Published in final edited form as:
J Biopharm Stat. 2010 July ; 20(4): 850–869. doi:10.1080/10543401003618884.

EXACT INFERENCE FOR COMPLEX CLUSTERED DATA USING WITHIN-CLUSTER RESAMPLING

Dean Follmann,
Michael Fay
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA

Abstract

This paper introduces exact permutation methods for use when there are independent clusters of data with arbitrary within-cluster correlation. To eliminate the problem of clustering, we randomly select a data point from each cluster and for this now independent data, and calculate our test statistic and the associated support points for all possible permutations. While clearly valid, this is also inefficient. We repeat this process until all possible independent data sets have been created and use the support points averaged over the randomly created data sets as our reference distribution for the averaged test statistic. This approach uses all of the data and is a permutation extension of within-cluster resampling (WCR). We discuss both exact and Monte Carlo versions of the approach and apply it to several data sets. WCR permutation can be applied in quite general settings when within cluster correlation is a nuisance and exact inference is necessary.

Keywords

Clustering; Correlated data; Multiple outputation; Permutation test; Within-cluster resampling

1. INTRODUCTION

The use of permutation tests is well established and a popular option for statistical inference when exact p values are desired. Permutation tests are typically applied to statistically independent data points. In some settings, however, observations will be clustered and application of permutation methods will not be obvious. In a recent collaboration (Di Mascio et al., 2004), repeated measurements of CD4 cell counts and viral load were obtained on human immunodeficiency virus (HIV)-infected individuals. Many of the viral loads were below the limit of detection; detectable values were called “blips.” The researchers, who were comfortable with permutation tests, wanted to determine whether the relative change in CD4 was the same in consecutive “nonblip, blip” visits as for consecutive “nonblip, nonblip” visits. If each individual provided a single pair of such couples, a permutation t-test or Wilcoxon signed rank test could be performed using the pair of relative changes. However, each individual provided many such couples.

Address correspondence to Dean Follmann, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive MSC 7609, Bethesda, MD 20892, USA; dfollmann@niaid.nih.gov.

Permutation methods for correlated data have been discussed by a variety of authors in specific settings. Fay and Shih (1998) introduced distribution permutation tests (DPT) for the situation where each cluster provides multiple responses. Their work generalizes Wilcoxon rank sum and permutation *t*-tests to clustered data. Generalizations of the Wilcoxon rank sum test to clustered data that apply asymptotically have been derived (Datta and Satten, 2005; Rosner et al., 2003). Gail et al. (1996) discuss and evaluate permutation tests for group randomized trials. Braun and Feng (2001) introduce optimal permutation tests for group randomized trials by postulating models for the within-cluster observations. Finally, Cai and Shen (2000) apply permutation tests for clustered survival data by creating test statistics that pretend the data are independent, but then permute cluster labels rather than individual labels.

This paper introduces a general approach to construct permutation tests for clustered data with a common cluster-wide covariate that is to be permuted. The idea extends the within-cluster resampling of Hoffman et al. (2001) (see also Follmann et al., 2003). The basic idea is that if each cluster provided a single data point, standard permutation methods could be used. Thus, we randomly select one data point from each cluster. Based on this now independent data, we form both the test statistic and the b support points of the permutation distribution. We repeat the process for all m possible within-cluster (WC) resamples or ways we can do this, and average the test statistics for each of the b permutations over these m resamples. The b support points of the averaged test statistic provide the null distribution. We call this exhaustive WCR permutation (WCRP). We also describe a Monte Carlo approach where we randomly sample from the set of permutations and resamples. Note that WCR averaging of a test statistic derived from estimating equations has been used by Williamson et al. (2003) and Datta and Satten (2005, 2008) but they relied on asymptotic approximations for inference.

We begin with a review of permutation tests and then define WCRP, showing how it works in simple settings. We next describe Monte Carlo WCRP and show how to come close to the (exhaustive) WCRP result by using an ad hoc “6-tens” algorithm. A small simulation is conducted to illustrate key features of the algorithm. We finish by applying the proposed methods to some data sets that motivated its development.

2. EXHAUSTIVE PERMUTATION

We first review the situation where each of n clusters provides one data point. In this situation, the outcomes are given by the n-vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. Associated with the ith cluster is a variable Z_i, e.g., a group indicator, or the dose of a drug. Under the null hypothesis H_0, all permutations of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ are equally likely. We form the test statistic on the observed data, $t_0 = t(\boldsymbol{X}, \boldsymbol{Z})$. We calculate t for all permutations of $\boldsymbol{Z}$, denoting the kth permutation as $t_k = t\{\boldsymbol{X}, \boldsymbol{Z}(\boldsymbol{\pi}_k)\}$, where $\boldsymbol{\pi}_k$, $k = 1, \ldots, b$ is a listing of all possible nonredundant permutations, with $\boldsymbol{\pi}_0$ corresponding to the unpermuted data: $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{\pi}_0)$. For example, if $\boldsymbol{Z} = 1.1, 2.2, 3.3, 4.4$, corresponding to 4 doses of a drug, and $\boldsymbol{\pi}_k =$ (4, 1, 3, 2), then $b = 4!$ and $\boldsymbol{Z}(\boldsymbol{\pi}_k) = 4.4, 1.1, 3.3, 2.2$. The one-sided lower p value is $p_\ell = b^{-1}\sum_{k=1}^{b} I(t_k \le t_0)$ where $I(\,)$ is the indicator function.

As a simple example, consider the made-up data of six clusters with a single response in each cluster: $\boldsymbol{X}$ = [3.3, 3.1, 0.8, 1.1, 1.5, 2.3] and two groups, $\boldsymbol{Z}$ = [0, 0, 0, 1, 1, 1]. We can test equality of the distributions $F(x|Z = 0)$ and $F(x|Z = 1)$ by using the permutation t-test whose test statistic is the mean difference:

$$t_0 = t(\boldsymbol{X}, \boldsymbol{Z}) = \frac{\sum_{i=1}^{n} Z_i X_i}{3} - \frac{\sum_{i=1}^{n}(1 - Z_i)X_i}{3}$$

Here, $b = \binom{6}{3} = 20$ and $p_l = 4/20$.

