Gaussian prepivoting for finite population causal inference

Peter L. Cohen | Colin B. Fogarty

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence
Peter L. Cohen, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
Email: pclcohen@mit.edu

Abstract
In finite population causal inference exact randomization tests can be constructed for sharp null hypotheses, hypotheses which impute the missing potential outcomes. Oftentimes inference is instead desired for the weak null that the sample average of the treatment effects takes on a particular value while leaving the subject-specific treatment effects unspecified. Tests valid for sharp null hypotheses can be anti-conservative should only the weak null hold. We develop a general framework for unifying modes of inference for sharp and weak nulls, wherein a single procedure simultaneously delivers exact inference for sharp nulls and asymptotically valid inference for weak nulls. We employ randomization tests based upon prepivoted test statistics, wherein a test statistic is first transformed by a suitably constructed cumulative distribution function and its randomization distribution assuming the sharp null is then enumerated. For a large class of test statistics, we show that prepivoting may be accomplished by employing the push-forward of a sample-based Gaussian measure based upon a suitable covariance estimator. The approach enumerates the randomization distribution (assuming the sharp null) of a p-value for a large-sample test known to be valid under the weak null, and uses the resulting randomization distribution for inference. The versatility of the method is demonstrated.
In finite population causal inference two distinct hypotheses of ‘no treatment effect’ are commonly tested: Fisher’s sharp null and Neyman’s weak null. Fisher’s sharp null of no effect refers to the null that the responses under treatment and under control are the same for all individuals in the study (Rosenbaum, 2002). The sharp null imputes the missing values of the potential outcomes for all individuals, so doing facilitating the use of randomization tests to provide exact inference with randomization alone acting as the basis for inference (Fisher, 1935). Neyman’s weak null instead specifies only that the average of the treatment effects for the individuals in the experiment equals zero, while allowing for heterogeneity in the unit-specific effects. The missing potential outcomes are no longer imputed under the weak null, such that the randomization distribution under the weak null remains unknown. Consequently, inference for the weak null has historically proceeded using asymptotically conservative analytical approximations to the limiting distribution of the treated-minus-control difference in means.

While the exactness attained under the sharp null is appealing, randomization tests have been criticized for the seemingly restricted nature of the conclusions to which the researcher is entitled should the null be rejected (Caughey et al., 2017). While the researcher may suggest that the treatment effect is not zero for all individuals, generally one is not entitled to a statement of whether the treatment effect is positive or negative on average for the individuals in the study. To address this, a recent literature has emerged on how randomization tests may be modified to maintain asymptotic validity under the weak null hypothesis. The resulting methods provide a single testing procedure that is asymptotically valid for the weak null hypothesis, while maintaining exactness should the sharp null also be true (Ding, 2017; Ding & Dasgupta, 2018; Fogarty, 2020; Loh et al., 2017; Wu & Ding, 2018).

The existing literature attains this unified mode of inference largely on a case-by-case basis: for a given experimental design, a specially catered test statistic is constructed so that the corresponding randomization test under the sharp null maintains asymptotic validity under the weak null. In this work, we provide both general conditions under which the unification may be achieved and a general methodology for attaining it. The central idea is to leverage prepivoting, an idea introduced in Beran (1987, 1988). For most commonly employed experimental designs and test statistics, the reference distribution generated by the prepivoted statistic under the assumption of the sharp null asymptotically stochastically dominates the true, but unknowable, randomization distribution under the weak null, yielding asymptotically conservative inference for the weak null while maintaining exactness under the sharp null hypothesis. As we demonstrate, prepivoting succeeds in many scenarios where other common resolutions such as studentization prove inadequate.

At a high level, prepivoting takes a test statistic $T_0$ and composes it with a cumulative distribution function $\hat{F}$ constructed from the observed data, forming the new test statistic $T_1 = \hat{F}(T_0)$.
If \( \hat{F} \) were a consistent estimate of \( T_0 \)'s limit distribution, \( \hat{F}(T_0) \) would, through an asymptotic application of the probability integral transform, tend to a standard uniform. Under the weak null hypothesis, the true distribution function for common test statistics \( T_0 \) cannot generally be consistently estimated. Fortunately, as developed in Section 5, a distribution function for a random variable that asymptotically stochastically dominates \( T_0 \) may be constructed. For most common test statistics for the weak null hypothesis, under conditions outlined in Section 5 this dominating distribution function amounts to a suitable pushforward of a multivariate Gaussian measure constructed using a conservative covariance estimator. Using this estimated distribution function, \( T_1 = \hat{F}(T_0) \) is instead stochastically dominated by a standard uniform in the limit. Observe that through this construction, the prepivoted test \( T_1 = \hat{F}(T_0) \) is precisely one minus the large sample \( p \)-value for the test statistic \( T_0 \) leveraging the central limit theorem. Rather than using this \( p \)-value to reach a conclusion by comparing its value to the desired \( \alpha \), we instead use the reference distribution of this large-sample \( p \)-value enumerated over all possible randomizations assuming the sharp null holds. This reference distribution generally converges pointwise to the standard uniform distribution function for commonly used covariance estimators. As a result, inference is guaranteed to be asymptotically conservative under the weak null while maintaining exactness under the sharp null. The general takeaway is that rather than looking at the randomization distribution of a test statistic itself under the sharp null, one should instead enumerate the randomization distribution of one minus an asymptotically valid \( p \)-value to restore validity of Fisher randomization tests when only the weak null holds.

In Section 2, we introduce notation for finite population causal inference and detail some standard assumptions. Section 3 defines the reference distribution assuming the truth of Fisher’s sharp null and juxtaposes it with its true though unknowable randomization distribution under Neyman’s weak null of no effect on average. After an overview of useful asymptotic results on completely randomized designs in Section 4, Section 5 introduces Gaussian prepivoting in the context of suitably constructed functions of treated-minus control difference in means. Section 6 provides examples of and insight into prepivoting using Gaussian measure. Section 7 extends these results to other asymptotically linear estimators including regression-adjusted estimators, while Section 8 provides simulation studies highlighting the benefits of Gaussian prepivoting.

### 2 | NOTATION AND REVIEW

#### 2.1 | Notation for finite population causal inference

While the developments in this work apply quite generally across common experimental designs and with two or more levels of the treatment, in this work we focus on completely randomized experiments and rerandomized experiments with two treatments; see the web-based supporting materials for extensions to paired designs and to completely randomized designs with multi-valued treatments. Consider a collection of \( N \) individuals, where \( n_1 \) receive treatment and \( n_0 = N - n_1 \) receive the control. For the \( i \)th individual, the random variable \( Z_i \) is the treatment indicator, taking the value 1 if the \( i \)th individual receives treatment and 0 otherwise. We assume that the stable unit treatment value assumption holds, such that there is no interference and that there are no hidden levels of the treatment (Rubin, 1980). The \( i \)th individual has two deterministic potential outcomes: \( y_i(1) \), the \( d \)-dimensional outcome under treatment, and \( y_i(0) \) the \( d \)-dimensional outcome under control. Furthermore, the \( i \)th unit has deterministic
covariates $x_i \in \mathbb{R}^k$. The $j$th coordinate of $y_i(z)$ is $y_{ij}(z)$, and the analogous statement holds for $x_{ij}$. The random vector $Z$ represents $(Z_1, \ldots, Z_N)^T$; likewise $y(1) = (y_1(1), \ldots, y_N(1))^T$ and $y(0) = (y_1(0), \ldots, y_N(0))^T$. Under the finite population model, the potential outcomes are viewed as fixed across randomizations, and the only randomness enters through $Z$, the treatment allocation. For a discussion of the finite population inference framework, we suggest Ding et al. (2017). The observed outcome-vector for individual $i$ is $y_i(Z_i)$ and the collection of these is denoted $y(Z)$. Causal inference with multiple outcomes is becoming increasingly common in modern applications ranging from drug repurposing studies to A/B tests assessing the impact of competing web page designs on various user engagement metrics. See Teixeira-Pinto et al. (2009) and Teixeira-Pinto and Mauri (2011) for concrete examples of causal inference with multiple endpoints in biomedical sciences, and see Ding et al. (2019) for a reference on the underlying mathematics of multivariate potential outcome models.

The vector of treatment effects for the $i$th individual is $\tau_i = y_i(1) - y_i(0)$. The average treatment effect for the individuals in the experiment is $\overline{\tau} = N^{-1} \sum_{i=1}^N \tau_i$. As the two potential outcomes are not jointly observable, $\tau_i$ is unknown for all individuals. Neyman’s weak null of no treatment effect on average is $H_N$: $\overline{\tau} = 0$, while Fisher’s sharp null further stipulates $H_F$: $\tau_i = 0 (i = 1, \ldots, N)$ such that the treatment made no difference among any of the $d$ outcomes measured. We implicitly define the alternative hypothesis as that which complements the null, so for $H_N$ the alternative is $H_A$: $\tau_i \neq 0$ and for $H_F$ the alternative is $H_A$: $\exists i$ s.t. $\tau_i \neq 0$. Consequently, our tests are non-directional; this differs from the one-sided bounded alternatives tested by Caughey et al. (2017). Furthermore, the one-sided bounded alternatives of Caughey et al. (2017) bound each individual’s treatment effect, whereas we are interested in unifying inference for both individual effects and aggregate effects.

For any matrix $r \in \mathbb{R}^{N \times d}$ and any binary vector $W$ with $\sum_{i=1}^N W_i = n_1$, we define the function

$$\widehat{r}(r, W) = \frac{1}{n_1} \sum_{i=1}^N W_i r_i - \frac{1}{n_0} \sum_{i=1}^N (1 - W_i) r_i.$$

Using this notation, the observed treated-minus-control difference in means for the outcome variables is $\widehat{r}(y(Z), Z)$ and is often denoted by $\widehat{r}$ as shorthand. In general, ‘hats’ are used to denote functions of observed quantities whose limiting properties will eventually be studied herein. Define $\overline{y}(0) = N^{-1} \sum_{i=1}^N y_i(0)$ and $\overline{y}(1) = N^{-1} \sum_{i=1}^N y_i(1)$ to be the average potential outcomes for the $N$ individuals in the study population. Likewise, we define the covariance matrices

$$\Sigma_y = (N - 1)^{-1} \sum_{i=1}^N (y_i(z) - \overline{y}(z))(y_i(z) - \overline{y}(z))^T, z \in \{0, 1\};$$

$$\Sigma_r = (N - 1)^{-1} \sum_{i=1}^N (r_i - \overline{\tau})(r_i - \overline{\tau})^T.$$

To emphasize the distinction between functions of observed outcomes and functions of covariates, we define the function $\widehat{\delta}(x, W)$ with binary $W$ such that $\sum_{i=1}^N W_i = n_1$ as

$$\widehat{\delta}(x, W) = \frac{1}{n_1} \sum_{i=1}^N W_i x_i - \frac{1}{n_0} \sum_{i=1}^N (1 - W_i) x_i.$$
The function $\hat{\delta}(x, W)$ is a special case of $\hat{r}(r, W)$. The observed difference in means for covariates is $\hat{\delta}(x, Z)$, abbreviated as $\hat{\delta}$. The finite population mean of the covariates is $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$. The finite population covariance matrix for the covariates is $\Sigma_x$, defined by simply replacing $y_i(z)$ with $x_i$ and $\bar{y}(z)$ with $\bar{x}$ in the definition of $\Sigma_y(z)$. The finite population covariance between potential outcomes and covariates is $\Sigma_{y|x}$ for $z = 0, 1$, and the covariance between treatment effects and covariates is $\Sigma_{x|x}$. Asymptotic arguments that follow will imagine a single sequence of finite populations of increasing size, with $N \to \infty$. As a result, quantities such as $\Sigma_x$ themselves vary as $N \to \infty$ and should be denoted by $\Sigma_{x,N}$ to reflect this. Generally the dependence is suppressed to reduce notational clutter; however, we do employ the notation $\Sigma_{x,\infty}$ to denote the limiting value of $\Sigma_{x,N}$ as $N \to \infty$, and likewise for other finite population quantities. For more on the finite population model for causal inference, see Imbens and Rubin (2015) and Ding et al. (2017) among many.

