Permutation Tests Using Arbitrary Permutation Distributions

Aaditya Ramdas
Carnegie Mellon University, Pittsburgh, USA
Rina Foygel Barber
University of Chicago, Chicago, USA
Emmanuel J. Candès
Stanford University, Stanford, USA
Ryan J. Tibshirani
University of California Berkeley, Berkeley, USA

Abstract

Permutation tests date back nearly a century to Fisher's randomized experiments, and remain an immensely popular statistical tool, used for testing hypotheses of independence between variables and other common inferential questions. Much of the existing literature has emphasized that, for the permutation p-value to be valid, one must first pick a subgroup G of permutations (which could equal the full group) and then recalculate the test statistic on permuted data using either an exhaustive enumeration of G, or a sample from G drawn uniformly at random. In this work, we demonstrate that the focus on subgroups and uniform sampling are both unnecessary for validity—in fact, a simple random modification of the permutation p-value remains valid even when using an arbitrary distribution (not necessarily uniform) over any subset of permutations (not necessarily a subgroup). We provide a unified theoretical treatment of such generalized permutation tests, recovering all known results from the literature as special cases. Thus, this work expands the flexibility of the permutation test toolkit available to the practitioner.

AMS (2000) subject classification. Primary 62G10; Secondary 62H15, 62F03.
Keywords and phrases. Generalized permutation tests, Subgroups, Randomized inference

1 Introduction

Suppose we observe data $X_1, \ldots, X_n \in \mathcal{X}$, and would like to test the null hypothesis

$$H_0 : X_1, \ldots, X_n \text{ are exchangeable.} \tag{1.1}$$

(Note that the hypothesis that the X_i's are i.i.d., is a special case of this null.) We assume that we have a pre-specified test statistic, which is a function $T : \mathcal{X}^n \to \mathbb{R}$, where, without loss of generality, we let larger values of $T(X) = T(X_1, \ldots, X_n)$ indicate evidence in favor of an alternative hypothesis.

Since the null distribution of the X_i's is not specified exactly, we usually do not know the null distribution of $T(X)$. The *permutation test* avoids this difficulty by comparing $T(X)$ against the same function applied to permutations of the data. To elaborate, let $\mathcal{S}_n$ denote the set of all permutations on $[n] := \{1, \ldots, n\}$, and define

$$x_\sigma := (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

for any $x \in \mathcal{X}^n$ and any $\sigma \in \mathcal{S}_n$. Then, we can compute a p-value

$$P = \frac{\sum_{\sigma \in \mathcal{S}_n} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{n!}, \tag{1.2}$$

which ranks $T(X)$ amongst $\{T(X_\sigma)\}_{\sigma \in \mathcal{S}_n}$ sorted in decreasing order. Then, under the null hypothesis H_0, P is a valid p-value, meaning $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$.[1]

As an example, suppose that the observed data set actually consists of pairs (X_i, Y_i), which are assumed to be i.i.d. from some joint distribution. If we are interested in testing $H_0' : X \perp\!\!\!\perp Y$, we can reframe this question as testing whether $X_1, \ldots, X_n$ are i.i.d. conditional on $Y_1, \ldots, Y_n$—in particular, under H_0', it holds that X follows an exchangeable distribution *conditional* on Y. Our test statistic T might be chosen as

$$T(X) = \left| \text{Corr}\big((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)\big) \right|.$$

In order to see whether the observed correlation is sufficiently large to be statistically significant, we would compare $T(X, Y)$ to the correlations computed on permuted data,

$$T(X_\sigma) = \left| \text{Corr}\big((X_{\sigma(1)}, \ldots, X_{\sigma(n)}), (Y_1, \ldots, Y_n)\big) \right|.$$

The resulting p-value computed as in Eq. 1.2 is then a valid p-value under the null hypothesis H_0'. In addition to testing independence, permutation

---

[1] Note that we always have $P > 0$, because of the identity permutation $\sigma = \text{Id} \in \mathcal{S}_n$ (for which $X_\sigma = X$ and thus $\mathbb{1}\{T(X_\sigma) \geq T(X)\} = 1$). Other permutation p-values in this paper, like Eq. 1.3, may explicitly include a "1+" term in the numerator and denominator, but their similarity to the above formula can be intuitively justified by thinking of the extra "1+" as resulting from the identity permutation.

tests are also commonly used for testing other hypotheses, such as whether two samples follow the same distribution.[2]

The p-value P computed in Eq. 1.2 requires computing $T(X_\sigma)$ for every $\sigma \in \mathcal{S}_n$. One may naturally be interested in reducing the computational cost of this procedure, since computing $T(X_\sigma)$ for $|\mathcal{S}_n| = n!$ many permutations may be computationally prohibitive for even moderately large n. As is well known, we can obtain valid p-values by *uniformly* randomly sampling permutations from $\mathcal{S}_n$ and computing

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m}) \geq T(X)\right\}}{1 + M}, \tag{1.3}$$

in which $\sigma_1, \ldots, \sigma_M$ are i.i.d. uniform draws from $\mathcal{S}_n$.

In a different direction, one can also reduce the set of permutations σ to subsets of $\mathcal{S}_n$. Specifically, let $G \subseteq \mathcal{S}_n$ be any subset, and define

$$P = \frac{\sum_{\sigma \in G} \mathbb{1}\left\{T(X_\sigma) \geq T(X)\right\}}{|G|}, \tag{1.4}$$

where $|G|$ is the cardinality of G.

Clearly, if G does not contain the identity permutation, then P cannot be a p-value because it could potentially take on the value zero. However, including the identity permutation is not sufficient. The literature repeatedly emphasizes that P defined in Eq. 1.4 is a valid p-value only if the subset $G \subseteq \mathcal{S}_n$ is in fact a *subgroup*[3] (Hemerik and Goeman, 2018, Theorem 1).

The subgroup G mentioned above may be chosen strategically to balance between computational efficiency and the power of the test (see, e.g., Hemerik and Goeman (2018) & Koning and Hemerik (2022)). In case G has

---

[2]Permutation tests are a special case of "invariance-based testing" (Lehmann et al., 2005, Chapter 6).