3. WITHIN-CLUSTER RESAMPLING PERMUTATION

3.1. Basic Definition

Now consider the setting where the ith cluster has m_i data points, with $m_i > 1$ for some i. Denote the responses by the long vector

$$\boldsymbol{X} = \left(X_{11}, \ldots, X_{1m_1}, X_{21}, \ldots, X_{2m_2}, \ldots, X_{n1}, \ldots, X_{nm_n}\right)$$

where X_{ij} is the jth data point in the ith cluster. Importantly, the covariate of interest $\boldsymbol{Z}$ is not allowed to change within a cluster, so $\boldsymbol{Z}$ has dimension n.

For a single within-cluster resample we randomly select a single data point from each cluster. Table 1 provides a diagram that should aid understanding of this notation. Let the chosen index for cluster i be $j(i)$, where $j(i) \in \{1, \ldots, m_i\}$. Let $\boldsymbol{j} = \{j(1), \ldots, j(n)\}$, define $\boldsymbol{X}(\boldsymbol{j}) = X_{1j(1)}, \ldots, X_{nj(n)}$, and set $X_i(\boldsymbol{j}) = X_{ij(i)}$. For the jth WCR, we define our test statistic as $t_{j0} = t\{\boldsymbol{X}(\boldsymbol{j}_j), \boldsymbol{Z}\}$. To obtain the exact permutation distribution for this jth resample, we calculate all b possible permutations of $\mathbf{Z}$, $\boldsymbol{Z}(\boldsymbol{\pi}_1), \ldots, \boldsymbol{Z}(\boldsymbol{\pi}_b)$, and for each, the associated test statistic. (We generalize this to allow uncountable sets of permutations in section 5.) Define the test statistic based on the jth resample and kth permutation as $t_{jk} = t\{\boldsymbol{X}(\boldsymbol{j}_j), \mathbf{Z}(\boldsymbol{\pi}_k)\}$. As shown in Table 2, exhaustive WCRP then averages each column (unique permutation) of test statistics over all within-cluster resamples (the rows). The exhaustive WCR permutation distribution is given by the final row of Table 2 and a lower p value is given by $p_\ell = b^{-1}\sum_{k=1}^{b} I(\bar{t}_{.k} \le \bar{t}_{.0})$. Note that there are $m = m_1 \times \cdots \times m_n$ unique ways to select one observation from each person.

As another aid to understanding, refer back to the made-up data from the previous section, but now suppose that person 3 gives us an additional data point, say 2.7 in addition to her old 0.8. In this case there are two possible WC resamples $\boldsymbol{j}_1$ = (1, 1, 1, 1, 1, 1) and $\boldsymbol{j}_2$ = (1, 1, 2, 1, 1, 1). As before, for each resample there are 20 permutations so $b = 20$ and $m = 2$. The test statistics are given by

$$t_{j0} = t\left\{\boldsymbol{X}\left(\boldsymbol{j}_j\right), \boldsymbol{Z}\right\} = \frac{\sum_{i=1}^{n} Z_i X_i\left(\mathbf{j}_j\right)}{3} - \frac{\sum_{i=1}^{n}(1 - Z_i)X_i\left(\boldsymbol{j}_j\right)}{3}$$

for $j = 1, 2$. For each possible permutation $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{20}$, we can calculate the pair of test statistics for the two WC resamples, say $(t_{11}, t_{21}), \ldots, (t_{1,20}, t_{2,20})$. Figure 1 plots these pairs of test statistics along their projections onto the axes that correspond to the permutation distributions for $j = 1(2)$ for the horizontal (vertical) projections. The associated lower p values for the two WC resamples are 4/20 and 1/20, respectively. So the additional data point gives some additional evidence against the null hypothesis, but it is unclear how to make this precise.

Under the null hypothesis that all permutations of $\boldsymbol{Z}$ are equally likely, all two-dimensional points in Fig. 1 are equally likely. We can thus define a rejection region by some shape in two-dimensional space other than the vertical or horizontal lines. A natural test is to take the average of the two test statistics. This test is proportional to the projection of each point onto the 45° line as shown in Fig. 1. The p values are determined by the number of projected support points that are more extreme than the projected test statistic. We see that the lower p value based on this maneuver is $p_\ell = 2/20$.

With standard permutation tests, inference is unaffected by any monotone transformation of the test statistic. However, with WCR permutation, this is not the case. Suppose that instead of using the usual test statistic, say t, we used $\log(t)$. With WCR permutation OP we would average either t or $\log(t)$ and the p values could differ. To ensure that inference is unaffected by monotone transformations of the test statistic one can use ranks. Specifically, let R_{jk} be the rank of t_{jk} among $t_{j1}, \ldots, t_{jb}$. For each resample we replace the t_{jk}s with these associated ranks. A “rank” WCR permutation p value rank is the the percentile of $\bar{R}_{\cdot 0}$ with respect to the empirical distribution of $\bar{R}_{\cdot 1}, \cdots, \bar{R}_{\cdot b}$.

WCR permutation cannot be applied to permutation tests where the permutation set $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_b$ depends on $\boldsymbol{j}_j$. If this happens then we cannot line up the columns of the matrix of Table 2 correctly. For example, in Fisher’s exact test, we condition on the marginal totals, and if these change with different resamples we cannot apply the procedure. Another example where the permutation set can depend on $\boldsymbol{\pi}$ is when Z changes within a cluster. An example is a crossover study with $Z = (0, 1)$ or $(1, 0)$ denoting the treatment assignment sequence over the two periods for each person (cluster). Here, one might have an WC resample where only $Zs = 0$ are selected and a test statistic could not even be calculated.

Since WCRP is a permutation test, it is formally testing the strong null hypothesis that z is a meaningless label. This strong null is a composite that the distribution of cluster sizes and the joint distribution of outcomes given a cluster size are both free of z. In symbols, we test

$$H_0: G(m \mid z) = G(m) \cap F_m(x_1, \ldots, x_m \mid z) = F_m(x_1, \ldots, x_m) \text{ for all } m$$

where $G(\ )$ is the distribution function of the cluster sizes and $F_m(\ )$ is the joint distribution $X_1, \ldots, X_m$. Note that this does allow for informative cluster sizes as $F_m(\ )$ can depend on m.

