal{Z}|$, where $|\mathcal{Z}|$ equals

$$N := \frac{n}{n/2} = \frac{n!}{(n/2)!(n/2)!}.$$

An exact $p$-value is

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

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

Since Theorem 3 applies, the test essentially does not rely on a group structure. Hence, we may consider sampling $Z$ from a set which does not correspond to a group. In a different context, this is also done in Onghena and Edgington (1994), 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 ABBBBAA. In our setting, if $n = 8$, instead of taking $\mathcal{Z}$ to be the set of all permutations of $(0,0,0,0,1,1,1,1)$ we could sample $Z$ from a subset of these permutations which does correspond to a group, and still obtain an exact test (for certain $\alpha$). As Onghena and Edgington (1994) illustrates, this may be useful in some settings. However, in a typical clinical 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 is to draw $Z$ from a set that is strictly larger than the set in (5), for example, from the set of all possible labelings, $\{0,1\}^n$. Indeed, if the standard randomization test is used, the smallest possible $p$-value that can be obtained is $1/N$, due 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 $\alpha$ are often used nowadays, for example due to multiple testing. The discreteness of the permutation $p$-value is a well-known downside of permutation-based tests (Berger, 2000).

If we take $\mathcal{Z} = \{0,1\}^n$, however, then $|\mathcal{Z}| = 2^8$, 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 randomization test. Under $H_0$, if $\alpha$ is a multiple of $1/2^8$, the test with $\mathcal{Z} = \{0,1\}^n$ rejects with probability exactly $\alpha$. Otherwise the test rejects with probability less than $\alpha$.
under $H_0$. For $\mathcal{Z} = \{0,1\}^n$, 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

$$T(Z, Y) = \sum_{\{i : Z_i = 1\}} (Y_i - \bar{Y}) - \sum_{\{i : Z_i = 0\}} (Y_i - \bar{Y}),$$

where $\bar{Y} = n^{-1}(Y_1 + \ldots + Y_n)$. Using this test statistic ensures that under $H_0$, the expected value of $T(Z, Y)$ does not depend on the random labelling $Z$.

That it is possible to take $\mathcal{Z} = \{0,1\}^n$ has been noted by several authors (Pocock, 1979; Kalish and Begg, 1985; Lachin, 1988a; Wei and Lachin, 1988; Suresh, 2011; Rosenberger et al., 2019). They do not recommend this approach, but merely mention it as a possibility, while focusing on more common randomization schemes. Their main argument against taking $\mathcal{Z} = \{0,1\}^n$ seems to be that it is “inefficient” (Pocock, 1979, p.188). While this is true when $\alpha$ is large enough, the opposite is true when $\alpha$ is rather small. The idea that using $\mathcal{Z} = \{0,1\}^n$ leads to higher-resolution $p$-values, is not mentioned by these authors. 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 $\mathcal{Z} = \{0,1\}^n$. Then, if we happen to draw $Z = (0,\ldots,0)$ or $Z = (1,\ldots,1)$, 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 $\mathcal{Z}$. We leave the question of how to choose $\mathcal{Z}$ for future research. In any case, if $\alpha < 1/N$, it can be useful to consider a test with $|\mathcal{Z}|$ larger than $N$. Note that in practice, we should choose $\mathcal{Z}$ before administering the treatments. Once the treatments have been given, we cannot change our minds about $\mathcal{Z}$. The test based on $\mathcal{Z} = \{0,1\}^n$ is further studied with simulations in Section 4.

3.4 Randomization 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. 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 $Z$ be uniformly distributed on $\mathcal{Z} = \{0,1\}^n$ or some subset thereof, as before. Then we could draw from the two populations as indicated by $Z$, i.e., for every $1 \leq i \leq n$, the $i$-th individual is drawn from the first population if $Z_i = 0$ and from the second population if $Z_i = 1$. For $1 \leq i \leq n$, 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 $\mathcal{Z}$ as in (5), then the test will be equivalent to a standard permutation test. For many other choices of $\mathcal{Z}$, we obtain a novel type of test. If we take for instance $\mathcal{Z} = \{0,1\}^n$, 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 the two tests in Section 3.3: a standard randomization test and the alternative test that provides higher-resolution \( p \)-values. The data were 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 \( Z_i = 0 \); if \( Z_i = 1 \) it had the same distribution, but with an increase in mean of \( \eta \geq 0 \). Under the null hypothesis, \( \eta = 0 \). The first test considered was the standard randomization test. This test uses \( N = (n!) / ((n/2)!(n/2)!)) = 70 \) permutations. The second test was the one based on all \( 2^n = 256 \) relabellings in \( \{0,1\}^n \). 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 \( \alpha \) is multiple of \( 1/256 \).

In Table 1, for different values of the significance level \( \alpha \), 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 randomization test had no power for \( \alpha < 1/70 \), due to the fact that only 70 relabellings are available with this approach. The test based on all 256 relabellings in \( \{0,1\}^n \), however, did have substantial power, as explained in Section 3.3. In the table, the estimated size for \( \alpha = 1/256 \) is 0.0041, which is approximately the true size \( 1/256 \). Note that for \( \alpha = 0.005 \), the size and power are the same as for \( \alpha = 1/256 \). The reason is the discreteness of the \( p \)-value: 0.005 lies between \( 1/256 \) and \( 2/256 \).

| \( \alpha \) | test 1/256 | .005 | .01 | .02 | .05 |
|-----------|-------------|------|-----|-----|-----|
| size      |             |      |     |     |     |
| test 1    | 0           | 0    | 0   | .0122 | .0418 |
| test 2    | .0041       | .0041 | .0070 | .0188 | .0456 |
| power     |             |      |     |     |     |
| test 1    | 0           | 0    | 0   | .9011 | .9725 |
| test 2    | .5415       | .5415 | .7010 | .8463 | .9257 |

5 Discussion

In this paper, we have distinguished between two types of permutation-based tests: tests which fundamentally rely on a group structure and tests based on treatment randomization, which do not require a group structure. We have discussed that in settings where treatments are randomly assigned, it can be useful to consider a randomization scheme which does not correspond to a group. In particular, this allows obtaining higher-resolution exact \( p \)-values than are possible with standard randomization 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, since 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 “randomization tests” in the sense of Onghena (2018) and
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 “randomization 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 “randomization tests” only when there is some form of treatment randomization. This is in line with Kempthorne and Doerfler (1969), Edgington and Onghena (2007), Onghena (2018) and Rosenberger et al. (2019).