[3]For completeness, a group is a set paired with an operation that takes any two elements of the set and produces a third, such that the operation is associative, an identity element exists, and every element has an inverse. A subgroup is just a subset of the original group that maintains the same properties—in particular, any subgroup must contain the identity element. The group $\mathcal{S}_n$ is called the symmetric group; its elements are the $n!$ permutations over n objects. The operation is denoted $\circ$, sometimes called "composition"; its action is to compose any two permutations σ, σ' to yield a third one $\nu := \sigma \circ \sigma'$ which is given by $\nu(i) = \sigma(\sigma'(i))$ for $i \in [n]$. The inverse of σ, denoted σ^{-1}, is defined by setting $\sigma^{-1}(i) = j$ if $\sigma(j) = i$, so that $\sigma \circ \sigma^{-1}$ always equals the identity permutation introduced earlier. Note that $\mathcal{S}_n$ is not an Abelian group, meaning that $\circ$ is not commutative, since usually, $\sigma \circ \sigma' \neq \sigma' \circ \sigma$. For a subset $G \subseteq \mathcal{S}_n$, we can verify that G is a subgroup if it is closed under composition (i.e., $\sigma \circ \sigma' \in G$ for any $\sigma, \sigma' \in G$).

a large cardinality, the aforementioned references show that sampling permutations *uniformly at random* from G also yields valid p-values—that is, the randomized p-value P from Eq. 1.3 is valid if $\sigma_1, \ldots, \sigma_M$ are i.i.d. samples drawn uniformly from G rather than from $\mathcal{S}_n$. Again, choosing G to be a subgroup (rather than an arbitrary subset), and sampling uniformly rather than from an arbitrary distribution, are both important for the validity of the resulting p-value.

1.1. Contributions The background above naturally leads to the following question: while it is indeed correct that P from Eq. 1.4 is not a p-value for general subsets $G \subseteq \mathcal{S}_n$, is it possible to slightly modify the definition of P so that it retains its validity for subsets G that are not subgroups? Further, while sampling the σ_m's nonuniformly from G would destroy the validity of P from Eq. 1.3, can we modify the definition of P so that nonuniform sampling from a set is allowed? The first question is addressed by Hemerik and Goeman (2018), as we will describe below; to our knowledge, the second question has not been addressed in the literature.

In this paper, we will broaden the applicability of permutation tests and present generalizations, which yield valid p-values even when we sample permutations—with or without replacement—from a non-uniform distribution over all permutations or from an arbitrary subset of permutations. In doing so, we shall carefully explain how this generalization relates and extends all previous options. This generalization yields new and more flexible permutation test methods; we leave a detailed study of pros and cons of these generalizations (such as how they trade off the two types of errors) to future work.

2 A Generalized Permutation Test

2.1. Testing with an Arbitrary Distribution We now present our first generalization of the permutation test. It allows us to use any (not necessarily uniform) distribution over $\mathcal{S}_n$ in order to construct our permutation p-value.

THEOREM 2.1. *Let q be any distribution over $\sigma \in \mathcal{S}_n$. Let $\sigma_0 \sim q$ be a random draw, and define*

$$P = \sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X)\right\}. \tag{2.1}$$

Then P is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0,1]$.

In this theorem, validity is retained when conditioning on the order statistics of the data, meaning that $\mathbb{P}_{H_0}\{P \leq \alpha | X_{(1)}, \ldots, X_{(n)}\} \leq \alpha$, where

$X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of $X = (X_1, \ldots, X_n)$.[4] The reason that this holds is simply because H_0 remains true even conditional on the order statistics—that is, if X is exchangeable, then $X \mid (X_{(1)}, \ldots, X_{(n)})$ is again exchangeable. The same conditional validity holds for all results to follow, as well. However, one cannot condition on σ_0; the result only holds marginally over σ_0, and this external randomization is key to retaining validity.

We defer the proof to Section 2.3—we will first discuss connections to the existing literature in order to provide more context and intuition for the above theorem.

Uniform Distribution Over a Subgroup   To begin with, assume q is the uniform distribution over a fixed subgroup G of $\mathcal{S}_n$. Then in this case, the p-value in Eq. 2.1 takes the special form

$$P = \sum_{\sigma \in G} \frac{1}{|G|} \cdot \mathbb{1}\left\{ T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X) \right\} = \sum_{\sigma \in G} \frac{1}{|G|} \cdot \mathbb{1}\left\{ T(X_\sigma) \geq T(X) \right\},$$

where the second equality holds because a subgroup G is closed under inverses and composition, so $\{\sigma \circ \sigma_0^{-1} : \sigma \in G\} = G$ for any $\sigma_0 \in G$. This simple observation recovers a well-known fact we discussed earlier; namely, one can restrict the set of permutations to an arbitrary subgroup, and the p-value P defined in Theorem 2.1 will then coincide with our earlier definition Eq. 1.4 (proved to be a valid p-value in (Hemerik and Goeman, 2018, Theorem 1)).

Uniform Distribution Over a Subset   Consider now a uniform distribution over an arbitrary subset S that is not a subgroup. In this case, the definition of P in Theorem 2.1 is equal to

$$P = \frac{\sum_{\sigma \in S} \mathbb{1}\left\{ T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X) \right\}}{|S|}, \tag{2.2}$$

as proposed earlier by Hemerik and Goeman (2018). This is, in general, not the same as

$$P' = \frac{\sum_{\sigma \in S} \mathbb{1}\left\{ T(X_\sigma) \geq T(X) \right\}}{|S|} \tag{2.3}$$

---

[4]The notation of the order statistics implicitly assumes $\mathcal{X} = \mathbb{R}$. More generally, for an arbitrary space $\mathcal{X}$, the validity of P is retained when conditioning on the unordered observed data, i.e., the multiset $\{X_1, \ldots, X_n\}$.

(which is equivalent to the quantity defined in Eq. 1.4 earlier, with the subset S in place of a subgroup G). As we shall see below, P' is generally not a p-value, a fact which can cause large issues in practice, as has been frequently emphasized. For example, consider the tool of *balanced permutations*—in the setting of testing whether a randomly assigned treatment has a zero or nonzero effect, this method has been proposed as a variant of the permutation test in this setting, where the subset S consists of all permutations such that the permuted treatment group contains exactly half of the original treatment group, and half of the original control group. Southworth et al. (2009) show that the quantity P' computed in Eq. 2.3 for this choice of subset S can be substantially anti-conservative, i.e., $\mathbb{P}\{P \leq \alpha\} > \alpha$, particularly for low significance levels α. (See also Hemerik and Goeman (2018) for additional discussion of this issue.)

A simple example may help to illustrate this point, and to give intuition for the role of the random permutation σ_0.

EXAMPLE 2.1. Let $n = 4$, and consider the set

$$S = \{\mathrm{Id}, \sigma_{1\leftrightarrow3,2\leftrightarrow4}, \sigma_{1\leftrightarrow4,2\leftrightarrow3}\},$$

where, e.g., $\sigma_{1\leftrightarrow3,2\leftrightarrow4}$ is the permutation swapping entries 1 and 3 and also swapping 2 and 4. Let $X_1, X_2, X_3, X_4 \overset{iid}{\sim} \mathcal{N}(0,1)$ be standard normal random variables (so that the null hypothesis of exchangeability, H_0, is satisfied), and set $T(X) = X_1 + X_2$. Then the quantity P' defined in Eq. 2.3 is equal to

$$P' = \frac{\mathbb{1}\left\{T(X_{\mathrm{Id}}) \geq T(X)\right\} + \mathbb{1}\left\{T(X_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}}) \geq T(X)\right\} + \mathbb{1}\left\{T(X_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}}) \geq T(X)\right\}}{3}.$$

This gives

$$P' = \begin{cases} \frac{1+0+0}{3} = \frac{1}{3}, & \text{if } X_3 + X_4 < X_1 + X_2, \\ \frac{1+1+1}{3} = 1, & \text{otherwise} \end{cases}$$

and, therefore,

$$P' = \begin{cases} \frac{1}{3}, & \text{with probability} \frac{1}{2}, \\ 1, & \text{with probability} \frac{1}{2}. \end{cases}$$

We can thus see that P' is anti-conservative at the threshold $\alpha = \frac{1}{3}$.