3.2. Example: Permutation *t*-Test

In some simple cases, we can calculate the final row of Table 2 directly. Consider the two-sample *t*-test setting where $Z_i = 0(1)$ identifies the control (treatment) group, and person *i* gives clustered outcomes $X_{i1}, \ldots, X_{im(i)}$. One can show that exhaustive WCR permutation corresponds to taking the sample average for each person and using these averages in a permutation *t*-test. That is, the averaged test statistic is

$$\bar{t}_{\cdot 0} = \frac{\sum_{i=1}^{n} Z_i \bar{X}_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} (1 - Z_i)\bar{X}_i}{\sum_{i=1}^{n} (1 - Z_i)} \tag{1}$$

where $\bar{X}_i$ is the within-cluster mean. The *b* support points of the permutation distribution are

$$\bar{t}_{\cdot k} = \frac{\sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\bar{X}_i}{\sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)} - \frac{\sum_{i=1}^{n} \{1 - Z_i(\boldsymbol{\pi}_k)\}\bar{X}_i}{\sum_{i=1}^{n} \{1 - Z_i(\boldsymbol{\pi}_k)\}}$$

for $k = 1, \ldots, b$, where $Z_i(\boldsymbol{\pi}_k)$ is the *i*th element of $\mathbf{Z}(\boldsymbol{\pi}_k)$. Note that this does allow for informative cluster sizes as $F_m(\ )$ can depend on *m*.

3.3. Example: Linear Permutation Tests

Suppose the test statistic for WC resample $\boldsymbol{j}$ can be written as

$$t\left\{X(\boldsymbol{j}), Z\right\} = \sum_{i=1}^{n} Z_i w_{\boldsymbol{X}}\{X_i(\boldsymbol{j})\}$$

where $w_{\boldsymbol{X}}(\ )$ is a function that may depend on the entire vector $\boldsymbol{X}$ and possibly other covariates, but cannot depend on $\boldsymbol{Z}_i$. Then for permutation *k* we can write

$$\begin{aligned}
\bar{t}_{\cdot k} &= \sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\frac{1}{m}\sum_{j=1}^{m} w_X\left\{X_i\left(\boldsymbol{j}_j\right)\right\} \\
&= \sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\frac{1}{\prod_{\ell=1}^{n} m_\ell}\sum_{\ell_1=1}^{m_1} \cdots \sum_{\ell_n=1}^{m_n} w_X\left(X_{i\ell_i}\right) \\
&= \sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\frac{\prod_{h \neq i} m_h}{\prod_{\ell=1}^{n} m_\ell}\sum_{\ell_i=1}^{m_i} w_X\left(X_{i\ell_i}\right) \\
&= \sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\frac{1}{m_i}\sum_{\ell=1}^{m_i} w_X(X_{i\ell}) = \sum_{i=1}^{n} Z_i(\boldsymbol{\pi}_k)\bar{w}_i
\end{aligned}$$

This form covers many tests. The permutation *t*-test has $\bar{w}_i = \bar{X}_i$. An analysis of covariance (ANCOVA) permutation test statistic can be fashioned based on the model

$$X_{ij} = \beta_0 + \beta_1 W_i + \Delta Z_i + \epsilon_{ij}$$

where $\boldsymbol{\epsilon}_{ij}$ are mean 0 random variables free of Z_i. We then obtain

$$\overline{w}_i = \frac{(\overline{X}_i - \overline{X}) - (\overline{W} - W_i)\hat{\beta}_1(\overline{X})}{n}$$

where $\hat{\beta}_1(\overline{X})$ is the slope of the regression of $\overline{X}_i$ on W_i and $\overline{X} = \sum_i \overline{X}_i/n$. The Wilcoxon rank sum test has $\overline{w}_i = \overline{R}_i(\boldsymbol{X}) = m^{-1}\sum_{i=1}^{m} R\left\{X_i\left(\boldsymbol{j}_j\right)\right\}$, where $R\{X_i(\boldsymbol{j}_j)\}$ is the rank of $X_i(\boldsymbol{j}_j)$ among $X_1(\boldsymbol{j}_j), \ldots, X_n(\boldsymbol{j}_j)$. Additionally, the distribution permutation tests of Fay and Shih (1998) have $\overline{w}_i = n^{-1}\sum_{h=1}^{n} \phi(\hat{F}_i, \hat{F}_h)$, where $\phi(\ )$ is some function and $\hat{F}_i$ is the empirical distribution function for the responses $X_{i1}, \ldots, X_{im_i}$ of cluster i.

4. MONTE CARLO WCR PERMUTATION

4.1. Basic Setup

In many settings, b and m will be too large for exhaustive WCR permutation to be feasible and we must approximate the exact p value by Monte Carlo methods. In this section we discuss how to choose the number of randomly selected Monte Carlo resamples and permutations so that the resultant p value is close to the exhaustive WCR permutation p value.

Let $y_k = I(\bar{t}_{.k} - \bar{t}_{.0} \geq 0)$, where $k = 1, \ldots, b$, so that $\bar{y} = p_u$ is the exhaustive WCRP upper p value. To approximate the permutation distribution, we randomly draw a large number, say B, of permutations. Denote the associated indicator functions by $y_K = I(\bar{t}_{.K} - \bar{t}_{.0} \geq 0)$, where upper case $K = 1, \ldots, B$ are the indices of a random (with replacement) sample of the integers $1, \ldots, b$. Note that these $y_1, \ldots, y_B$ terms are iid Bernoullis with probability p_u. By the Central Limit Theorem (CLT) we know that

$$\frac{1}{B}\sum_{K=1}^{B} y_K \approx N\left(p_u, \frac{p_u(1 - p_u)}{B}\right)$$

One could choose B to achieve any desired level of accuracy.