### 2.2 Rerandomized designs and balance criterion

The set of all possible treatment assignments $Z$ is denoted by $\Omega$, and is determined by the experimental design. In completely randomized experiments, covariates are not used to inform the chosen treatment assignment and $\Omega_{\text{CRE}} = \{z: \sum_{i=1}^{N} z_i = n_1\}$. To mitigate the risk of significant covariate imbalance, Morgan and Rubin (2012) suggest instead building covariate balance into the treatment allocation process through rerandomization. The study is conducted by collecting covariate data for the study participants, determining a measure of imbalance and a threshold for deciding what imbalances are acceptable, and selecting a treatment allocation uniformly over the set of allocations satisfying the balance criterion (Li et al., 2018). Stringent balance criterion reduce the cardinality of $\Omega$ by eliminating undesirable assignments, with the hopes of improving precision as a consequence. Naturally, randomization inference must take into account that the allowable realizations of $Z$ depend upon the condition that covariate balance is met.

A **balance criterion** is a Boolean-valued function $\phi(\cdot)$, where $\phi(\sqrt{N}\hat{\delta}) = 1$ is taken to mean that the treatment allocation $Z$ which results in the particular realization of $\hat{\delta}$ under consideration satisfies appropriate covariate balance. We impose the following restriction on $\phi$:

**Condition 1.** $\phi: \mathbb{R}^k \mapsto \{0, 1\}$ is an indicator function such that the set $M = \{b: \phi(b) = 1\}$ is closed, convex, mirror-symmetric about the origin (i.e. $b \in M \iff -b \in M$) with non-empty interior.

### 2.3 Regularity conditions

We make the following assumptions about the structure of the finite populations and experimental designs as $N$ goes to infinity. These assumptions are for the most part standard in the literature; see, for instance, Wu and Ding (2018).

**Assumption 1.** The proportion $n_1/N$ limits to $p \in (0, 1)$ as $N \to \infty$.

**Assumption 2.** All finite population means and covariances having limiting values for both the potential outcomes and the covariates. For instance, $\lim_{N \to \infty} \bar{y}(z) = \bar{y}_{\infty}(z)$ for $z \in \{0, 1\}$ and $\lim_{N \to \infty} \Sigma_y(1) = \Sigma_y(1, \infty)$. 
Assumption 3. There exists some $C < \infty$ for which, for all $z \in \{0, 1\}$, all $j = 1, \ldots, d$ and all $N$,

$$
\frac{\sum_{i=1}^{N} (y_{ij}(z) - \bar{y}_j(z))^4}{N} < C,
$$

where $\bar{y}_j(z)$ denotes the $j$th coordinate of $y(z)$. Further, the above holds for the covariates with $x_{ij}$ replacing $y_{ij}(z)$ above for $j = 1, \ldots, k$.

Assumption 3 is used to obtain finite population inference strong laws of large numbers for mean and variance estimators. Such an assumption is made at times for mathematical convenience to simplify the analysis of certain random distributions and may hold under weaker assumptions. Assumption 3 is commonplace in the literature on finite population causal inference; see, for instance, Wu and Ding (2018); Lin (2013); Freedman (2008a, b).

2.4 | A technical note on the convergence of random measures

A random sequence of probability measures $\hat{\mu}_N$ on $S$ converges weakly in probability to a deterministic probability measure $\mu$ if $\int_S f d\hat{\mu}_n \to \int_S f d\mu$ for all continuous bounded functions $f: S \to \mathbb{R}$ (Dümbgen & Del Conte-Zerial, 2013, Section 2). Aspects of the Portmanteau Theorem (van der Vaart & Wellner, 1996, Theorem 1.3.4) extend to weak convergence in probability of random measures (Crauel, 2002; Dümbgen & Del Conte-Zerial, 2013). Most importantly for our purposes is that if $\{\hat{F}_N\}$ are random cumulative distribution functions and $F$ is a fixed cumulative distribution function, then their associated measures converge weakly in probability if and only if $\hat{F}_N(t) \xrightarrow{p} F(t)$ for all $t$ which are continuity points of $F$; we take this as the definition of weak convergence in probability for random cumulative distribution functions.

3 | RANDOMIZATION DISTRIBUTIONS AND TESTS

3.1 | Randomization distributions

Consider a scalar test statistic $T(y(Z), Z)$, a function of the observed responses and the treatment assignment received. The randomization distribution for the test statistic $T$ is

$$
\mathcal{R}_T(t) = \frac{1}{|\Omega|} \sum_{w \in \Omega} 1\{ T(y(w), w) \leq t \}. \tag{1}
$$

$\mathcal{R}_T$ is the true cumulative distribution function of $T(y(Z), Z)$ with respect to the randomness in the treatment allocation $Z$, distributed uniformly over $\Omega$. If we had access to $\mathcal{R}_T$ under the null hypothesis in question, we could make direct use of it to provide inference that is exact in finite samples, proceeding without dependence on asymptotics. Under Fisher’s sharp null hypothesis, $\mathcal{R}_T$ is specified by the observed outcomes as $y(Z) = y(w)$ for any $w \in \Omega$. Unfortunately, the distribution is generally unknown under the weak null, as the weak null merely constrains the missing potential outcomes without determining them.
3.2 Randomization tests assuming the sharp null

In practice, an experimenter draws a single realization of \( Z \), in so doing only revealing the values of the potential outcomes corresponding to the observed assignment. Suppose that regardless of whether or not Fisher’s sharp null hypothesis actually holds, the researcher considers use of the randomization distribution to which she or he would be entitled if the sharp null were true. This reference distribution takes the form

\[
\hat{\mathcal{P}}_T(t) = \frac{1}{|\Omega|} \sum_{w \in \Omega} 1\{T(y(Z), w) \leq t\}. \tag{2}
\]

While \( \mathcal{R}_T = \hat{\mathcal{P}}_T \) under the sharp null, under the weak null \( \hat{\mathcal{P}}_T \) is a random distribution function as it varies with \( Z \). Inference using \( \hat{\mathcal{P}}_T \) proceeds as though \( y(Z) \) would have been the observed response for any \( w \in \Omega \). As the true response \( y(w) \) under assignment \( w \) need not align with \( y(Z) \), \( \hat{\mathcal{P}}_T \) does not actually reflect the true randomization distribution under the weak null. This gives rise to potentially anti-conservative inference should \( \hat{\mathcal{P}}_T \) be used to test the weak null hypothesis.

For \( \alpha \in (0, 1) \) define the Fisher randomization test of nominal level \( \alpha \) by

\[
\varphi_T(\alpha) = 1 \left\{ T(y(Z), Z) \geq \hat{\mathcal{P}}_T^{-1}(1-\alpha) \right\}. \tag{3}
\]

Under the sharp null, \( \mathbb{E}\{\varphi_T(\alpha)\} \leq \alpha \) for any sample size as \( \hat{\mathcal{P}}_T = \mathcal{R}_T \). Throughout this paper, we examine the extent to which certain choices of test statistics entitle us to asymptotic Type I error control at \( \alpha \) when \( \varphi_T(\alpha) \) is used to conduct inference but only the weak null holds. For a given test statistic \( T \), we will often proceed by juxtaposing its true limiting behaviour under the randomization distribution \( \mathcal{R}_T \) with the limiting behaviour of \( \hat{\mathcal{P}}_T \), the randomization distribution if we (incorrectly) assumed that the sharp null held.

3.3 Towards a unified mode of inference

Suppose that for a test statistic \( T(y(Z), Z) \) based upon the observed outcomes \( y(Z) \) and the treatment allocation \( Z \),

1. \( \hat{\mathcal{P}}_T \) converges weakly in probability to a fixed distribution \( \mathcal{P}_{T,\infty} \) as \( N \to \infty \); and
2. \( \mathcal{R}_T \) converges pointwise to a fixed distribution \( \mathcal{R}_{T,\infty} \) at all continuity points of \( \mathcal{R}_{T,\infty} \). Formally, \( \mathcal{R}_T(t) \to \mathcal{R}_{T,\infty}(t) \ \forall t \in \text{cont}(\mathcal{R}_{T,\infty}) \), where \( \text{cont}(\mathcal{R}_{T,\infty}) \) is the set of continuity points of \( \mathcal{R}_{T,\infty} \).

The test statistic \( T(y(Z), Z) \) is called asymptotically sharp-dominant if, under \( H_N \), \( \mathcal{P}_{T,\infty}(t) \leq \mathcal{R}_{T,\infty}(t) \) for any scalar \( t \). This implies that the \( (1 - \alpha) \) quantile of \( \mathcal{P}_{T,\infty} \) is at or above the \( (1 - \alpha) \) quantile of \( \mathcal{R}_{T,\infty} \). If \( T(y(Z), Z) \) is asymptotically sharp-dominant, then inference based upon the reference distribution \( \hat{\mathcal{P}}_T \) will be asymptotically conservative even if only \( H_N \) holds (Wu & Ding, 2018, Proposition 4), satisfying \( \lim \sup \mathbb{E}\{\varphi_T(\alpha)\} \leq \alpha \) as \( N \to \infty \) while maintaining exactness should the sharp null be true.
Many common test statistics are not asymptotically sharp-dominant over all elements of the weak null. For instance, with univariate potential outcomes and under a completely randomized design with imbalanced treated and control groups, the absolute difference in means \( T(y(Z), Z) = \sqrt{N} |\bar{y}(Z)/\bar{Z} - \bar{y}(Z)/\bar{Z}| \) is not generally asymptotically sharp-dominant as there exist sequences of potential outcomes satisfying the weak null such that \( \lim\inf \mathbb{E}(\varphi_T(\alpha)) > \alpha \); see Ding (2017), Wu and Ding (2018 Cor. 3), or Loh et al. (2017) for details. For this test statistic, simply studentizing by the usual standard error estimator ensures sharp dominance. However, studentization fails to generalize to other more complicated test statistics and complex experimental designs. Significant efforts have recovered appropriate studentization techniques for some test statistics, but each test statistic requires its own separate analysis (Wu & Ding, 2018). For some experimental designs, studentizing the difference in means is not sufficient to regain asymptotically valid inference even in the univariate case; we explore this topic in Section 5.2 and 8.1 in the context of rerandomization. In Section 5, we present a general method called Gaussian prepivoting which both recovers studentization when it alone would be sufficient, but also yields asymptotic sharp-dominance in circumstances where studentization would be insufficient. Before describing the method, we recall a few important results on the difference in means in completely randomized designs which underpin the success of Gaussian prepivoting.