Next, we will see how the correction in Eq. 2.2 fixes the failure described above. Denote by P_σ the p-value in Eq. 2.2 calculated conditional on the random σ_0 being equal to σ, so that

$$P = \begin{cases} P_{\mathrm{Id}}, & \text{w.p. } 1/3, \\ P_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}}, & \text{w.p. } 1/3, \\ P_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}}, & \text{w.p. } 1/3. \end{cases} \tag{2.4}$$

Then, the calculation that was previously performed effectively shows that

$$P_{\mathrm{Id}} = \begin{cases} \frac{1}{3}, & \text{if } X_3 + X_4 < X_1 + X_2, \\ 1, & \text{otherwise.} \end{cases}$$

A similar straightforward calculation then yields

$$P_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}} = P_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}} = \begin{cases} \frac{2}{3}, & \text{if } X_3 + X_4 < X_1 + X_2, \\ 1, & \text{otherwise.} \end{cases}$$

Put together, we obtain

$$P = \begin{cases} \frac{1}{3}, & \text{w.p. } 1/6, \\ \frac{2}{3}, & \text{w.p. } 1/3, \\ 1, & \text{w.p. } 1/2. \end{cases} \tag{2.5}$$

This is indeed stochastically larger than uniform, and is thus a valid p-value.

The role of σ_0 To better understand the role of the random permutation σ_0, let us consider Example 2.1 again, and look more closely at what goes wrong there. We observe that P' compares the observed statistic $T(X)$ against the set $\{T(X_\sigma)\}_{\sigma\in S} = \{T(X_{\mathrm{Id}}), T(X_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}}), T(X_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}})\}$. For P' to be a valid p-value, given the (unordered) set of potential data vectors $\{X_{\mathrm{Id}}, X_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}}, X_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}}\}$, it suffices that the actual observed data X is equally likely to be any one of these three. Now suppose this set is equal to $\{(0.8, 0.5, 0.2, 1.0), (1.0, 0.2, 0.5, 0.8), (0.5, 0.8, 1.0, 0.2)\}$, in no particular order. Each of these three vectors have equal likelihood under H_0 (due to exchangeability). Counterintuitively, however, our knowledge of the subset of permutations S implies that we *must* have $X = (1.0, 0.2, 0.5, 0.8)$—otherwise we could not have obtained this particular set. For instance, if $X = (0.8, 0.5, 0.2, 1.0)$, then we would have $X_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}} = (0.2, 1.0, 0.8, 0.5)$—but this does not lie in our set, so it cannot be the correct value of X. In other words, if we condition on the unordered set $\{X_\sigma\}_{\sigma\in S}$, which is the *orbit* of the data X under the actions of permutations $\sigma \in S$, our intuition tells us that X is equally likely to be any element of this orbit—but in fact, for a non-subgroup S, X might be uniquely identified from its orbit.

Now consider what happens if we compute the corrected p-value P in Eq. 2.2 and let us once more examine the question of identifying the data X from its orbit. The p-value P compares the observed statistic $T(X)$ against the set $\{T(X_{\sigma\circ\sigma_0^{-1}})\}_{\sigma\in S}$, and so now the question is whether we can identify

X from the set $\{X_{\sigma\circ\sigma_0^{-1}}\}_{\sigma\in S}$, which is the orbit of $X_{\sigma_0^{-1}}$ for a randomly drawn $\sigma_0 \in S$. Identifying X is no longer possible because of the random σ_0. For instance, working again with Example 2.1, suppose this set $\{X_{\sigma\circ\sigma_0^{-1}}\}_{\sigma\in S}$ is equal to $\{(0.8, 0.5, 0.2, 1.0), (1.0, 0.2, 0.5, 0.8), (0.5, 0.8, 1.0, 0.2)\}$, in no particular order. We can identify that this is the orbit of $x = (1.0, 0.2, 0.5, 0.8)$ under S—that is, this set is equal to $\{x_\sigma\}_{\sigma\in S}$. Then the following three possibilities are equally likely:

- $\sigma_0 = \text{Id}$ and so $X = x_{\text{Id}^{-1}} = x = (1.0, 0.2, 0.5, 0.8)$;

- $\sigma_0 = \sigma_{1\leftrightarrow3,2\leftrightarrow4}$ and so $X = x_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}^{-1}} = (0.5, 0.8, 1.0, 0.2)$;

- $\sigma_0 = \sigma_{1\leftrightarrow4,2\leftrightarrow3}$ and so $X = x_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}^{-1}} = (0.8, 0.5, 0.2, 1.0)$.

In other words, X is now equally likely to be any of the three values in our set, and validity is restored.

2.2. Random samples from an arbitrary distribution Our second generalization concerns permutations that are randomly chosen from an arbitrary distribution.

THEOREM 2.2. *Let q be any distribution over $\sigma \in \mathcal{S}_n$. Let $\sigma_0, \sigma_1, \ldots, \sigma_M \overset{iid}{\sim} q$, and define*

$$P = \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T(X_{\sigma_m\circ\sigma_0^{-1}}) \geq T(X)\right\}}{1 + M}. \tag{2.6}$$

Then P is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0,1]$.

This result is closely related to Besag and Clifford (1989)'s well known construction for obtaining exchangeable samples from Markov chain Monte Carlo (MCMC) sampling—the details are deferred to Section 4.2.

Just as before, some special cases of this result are well known to statisticians.

Random permutations from $\mathcal{S}_n$ In the simple case where q is the uniform distribution over $\mathcal{S}_n$, Theorem 2.2 states that

$$P = \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T(X_{\sigma_m\circ\sigma_0^{-1}}) \geq T(X)\right\}}{1 + M} \overset{d}{=} \frac{1 + \sum_{m=1}^M \mathbb{1}\{T(X_{\sigma_m}) \geq T(X)\}}{1 + M} \tag{2.7}$$

is a valid p-value. The equality in distribution above holds because the $\sigma_m \circ \sigma_0^{-1}$'s are i.i.d. draws from $\mathcal{S}_n$. Hence, this recovers the most commonly implemented form of the permutation test.