If $m = m_1 \times \cdots \times m_n$ was small, we could calculate $\bar{t}_{.K}$ exactly. If m is large, we will need to estimate $\bar{t}_{.K}$. Since we only care if $\bar{t}_{.K} \geq \bar{t}_{.0}$, the precision of our estimate of $\bar{t}_{.K}$ should depend on how far it is from $\bar{t}_{.0}$. Thus we will allow M_K, the number of randomly selected WC resamples for permutation Π_K to vary with K. Let $\boldsymbol{J}_J$ be the Jth Monte Carlo draw from the full set of possible resamples, $\boldsymbol{j}_1, \ldots, \boldsymbol{j}_m$. Also let Π_K be the Kth Monte Carlo draw from the full set of possible permutations, $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_b$. We write the test statistic associated with the Jth resample and Kth permutation as $T_{JK} = T\{\boldsymbol{X}(\boldsymbol{J}_J), \mathbf{Z}(\boldsymbol{M}_K)\}$. We use upper case T, $\boldsymbol{J}$, J, Π, K to emphasize that they are random variables and not fixed constants as for the exhaustive case. For the Kth selected permutation, we estimate y_K by first drawing M_K WC resamples for this permutation, say, $\boldsymbol{J}_{K1}, \ldots, \boldsymbol{J}_{KM_K}$. We then form $\bar{T}_{.K} = M_K^{-1}\sum_{J=1}^{M_K} T\{\boldsymbol{X}(\boldsymbol{J}_{KJ}), \boldsymbol{Z}(\boldsymbol{\Pi}_K)\}$, and also estimate $\bar{t}_{.0}$ for these same resamples,

$\bar{T}_{.0} = M_K^{-1} \sum_{J=1}^{M_K} T\{\boldsymbol{X}(\boldsymbol{J}_{KJ}), \boldsymbol{Z}(\boldsymbol{\pi}_0)\}$. We then form $Y_K = I(\bar{T}_{.K} - \bar{T}_{.0} \geq 0)$. Note that for each support point we reestimate $\bar{t}_{.0}$. If we used a common estimate for all resamples, then $Y_1, \ldots, Y_B$ would not be independent. Although inefficient, using a different $\bar{T}_{.0}$ for each response ensures that the $Y_1, \ldots, Y_B$ are independent and thus allows easy approximation of $\mathrm{var}(\bar{Y})$.

We define the (upper) Monte Carlo WCRP permutation p value as $\hat{p}_u = B^{-1} \sum_{K=1}^{B} Y_K$. In practice, we will often use the asymptotically (on B) equivalent form $\tilde{p}_u = (1 + \sum Y_K)/(B+1)$ because it bounds the p value away from zero and ensures proper size (Fay and Follmann, 2002). We want to select each M_K so that each Y_K has a high probability of equaling y_K.

Note that for a fixed K we can think of $\bar{T}_{.K}$ as a sample mean based on M_K random draws from the column of Table 2 where $\boldsymbol{\pi} = \boldsymbol{\Pi}_K$. That is, from the discrete distribution with support points $t_{1K}, \ldots, t_{mK}$. This also applies to $\bar{T}_{.0}$. Without loss of generality, let $\boldsymbol{\Pi}_K = \boldsymbol{\pi}_k$. By the CLT, as $M_K \to \infty$, $\bar{T}_{.K} - \bar{T}_{.0}$ is approximately normal with mean $\bar{t}_{.k} - \bar{t}_{.0}$ and variance σ_{k0}^2 / M_K, where $\sigma_{k0}^2 = \sum_{j=1}^{m} \left\{ t_{jk} - t_{j0} - (\bar{t}_{.k} - \bar{t}_{.0}) \right\}^2 / m$.

For each K we only need to know if $\bar{t}_{.K} - \bar{t}_{.0}$ is positive or negative. To achieve this with high probability, we propose the 6-tens algorithm, an ad hoc procedure that chooses M_K so that

$$|Z_K| \equiv \left| \frac{\bar{T}_{.K} - \bar{T}_{.0}}{\sqrt{\hat{\sigma}_{K0}^2 / M_K}} \right|$$

is large, and thus the sign of $\bar{T}_{.K} - \bar{T}_{.0}$ equals that of $\bar{t}_{.K} - \bar{t}_{.0}$ with high probability.

4.2. 6-Tens Algorithm

1. Set $M_K = 10$ for $K = 1, \ldots, B$. For each of the B selected permutations $\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_B$, randomly draw 10 resamples, say $\boldsymbol{J}_{K1}, \ldots, J_{K10}$, and calculate

   $$D_{K\ell} = T\{\boldsymbol{X}(\boldsymbol{J}_{K\ell}), \boldsymbol{Z}(\boldsymbol{\Pi}_K)\} - T\{\mathbf{X}(\boldsymbol{J}_{K\ell}), \boldsymbol{Z}(\boldsymbol{\Pi}_0)\}$$

   for $\ell = 1, \ldots, 10$. Calculate the mean and sample variance of the Ds, say $\bar{D}_K$ and S_{K0}^2, and form

   $$Z_K = \frac{\bar{D}_K}{\sqrt{S_{K0}^2 / 10}}$$

   Let $\mathscr{K}_6$ be the set of Ks for which $|Z_K| < 6$.

2. For each $K \in \mathscr{K}_6$ set $M_K = M_K \times 10$, randomly generate M_K differences $D_{K\ell}$ and form

$$Z_K = \frac{\bar{D}_K}{\sqrt{S_{K0}^2 / M_K}}$$

   Again let $\mathscr{K}_6$ be the set of Ks for which $|Z_K| < 6$.

3. Repeat step 2 until either $\mathscr{K}_6$ is empty, or $M_K > M_{\max}$, some prespecified maximum value.

4. Let $Y_K = I(Z_K > 0)$, where Z_K is the final value from step 3, and estimate the p value as $\tilde{p}_u = (B + 1)^{-1}(1 + \sum_{K=1}^{B} Y_K)$.

One can show that the expected Monte Carlo p value does not equal the exact p value or $E[\bar{Y}] \neq \bar{y}$, so bias can be a problem in principle. With the 6-tens algorithm we try to essentially eliminate bias by making each Y_K a very good estimate of y_K. In the Appendix the bias of $\bar{Y}$ is explored in more detail, along with a more formal justification of the 6-tens algorithm.

4.3. Evaluation of the 6-Tens Algorithm

To empirically evaluate the 6-tens algorithm, we conducted a simple simulation. We assumed that the true support points $\bar{t}_{\cdot k}$ followed a normal$(0, \tau^2)$ distribution and that the variance of $T_{jk} \mid \bar{t}_{\cdot k}$ was σ^2. We fixed $\bar{t}_{\cdot 0} > 0$ so the WCRP upper p value was $\Phi(-\bar{t}_0/\tau)$. We pretended that σ^2 was known and did not estimate it. We randomly selected $B = 1000$ permutations (true support points) and calculated $\bar{y}$, the “true” p value for that choice of permutations/support points. Note that $\bar{y} \approx \Phi(-\bar{t}_0/\tau)$. This approximation becomes exact as $B \to \infty$. We fixed this set of support points/permutations and then repeated the 6-tens algorithm 10,000 times. For each of these 10,000 repeats, we calculated $\bar{Y}$ and $\widehat{\text{var}}[\bar{Y}]$. The objective was to see whether the 6-tens algorithm worked well in terms of bias, mean squared error (MSE), and accuracy of the variance approximation (10), for a specific set of permutations.