4 USEFUL RESULTS FOR THE DIFFERENCE-IN-MEANS IN COMPLETELY RANDOMIZED DESIGNS

4.1 Asymptotic normality and conservative covariance estimation for the randomization distribution

Consider the distribution of \( \sqrt{N}(\hat{\tau} - \bar{\tau}, \hat{\delta})^T \) in a completely randomized design. Under Assumptions 1–3, a finite population central limit theorem applies (Li & Ding, 2017), and \( \sqrt{N}(\hat{\tau} - \bar{\tau}, \hat{\delta})^T \) converges in distribution to a mean-zero multivariate Gaussian with covariance matrix \( V \) of the form

\[
V = \begin{pmatrix}
V_{\tau\tau} & V_{\tau\delta} \\
V_{\delta\tau} & V_{\delta\delta}
\end{pmatrix};
\]

\[
V_{\tau\tau} = p^{-1} \Sigma_{y(1),\infty} + (1 - p)^{-1} \Sigma_{y(0),\infty} - \Sigma_{\tau,\infty};
\]

\[
V_{\delta\delta} = (p(1 - p))^{-1} \Sigma_{x,\infty};
\]

\[
V_{\tau\delta} = p^{-1} \Sigma_{y(1)x,\infty} + (1 - p)^{-1} \Sigma_{y(0)x,\infty} = V_{\delta\tau}^T.
\]

While \( V_{\delta\delta} \) and \( V_{\tau\delta} \) can be consistently estimated, \( V_{\tau\tau} \) cannot be in the presence of effect heterogeneity due to its dependence on \( \Sigma_{\tau} \), the covariance of the unobserved treatment effects. Consequently, one cannot consistently estimate the probability that \( \sqrt{N}(\hat{\tau} - \bar{\tau}) \) falls within a given region \( B \). While consistent variance estimates are not available, there are several covariance estimators \( \hat{V}_{\tau\tau}(y(Z), Z) \) for \( V_{\tau\tau} \) satisfying \( \hat{V}_{\tau\tau} \rightarrow V_{\tau\tau} \) for some \( \Delta \geq 0 \) under Assumptions 1–3 in completely randomized designs. These estimators typically have the property that \( \Sigma_{\tau\tau} = 0 \) implies consistency, rather than asymptotic conservativeness; see Ding et al. (2019) for more details. So while the matrix \( V \) cannot generally be consistently estimated, one can construct an estimate converging in probability to a matrix \( \hat{V} \) of the form
\[
\bar{V} = \begin{pmatrix}
V_{rr} + \Delta & V_{r\delta} \\
V_{\delta r} & V_{\delta\delta}
\end{pmatrix}
\]

with \( \Delta \geq 0 \).

As an illustration, consider the conventional covariance estimator for the difference in means in a two-sample problem, \( \hat{V}_{rr} = N(\hat{\Sigma}_{y(1)}/n_1 + \hat{\Sigma}_{y(0)}/n_0) \) with

\[
\hat{\Sigma}_{y(1)} = \frac{1}{n_1 - 1} \sum_{i=1}^{N} Z_i \left( y_i(1) - \frac{1}{n_1} \sum_{i=1}^{N} Z_i y_i(1) \right) \left( y_i(1) - \frac{1}{n_1} \sum_{i=1}^{N} Z_i y_i(1) \right)^T
\]

and the analogous for \( \hat{\Sigma}_{y(0)} \). Under both completely randomized experiments and rerandomized experiments with balance criterion satisfying Condition 1, this estimator satisfies \( \hat{V}_{rr} - V_{rr} \xrightarrow{p} \Sigma_{r,\infty} \geq 0 \) under Assumptions 1–3.

### 4.2 Limiting behaviour of the reference distribution

Suppose we have a completely randomized design, and consider the random variable

\[
\{ \sqrt{N} \hat{\tau}(\hat{y}(Z), W), \sqrt{N} \hat{\delta}(x, W) \}^T
\]

\[
= \sqrt{N} \left\{ \frac{1}{n_1} \sum_{i=1}^{N} W_i \hat{y}_i(Z_i) - \frac{1}{n_0} \sum_{i=1}^{N} (1-W_i) \hat{y}_i(Z_i), \frac{1}{n_1} \sum_{i=1}^{N} W_i x_i - \frac{1}{n_0} \sum_{i=1}^{N} (1-W_i) x_i \right\}^T
\]

where \( Z \) and \( W \) are independent, identically distributed and drawn uniformly from \( \Omega \) and \( \hat{y}_i(Z_i) = y_i(Z_i) - Z_i \bar{r} \), such that \( \hat{y}(Z) = y(Z) - Z \bar{r}^T \).

**Proposition 1.** Subject to Assumptions 1–3, under a completely randomized design the distribution of \( \{ \sqrt{N} \hat{\tau}(\hat{y}(Z), W), \sqrt{N} \hat{\delta}(x, W) \}^T \mid Z \) converges weakly in probability to a multivariate Gaussian measure, with mean zero and covariance \( \bar{V} \) of the form

\[
\bar{V} = \begin{pmatrix}
\bar{V}_{rr} & \bar{V}_{r\delta} \\
\bar{V}_{\delta r} & \bar{V}_{\delta\delta}
\end{pmatrix}.
\]

\[
\bar{V}_{rr} = (1 - p)^{-1} \Sigma_{y(1),\infty} + p^{-1} \Sigma_{y(0),\infty};
\]

\[
\bar{V}_{\delta\delta} = (p(1-p))^{-1} \Sigma_{x,\infty};
\]

\[
\bar{V}_{r\delta} = (1 - p)^{-1} \Sigma_{y(1)x,\infty} + p^{-1} \Sigma_{y(0)x,\infty} = \bar{V}_{\delta r}^T.
\]

The proof of this statement is contained within the proof of Theorem 1 in Wu and Ding (2018) and is omitted. Under the sharp null, \( \bar{V} = V \) as \( \hat{y}_i(1) = y_i(0) \) for all \( i \). Under the weak null however, while \( \bar{V}_{\delta\delta} = V_{\delta\delta} \) generally \( \bar{V}_{rr} \neq V_{rr} \) and \( \bar{V}_{\delta r} \neq V_{\delta r} \). The divergence between \( V \) and \( \bar{V} \) can render randomization tests for the weak null hypothesis anti-conservative; examples are given in Section 5.2. We now describe how pre pivoting may be used to guarantee asymptotic correctness when inference for the weak null hypothesis is conducted using a reference distribution generated under the sharp null.
5 | GAUSSIAN PREPIVOTING

5.1 | Prepivoting with an estimated pushforward measure

Consider functions \( f_\eta: \mathbb{R}^d \rightarrow \mathbb{R} \) subject to the following requirement:

**Condition 2.** For any \( \eta \in \Xi, f_\eta(\cdot): \mathbb{R}^d \mapsto \mathbb{R}_+ \) is continuous, quasi-convex and nonnegative with \( f_\eta(t) = f_\eta(-t) \) for all \( t \in \mathbb{R}^d \). Furthermore, \( f_\eta(t) \) is jointly continuous in \( \eta \) and \( t \).

We begin with statistics for \( H_N \) of the form

\[
T(y(Z), Z) = f_\xi(\sqrt{N}\hat{t}),
\]

where \( \hat{\xi} = \hat{\xi}(y(Z), Z) \) satisfies the following condition for some set \( \Xi \):

**Condition 3.** With \( W, Z \) independent and each uniformly distributed over \( \Omega \),

\[
\hat{\xi}(y(Z), Z) \overset{p}{\rightarrow} \xi; \quad \hat{\xi}(y(Z), W) \overset{p}{\rightarrow} \xi,
\]

for some \( \xi, \hat{\xi} \in \Xi \).

As will be shown in Section 5.2, several commonly encountered statistics for Neyman’s null are of this form. A detailed discussion of Condition 2 is included in the web-based supporting material. Suppose further that one employs a covariance estimator \( \hat{V}(y(Z), Z) \) with the following property:

**Condition 4.** With \( W, Z \) independent, both uniformly distributed over \( \Omega \), and for some \( \Delta \geq 0, \Delta \in \mathbb{R}^{d \times d} \),

\[
\hat{V}(y(Z), Z) - V \overset{p}{\rightarrow} \left( \Delta - O_{d,k} \Delta k \right); \quad \hat{V}(y(Z), W) - \hat{V} \overset{p}{\rightarrow} O_{(d+k),(d+k)}.
\]

As a concrete example satisfying Conditions 2–4, suppose that \( f_\eta(t) = t^T\eta^{-1}t^T \) and \( \hat{\xi}(y(Z), W) = \hat{V}_{\text{Neyman}}(y(Z), W) \) with \( \hat{V}_{\text{Neyman}} \) denoting the usual Neyman variance estimator of Neyman (1990); numerous other examples are included in Section 5.2. Observe that when assuming the weak null for the purpose of testing \( y(Z) = y(Z) \) and \( \hat{\tau} - \bar{\tau} = \hat{\tau} \), Gaussian prepivoting transforms the test statistic \( T(y(Z), W) = f_\xi(\sqrt{N}\hat{t}(y(Z), W)) \) into a new statistic of the form

\[
G(y(Z), W) = \frac{\gamma^{(d+k)}_{0,\hat{V}(y(Z), W)} \{ (a, b)^T : f_\xi(a) \leq T(y(Z), W) \land \phi(b) = 1 \}}{\gamma^{(k)}_{0,\hat{V}_{\delta}} \{ b : \phi(b) = 1 \}}
\]

where \( \gamma^{(p)}_{\mu,\Sigma}(B) \) is the \( p \)-dimensional Gaussian measure of a set \( B \) with mean parameter \( \mu \) and covariance \( \Sigma \), i.e.

\[
\gamma^{(p)}_{\mu,\Sigma}(B) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \int_{x \in B} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \, dx.
\]
For \((\mathbf{A}, \mathbf{B})^T\) jointly multivariate normal with mean zero and covariance \(\tilde{V}\), \(\mathbf{A} \in \mathbb{R}^d\), \(\mathbf{B} \in \mathbb{R}^k\), \(G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) represents the \(f^*_\xi\)-pushforward measure of \(\mathbf{A}|\phi(\mathbf{B})=1\) evaluated on the set \((-\infty, T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}))\). That is, \(G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) treats \(f^*_\xi\) and \(\tilde{V}\) as fixed and computes the conditional probability that \(f^*_\xi(\mathbf{A})\) falls at or below the observed value for \(T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})=f^*_\xi(\sqrt{N}\tilde{r})\) given that \(\phi(\mathbf{B})=1\). From the perspective of hypothesis testing, \(G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) is 1 minus the large-sample \(p\)-value for \(T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) leveraging the finite population central limit theorem and the estimated covariance \(\tilde{V}\).