Random permutations from a subgroup The distributional equality Eq. 2.7 extends to any uniform distribution q over a subgroup G of $\mathcal{S}_n$ since in this case, as before, the random variables $\sigma_m \circ \sigma_0^{-1}$ are i.i.d. draws from G. This gives the following well-known result (see, e.g., Hemerik and Goeman (2018, Theorem 2)):

COROLLARY 2.1. *Let $G \subseteq \mathcal{S}_n$ be a subgroup, and sample $\sigma_1, \ldots, \sigma_M \overset{iid}{\sim} Unif(G)$. Then*

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m}) \geq T(X)\right\}}{1 + M} \tag{2.8}$$

is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$.

Random permutations from a subset Consider now the case where q is a uniform distribution over an arbitrary subset S. When S is not a subgroup, the p-value

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1 + M}$$

may not have the same distribution as

$$P' = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m}) \geq T(X)\right\}}{1 + M}.$$

Here, Theorem 2.2 gives:

COROLLARY 2.2. *Let $S \subseteq \mathcal{S}_n$ be any fixed subset of permutations. Sample $\sigma_0, \sigma_1, \ldots, \sigma_M \overset{iid}{\sim} Unif(S)$. Then*

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1 + M} \tag{2.9}$$

is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$.

To the best of our knowledge, this statement had not been recorded in the literature. As a variant, the same result holds if we instead draw permutations without replacement.

COROLLARY 2.3. *Consider the variant of Corollary 2.2 in which the permutations are drawn without replacement. Then the p-value P defined in Eq. 2.9 is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$. In the special case where S is a subgroup, the same conclusion also applies for the p-value P defined in Eq. 2.8.*

PROOF. Let $S' \subseteq S$ be a subset of size $M+1$ chosen uniformly at random. Let $\sigma_0, \sigma_1, \ldots, \sigma_M$ be a random ordering of S'—in particular, this means that σ_0 is drawn uniformly from S'. Then by Theorem 2.1, applied with q taken to be the uniform distribution over S', P is a valid p-value.

The second claim follows from the fact that

$$(\sigma_1 \circ \sigma_0^{-1}, \ldots, \sigma_M \circ \sigma_0^{-1}) \overset{d}{=} (\sigma_1, \ldots, \sigma_M),$$

whenever S is a subgroup.

To guide the reader, Fig. 1 summarizes the connections between all the results presented thus far in the paper. Interestingly, as highlighted in the figure, Theorems 2.1 and 2.2 can be derived from each other; we will elaborate on this connection below.

Finally, we present another simple example to highlight the necessity of the σ_0 term, in the case of nonuniform sampling. Indeed, even "intuitive" modifications of the uniform sampling scheme may fail to produce valid p-values.

EXAMPLE 2.2. If one considers $P' = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(X_{\sigma_m}) \geq T(X)\}}{1 + M}$ to be a Monte Carlo estimate of the p-value $P = \frac{\sum_{\sigma \in \mathcal{S}_n} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{n!}$ computed in Eq. 1.2, then a lower-variance estimate may be obtained by "antithetic sampling"—that is, pairing a random draw $\sigma_m \in \mathcal{S}_n$ with its reverse $\text{Rev}(\sigma_m) = (\sigma_m(n), \ldots, \sigma_m(1))$ (see, e.g., Mitchell et al. (2022) for an example of this variance reduction technique). However, using antithetic sampling can lead to an invalid p-value—specifically, if $\sigma_1, \ldots, \sigma_{M/2}$ are drawn uniformly at random from $\mathcal{S}_n$ (or from some subgroup $G \subseteq \mathcal{S}_n$), and we
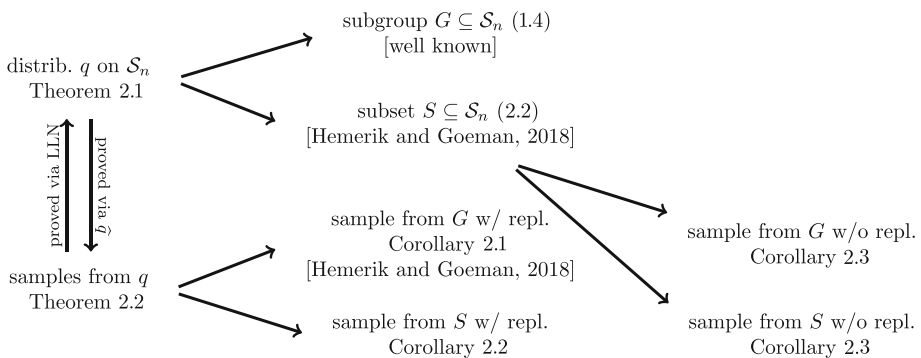

Figure 1: A flowchart to illustrate the connections between (most of) the results presented in this paper. Arrows point from more general results to their special cases

then set $\sigma_{M/2+i} = \text{Rev}(\sigma_i)$ for each $i = 1, \ldots, M/2$, then the quantity P' may not be a valid p-value. For instance, suppose we take $M = 2$, so that $\sigma_2 = \text{Rev}(\sigma_1)$ where σ_1 is drawn uniformly from $\mathcal{S}_n$. Take $T(x) = X_1 + X_n$, and draw $X_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Then

$$P' = \frac{1 + \mathbb{1}\left\{X_{\sigma_1(1)} + X_{\sigma_1(n)} \geq X_1 + X_n\right\} + \mathbb{1}\left\{X_{\sigma_2(1)} + X_{\sigma_2(n)} \geq X_1 + X_n\right\}}{3}$$

$$= \frac{1 + 2 \cdot \mathbb{1}\left\{X_{\sigma_1(1)} + X_{\sigma_1(n)} \geq X_1 + X_n\right\}}{3}.$$

Then we can verify that, conditional on σ_1, if $\{\sigma_1(1), \sigma_1(n)\} = \{1, n\}$ then $P = 1$, while if $\{\sigma_1(1), \sigma_1(n)\} \neq \{1, n\}$ then $P' = \frac{1}{3}$ or $P' = 1$ each with probability $\frac{1}{2}$, which yields

$$\mathbb{P}\left\{P' = \frac{1}{3}\right\} = \frac{1}{2} - \frac{1}{n(n-1)} > \frac{1}{3},$$

with the last step holding if $n > 3$. We can thus see that P' is anti-conservative at the threshold $\alpha = \frac{1}{3}$.