Table 3 shows an unsurprising decrease in bias as max M_K increases. Importantly, the average estimated variance appears quite close to the actual sample variance, suggesting the approximation of (10) will be useful in setting sample size. While the scenarios here are limited, a maximum of 10, 000 seems to be a reasonable choice for the 6-tens algorithm.

5. A MORE GENERAL FORM OF THE TEST

Although we have spoken only about conditional permutation tests, we can generalize to other tests that are permutation-like. In this section, $\boldsymbol{Z}$ denotes a random vector from a distribution $f_{\boldsymbol{Z}}(\boldsymbol{z})$ where $\mathbf{z}$ is a realization of this variable. First we write the permutation tests in a form that will be easily generalized. Let $\bar{t}(\boldsymbol{z}) = m^{-1}\sum_{j=1}^{m} t\{\boldsymbol{X}(\boldsymbol{j}_j), \boldsymbol{z}\}$, so that $\bar{t}_{\cdot k} = \bar{t}\{\boldsymbol{z}(\boldsymbol{\pi}_k)\}$. Then we generalize the lower WCRP p value by writing it as

$$E_{\boldsymbol{Z}}[I\{\bar{t}(\boldsymbol{Z}) \le \bar{t}(z_0)\}] \tag{2}$$

where $E_{\boldsymbol{Z}}$ represents expectation over the n-dimensional random vector $\boldsymbol{Z}$, and $\boldsymbol{z}_0$ is the observed value of $\boldsymbol{z}$. In the usual setting, the distribution of $\boldsymbol{Z}$ has b equally likely support points, and $f_{\boldsymbol{Z}}$ is the permutation distribution on $\boldsymbol{Z}$, such that

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \begin{cases} \frac{1}{b} & \text{if } \boldsymbol{z} = \boldsymbol{z}_0(\boldsymbol{\pi}_k) \text{ for } k = 1, \ldots, b \\ 0 & \text{otherwise} \end{cases}$$

In that case we condition on the observed value $\boldsymbol{Z} = \boldsymbol{z}_0$ and consider permutations of that observed value. We can generalize to unconditional tests by allowing the null hypothesis distribution of $\boldsymbol{Z}$ to not be conditional on $\boldsymbol{z}_0$. For example, one could let $f_{\boldsymbol{Z}}$ be an n-dimensional continuous distribution with an infinite number of support points. In section 7.3 we give an example where $\mathbf{z}_0$ is an n-dimensional zero vector, and $f_{\boldsymbol{Z}}$ is continuous with

$$f_{\boldsymbol{Z}}(\boldsymbol{Z}) = \begin{cases} \frac{1}{2\pi n} & \text{if } Z_i \in [0, 2\pi] \text{ for all } i \in \{1, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

where $\boldsymbol{\pi} \approx 3.14 \ldots$ is the irrational number, not a permutation. The lower p value is still represented by Eq. (2).

6. SIMULATION

To evaluate the performance of different tests for clustered data, we conducted a small simulation for the two-sample setting. We generated data under a normal mixture model: $X_{ij} = \mu Z_i + b_i + \boldsymbol{\epsilon}_{ij}$, where $i = 1, \ldots, 2n$ indexes clusters, $j = 1, \ldots, m_i$ indexes observations within clusters, $Z_i = 0, 1$ is the treatment group indicator for cluster i, $\boldsymbol{m}_Z$ the cluster sizes for group $Z = 0, 1$, and there are n clusters per group. We generated $b_i \sim N\left(0, \sigma_b^2\right)$ and $\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$ with all b's and $\boldsymbol{\epsilon}$'s independent. We considered 36 different situations as follows:

- $\left(\sigma_b^2, \sigma_e^2\right) = (0, 1), (1, 1)$, and $(1, 0)$. These correspond to within-cluster correlations of 0, 0.5, and 1.
- $\mu = 0$ or a where a was chosen to so that EOP had approximately 80% power.
- Three types of cluster sizes were considered: rectangular: $m_i = 5$, triangular: $m_i = i$, and informative: $m_i = 1 \times I(b_i \le 0) + 10 \times I(b_i > 0)$
- Number of clusters per group = 5 or 15.

We evaluated three tests, single WCR permutation (SWCRP), exhaustive WCR permutation (EWCRP), and the Wilcoxon rank sum test of Datta and Satten (2005) (DS). For SWCRP the test statistic is the difference means of the first observation in each cluster;

$$T_0^S = \frac{\sum_{i=1}^{2n} Z_i X_{i1}}{\sum_{i=1}^{2n} Z_i} - \frac{\sum_{i=1}^{2n} (1 - Z_i) X_{i1}}{\sum_{i=1}^{2n} (1 - Z_i)}$$

while for EWCRP the test statistic is the difference in cluster means,

$$T_0^E = \frac{\sum_{i=1}^{2n} Z_i \overline{X}_i}{\sum_{i=1}^{2n} Z_i} - \frac{\sum_{i=1}^{2n} (1 - Z_i) \overline{X}_i}{\sum_{i=1}^{2n} (1 - Z_i)}$$

The test of Datta and Satten is a WCR version of the Wilcoxon test and is based on

$$W_{\boldsymbol{J}}^* = \frac{1}{M} \sum_{i=1}^{n} Z_i \mathrm{Rank}\left(X_{i\boldsymbol{J}(i)}\right)$$

the sum of the ranks for a randomly selected WCR, $\boldsymbol{J}$. The test is based on forming $E\left[W_{\boldsymbol{J}}^*\right]$, where the expectation is over $\boldsymbol{J}$, and then standardizing $E\left[W_{\boldsymbol{J}}^*\right]$ and comparing the standardized test statistic to an asymptotically valid standard normal null distribution.