We now describe how to use the prepivoted statistic \(G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) to provide a single procedure that is both exact for \(H_F\) and asymptotically conservative for \(H_N\). In order to provide a precise implementation of this, we give detailed pseudocode in Algorithm 1 and provide example code through the web-based supporting material. First, we compute the prepivoted test statistic \(G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\) given the observed data; this proceeds according to Equation (5) and is Step 1 of Algorithm 1. Next, we construct the reference distribution \(\hat{\mathcal{P}}_{G}(\cdot)\), the construction of which requires imputing counterfactual outcomes as if Fisher’s sharp null held; this is Step 2 of Algorithm 1. Finally, the \(p\)-value for testing \(H_N\) is computed and we reject the null when this lies below or at the nominal level \(\alpha\).

---

**Algorithm 1: Inference for the weak null through Gaussian prepivoting**

**Input:** An observed treatment allocation \(\mathbf{z}\), with observed responses \(\mathbf{y}(\mathbf{z})\), test statistic \(T(\mathbf{y}(\mathbf{z}), \mathbf{z}) = f^*_\xi(\sqrt{N}\tilde{r}_{obs})\) and covariance estimator \(\hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{z})\)

**Result:** The \(p\)-value for the Gaussian prepivoted test statistic

**Step 1: The observed prepivoted statistic**

Compute \(f^*_\xi(\mathbf{y}(\mathbf{z}), \mathbf{z})(\cdot); \hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{z})\). Compute

\[
g_{\mathbf{z}} = \frac{N^{(d+k)}}{\gamma_{0, \hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{z})}} \left\{ \langle \mathbf{a}, \mathbf{b} \rangle^T : f^*_\xi(\mathbf{y}(\mathbf{z}), \mathbf{z})(\mathbf{a}) \leq T(\mathbf{y}(\mathbf{z}), \mathbf{z}) \land \phi(\mathbf{b}) = 1 \right\}
\]

**Step 2: The reference distribution \(\hat{\mathcal{P}}_{G}\)**

for \(w \in \Omega\) do

\[
g_{\mathbf{w}} = \frac{N^{(d+k)}}{\gamma_{0, \hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{w})}} \left\{ \langle \mathbf{a}, \mathbf{b} \rangle^T : f^*_\xi(\mathbf{y}(\mathbf{z}), \mathbf{w})(\mathbf{a}) \leq T(\mathbf{y}(\mathbf{z}), \mathbf{w}) \land \phi(\mathbf{b}) = 1 \right\}
\]

end

return

\[
p_{val} = \frac{1}{|\Omega|} \sum_{\mathbf{w} \in \Omega} \mathbb{I}(g_{\mathbf{w}} \geq g_{\mathbf{z}});
\]

\[
\varphi_G(\alpha) = \mathbb{I}(p_{val} \leq \alpha).
\]
Observe that $1 - g_z$ defined within Algorithm 1 is the usual large-sample $p$-value based upon a Gaussian approximation and using the covariance estimator $\hat{\nu}$. The large-sample test compares $1 - g_z$ to $a$, the desired Type I error rate, and rejects if $1 - g_z \leq a \Leftrightarrow g_z \geq 1 - a$. The Gaussian prepivoted randomization test instead rejects if $g_z \geq \hat{P}_G^{-1}(1 - a)$. The following Theorem, in concert with Lemma 11.2.1 of Lehmann and Romano (2005), show under our assumptions $\hat{P}_G^{-1}(1 - a) \overset{p}{\rightarrow} 1 - a$, such that the prepivoted randomization test is asymptotically equivalent to large sample test under the weak null. By using $\hat{P}_G^{-1}(1 - a)$ instead of $1 - a$, exactness under the sharp null is preserved.

**Theorem 1.** Suppose we have either a completely randomized design or a rerandomized design with balance criterion $\phi$ satisfying Condition 1. Suppose $T(y(Z), Z)$ is of the form (4) for some $f_z$ and $\hat{\xi}$ satisfying Conditions 2 and 3. Suppose further that we employ a covariance estimator $\hat{\nu}$ satisfying Condition 4 when forming the prepivoted test statistic $G(y(Z), Z)$. Then, under $H_{N}: \bar{\tau} = 0$ and under Assumptions 1–3, $G(y(Z), Z)$ converges in distribution to a random variable $\hat{U}$ taking values in $[0,1]$ satisfying

$$\mathbb{P}(\hat{U} \leq t) \geq t,$$

for all $t \in [0,1]$. Furthermore, the distribution $\hat{P}_G(t)$ satisfies $\hat{P}_G(t) \overset{p}{\rightarrow} t$ for all $t \in [0,1]$.

**Corollary 1.** Under the conditions of Theorem 1; the prepivoted test statistic $G(y(Z), Z)$ is asymptotically sharp dominant regardless of whether the base statistic $T(y(Z), Z)$ was. Consequently, $p$-values derived under $\hat{P}_G$ via Algorithm 1 are guaranteed to be exact under $H_F$ and asymptotically conservative under just $H_N$.

Theorem 1 states that under the weak null, $G(y(Z), Z)$ converges in distribution to a random variable which is stochastically dominated by the standard uniform. Meanwhile, the reference distribution for $G(y(Z), Z)$ constructed assuming (incorrectly) that the sharp null holds converges pointwise to the distribution function of a standard uniform. As a result, the randomization distribution for $G(y(Z), Z)$ is asymptotically sharp-dominant: the reference distribution generated in this manner yields asymptotically conservative inference for the weak null hypothesis, while maintaining exactness should the sharp null also hold. By exploiting the duality between hypothesis testing and confidence sets Theorem 1 provides the basis for generating exact and asymptotically conservative confidence sets for treatment effect; this is explored in the web-based supporting material.

**Remark 1.** Consider the function

$$\hat{F}(t) = \frac{\gamma_{0,\hat{\nu}}^{(d+k)} \left\{ (a, b)^T : f_z(a) \leq t \land \phi(b) = 1 \right\}}{\gamma_{0,\hat{\nu}_{65}}^{(k)} \left\{ b : \phi(b) = 1 \right\}}$$

the estimated distribution function for $f_z(\sqrt{N}\hat{\tau})\phi(\sqrt{N}\hat{\delta}) = 1$ based upon a finite population central limit theorem. In special cases, the function $\hat{F}(t)$ may have a known closed form. This is true of the test statistics which are sharp-dominated by a $\chi^2$ distribution considered in Wu and Ding (2018), for example. Should this not be the case, one can approximate $\hat{F}(\cdot)$ by way of Monte-Carlo approximation, replacing the measures $\gamma_{0,\hat{\nu}}$ and $\gamma_{0,\hat{\nu}_{65}}$ with estimates based upon $B$ draws.
from a multivariate normal with mean 0 and covariance \( \hat{V} \) when enumerating the reference distribution. Importantly, such Monte-Carlo approximation does not corrupt finite-sample exactness under Fisher’s sharp null.

### 5.2 Examples of Gaussian prepivoting

Through a series of examples, we now provide illustrations of the transformations achieved by (5). As will be demonstrated, the form recovers several randomization tests previously known to be valid for weak null hypotheses in the literature while providing a basis for new randomization tests for weak nulls using other test statistics. These examples serve four objectives: (i) unify previous ad hoc solutions under the framework of Gaussian prepivoting; (ii) provide an alternative approach to already valid procedures; (iii) highlight that prepivoting can succeed even where studentization fails; and (iv) extend randomization inference for \( H_F \) and \( H_N \) to new experimental designs.

**Example 1.** (Absolute difference in means). Let \( \sqrt{N}\hat{\tau} \) be univariate, consider a completely randomized design with no rerandomization, and let \( T_{D_{IM}}(y(Z), Z) = \sqrt{N}|\hat{\tau}| \), with \( f_\eta(t) = |t| \) and \( \hat{\xi} = 1 \). The randomization distribution for \( T_{D_{IM}}(y(Z), Z) \) is not asymptotically sharp-dominant, such that employing the reference distribution assuming that the sharp null holds may lead to anti-conservative inference. The conventional fix is to studentize \( \sqrt{N}|\hat{\tau}| \) using a variance estimator satisfying Condition 1, forming instead \( T_{Su}(y(Z), Z) = \sqrt{N}|\hat{\tau}|/\sqrt{\hat{V}_{rr}} \) (Loh et al., 2017).

As \( \phi(\cdot) = 1 \) deterministically in a completely randomized design, Gaussian prepivoting via (5) yields the test statistic

\[
G_{D_{IM}}(y(Z), Z) = \gamma_{0,\hat{V}_{rr}}^{(1)} \{ \alpha: |\alpha| \leq \sqrt{N}|\hat{\tau}| \} = 1 - 2\Phi \left( -\frac{\sqrt{N}|\hat{\tau}|}{\sqrt{\hat{V}_{rr}}} \right),
\]

where \( \Phi(\cdot) \) is the standard normal distribution function. For any \( Z \), the pairs \( \{ G_{D_{IM}}(y(Z), w), T_{D_{IM}}(y(Z), w) \} \) have rank correlation equal to 1 when computed for all \( w \in \Omega \). As a result, the reference distribution using the studentized difference in means assuming the sharp null holds may lead to anti-conservative inference. The conventional fix is to studentize \( \sqrt{N}|\hat{\tau}| \) using a variance estimator satisfying Condition 1, forming instead \( T_{Su}(y(Z), Z) = \sqrt{N}|\hat{\tau}|/\sqrt{\hat{V}_{rr}} \) (Loh et al., 2017).