  2.3. *Proof of* Theorem 2.1 First, for any fixed $\sigma' \in \mathcal{S}_n$, we have

$$\mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\{T(X_\sigma) \geq T(X_{\sigma'})\} \leq \alpha\right\} = \mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\{T(X_{\sigma \circ \sigma'^{-1}}) \geq T(X)\} \leq \alpha\right\}$$

$$(2.10)$$

because $X \overset{d}{=} X_{\sigma'}$ under H_0 (and note that $(X_{\sigma'})_{\sigma \circ \sigma'^{-1}} = X_\sigma$). Next, we will apply a deterministic inequality by Harrison (2012): for all $t_1, \ldots, t_N \in [-\infty, \infty]$ and all $\alpha, w_1, \ldots, w_N \in [0, \infty]$,

$$\sum_{k=1}^{N} w_k \mathbb{1}\left\{\sum_{i=1}^{N} w_i \mathbb{1}\{t_i \geq t_k\} \leq \alpha\right\} \leq \alpha.$$

Applying this bound with $q(\sigma)$'s in place of the w_i's, and $T(X_\sigma)$'s in place of the t_i's, we obtain

$$\sum_{\sigma' \in \mathcal{S}_n} q(\sigma') \cdot \mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\{T(X_\sigma) \geq T(X_{\sigma'})\} \leq \alpha\right\}$$

$$= \mathbb{E}\left[\sum_{\sigma' \in \mathcal{S}_n} q(\sigma') \cdot \mathbb{1}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\{T(X_\sigma) \geq T(X_{\sigma'})\} \leq \alpha\right\}\right] \leq \alpha.$$

$$(2.11)$$

Finally, we have

$$
\begin{aligned}
\mathbb{P}\left\{P \leq \alpha\right\} &= \mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X)\right\} \leq \alpha\right\} \\
&= \sum_{\sigma' \in \mathcal{S}_n} \mathbb{P}\left\{\sigma_0 = \sigma' \text{ and } \sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma'^{-1}}) \geq T(X)\right\} \leq \alpha\right\} \\
&= \sum_{\sigma' \in \mathcal{S}_n} q(\sigma') \cdot \mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma'^{-1}}) \geq T(X)\right\} \leq \alpha\right\} \\
&= \sum_{\sigma' \in \mathcal{S}_n} q(\sigma') \cdot \mathbb{P}\left\{\sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\left\{T(X_\sigma) \geq T(X_{\sigma'})\right\} \leq \alpha\right\} \leq \alpha,
\end{aligned}
$$

where the third step holds since $\sigma_0 \sim q$ is drawn independently of the data X, while the last two steps apply Eq. 2.10 and Eq. 2.11.

2.4. *Connecting* Theorems 2.1 *and* 2.2 As mentioned earlier, Theorems 2.1 and 2.2 can be derived from each other. We now give these proofs to show the connection.

ALTERNATIVE PROOF OF THEOREM 2.1 (VIA THEOREM 2.2). Let $\sigma_0, \sigma_1, \sigma_2, \ldots \overset{iid}{\sim} q$, and for any fixed M, define

$$
\begin{aligned}
P_M &= \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1 + M} = \frac{\sum_{m=0}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1 + M} \\
&= \sum_{\sigma \in \mathcal{S}_n} \frac{\sum_{m=0}^{M} \mathbb{1}\left\{\sigma_m = \sigma\right\}}{1 + M} \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X)\right\}.
\end{aligned}
$$

By the Law of Large Numbers, we see that $\frac{\sum_{m=0}^{M} \mathbb{1}\{\sigma_m = \sigma\}}{1+M} \to q(\sigma)$ almost surely for all $\sigma \in \mathcal{S}_n$, and therefore, $P_M \to P$ almost surely, where P is the p-value defined in Eq. 2.1. In particular, this implies that P_M converges to P in distribution, and therefore

$$
\mathbb{P}\left\{P \leq \alpha\right\} = \lim_{M \to \infty} \mathbb{P}\left\{P_M \leq \alpha\right\} \leq \alpha,
$$

where the last step holds since, for every $M \geq 1$, P_M is a valid p-value by Theorem 2.2.

PROOF OF THEOREM 2.2 (VIA THEOREM 2.1). Let $\sigma_0, \sigma_1, \ldots, \sigma_M \overset{iid}{\sim} q$, and define the empirical distribution

$$
\widehat{q} = \frac{1}{M+1} \sum_{m=0}^{M} \delta_{\sigma_m},
$$

where δ_σ is the point mass at σ. Now we treat $\widehat{q}$ as fixed. Let k be drawn uniformly from $\{0, \ldots, M\}$ (that is, σ_k is drawn at random from $\widehat{q}$). Applying Theorem 2.1 with $\widehat{q}$ in place of q, we then see that

$$P = \sum_{\sigma \in \mathcal{S}_n} \widehat{q}(\sigma) \cdot \mathbb{1}\left\{T(X_{\sigma \circ \sigma_k^{-1}}) \geq T(X)\right\} = \frac{\sum_{m=0}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_k^{-1}}) \geq T(X)\right\}}{1 + M}$$

is a valid p-value conditional on $\widehat{q}$, and therefore also valid after marginalizing over $\widehat{q}$. Since $\sigma_0, \ldots, \sigma_M$ are drawn i.i.d. and are therefore in a random order, we see that

$$
\begin{aligned}
P &= \frac{\sum_{m=0}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_k^{-1}}) \geq T(X)\right\}}{1+M} \overset{d}{=} \frac{\sum_{m=0}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1+M} \\
&= \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1+M},
\end{aligned}
$$

which is the desired p-value.

2.5. Another perspective: exchangeable permutations Many of the results described above can be viewed through the lens of exchangeability—not on the data X (which we assume to be exchangeable under the null hypothesis H_0), but on the collection of permutations used to define the p-value P.

THEOREM 2.3. *Let $\sigma_0, \sigma_1, \ldots, \sigma_M \in \mathcal{S}_n$ be a random set of permutations, which are exchangeable, i.e.,*

$$(\sigma_0, \sigma_1, \ldots, \sigma_M) \overset{d}{=} (\sigma_{\pi(0)}, \sigma_{\pi(1)}, \ldots, \sigma_{\pi(M)})$$

for any fixed permutation π on $\{0, \ldots, M\}$. Then

$$P = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(X_{\sigma_m \circ \sigma_0^{-1}}) \geq T(X)\right\}}{1 + M}$$

is a valid p-value, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$.

Many of the results stated earlier can be viewed as special cases—in particular, the results for a subgroup G, or for a subset S, as well as our more general result Theorem 2.2 for permutations drawn i.i.d. from q.

PROOF. To be clear, this theorem is essentially just a new perspective, and can be proved as a corollary to Theorem 2.1. To see why, let $\sigma_0, \ldots, \sigma_M$ be exchangeable, and let $\widehat{q} = \frac{1}{M+1} \sum_{m=0}^{M} \delta_{\sigma_m}$ be the empirical distribution

induced by the *unordered* set of drawn permutations. Then since $\sigma_0, \ldots, \sigma_M$ is exchangeable, conditional on $\widehat{q}$ it holds that σ_0 is a random draw from $\widehat{q}$. Applying Theorem 2.1 with $\widehat{q}$ in place of q gives the conclusion.