For all permutation-based test statistics, we simulated the permutation distribution by scrambling the Z_is 299 times and rejected the one-sided null at α = .05 if the simulated p value $\{1 + \sum_{p=1}^{299} I\left(T_p \geq T_0\right)\}/300$ was smaller than .05, where T_p is the pth permuted test statistic. Note that T_0^E exhausts all within-cluster resamples, but uses Monte Carlo approximation for the permutation distribution.

From Table 4, we see that both EWCRP and SWCRP always control the type I error rate, as they must. For Datta and Satten, there is some minor inflation for n = 15 and more substantial inflation with n = 5. This is entirely expected as Datta and Satten's null distribution is based on an asymptotic argument. Thus, as always, if control of the type I error rate is a major concern, exact methods such as EWCRP have appeal. For n = 15, the power of Datta and Satten's test is quite similar to EWCRP under rectangular and triangular data structure for the m_is. Under an informative cluster size, EWCRP is more powerful for ρ = .5 and 1. For n = 5, the EWCRP has similar or somewhat less power than Datta and Satten for the rectangular and triangular data structures—presumably a consequence of the coarseness of the permutation distribution for small n, and the modest type I error rate inflation of Datta and Satten. For the informative cluster size setting, the EWCRP has better power for ρ = .5 and 1.

7. EXAMPLES

7.1. Viral Blips and HIV Infection

The introduction of highly active antiretroviral therapy (HAART) has had a profound impact on the treatment of patients with HIV. Potent combinations of drugs can often render the load of virus (VL) circulating in the blood below the limit of detection of common assays, currently 50 copies/ml. However, even during periods of sustained viral suppression, sometimes “blips” occur—occasions when the amount of virus is above the limit of detection. The meaning of such blips is controversial and under investigation (Di Mascio et al., 2004). Part of their investigation focused on whether the occurrence of blips was associated with a change in the number of CD4 cells, cells of the adaptive immune system that both fight and are infected by HIV.

One approach to this problem would be to build a fairly complicated model for the repeated CD4 and VL data. If such a model could be correctly specified it could then be used to draw conclusions about various hypotheses, including the effect of blips on CD4 and CD8 cells. A different approach is to test specific hypotheses using simple statistics that are easily described and understood by a medical audience.

If the increase in virions during a blip impairs the recovery of CD4 cells, the relative change in CD4 counts on two successive visits starting with a nonblip and ending with a blip, say, $R_{01} = (\mathrm{CD4}_j - \mathrm{CD4}_{j-1})/\mathrm{CD4}_{j-1}$, should tend to be lower than the analogous relative change ending in a nonblip, say, R_{00}. If each patient provided a single pair R_{01}, R_{00}, then one could form $D = R_{01} - R_{00}$. The associated $D_1, \ldots, D_n$ could be the data for a paired difference t-test. However, patients generally have sustained periods of viral suppression and thus many R_{00}s and several R_{01}s. In our dataset there are 44 patients with acute HIV infection with both 00 and 01 couples. The total number of couples was 1094 and the mean (SD) number of couples per person was 24.5 (9.9). The mean (SD) number of 01 couples was 2.5 (1.9).

Here exhaustive WCR permutation corresponds to taking the sample average for each patient and using these averages in a permutation test. Thus we form $D = \bar{R}_{01} - \bar{R}_{00}$ for each patient and then take the average of these average differences to form our test statistic:

$$\bar{t}_{\cdot k} = \sum_{i=1}^{n} \frac{Z_i(\boldsymbol{\pi}_k) D_i}{n}$$

where $Z_i = \pm 1$ and $Z(\boldsymbol{\pi}_1), \ldots, Z(\boldsymbol{\pi}_b)$ enumerates the $b = 2^n$ possible signs of an n-vector. There were 44 patients who had both 00 and 01 couples during acute infection. The exact upper p value here is .1793, and took about 30 s to calculate on a desktop computer.

A random effects model could also be fit to these data, where

$$R_{ij} = b_i + \Delta Z_{ij} + \epsilon_{ij}$$

where b_i, ϵ_{ij} are independent normals with mean zero and variances τ^2 and σ^2, respectively, R_{ij} is the relative change in CD4 count for patient i for the jth couple, and Z_{ij} is the indicator for whether this couple started with a blip in viremia. This approach makes parametric assumptions and the p value for testing $\Delta = 0$ is .14.

7.2. Heart Function in HIV-Infected Children

A retrospective study of 133 HIV-infected children was conducted at the National Institutes of Health to assess the contributions of progressive HIV disease (as measured by CD4 counts and percentages), vertical transmission, and azidothymidine (AZT) on cardiac function (Domanski et al., 1995). Cardiac function was measured by the fractional shortening of the left ventricle, which is essentially the fractional decrease in volume during a heartbeat, and larger values are desirable. Table 5 provides some descriptive statistics for the variables of interest, including averages of all the measurements and averages of each patient's values.

Of primary interest was the effect of the covariates on fractional shortening:

$$E\left[\mathrm{FS}_{ij}\right] = \beta_0 + \ \mathrm{age}_{ij}\beta_1 + \mathrm{CD}_{ij}\beta_2 + \%\mathrm{CD}_{ij}\beta_3 + I(\mathrm{AZT})_{ij}\beta_4 + I(\mathrm{vertical})_i\Delta \qquad (3)$$

where $i = 1, \ldots, n = 133$ and $j = 1, \ldots, m_i$. Thus i, j is the jth visit of the ith patient and age_{ij}, $\mathrm{CD4}_{ij}$ are the age and CD4 count, % $\mathrm{CD4}_{ij}$ is the percentage of CD4 cells among the white blood cells, $I(\mathrm{AZT})_{ij}$ is 1 if on AZT, and $I(\mathrm{vertical})_i$ is 1 if infected by the mother.