**Example 2.** (Multivariate studentization). Let \( \sqrt{N}\hat{\tau} \) now be multivariate and suppose we have a completely randomized design. Wu and Ding (2018) suggest the test statistic

\[
T_{\chi^2}(y(Z), Z) = \left( \sqrt{N}\hat{\tau} \right)^T \hat{V}_{rr}^{-1} \left( \sqrt{N}\hat{\tau} \right),
\]

with \( \hat{V}_{rr} = \frac{N}{n_1} \hat{\Sigma}_{1(1)} + \frac{N}{n_0} \hat{\Sigma}_{0(0)} \). For this test statistic, \( f_\eta(t) = t^T \eta^{-1} t \) and \( \hat{\xi} = \hat{V}_{rr} \). Wu and Ding (2018) show that under our assumptions, under the weak null this test statistic converges in distribution
to \( \sum_{i=1}^{d} w_i \xi_i^2 \) where \( w_i \in [0, 1] \) are weights and \( \xi_1, \ldots, \xi_d \overset{iid}{\sim} \mathcal{N}(0,1) \) while the reference distribution of \( T_{\chi^2} (y(Z), Z) \) attained assuming that the sharp null holds converges weakly in probability to the \( \chi_d^2 \) distribution. As a result, \( T_{\chi^2} (y(Z), Z) \) is asymptotically sharp-dominant, and its reference distribution assuming the sharp null may be used for inference for the weak null hypothesis. Here, Gaussian prepivoting produces

\[
G_{\chi^2} (y(Z), Z) = \gamma^{(d)}_{0, \hat{\nu}_{\tau\tau}} \left\{ a : a^T \hat{\nu}_{\tau\tau}^{-1} a \leq T_{\chi^2} (y(Z), Z) \right\} = F_d \left\{ T_{\chi^2} (y(Z), Z) \right\},
\]

where \( F_d (\cdot) \) is the distribution function of a \( \chi_d^2 \) random variable. For any \( Z \), the pairs \( \{ G_{\chi^2} (y(Z), w), T_{\chi^2} (y(Z), w) \} \) have rank correlation equal to 1 when computed for all \( w \in \Omega \), such that Gaussian prepivoting yields equivalent inference to that attained using the distribution of \( T_{\chi^2} (y(Z), Z) \) under the sharp null. This demonstrates objective (ii).

Suppose instead that, erroneously, a practitioner proceeded with the more typical form of Hotelling’s \( T^2 \) statistic employing a pooled covariance estimator,

\[
T_{\text{Pool}} (y(Z), Z) = \left( \sqrt{N} \hat{\tau} \right)^T \left( \hat{\nu}_{\text{Pool}} \right)^{-1} \left( \sqrt{N} \hat{\tau} \right); \\
\hat{\nu}_{\text{Pool}} = \left( \frac{n}{n_0} + \frac{N}{n_1} \right) \left( \frac{(n_1 - 1) \hat{\Sigma}_{y(1)} + (n_0 - 1) \hat{\Sigma}_{y(0)}}{n_1 + n_0 - 2} \right).
\]

For this test statistic, \( f_\eta (t) = t^T \eta^{-1} t \) as before, but \( \hat{\xi} = \hat{\nu}_{\text{Pool}}. \) In this case, \( T_{\text{Pool}} (y(Z), Z) \) is not asymptotically sharp-dominant, such that the reference distribution using this statistic and assuming the sharp null may yield invalid inference. Gaussian prepivoting returns the test statistic

\[
G_{\text{Pool}} (y(Z), Z) = \gamma^{(d)}_{0, \hat{\nu}_{\tau\tau}} \left\{ a : a^T \hat{\nu}_{\text{Pool}}^{-1} a \leq T_{\text{Pool}} (y(Z), Z) \right\}.
\]

Importantly, \( G_{\text{Pool}} (y(Z), Z) \) continues to use the Gaussian measure computed with the covariance matrix \( \hat{\nu}_{\tau\tau} \) in forming the suitable transformation, despite the fact that the pooled covariance matrix is used in forming \( T_{\text{Pool}} (y(Z), Z) \). For fixed \( Z, G_{\text{Pool}} (y(Z), w) \) generally will not have perfect rank correlation with \( T_{\text{Pool}} (y(Z), w) \) when computed over \( w \in \Omega \), such that the two randomization tests assuming the sharp null no longer furnish identical \( p \)-values. This divergence is necessary: while \( T_{\text{Pool}} (y(Z), Z) \) is not asymptotically sharp-dominant, Theorem 1 asserts that \( G_{\text{Pool}} (y(Z), Z) \) is, such that the reference distribution for \( G_{\text{Pool}} (y(Z), Z) \) assuming the sharp null yields asymptotically conservative inference for the weak null. Gaussian prepivoting can thus restore asymptotic validity to a test statistic employing improper studentization, illustrating objective (iii).

Example 3. (Max absolute \( t \)-statistic). Consider again multivariate \( \sqrt{N} \hat{\tau} \) and a completely randomized design, and consider the test statistic

\[
T_{|\text{max}|} (y(Z), Z) = \max_{1 \leq j \leq d} \left| \frac{\sqrt{N} |\hat{\tau}_j|}{\sqrt{\hat{\nu}_{\tau\tau, jj}} \hat{\xi}} \right|
\]

where \( \hat{\nu}_{\tau\tau, jj} \) is the \( jj \) element of \( \hat{\nu}_{\tau\tau} \). For this statistic, \( f_\eta (t) = \max_{1 \leq j \leq d} |t_j| / \eta_j \), and \( \hat{\xi} = (\hat{\nu}_{\tau\tau, 11}, \ldots, \hat{\nu}_{\tau\tau, dd})^T \). For \( d \geq 2 \), \( T_{|\text{max}|} (y(Z), Z) \) is not asymptotically sharp-dominant
under the weak null: the reference distribution generated under the sharp null depends upon the correlation matrix corresponding to \( \bar{V} \), while the true randomization distribution is governed by the correlations encoded within \( V \). The Gaussian prepivoted correction takes the form

\[
G_{\text{max}}(y(Z), Z) = \gamma^{(d)}_{\theta, \bar{V}_{\tau \tau}} \left\{ a: \max_{1 \leq j \leq d} \frac{|a_j|}{\sqrt{\bar{V}_{\tau \tau,j}}} \leq \max_{1 \leq j \leq d} \frac{\sqrt{N}|\hat{t}_j|}{\sqrt{\bar{V}_{\tau \tau,j}}} \right\},
\]

which composes \( T_{\text{max}}(y(Z), Z) \) with the distribution function for \( \max |A_j|/\sqrt{\bar{V}_{\tau \tau,j}}, j = 1, ..., d \), when \( A \) is multivariate Gaussian with mean zero and covariance \( \bar{V}_{\tau \tau} \). Gaussian prepivoting rectifies the insufficiency of the studentization in \( T_{\text{max}} \), thereby providing an example of objective (iii).

**Example 4.** (Rerandomization). Let \( \sqrt{N}\hat{t} \) be univariate and suppose we now consider a rerandomized design with balance criterion \( \phi \) satisfying Condition 1. Consider the absolute difference in means, \( f_{\xi}(\sqrt{N}\hat{t}) = \sqrt{N}|\hat{t}| \), such that \( \hat{\xi} = 1 \). Gaussian prepivoting yields the test statistic

\[
G_{\text{Re}}(y(Z), Z) = \gamma^{(1+k)}_{\theta, \bar{V}} \left\{ (a, b)^T: |a| \leq \sqrt{N}|\hat{t}| \land \phi(b) = 1 \right\}
\]

\[
\gamma^{(k)}_{\theta, \bar{V}_{\delta \delta}} \left\{ b: \phi(b) = 1 \right\}
\]

For completely randomized designs with \( \phi(\cdot) = 1 \) deterministically, Gaussian prepivoting is equivalent to studentizing as described in Example 1. In general rerandomized designs however, observe that the transformation depends upon the particular form of the balance criterion \( \phi \), and that the reference distribution will depend upon the relationship between the potential outcomes and the covariates used in the balance criterion. As a result, it will generally not be the case that the reference distribution of \( G_{\text{Re}}(y(Z), Z) \) under the sharp null yields equivalent inference to that attained using \( \sqrt{N}|\hat{t}|/\sqrt{\bar{V}_{\tau \tau}} \). This suggests that in rerandomized designs, studentization alone is insufficient for attaining an asymptotically sharp-dominant test statistic. In Section 8.1, we show this through an example in the case of Mahalanobis rerandomization. Lemmas A15 and A16 of Li et al. (2018) show that under our conditions, probability limits for estimators \( \bar{V} \) derived under complete randomization are generally preserved under rerandomized designs. Once again, Theorem 1 ensures that \( G_{\text{Re}}(y(Z), Z) \) will be asymptotically sharp-dominant, such that the randomization distribution assuming the sharp null may be employed for inference for the weak null. The development of a finite sample exact method for testing \( H_P \) which is asymptotically valid for testing \( H_N \) in rerandomized designs is novel, but its construction is extremely simple within the framework of Gaussian prepivoting; this highlights Gaussian prepivoting’s portability to designs outside of just completely randomized experiments. In the web-based supporting material we provide two more examples of this portability: one for matched-pair designs and one for experiments with any finite number of treatment arms. This highlights objective (iv).

For the interested reader, in the web-based supporting material we include this same collection of examples written directly in the form of Gaussian integrals, and we include verification of Conditions 2–4.
6 | GAUSSIAN COMPARISON, STOCHASTIC DOMINANCE AND THE PROBABILITY INTEGRAL TRANSFORM

6.1 | Gaussian comparison and Anderson’s Theorem

We now highlight the essential technical ingredients underpinning the success of Gaussian prepivoting. Consider two mean-zero multivariate Gaussian vectors \((A_1, B_1)^T\) and \((A_2, B_2)^T\), with covariances \( \Lambda(2) - \Lambda(1) \geq 0 \) and \( \Lambda_{bb} > 0 \); the inequalities are stated with respect to the Loewner partial order on positive semidefinite matrices. Let the dimensions of \(A_j\) and \(B_j\) be \(d\) and \(k\), respectively, for \(j = 1, 2\). Compare the tail probabilities for

\[
\begin{align*}
\mathbb{P}(\phi(B_1) = 1) &= f(A_1) \geq v, \\
\mathbb{P}(\phi(B_2) = 1) &= f(A_2) \geq v,
\end{align*}
\]

where \(\phi\) and \(f\) satisfy Conditions 1 and Condition 2, respectively. The following result is a straightforward corollary of Anderson’s (1955) theorem for multivariate Gaussians; see also Theorem 4.2.5 of Tong (1990).