However, we can also prove this result in a more intuitive way, using the framework of exchangeability:

ALTERNATIVE PROOF OF THEOREM 2.3. Since the sequence $\sigma_0, \sigma_1, \ldots,$ σ_M is exchangeable,

$$T(X_{\sigma_0}), T(X_{\sigma_1}), \ldots, T(X_{\sigma_M})$$

is also exchangeable conditional on X. It is thus still exchangeable after marginalizing over X. Therefore, under the null hypothesis H_0, the test statistic values

$$T(X) = T(X_{\sigma_0 \circ \sigma_0^{-1}}), \; T(X_{\sigma_1 \circ \sigma_0^{-1}}), \; \ldots, \; T(X_{\sigma_M \circ \sigma_0^{-1}}) \qquad (2.12)$$

are also exchangeable—this follows immediately from the previous line because $X \overset{d}{=} X_{\sigma_0^{-1}}$ under H_0. This shows that the p-value P defined in Theorem 2.3 is valid.

3 Averaging to reduce variance

The p-value P defined in Eq. 2.1 can equivalently be written as

$$P = \mathbb{P}_{\sigma \sim q} \left\{ T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X) | X, \sigma_0 \right\}.$$

It is clear that P is random even if we condition on the observed data X, because of the randomness due to σ_0. Consequently, in some settings P may be quite variable conditional on the data X, and this may be undesirable.

To address this issue, we can also consider averaging over σ_0 (in addition to averaging over σ) in the calculation of P. This alternative definition is now a deterministic function of the observed data X, but may no longer be a valid p-value. Nonetheless, the following theorem shows a bound on the Type I error.

THEOREM 3.1. *Let q be any distribution over $\sigma \in \mathcal{S}_n$. Define*

$$\bar{P} = \sum_{\sigma, \sigma_0 \in \mathcal{S}_n} q(\sigma) q(\sigma_0) \cdot \mathbb{1} \left\{ T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X) \right\},$$

$$(3.1)$$

or equivalently,

$$\bar{P} = \mathbb{P}_{\sigma,\sigma_0 \overset{iid}{\sim} q} \left\{ T(X_{\sigma \circ \sigma_0^{-1}}) \geq T(X) \Big| X \right\}.$$

Then $\bar{P}$ is a valid p-value up to a factor of 2, i.e., $\mathbb{P}_{H_0} \{\bar{P} \leq \alpha\} \leq 2\alpha$ for all $\alpha \in [0, 1]$. In other words, the quantity $\min\{2\bar{P}, 1\}$ is a valid p-value.

PROOF. Draw $\sigma_0^{(1)}, \sigma_0^{(2)}, \dots \overset{iid}{\sim} q$. Let

$$P_m = \sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1} \left\{ T(X_{\sigma \circ \sigma_0^{(m)-1}}) \geq T(X) \right\},$$

for each $m \geq 1$. Then by Theorem 2.1, each P_m is a valid p-value. It is known (Rüschendorf, 1982; Vovk and Wang, 2020) that the average of valid p-values is a valid up to a factor of 2, i.e., for any $M \geq 1$ the average $\bar{P}_M = \frac{1}{M} \sum_{m=1}^{M} P_m$ satisfies $\mathbb{P}\{\bar{P}_M \leq \alpha\} \leq 2\alpha$ for all $\alpha \in [0, 1]$. We can equivalently write

$$\bar{P}_M = \sum_{\sigma' \in \mathcal{S}_n} \frac{\sum_{m=1}^{M} \mathbb{1}\left\{\sigma_0^{(m)} = \sigma'\right\}}{M} \cdot \sum_{\sigma \in \mathcal{S}_n} q(\sigma) \cdot \mathbb{1}\{T(X_{\sigma \circ \sigma'^{-1}}) \geq T(X)\}.$$

By the Law of Large Numbers, $\bar{P}_M$ converges almost surely to the p-value $\bar{P}$ defined in Eq. 3.1, which completes the proof.

Returning to Example 2.1, we see that while P was a *mixture* of P_{Id}, $P_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}}, P_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}}$, we now have that $\bar{P}$ is an *average* of these, meaning $\bar{P} = \frac{1}{3}(P_{\mathrm{Id}} + P_{\sigma_{1\leftrightarrow4,2\leftrightarrow3}} + P_{\sigma_{1\leftrightarrow3,2\leftrightarrow4}})$. Simplifying, we get

$$\bar{P} = \begin{cases} \frac{5}{9}, & \text{w.p. } 1/2, \\ 1, & \text{w.p. } 1/2. \end{cases}$$

It is worth noting that this new quantity $\bar{P}$ is neither more conservative nor more anti-conservative than the p-value P in Eq. 2.5 from earlier. This is perhaps a more general phenomenon: the average of p-values need not in general be anti-conservative, and indeed it could often be more conservative, than the original p-values.

Analogously, the p-value in Theorem 2.2, computed via random samples from q, can also be averaged to reduce variance.

THEOREM 3.2. *Let q be any distribution over $\sigma \in \mathcal{S}_n$. Let $\sigma_0, \sigma_1, \dots, \sigma_M \overset{iid}{\sim} q$, and define*

$$\bar{P} = \frac{\sum_{m=0}^{M} \sum_{m'=0}^{M} \mathbb{1}\left\{ T(X_{\sigma_m \circ \sigma_{m'}^{-1}}) \geq T(X) \right\}}{(1 + M)^2}. \tag{3.2}$$

Then P is a valid p-value up to a factor of 2, i.e., $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq 2\alpha$ for all $\alpha \in [0,1]$. Thus, as before, the quantity $\min\{2\bar{P}, 1\}$ is a valid p-value.

The proof is similar to that of Theorem 3.1, and we omit it for brevity.

4   Connections to the Literature

We next mention a few connections to the broader literature.

4.1.   Permutation tests vs randomization tests Hemerik and Goeman (2021) describe the difference between two testing frameworks, permutation tests (as studied in our present work) versus randomization tests. The difference is subtle, because randomization tests may still use permutations. Specifically, Hemerik and Goeman (2021) highlight

> 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.

To better understand this distinction, we can consider a scenario where a fixed subset $S \subseteq \mathcal{S}_n$, which is not a subgroup, is used for a randomization test rather than a permutation test. Consider a study comparing a treatment versus a placebo, with $n/2$ many subjects assigned to each of the two groups. We can use a permutation σ to denote the treatment assignments, with $\sigma(i) \leq n/2$ indicating that subject i receives the treatment, and $\sigma(i) > n/2$ indicating that subject i receives the placebo. Now we switch notation, to be able to compare to permutation tests more directly—writing $X = (1, \ldots, 1, 0, \ldots, 0)$, suppose that we will assign treatments via the permuted vector X_σ, i.e., for each subject $i = 1, \ldots, n$, under this permutation σ the ith subject will receive the treatment if $X_{\sigma(i)} = 1$, or the placebo if $X_{\sigma(i)} = 0$.