The purpose of this paper has not been to identify the first permutation test, which would not be straightforward (Berry et al., 2014). In any case, it is clear that, once the concepts of randomization of treatments and random sampling from populations had been established in the 1920’s (Rubin, 1990; Fisher, 1925; Neyman and Pearson, 1928), the way was paved for the theoretical development of permutation-based tests. However, until the 1980’s, 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).

References

Airola, A., Pahikkala, T., Boberg, J., and Salakoski, T. Applying permutation tests for assessing the statistical significance of wrapper based feature selection. In 2010 Ninth International Conference on Machine Learning and Applications, pages 989–994. IEEE, 2010.

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

Anderson, M. J. and Robinson, J. Permutation tests for linear models. Australian & New Zealand Journal of Statistics, 43(1):75–88, 2001.

Berger, V. W. Pros and cons of permutation tests in clinical trials. Statistics in medicine, 19(10): 1319–1328, 2000.

Berry, K. J., Johnston, J. E., and Mielke Jr, P. W. A chronicle of permutation statistical methods. Cham: Springer, 2014.

Bi, J. and Kuesten, C. Revisiting Fishers Lady Tasting Tea from a perspective of sensory discrimination testing. Food quality and preference, 43:47–52, 2015.

Box, J. R. A. Fisher: The life of a scientist, 1978.

Braun, T. M. and Feng, Z. Optimal permutation tests for the analysis of group randomized trials. Journal of the American Statistical Association, 96(456):1424–1432, 2001.

Cox, D. Randomization in the design of experiments. International Statistical Review, 77(3):415–429, 2009.

Edgington, E. and Onghena, P. Randomization tests. Chapman and Hall/CRC, 2007.

Fan, J., Tam, P., Woude, G. V., and Ren, Y. Normalization and analysis of cdna microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. Proceedings of the National Academy of Sciences, 101(5):1135–1140, 2004.

Fisher, R. A. Statistical methods for research workers. Oliver & Boyd, 1925.

Fisher, R. A. The design of experiments. Oliver and Boyd, 1935a.

Fisher, R. A. The logic of inductive inference. Journal of the royal statistical society, 98(1):39–82, 1935b.

Fisher, R. A. “The coefficient of racial likeness” and the future of craniometry. Journal of the Anthropological Institute of Great Britain and Ireland, 66:57–63, 1936.
Freedman, D. and Lane, D. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

Golland, P., Liang, F., Mukherjee, S., and Panchenko, D. Permutation tests for classification. In *International Conference on Computational Learning Theory*, pages 501–515. Springer, 2005.

Harris, H. and Kalnus, H. The measurement of taste sensitivity to phenylthiourea (P.T.C). *Annals of Eugenics*, 15(1):24–31, 1949.

Hemerik, J. and Goeman, J. J. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155, 2018a.

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

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

Hoeffding, W. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952.

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

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

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

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

Lachin, J. M. Statistical properties of randomization in clinical trials. *Controlled Clinical Trials*, 9(4):289–311, 1988b.

Langsrud, Ø. Rotation tests. *Statistics and computing*, 15(1):53–60, 2005.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.

Meinshausen, N. and Bühlmann, P. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4):893–907, 2005.

Mielke, P. W. and Berry, K. J. *Permutation methods: a distance function approach*. Springer Science & Business Media, 2007.

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

Ojala, M. and Garriga, G. C. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.

Onghena, P. Randomization tests or permutation tests? A historical and terminological clarification. *Randomization, masking, and allocation concealment*, pages 209–227, 2018.

Onghena, P. and Edgington, E. S. Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy*, 32(7):783–786, 1994.

Pesarin, F. Some elementary theory of permutation tests. *Communications in Statistics-Theory and Methods*, 44(22):4880–4892, 2015.

Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):39, 2010.
Pitman, E. J. G. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

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

Rao, K., Drikvandi, R., and Saville, B. Permutation and bayesian tests for testing random effects in linear mixed-effects models. *Statistics in medicine*, 2019.

Rosenberger, W. F., Uschner, D., and Wang, Y. Randomization: The forgotten component of the randomized clinical trial. *Statistics in medicine*, 38(1):1–12, 2019.

Rosenblatt, J. D., Benjamini, Y., Gilron, R., Mukamel, R., and Goeman, J. J. Better-than-chance classification for signal detection. *arXiv preprint arXiv:1608.08873*, 2016.

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

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

Southworth, L. K., Kim, S. K., and Owen, A. B. Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638, 2009.

Suresh, K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences*, 4(1):8, 2011.

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

Wald, A. and Wolfowitz, J. Statistical tests based on permutations of the observations. *The Annals of Mathematical Statistics*, 15(4):358–372, 1944.

Wei, L. and Lachin, J. M. Properties of the urn randomization in clinical trials. *Controlled clinical trials*, 9(4):345–364, 1988.

Westfall, P. H. and Young, S. S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

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

Yates, F. Contingency tables involving small numbers and the $\chi^2$ test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.