Lemma 1. Under the stated conditions, for any scalar \(v\),

\[
\mathbb{P} \left\{ f(A_1) \geq v | \phi(B_1) = 1 \right\} \leq \mathbb{P} \left\{ f(A_2) \geq v | \phi(B_2) = 1 \right\}.
\]

The result follows immediately from Anderson’s theorem after noting that the set \(B_v = \{(a, b)^T : f(a) \leq v \land \phi(b) = 1\}\) is convex and mirror-symmetric for any \(v\). This can be seen through our assumption that \(f(\cdot)\) is quasi-convex and mirror-symmetric, such that its sublevel sets are convex and mirror symmetric. We further have that \(\mathbb{P}(\phi(B_1) = 1) = \mathbb{P}(\phi(B_2) = 1) > 0\) given the structure of the covariance matrices \(M_1\) and \(M_2\) and Condition 1, completing the proof.

6.2 | Stochastic dominance and the probability integral transform

For two real-valued random variables \(S\) and \(T\), \(S\) (first order) stochastically dominates \(T\) if \(F_S(a) \leq F_T(a)\) for all \(a \in \mathbb{R}\), where \(F_S\) and \(F_T\) are the distribution functions of \(S\) and \(T\), respectively.

Suppose now that \(S\) and \(T\) are continuous and that \(S\) stochastically dominates \(T\). By the probability integral transform, the distribution of \(F_T(T)\) would be standard uniform. The following proposition considers transforming the random variable \(T\) not by its own distribution function, but rather by the distribution function of \(S\), its stochastically dominating random variable.

Lemma 2. Suppose that \(S, T\) are continuous random variables and that \(S\) stochastically dominates \(T\). Then, \(F_S(T)\) is stochastically dominated by a standard uniform random variable.
Proof. For any $t \in [0,1]$, $\mathbb{P}\{F_S(T) \leq t\} = \mathbb{P}\{T \leq F_S^{-1}(t)\} \geq \mathbb{P}\{S \leq F_S^{-1}(t)\} = t$.

In the setup of Section 6.1, under Conditions 1 and 2, we have by Proposition 1 that $f(A_2|\phi(B_2)) = 1$ stochastically dominates $f(A_1|\phi(B_1)) = 1$. Consequently, composing $f(A_1|\phi(B_1)) = 1$ with the distribution function of $f(A_2|\phi(B_2)) = 1$ would yield a random variable that is stochastically dominated by a standard uniform.

6.3 A proof sketch for Theorem 1

While a formal proof of Theorem 1 is deferred to the web-based supporting materials, here we provide an informal sketch in light of Lemmas 1 and 2. Under Assumptions 1–3 and Condition 1, $\sqrt{N}(\hat{\tau} - \tau)$ converges in distribution to $A_1|\phi(B_1) = 1$, where $(A_1, B_1)^T$ are jointly multivariate normal with covariance $V$. Recall that $T(y(Z), Z) = f_{\hat{\xi}}(\sqrt{N}\hat{\tau})$ for some $f_{\eta}$ satisfying Condition 1 for all $\eta \in \Xi$, some $\hat{\xi}$ satisfying Condition 3, and with a balance criterion $\phi$ satisfying Condition 1. By Condition 3 and the assumption of the weak null, we have that $\hat{\xi}(y(Z), Z)$ converges in probability to $\xi$. Therefore, under the weak null, by Lemma 1 the limiting distribution of $T(y(Z), Z)$ would be stochastically dominated by that of $f_{\hat{\xi}}(A_2|\phi(B_2)) = 1$ for any $(A_2, B_2)^T$ multivariate Gaussian with covariance matrix

$$\sqrt{V} = V + \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix}$$

with $\Delta \geq 0$. The transformation

$$\overline{G}(y(Z), Z) = \frac{\gamma^{(d+k)}_{0,\sqrt{V}} \{ (a, b)^T : f_{\hat{\xi}}(a) \leq f_{\hat{\xi}}(\sqrt{N}\hat{\tau}) \land \phi(b) = 1 \}}{\gamma^{(k)}_{0,\sqrt{V}} \{ b : \phi(b) = 1 \}}$$

transforms $T(y(Z), Z)$ by the distribution function of a random variable which stochastically dominates its limiting distribution. By Lemma 2 and the continuous mapping theorem, asymptotically $\overline{G}(y(Z), Z)$ is stochastically dominated by a standard uniform. By Condition 4, the covariance estimator $\hat{V}$ used in forming $G(y(Z), Z)$ has a probability limit of the required form for stochastic dominance. Therefore, another application of the continuous mapping theorem yields that $G(y(Z), Z) - \overline{G}(y(Z), Z) = o_p(1)$, such that by Slutsky’s Theorem $G(y(Z), Z)$ is itself stochastically dominated by a standard uniform.

Meanwhile, Proposition 1 and Condition 1 yield that under the weak null the distribution of $\sqrt{N}(\hat{\tau}(y(Z), W))|Z$ converges weakly in probability to the distribution of $\hat{A}|\phi(B) = 1$, where $(\hat{A}, \hat{B})^T$ are jointly multivariate Gaussian with mean zero and covariance $\hat{V}$. The distribution of $f_{\hat{\xi}(\hat{\xi}(y(Z), W))|Z}$ is precisely $\hat{P}$, the reference distribution assuming the sharp null holds for the test statistic $T(y(Z), Z) = f_{\hat{\xi}}(\sqrt{N}\hat{\tau})$. By Condition 4, $\overline{V}(y(Z), W)$ converges in probability to $\hat{V}$ itself. Further, by Condition 3 $\hat{\xi}(y(Z), W)$ converges in probability to $\hat{\xi}$. Applying the continuous mapping theorem and Slutsky’s Theorem for randomization distributions (Chung & Romano, 2016, Lemmas A5-A6), one sees that Gaussian prepivoting furnishes a transformation that amounts to, asymptotically, an application of the probability integral transform. As a result, $\hat{P}_G(t)$ converges in probability to $t$, the distribution function of the standard uniform, for all $t \in [0, 1]$. 
7 | EXTENSIONS TO ASYMPTOTICALLY LINEAR ESTIMATORS

Theorem 1 may be extended to estimators other than the difference in means. Consider an estimator \( \hat{\tau}(y(Z), Z) \) such that

\[
\sqrt{N} \{ \hat{\tau}(y(Z), Z) - \bar{\tau} \} = \sqrt{N} \left( \frac{1}{n_1} \sum_{i=1}^{N} Z_i r_i(Z_i) - \frac{1}{n_0} \sum_{i=1}^{N} (1 - Z_i) r_i(Z_i) \right) + o_p(1)
\]

for some constants \( \{r_i(0), r_i(1)\}_{i=1}^{N} \) which may change with \( N \) and that satisfy \((1/N) \sum_{i=1}^{N} (r_i(1) - r_i(0)) = 0 \) along with Assumptions 2 and 3. Suppose further that \( \hat{\tau}(\hat{y}(Z), W), W \) independent from \( Z \) and drawn uniformly from \( \Omega \), satisfies

\[
\sqrt{N} \hat{\tau}(\hat{y}(Z), W) = \sqrt{N} \left( \frac{1}{n_1} \sum_{i=1}^{N} W_i \hat{r}_i(Z_i) - \frac{1}{n_0} \sum_{i=1}^{N} (1 - W_i) \hat{r}_i(Z_i) \right) + o_p(1)
\]

for potentially distinct constants \( \{\hat{r}_i(0), \hat{r}_i(1)\}_{i=1}^{N} \) which may change with \( N \) that satisfy \((1/N) \sum_{i=1}^{N} (\hat{r}_i(1) - \hat{r}_i(0)) = 0 \) along with Assumptions 2 and 3. Observe that the difference in means estimator satisfies these conditions with \( r_i(z) = \hat{r}_i(z) = y_i(z) - z\bar{\tau} \) for \( z \in [0, 1] \). Let \( \tau_i = r_i(1) - r_i(0) \). Let \( \Sigma_r(z), \Sigma_r z, \Sigma_r y(z)\) be the analogues of \( \Sigma y(Z), \Sigma_r x, \Sigma_r y(z)z \) and \( \Sigma r x \) for \( z \in [0, 1] \), and let the same hold with \( r \) replaced by \( \hat{r} \). Define \( V^{(r)}(V) \) as the analogues of \( V \) and \( \hat{V} \) computed now based upon \( r(z) \) and \( \hat{r}(z) \) instead of \( y(z) \) and \( \hat{y}(z) \) for \( z \in [0, 1] \).

Consider a test statistic for the weak null of the form \( \hat{T}(y(Z), Z) = f_\xi(\sqrt{N} \hat{\tau}) \) for some \( f_\xi \) satisfying Condition 2 and \( \hat{\xi} \) satisfying Condition 3, and suppose that there exists a covariance estimator \( \hat{V} \) satisfying Condition 4 with \( V \) and \( \hat{V} \) replaced by \( V^{(r)}(V) \). The Gaussian prepivoted test statistic is

\[
\tilde{G}(y(Z), Z) = \frac{\chi^{(d+k)}_{\theta, V}}{\chi^{(k)}_{\theta, V, \delta}} \left\{ (a, b)^T : f_\xi(a) \leq \hat{T}(y(Z), Z) \land \phi(b) = 1 \right\}
\]

**Theorem 2.** Suppose that Neyman's null, \( H_N: \bar{\tau} = 0 \), holds. Then, under the described restrictions on \( \hat{T}(y(Z), Z) \) and \( \hat{V} \) and under Assumption 1 and with Assumptions 2 and 3 applied to \( r_i(z), z \in [0, 1] \), \( \tilde{G}(y(Z), Z) \) converges in distribution to a random variable \( \tilde{U} \) taking values in \([0, 1]\) satisfying

\[
P(\tilde{U} \leq t) \geq t,
\]

for all \( t \in [0, 1] \). Furthermore, the distribution \( \tilde{\Phi}_G(t) \) satisfies \( \tilde{\Phi}_G(t) \rightarrow t \) for all \( t \in [0, 1] \).

In the web-based supporting material, we illustrate that the regression-adjusted average treatment effect estimator and its corresponding estimated variance presented in Lin (2013) can be viewed in this form. As a result, Theorem 2 provides justification for the use of the prepivoted randomization distribution of a regression-adjusted estimator.
8 SIMULATION STUDIES

8.1 Studentization and prepivoting in rerandomized designs

In the $b$th of $B$ iterations, we draw, for $i = 1, ..., N$, covariates iid as

$$
\mathbf{x}_i \overset{iid}{\sim} \mathcal{N} \left( 0, \begin{pmatrix} 1.0 & 0.8 & 0.2 \\ 0.8 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix} \right).
$$

Given these covariates, we draw $r_i(0)$ and $r_i(1)$ as

$$
r_i(0) = \mathbf{x}_i^T \beta_0 + \epsilon_i(0); \quad r_i(1) = \mathbf{x}_i^T \beta_1 + \epsilon_i(1),
$$

where $\beta_0 = (-6.4, -4.0, -2.4)$, $\beta_1 = c(0.2, 0.4, 0.6)^T$, $\epsilon_i(0) \sim \mathcal{E}(1) + 1$, $\epsilon_i(1) \sim \mathcal{E}(1/10) + 10$, $\epsilon_i(0)$ independent of $\epsilon_i(1)$, and $\mathcal{E}(\lambda)$ representing an exponential distribution with rate $\lambda$.