Now suppose that we draw a random treatment assignment $\sigma_{\mathrm{asgn}} \sim \mathrm{Unif}(S)$, from a fixed subset $S \subseteq \mathcal{S}_n$ (for example, S may be chosen to restrict to treatment assignments that are equally balanced across certain subpopulations). After the treatments are administered, the measured response variable is given by $Y = (Y_1, \ldots, Y_n)$. Fix any test statistic $T(X) = T(X, Y)$ (we will implicitly condition on Y), and compute

$$P = \frac{\sum_{\sigma \in S} \mathbb{1}\left\{T(X_\sigma) \geq T(X_{\sigma_{\mathrm{asgn}}})\right\}}{|S|}. \tag{4.1}$$

Since σ_{asgn} was drawn uniformly from S, this quantity P is a valid p-value. In the terminology of Hemerik and Goeman (2021), this test is a randomization test, not a permutation test. While the set of possible treatment assigments $\{X_\sigma : \sigma \in S\}$ happens to be indexed by permutations σ, the group structure of permutations is not used in any way, and we do not rely on any invariance properties.

Comparing to the invalid p-value $P = \frac{\sum_{\sigma \in S} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{|S|}$ considered in Eq. 2.3, we can easily see the distinction: for a randomization test, the observed statistic is $T(X_{\sigma_{\mathrm{asgn}}})$ for a randomly drawn $\sigma_{\mathrm{asgn}} \sim \mathrm{Unif}(S)$, while in the permutation test in Eq. 2.3, the observed statistic is $T(X)$ (i.e., using the *fixed* permutation Id in place of a randomly drawn σ_{asgn}). For this reason, the randomization test p-value in Eq. 4.1 is valid, while the permutation test calculation in Eq. 2.3 is not valid in general.

Now we again consider Hemerik and Goeman (2018)'s method using a fixed subset. This test in Eq. 2.2 is a permutation test, not a randomization test—the observed data X, and its corresponding statistic $T(X)$, do not arise from a random treatment assignment. More generally, our proposed test in Eq. 2.1 using an arbitrary distribution q on $\mathcal{S}_n$ is again a permutation test rather than a randomization test—that is, the observed data is given by X itself, not by a randomly chosen treatment assignment $X_{\sigma_{\mathrm{asgn}}}$ for $\sigma_{\mathrm{asgn}} \sim q$. Nonetheless, we are able to produce a valid p-value without assuming an underlying group structure or uniform sampling for the permutations considered by the test.

4.2.   Exchangeable MCMC The result of Theorem 2.2, which allows for random samples drawn from an arbitrary distribution q on $\mathcal{S}_n$, is closely connected to Besag and Clifford (1989)'s well known construction for obtaining exchangeable samples from Markov chain Monte Carlo (MCMC) sampling.

Consider a distribution Q_0 on $\mathcal{Z}$, and suppose we want to test

$$H_0 : \; Z \sim Q_0$$

with some test statistic $T(Z)$. To find a significance threshold for $T(Z)$, we would ideally like to draw from the null distribution, i.e., compare $T(Z)$ against $T(Z_1), \ldots, T(Z_M)$ for $Z_1, \ldots, Z_M \overset{iid}{\sim} Q_0$. However, in many settings, sampling directly from Q_0 is impossible, but we instead have access to a Markov chain whose stationary distribution is Q_0. If we run the Markov chain initialized at Z to obtain draws $Z_1, \ldots, Z_M$ (say, running the Markov chain for some fixed number of steps s between each draw), then dependence among these sequentially drawn samples means that $Z, Z_1, \ldots, Z_M$ are not

i.i.d., and are not even exchangeable. Without studying the mixing properties of the Markov chain, we cannot determine how large the number of steps needs to be for the dependence to become negligible. Instead, Besag and Clifford ([1989](#)) propose a construction where the samples are drawn in parallel (rather than sequentially), which ensures exchangeability:

THEOREM 4.1 (Besag and Clifford ([1989](#), Section 2)). *Let Q_0 be any distribution on a probability space $\mathcal{Z}$. Construct a Markov chain on $\mathcal{Z}$ with stationary distribution Q_0, whose forward and backward transition distributions (initialized at $z \in \mathcal{Z}$) are denoted by $Q_\rightarrow(\cdot|z)$ and $Q_\leftarrow(\cdot|z)$. Let $Q_\rightarrow^s(\cdot|z)$ and $Q_\leftarrow^s(\cdot|z)$ denote the forward and backward transition distributions after running s steps of the Markov chain, for some fixed $s \geq 1$. Given an initialization Z, suppose we generate data as in the left plot of* Fig. [2](#):*

$$\begin{cases} First, \ draw \ Z_* \sim Q_\leftarrow^s(\cdot|Z); \\ Then, \ draw \ Z_1, \ldots, Z_M \overset{iid}{\sim} Q_\rightarrow^s(\cdot|Z_*). \end{cases}$$

If it holds marginally that $Z \sim Q_0$, then the draws $Z, Z_1, \ldots, Z_M$ are exchangeable.

Given this exchangeability property, the quantity $P = \frac{1 + \sum_{m=1}^M \mathbb{1}\{T(Z_m) \geq T(Z)\}}{1+M}$ is then a valid p-value for testing $H_0 : Z \sim Q_0$.

Now we will see how Theorem [2.2](#) is related to this result. Let $\mathcal{Z} = \mathcal{X}^n$, and let Q_0 be any exchangeable distribution. In the setting of this paper, we do not know Q_0 precisely, which makes it a bit different from a typical setting where Besag and Clifford ([1989](#))'s method is applied. However, we
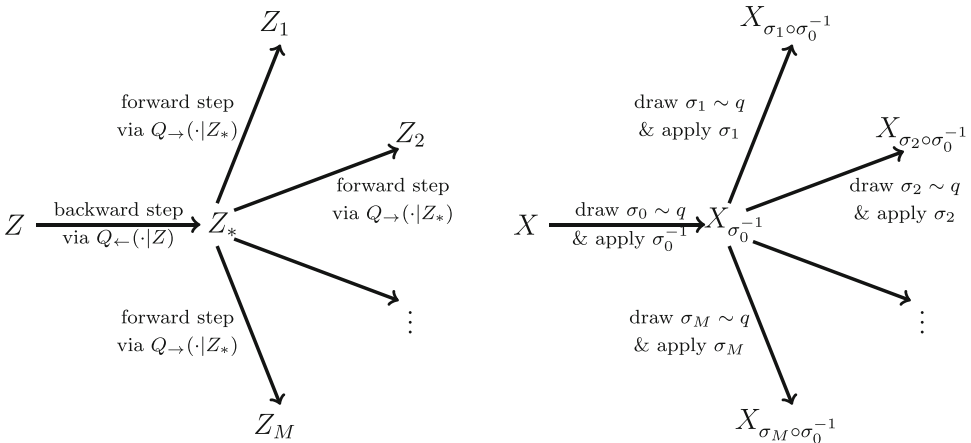

Figure 2: Left: Besag and Clifford ([1989](#))'s parallel construction (with $s = 1$). Right: the construction used in Theorem [2.2](#)

will work with a Markov chain for which *any* exchangeable distribution Q_0 is stationary, and in fact, Theorem 4.1 holds regardless of whether Q_0 is the unique stationary distribution for the Markov chain.