For regression, there are different approaches to permutation analysis (see Edgington, 1995; Kennedy and Cade, 1996; Manly, 1997). We use "permutation under the reduced model" (Freedman and Lane, 1983) as recommended by Anderson and Legendre (1999) to evaluate the effect of I(vertical). Permutation methods can be more powerful than the normal t-test approach if the errors are non-normal (Anderson and Legendre, 1999). Using an obvious notation, we write the regression model of Eq. (3), for a single resample $\boldsymbol{J}$ of n independent data points as

$$X_{J(i)} = \boldsymbol{W}'_{J(i)}\boldsymbol{\beta} + Z_{J(i)}\Delta + \epsilon_{J(i)} \qquad (4)$$

for $i = 1, \ldots, n$, where Z is I(vertical), and $\epsilon_{J(i)}$ is an error term. As a test statistic, we use the usual t-statistic for testing $\Delta = 0$ based on least-squares regression of X on $\boldsymbol{W}$, Z:

$$T\{\boldsymbol{X}(\boldsymbol{J}), \boldsymbol{Z}(\boldsymbol{\Pi}_0)\} = \frac{\widehat{\Delta}(\boldsymbol{J}, \ \boldsymbol{\Pi}_0)}{\sqrt{\widehat{\mathrm{var}}\{\widehat{\Delta}(\boldsymbol{J}, \ \boldsymbol{\Pi}_0)\}}} \qquad (5)$$

To obtain a permutation distribution for this test, we randomly select a permutation Π and perform an initial regression without Z:

$$X_{J(i)} = \boldsymbol{W}'_{J(i)}\boldsymbol{\beta}^r + \epsilon^r_{J(i)} \qquad (6)$$

The permuted residuals, $\hat{\epsilon}^r_{J(\Pi(i))}$ are added to the predicted Xs from this regression to form X*

$$X^*_{J(i)} = \boldsymbol{W}'_{J(i)}\widehat{\boldsymbol{\beta}}^r + \hat{\epsilon}^r_{J(\Pi(i))} \qquad (7)$$

where $\Pi(i)$ is the ith element of the permutation vector $\boldsymbol{\Pi}$. Then the newly created X^* is regressed on the original $\boldsymbol{W}$, Z:

$$X^*_{J(i)} = \boldsymbol{W}'_{J(i)}\boldsymbol{\beta} + Z_{J(i)}\Delta + \epsilon_{J(i)} \tag{8}$$

and based on this regression, the usual t-statistic for testing $\Delta = 0$ is formed: $T\{\boldsymbol{X}(\boldsymbol{J}), \boldsymbol{Z}(\Pi)\}$ from Eq. (5). For Monte Carlo WCR with the 6-tens algorithm, the preceding steps (6), (7), and (8) are repeated M_K times for both $\boldsymbol{\Pi}_K$ and $\boldsymbol{\Pi}_0$. One can show that the $T\{\boldsymbol{X}(\boldsymbol{J}), \boldsymbol{Z}(\Pi_0)\}$ as defined in Eq. (5) based on Eq. (4) also obtains from the procedure defined by Eqs. (6), (7), and (8) with $\boldsymbol{\Pi} = \boldsymbol{\Pi}_0$.

Using the preceding methods we tested the relationship between fractional shortening and vertical transmission while controlling for the effects of the four confounders. We used the 6-tens algorithms with $B = 1999$ and $M_{Max} = 10^5$ and obtained a lower p value of $p = .4745$ with standard error of 0.011, where 99.8% of the variance estimate is due to the second term of (10):$\bar{Y}(1 - \bar{Y})/\boldsymbol{B}$. This calculation took approximately 13 h on a PC.

One can also analyze these data using GEE (Liang and Zeger, 1986). We postulate a working independence correlation matrix and calculate a Wald statistic for Δ of −.32, which corresponds to a p value of .37. This method uses an asymptotic null distribution in contrast to WCRP, which can be liberal if the number of clusters is small or the covariates are not balanced (Fay and Graubard, 2001).

7.3. Correlated Angular Measurements

Follmann and Proschan (1999) discuss the problem of testing uniformity with correlated angular data. Of interest was whether times of seizures have a circadian pattern, or whether they are uniformly distributed on the 24-hour clock. Data from 12 patients were provided; one patient had one seizure, while another had 36 with a cluster of seizures a little before midnight.

Follmann and Proschan (1999) introduced a definition of uniformity called rotation invariance and provided several tests for rotation invariance that explicitly allowed for arbitrary clustering/correlation of angles within a cluster. Developing new methodology can be time-consuming, and it is also nice to have simple tools to attack complicated problems. The basic data here is the long vector

$$\boldsymbol{X} = \left(X_{11}, \ldots, X_{1m_1}, X_{21}, \ldots, X_{2m_2}, \ldots, X_{n1}, \ldots, X_{nm_n}\right),$$

where X_{ij} is the jth seizure time on the 24-hour clock (in radians) for individual i. A seizure at 6:00 a.m. = 0 radians and a seizure at noon = $3\boldsymbol{\pi}/2$ radians.

A standard test of uniformity for angular data is the Rayleigh test. For a single “resample” $\boldsymbol{J}$, the Rayleigh test statistic is

$$T\{\boldsymbol{X}(\boldsymbol{J})\} = \left[\frac{\sum \cos\{X_i(\boldsymbol{J})\}}{n}\right]^2 + \left[\frac{\sum \sin\{X_i(\boldsymbol{J})\}}{n}\right]^2$$

If $T\{\boldsymbol{X}(\boldsymbol{J})\}$ is close to 0, the angles are scattered, while if $T\{\boldsymbol{X}(\boldsymbol{J})\}$ is close to 1, the angles have a definite preferred direction. Mardia (1972) provides the exact null distribution of R and argues that for large n, $2nT\{\boldsymbol{X}(\boldsymbol{J})\}^2$ is approximately chi-square with 2 degrees of freedom. Though unnecessary here since the null distribution is known (Mardia, 1972), one could simulate the null distribution of $T\{\boldsymbol{X}(\boldsymbol{J})\}$ by forming

$$T\{\boldsymbol{X}(\boldsymbol{J}), \boldsymbol{Z}\} = \left[\frac{\sum \cos\{X_i(\boldsymbol{J}) + Z_i\}}{n}\right]^2 + \left[\frac{\sum \sin\{X_i(\boldsymbol{J}) + Z_i\}}{n}\right]^2$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, and the Z_i are independent and uniform $(0, 2\boldsymbol{\pi})$. One would generate many such $\boldsymbol{Z}$s and the associated Ts would comprise a simulated null reference distribution.