We form the potential outcomes under treatment and control in two distinct ways, one in which the sharp null holds and one in which only the weak null holds:

**Sharp Null:** $y_i(1) = y_i(0) = r_i(1)$

**Weak Null:** $y_i(1) = r_i(1)$; $y_i(0) = r_i(0) + \bar{r}(1) - \bar{r}(0)$

Of the $N$ individuals, $n_1 = 0.2N$ receive the treatment and $n_0 = 0.8N$ receive the control. We use a Mahalanobis-based rerandomized design, with criterion $\phi(\sqrt{N\hat{\delta}}) = 1\{(\sqrt{N\hat{\delta}})^TV^{-1}\sqrt{N\hat{\delta}} \leq 1\}$. This balance criterion reduces the cardinality of $\Omega$ by roughly 80% relative to a completely randomized design. For each $b$, we draw a single $\mathbf{Z} \in \Omega$, and proceed with inference using the reference distribution of the following test statistics under the incorrect assumption that the sharp null holds:

1. Absolute difference in means, unstudentized
2. Absolute difference in means, studentized
3. Gaussian prepivoting the absolute difference in means, studentized

The true reference distributions assuming the sharp null are replaced by Monte-Carlo estimates with 1000 draws from $\Omega$ for each $b$, and the desired Type I error rate is $\alpha = 0.05$. We also perform inference using the large-sample reference distribution for the absolute studentized difference in means in a rerandomized design; see Li et al. (2018) for more details. As a covariance estimator $\hat{\mathbf{V}}$, we use the conventional unpooled covariance estimator for $(\sqrt{N\hat{\delta}})^TV^{-1}\sqrt{N\hat{\delta}}$ in a two-sample design. For the generative models reflecting the sharp and weak nulls, we proceed with both $N = 50$ and $N = 1000$ to compare performance in small and large sample regimes. For each $N$, we conduct $B = 5000$ simulations.

Table 1 contains the results of the simulation study. Under the sharp null with $N = 50$, we see the benefits of using a randomization test: the randomization tests based upon the unstudentized, studentized and prepivoted absolute difference in means all resulted in a Type I error rate of 0.05 (up to noise from the Monte-Carlo simulation) as desired. Contrast this with the large-sample test, which had an estimated Type I error rate of 0.110 under the sharp null
hypothesis. Figure 1 explains the deficiency of the large-sample test by comparing the true distribution for the large-sample \( p \)-values to the standard uniform distribution. As is seen, at \( N = 50 \) small \( p \)-values are more likely to occur than what the standard uniform would predict at any point \( t \in [0, 1] \), resulting in inflated Type I error rates. By \( N = 1000 \), the asymptotic approximation performs much better, as the true distribution of \( p \)-values lies on top of the standard uniform. Gaussian prepivoting uses 1 minus these large-sample \( p \)-values as the test statistic whose randomization distribution is enumerated, such that the solid line in Figure 1 reflects 1 minus the randomization distribution of the Gaussian prepivoted test statistic. As Gaussian prepivoting uses a randomization test under the sharp null, the solid line also reflects the reference distribution employed for performing inference. That these coincide is a consequence of the sharp null holding, such that the randomization tests are exact tests for any sample size.

Under the weak null, we see in Table 1 that even at \( N = 1000 \), the unstudentized and studentized randomization tests erroneously assuming the sharp null have inflated Type I error rates. This pattern will persist even asymptotically, as in this simulation setup these test statistics are not asymptotically sharp-dominant. This may come as a surprise, as in completely randomized designs studentizing does furnish asymptotic sharp dominance. As evidenced here, the impact of covariates on the limiting distribution in rerandomized experiments invalidates studentization as a mechanism for attaining asymptotic sharp dominance. Figure 2 illustrates this in the case of the studentized test statistic. We see in the top panel that the true distribution function for the studentized test statistic lies below that of the reference distribution assuming the sharp null, such that the right-tail probabilities are larger for the true randomization distribution than they are for the reference distribution. This yields anti-conservative inference. We see in the bottom panel of Figure 2 that through use of Gaussian prepivoting, asymptotic conservativeness has been restored: the true randomization distribution of the prepivoted test statistic is stochastically dominated by the reference distribution assuming the sharp null, as predicted by Theorem 1. We further see that the cumulative distribution assuming the sharp null is converging to the distribution function of the standard uniform (a straight line between 0 and 1), again reflecting Theorem 1. Table 1 further shows that the Gaussian prepivoted test and the large-sample test have very similar rejection rates at \( N = 1000 \), reflecting the asymptotic equivalence of the two methods under the weak null.

|                  | Randomization test | Large-sample |
|------------------|--------------------|--------------|
|                  | No Stu. | Stu. | Pre. |                 |
| Sharp, \( N = 50 \) | 0.053  | 0.050 | 0.051 | 0.110          |
| Sharp, \( N = 1000 \) | 0.052 | 0.048 | 0.048 | 0.054          |
| Weak, \( N = 50 \) | 0.073  | 0.114 | 0.037 | 0.058          |
| Weak, \( N = 1000 \) | 0.070 | 0.083 | 0.018 | 0.019          |

The rows describe the simulation settings, which vary between the sharp and weak nulls holding and between small and large sample sizes. The first three columns represent the performance of randomization tests assuming the sharp null hypothesis and using the unstudentized absolute difference in means, absolute studentized difference in means, and Gaussian prepivoted absolute difference in means, respectively, to perform inference. The last column is a large-sample test which is asymptotically valid for the weak null, based upon Li et al. (2018). The desired Type I error rate in all settings is \( \alpha = 0.05 \).
8.2 | A comparison of multivariate tests

In each iteration $b = 1, \ldots, B$, we draw $(r_i(1))_{i=1}^N$ and $(r_i(0))_{i=1}^N$ independent from one another and iid from mean zero equicorrelated multivariate normals of dimension $k = 25$ with marginal variances one. The correlation coefficients governing $r_i(1)$ and $r_i(0)$ are 0 and 0.95, respectively. We will have two simulation settings, one each for the sharp and weak null:

- **Sharp Null**: $y_i(1) = y_i(0) = r_i(1)$.
- **Weak Null**: $y_i(1) = r_i(1)$; $y_i(0) = r_i(0) + \bar{r}(1) - \bar{r}(0)$.

In both settings, $n_1 = 0.2N$ individuals receive the treatment and $n_0 = 0.8N$ receive the control. We consider a completely randomized design, and proceed with inference using the reference distribution of the following test statistics under the (erroneous) assumption that the sharp null holds:

1. Hotelling’s $T$-squared, unpooled covariance
2. Hotelling’s $T$-squared, pooled covariance
3. Max absolute $t$-statistic, unpooled standard error
For each candidate test, we proceed with the randomization distribution both of the untransformed test statistic and the Gaussian prepivoted test statistic. These tests are conducted using Monte-carlo simulation to generate the reference distributions, with 1000 draws from $\Omega$ for each iteration $b$. In addition to the two types of randomization tests, we also compute a large-sample $p$-value for each test which is asymptotically valid under the weak null hypothesis. As a covariance estimator $\hat{\Sigma}$, we use the conventional unpooled covariance estimator for $\sqrt{N}\hat{\Sigma}$. For each test, we seek to maintain the Type I error rate at or below $\alpha = 0.05$. For the generative models reflecting the sharp and weak nulls we proceed with both $N = 300$ and $N = 5000$ to compare performance as $N$ increases. For each $N$, we conduct $B = 5000$ simulations.

Table 2 gives the estimated Type I error rates for the candidate tests. We first note the poor performance of the large-sample tests under both the sharp and weak null with $N = 300$. For instance, the large-sample $p$-values constructed using the unpooled, Hotelling procedure are attained using a $\chi^2_{25}$ distribution and have estimated Type I error rates of 0.321 under the sharp null for $N = 300$, and of 0.270 under the weak null for $N = 300$ despite the desired control at $\alpha = 0.05$. By $N = 5000$, the large-sample tests all have estimated Type I error rates approaching the nominal level under the sharp null, and below the nominal level under the weak null.
Naturally, all randomization tests attain (up to Monte-Carlo error) the desired Type I error rate under the sharp null at both $N = 300$ and $N = 5000$, highlighting the appeal of the randomization tests. Under the weak null, we see that the randomization test based upon the Hotelling $T$-statistic with a pooled covariance fails to control the Type I error rate even at $N = 5000$, reflecting that the test statistic is not asymptotically sharp-dominant. While the randomization test based on the max $t$-statistic controls the Type I error rate in these simulations, this is not guaranteed in general: in the web-based supporting materials we conduct this simulation at $\alpha = 0.25$, where anti-conservativeness of the max $t$-statistic arises. For both of these test statistics, applying Gaussian prepivoting restores guaranteed asymptotic conservativeness and results in test statistics whose performance closely aligns with the large-sample tests, a reflection of Theorem 1. For the test based upon Hotelling’s $T$-statistic with an unpooled covariance estimator, observe that the Type I error rates for the randomization tests with and without Gaussian prepivoting are identical in all four scenarios tested. As discussed in Example 2 of Section 5.2, this is because Gaussian prepivoting is unnecessary for this particular test statistic: Hotelling’s $T$ statistic with an unpooled covariance estimator is already asymptotically sharp-dominant as proven in Wu and Ding (2018). Applying Gaussian prepivoting recovers an equivalent randomization test, furnishing identical $p$-values for any observed outcomes $y(Z)$ for completely randomized designs.

In the web-based supporting materials, we provide a theoretical analysis of the statistical power of Gaussian prepivoting and include simulations to demonstrate the power in practice. We also provide analysis of real-world data from the Student Achievement and Retention experiment of Angrist et al. (2009).