Consider the Markov chain given by applying a randomly chosen permutation $\sigma \sim q$, that is, for $x = (x_1, \ldots, x_n)$,

$$Q_{\rightarrow}(\cdot|x) = \sum_{\sigma \in S} q(\sigma) \cdot \delta_{x_\sigma},$$

where δ_{x_σ} is the point mass at x_σ, while the backward transition probabilities are given by

$$Q_{\leftarrow}(\cdot|x) = \sum_{\sigma \in S} q(\sigma) \cdot \delta_{x_{\sigma^{-1}}}.$$

Then, to implement the test described in Theorem 2.2, we run Besag and Clifford (1989)'s method (with $s = 1$): we define $X_* = X_{\sigma_0^{-1}}$, and then define $X_m = (X_*)_{\sigma_m} = X_{\sigma_m \circ \sigma_0^{-1}}$ for $m = 1, \ldots, M$. This is illustrated on the right-hand side of Fig. 2. If X is exchangeable (that is, it is drawn from some exchangeable Q_0), then the exchangeability of $X, X_1, \ldots, X_M$ follows by Theorem 4.1, and this verifies that P is a valid p-value, thus completing the proof of Theorem 2.2.

Of course, we have only written out our method for the $s = 1$ case (where s is the number of steps of the Markov chain). New variants of our method can be constructed by taking $s > 1$ backward steps to the hidden node, and the same number s of forward steps to the permuted data. All of these are valid for the same reason as the $s = 1$ case.

5 Conclusion

We proposed a new method for permutation testing that generalizes previous methods. This idea naturally opens up new lines of theoretical and practical enquiry. In this work, we have focused on validity, but it is of course also important to examine the consistency and power of such methods. In particular, Dobriban (2022) & Kim et al. (2022) study the power of the permutation test when using the full permutation group $\mathcal{S}_n$; it would be interesting to examine this question in the context of using only a subset $S \subseteq \mathcal{S}_n$ or a nonuniform distribution over $\mathcal{S}_n$. In addition, the theoretical guarantees for all the permutation tests considered here ensure a p-value P that is valid in the sense of satisfying $\mathbb{P}_{H_0}\{P \leq \alpha\} \leq \alpha$, which means that P

could potentially be quite conservative under the null (for instance, we saw this behavior when 'fixing' the failure example in Section 2.1). It would also be interesting to understand which types of tests reduce overly conservative outcomes.

In conclusion, it is perhaps remarkable that one can still gain new understanding about classical permutation methods. In turn, this enhanced understanding can inform other areas of inference. As an example, the results from this paper were motivated by questions in conformal prediction (Vovk et al., 2005), a method for distribution-free predictive inference. Classically, conformal prediction has relied on exchangeability of data points (e.g., training and test data are drawn i.i.d. from the same unknown distribution), and thus the joint distribution of the data (including both training samples and a test point) is invariant under an arbitrary permutation. In contrast, in our recent work (Barber et al., 2023), we studied the problem of constructing prediction intervals when the data do not satisfy exchangeability; for instance, the distribution of observations may simply drift over time in an unknown fashion. Thus the data is no longer invariant under an arbitrary permutation, and so we instead restrict attention to a weighted distribution over simple permutations that only swap the test point with a random training point, which at least approximately preserve the distribution of the data. These swaps clearly do not form a subgroup of permutations, and are weighted non-uniformly; understanding how permutation tests operate in this setting, as in Theorem 2.1, is key to the findings in our aforementioned work.

Compliance with Ethical Standards.

Conflict of Interest. The authors have no conflict of interest to declare.

References

BARBER, RF, CANDÈS, EJ, RAMDAS, A and TIBSHIRANI, RJ (2023). Conformal prediction beyond exchangeability. The Annals of Statistics (to appear).

BESAG, J and CLIFFORD, P (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633–642.

DOBRIBAN, E (2022). Consistency of invariance-based randomization tests. *The Annals of Statistics* **50**, 4, 2443–2466.

HARRISON, MT (2012). Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika* **99**, 57–69.

HEMERIK, J and GOEMAN, J (2018). Exact testing with random permutations. *Test* **27**, 811–825.

HEMERIK, J and GOEMAN, J J (2021). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *Int Stat Rev* **89**, 367–381.

KIM, I, BALAKRISHNAN, S and WASSERMAN, L (2022). Minimax optimality of permutation tests. *Ann Stat* **50**, 225–251.

KONING, N W and HEMERIK, J (2022). More Efficient Exact Group-Invariance Testing: using a Representative Subgroup. arXiv:2202.00967.

LEHMANN, EL, ROMANO, JP and CASELLA, G (2005). Testing statistical hypotheses, volume 3. Springer.

MITCHELL, R, COOPER, J, FRANK, E and HOLMES, G (2022). Sampling permutations for shapley value estimation. *Journal of Machine Learning Research* **23**, 43, 1–46.

RÜSCHENDORF, L (1982). Random variables with maximum sums. *Adv Appl Probab* **14**, 623–632.

SOUTHWORTH, LK, KIM, SK and OWEN, AB (2009). Properties of balanced permutations. *J Comput Biol* **16**, 625–638.

VOVK, V, GAMMERMAN, A and SHAFER, G (2005). Algorithmic learning in a random world. Springer Science & Business Media.

VOVK, V and WANG, R (2020). Combining p-values via averaging. *Biometrika* **107**, 791–808.

AADITYA RAMDAS
DEPARTMENTS OF STATISTICS AND
MACHINE LEARNING, CARNEGIE MELLON
UNIVERSITY, PITTSBURGH, USA
E-MAIL: ARAMDAS@STAT.CMU.EDU

RINA FOYGEL BARBER
DEPARTMENT OF STATISTICS, UNIVERSITY
OF CHICAGO, CHICAGO, USA

EMMANUEL J. CANDÈS
DEPARTMENTS OF STATISTICS AND
MATHEMATICS, STANFORD UNIVERSITY,
STANFORD, USA

RYAN J. TIBSHIRANI
DEPARTMENT OF STATISTICS, UNIVERSITY
OF CALIFORNIA BERKELEY, BERKELEY,
USA