9 | DISCUSSION

9.1 | An open question: multivariate one-sided testing in finite population causal inference

The restrictions on the function $f_\phi$ outlined in Condition 2 require a quasi-convex, continuous function that is mirror-symmetric about the origin. This restriction results in convex, mirror-symmetric sublevel sets for $f_\phi$ and facilitates the application of Anderson’s theorem, such that

| TABLE 2 Inference in completely randomized designs with multiple outcomes |
|--------------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                         | Hotelling, Unpooled |                  | Hotelling, Pooled |                  | Max t-stat      |                  |                  |                  |
|                         | FRT | Pre. | LS   | FRT | Pre. | LS   | FRT | Pre. | LS   |
| Sharp, $N = 300$       | 0.050 | 0.050 | 0.321 | 0.052 | 0.047 | 0.086 | 0.051 | 0.050 | 0.068 |
| Sharp, $N = 5000$     | 0.044 | 0.044 | 0.053 | 0.047 | 0.042 | 0.045 | 0.046 | 0.045 | 0.048 |
| Weak, $N = 300$       | 0.117 | 0.117 | 0.270 | 0.975 | 0.166 | 0.157 | 0.020 | 0.006 | 0.008 |
| Weak, $N = 5000$     | 0.003 | 0.003 | 0.003 | 0.951 | 0.005 | 0.005 | 0.021 | 0.005 | 0.005 |

The rows describe the simulation settings, which vary between the sharp and weak nulls holding and between small and large sample sizes. There are three sets of columns, one corresponding to each of the three test statistics under consideration. For each set of columns, the column labelled ‘FRT’ represents the Fisher randomization test using that test statistic. The column labelled ‘Pre’. instead reflects the Fisher randomization test after applying Gaussian prepivoting to the original test statistic. The last column, labelled ‘LS’, is a large-sample test which is asymptotically valid for the weak null. The desired Type I error rate in all settings is $\alpha = 0.05$. 
dominance in the Loewner order on covariance matrices translates to the stochastic dominance under the weak null. While the restrictions on $f_j$ are sensible with two-sided alternatives, they preclude testing directional alternatives because of the mirror symmetry condition. For instance, suppose one wanted to test the null hypothesis $\overline{r}_i \leq 0$ for all $i = 1, \ldots, d$ versus the alternative that for at least one $i$ ($i = 1, \ldots, d$), $\overline{r}_i > 0$. In the univariate case, choosing $T(y(Z), Z) = \hat{r}/\hat{V}_{rr}^{1/2}$ does not provide a valid one-sided test for all $\alpha$. That said, it does provide a valid test for $\alpha \leq 0.5$, such that for any reasonable value for $\alpha$ to be deployed in practice a one-sided test is possible.

Suppose we have multivariate potential outcomes and consider the test statistic $T_{\max}(y(Z), Z) = \max_{1 \leq i \leq d} \max_{i \leq j \leq d} \hat{r}_{ij}/\hat{V}_{rr}^{1/2}$ with $\hat{V}_{rr}$ satisfying Condition 4. Consider the Gaussian prepivoted test statistic $G_{\max}(y(Z), Z)$. The following is, to the best of our knowledge, an open question: is it the case that, for any $\alpha \leq 0.5$, $G_{\max}$ is asymptotically sharp-dominant, in that $\limsup \mathbb{E}\{\varphi_{G_{\max}}(\alpha)\} \leq \alpha$? Under the assumptions imposed in this work, the answer would be true should the following conjecture on Gaussian comparisons hold:

**Conjecture 1.** Let $X = (X_1, \ldots, X_d)$, and $Y = (Y_1, \ldots, Y_d)$ be $d$-dimensional multivariate Gaussian vectors, with a common mean $\mu = (\mu_1, \ldots, \mu_d)$ but distinct covariances $\Sigma_X$ and $\Sigma_Y$, with ij entries $\sigma^X_{ij}$ and $\sigma^Y_{ij}$, respectively. Let $\gamma^X_{ij} = \mathbb{E}\{(X_i - X_j)^2\}$ and $\gamma^Y_{ij} = \mathbb{E}\{(Y_i - Y_j)^2\}$. Define $\text{med}(\max_i Y_i)$ as the median of $\max_i Y_i$, i.e. the value a such that $\mathbb{P}\left(\max_{1 \leq i \leq d} Y_i \leq a\right) = 0.5$. Suppose that $\sigma^Y_{ij} \geq \sigma^X_{ij}$ for all $i$ and that $\gamma^Y_{ij} \geq \gamma^X_{ij}$ for all $i, j$. Consider any point $c \geq \text{med}(\max_i Y_i)$. Then,

$$
\mathbb{P}\left(\max_{1 \leq i \leq d} X_i \leq c\right) \leq (\gamma) \mathbb{P}\left(\max_{1 \leq i \leq d} Y_i \geq c\right).
$$

The conjecture is true in the univariate case. Under the assumptions of this conjecture, the Sudakov–Fernique inequality (Adler & Taylor, 2009, Theorem 2.2.5) asserts that $\mathbb{E}\{\max_{1 \leq i \leq d} X_i\} \leq \mathbb{E}\{\max_{1 \leq i \leq d} Y_i\}$. Should we further assume $\sigma^X_{ii} = \sigma^Y_{ii}$, the result holds for all points $c$ through Slepian’s lemma (Slepian, 1962; Tong, 1990, Theorem 5.1.7). Unfortunately, a refined result about tail probabilities above the median does not appear to be available in the literature under the conditions outlined in the conjecture. A potential path forward may be a modification of the soft-max proof of the Sudakov–Fernique inequality found in Chatterjee (2005).

### 9.2 Summary

In this work, we present a general framework for designing randomization tests that are both exact for Fisher’s sharp null and are asymptotically conservative for Neyman’s weak null in completely randomized experiments and rerandomized designs. Loosely stated, the approach may be summarized as follows: if one has access to a large-sample test that is asymptotically conservative under Neyman’s weak null, then a Fisher randomization test using the $p$-value produced by that large-sample test will maintain asymptotic correctness under the weak null while additionally restoring exactness should the sharp null be true. As the Fisher randomization distribution of these $p$-values converges weakly in probability to a uniform, the resulting randomization test assuming the sharp will have the same large-sample performance under the weak null as large-sample test itself, and will further have the same asymptotic power under local alternatives as
the large-sample test. We show that Gaussian prepivoting exactly recovers several randomization tests known to be valid under the weak null, while providing a general approach to restore asymptotic correctness to randomization tests for a large class of test statistics. Importantly, our framework immediately provides valid randomization tests of the weak null hypothesis in rerandomized designs, absent from the literature until now.

10 ADDITIONAL MATERIALS

Supporting material containing proofs of the results may be found in the on-line version of this article. Example code to implement Algorithm 1 is available at https://github.com/PeterL_Cohen/PrepivotingCode.

ORCID

Peter L. Cohen © https://orcid.org/0000-0002-0569-5871
Colin B. Fogarty © https://orcid.org/0000-0002-5520-4639

REFERENCES

Adler, R.J. & Taylor, J.E. (2009) Random fields and geometry. Berlin: Springer Science & Business Media.
Anderson, T.W. (1955) The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. Proceedings of the American Mathematical Society, 6, 170–176.
Angrist, J., Lang, D. & Oreopoulos, P. (2009) Incentives and services for college achievement: evidence from a randomized trial. American Economic Journal: Applied Economics, 1, 136–63.
Beran, R. (1987) Prepivoting to reduce level error of confidence sets. Biometrika, 74, 457–468.
Beran, R. (1988) Prepivoting test statistics: a bootstrap view of asymptotic refinements. Journal of the American Statistical Association, 83, 687–697.
Caughey, D., Dafoe, A. & Miratrix, L. (2017) Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. arXiv preprint arXiv:1709.07339.
Chatterjee, S. (2005) An error bound in the Sudakov-Fernique inequality. arXiv preprint math/0510424.
Chung, E. & Romano, J.P. (2016) Multivariate and multiple permutation tests. Journal of Econometrics, 193, 76–91.
Crâuel, H. (2002) Random probability measures on Polish spaces, vol. 11 of Stochastics Monographs. London: Taylor & Francis.
Ding, P. (2017) A paradox from randomization-based causal inference. Statistical Science, 32, 331–345.
Ding, P. & Dasgupta, T. (2018) A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika, 105, 45–56.
Ding, P., Li, X. & Miratrix, L.W. (2017) Bridging finite and super population causal inference. Journal of Causal Inference, 5, 20160027. https://doi.org/10.1515/jci-2016-0027
Ding, P., Feller, A. & Miratrix, L. (2019) Decomposing treatment effect variation. Journal of the American Statistical Association, 114, 304–317.
Dümbgen, L. & Del Conte-Zerial, P. (2013) On low-dimensional projections of high-dimensional distributions. In: Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V. & Maathuis, M.H. (Eds.) From probability to statistics and back: high-dimensional models and processes, vol. 9 of Institute of mathematical statistics (IMS) collect, Beachwood, OH: Institute of Mathematical Statistics, pp. 91–104.
Fisher, R.A. (1935) The design of experiments. Edinburgh: Oliver & Boyd.
Fogarty, C.B. (2020) Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. Journal of the American Statistical Association, 115, 1518–1530.
Freedman, D.A. (2008a) On regression adjustments in experiments with several treatments. The Annals of Applied Statistics, 2, 176–196.
Freedman, D.A. (2008b) On regression adjustments to experimental data. Advances in Applied Mathematics, 40, 180–193.
Imbens, G.W. & Rubin, D.B. (2015) *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge: Cambridge University Press.

Lehmann, E.L. & Romano, J.P. (2005) *Testing statistical hypotheses*, 3rd edn. Springer Texts in Statistics. New York: Springer.

Li, X. & Ding, P. (2017) General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112, 1759–1769.

Li, X., Ding, P. & Rubin, D.B. (2018) Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115, 9157–9162.

Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman’s critique. *The Annals of Applied Statistics*, 7, 295–318.

Loh, W.W., Richardson, T.S. & Robins, J.M. (2017) An apparent paradox explained. *Statistical Science*, 32, 356–361.

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

Neyman, J. (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5, 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed.

Rosenbaum, P.R. (2002) *Observational studies*, 2nd edn. Springer Series in Statistics. New York: Springer-Verlag.

Rubin, D.B. (1980) Comment on “Randomization analysis of experimental data: the Fisher randomization test”. *Journal of the American Statistical Association*, 75, 591–593.

Slepian, D. (1962) The one-sided barrier problem for Gaussian noise. *The Bell System Technical Journal*, 41, 463–501.

Teixeira-Pinto, A. & Mauri, L. (2011) Statistical analysis of noncommensurate multiple outcomes. *Circulation: Cardiovascular Quality and Outcomes*, 4, 650–656.

Teixeira-Pinto, A., Siddique, J., Gibbons, R. & Normand, S.L. (2009) Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric Annals*, 39, 729–735.

Tong, Y.L. (1990) *The multivariate normal distribution*. Springer Series in Statistics. New York: Springer-Verlag.

van der Vaart, A.W. & Wellner, J.A. (1996) *Weak convergence and empirical processes*. Springer Series in Statistics. With applications to statistics New York: Springer-Verlag.

Wu, J. & Ding, P. (2018) Randomization tests for weak null hypotheses. *arXiv e-prints*, arXiv:1809.07419.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

---

**How to cite this article:** Cohen, P.L. & Fogarty, C.B. (2022) Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 295–320. [https://doi.org/10.1111/rssb.12439](https://doi.org/10.1111/rssb.12439)