We apply the idea of Monte Carlo WCR permutation using the general setup described in Section 5. We randomly generate B n-vectors, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_B$, where $\boldsymbol{Z}_K = (Z_{K1}, \ldots, Z_{Kn})$ and each Z_{Ki} is uniform $(0, 2\boldsymbol{\pi})$. For each $\boldsymbol{Z}_K$, we form

$$D_{K\ell} = \sum_{J=1}^{M_K} T\{\boldsymbol{X}(\boldsymbol{J}_J), \boldsymbol{Z}_K\} - T\{\boldsymbol{X}(\boldsymbol{J}_J), \boldsymbol{0}\}$$

We applied the 6-tens algorithm with $B = 1999$ and $M_{Max} = 10^5$ and obtained an upper p value of 1528/2000. The estimate of the standard deviation, from Eq. (10), of that p value is .011, with 98.7% of the variance estimate due to the second term of Eq. (10): $\bar{Y}(1 - \bar{Y})/B$. This calculation took approximately 10 h on a personal computer (PC). The tests of Follmann and Proschan for the null hypothesis of rotation invariance all provide p values greater than .40.

8. SUMMARY

This paper has introduced a general method for obtaining exact permutation results in the presence of arbitrary within cluster correlation by using within cluster resampling. A fixed set of permutations is selected, and then for each cluster a single outcome is randomly selected and a test statistic calculated for each permutation. The procedure of drawing a single outcome from each cluster is repeated many times, and the test statistics for each specific permutation are averaged over all WC resamples. These averaged test statistics are used as the null reference distribution for the averaged test statistic. In practice, Monte Carlo methods to approximate the permutation distribution may be required. An algorithm is proposed where computational effort is focused to determine whether the support points of the null distribution fall to the left or right of the test statistic, thus ensuring an accurate approximate p value. Different examples are used to illustrate the broad application of WCR permutation. WCR permutation is a handy and simple way to apply permutation methods when within-cluster correlation is a nuisance.

Acknowledgments

We thank Michael Proschan and Zonghui Hu for providing helpful comments on an earlier draft of this article.

APPENDIX: BIAS OF $\bar{Y}$

It is instructive to explore how bias becomes a substantial problem if one uses a small number of Monte Carlo resamples. This analysis also demonstrates how the easy approach of doing single resamples and averaging the *p* values over many resamples is generally conservative.

For this section, suppose that $M_K = M$ for all K, and define $p(B, M) = \bar{Y}$ as the *p* value based on B random permutations and M random resamples. For a fixed dataset, $p(B, M)$ is still a random variable dependent on the specific resamples and permutations that were selected. Since $E[p(B, M)]$ is free of B, we use $E\{p(\infty, M)\} = E(\bar{Y})$ to denote the expected Monte Carlo WCR *p* value based on M resamples. To see the problem for small samples, consider again the made-up data of section 3.1. Suppose that we set $M = 1$ but enumerate the 20 permutations. Thus the *p* value is either 1/20 or 4/20 and both occur with equal probability. Formally we can write

$$E[p(\infty, 1)] = \frac{1}{20}Pr(\boldsymbol{J} = \boldsymbol{j}_1) + \frac{4}{20}Pr(\boldsymbol{J} = \boldsymbol{j}_2) = 5/40$$

which is larger than $p(\infty, \infty) = 2/20$. If we set $M = 2$, then $E[p(\infty, 2)] = 9/80$, which is still larger than $p(\infty, \infty)$. One can show that $E[p(\infty, M)]$ is not monotone for this small dataset, and one can produce datasets where $E[p(\infty, M)] \le p(\infty, \infty) = \bar{y}$. Thus, the bias can go in either direction.

To get a handle on the form of the bias in large samples, we worked out some asymptotics for the two-group test statistic of Eq. (1). By the CLT, $\bar{T}_{.K} \approx N\left(\bar{t}_{.K}, \sigma_K^2/M\right)$ where $\sigma_K^2 = \sum_{j=1}^{m}\left(t_{jK} - \bar{t}_{.K}\right)^2/m$. As $n \to \infty$ the distribution of $\bar{t}_{.K}$, based on the difference in means statistic, approaches that of a normal distribution with mean 0 and some variance, say τ^2, by the permutational CLT (Sen, 1985). One can also show that $\sigma_K^2 = \sigma^2$ for the difference in means statistic. Thus, if we knew $\bar{t}_{.0}$ we would have the following approximation for the expected *p* value based on M WC resamples:

$$E\{p(\infty, M)\} = E\left\{I\left(\bar{T}_{.K} - \bar{t}_{.0}\right)\right\} \approx 1 - \Phi\left(\frac{\bar{t}_{.0}}{\sqrt{\tau^2 + \sigma^2/M}}\right) \qquad (9)$$

where the second expectation is over K, the random permutation index.

With Eq. (9), we see that the approximate asymptotic bias decreases with M and increases with σ^2/τ^2. To graphically appreciate this point, consider Fig. 2, which is a plot of the exhaustive and Monte Carlo WCR *p* values for a large dataset assuming normality of the $\bar{t}_{\cdot K}$s, normality of $\bar{T}_{\cdot K} \mid \bar{t}_{\cdot K}$ and knowledge of $\bar{t}_{\cdot 0}$. The area to the right of $\bar{t}_{\cdot 0}$ based on the $\bar{T}_{\cdot K}$s is approximately $E\{p(\infty, M)\}$, and this is larger than the area to the right of $\bar{t}_{\cdot 0}$ based on the $\bar{t}_{\cdot K}$s, which is $p(\infty, \infty)$. Thus, the average of single WC resample *p* values should be larger than the exhaustive WCR *p* value.

APPENDIX: JUSTIFICATION OF THE 6-TEN'S ALGORITHM

Note that for the fixed permutations $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_B$, we have Y_K as independent Bernoullis with

$$P(Y_K = 1) = P\left(\frac{\bar{T}_{\cdot K} - \bar{T}_{\cdot 0}}{\sqrt{\sigma_{K0}^2 / M_K}} > 0\right) \approx \Phi\left(\frac{\bar{t}_{\cdot K} - \bar{t}_{\cdot 0}}{\sqrt{\sigma_{K0}^2 / M_K}}\right) \approx \Phi(Z_K)$$

In the preceding, the first approximation is by the CLT and the second one is because we replace the parameters $\bar{t}_{\cdot K}$, $\bar{t}_{\cdot 0}$, and σ_{K0}^2 with estimates. This approximation works because for each K (except those with $\bar{t}_{\cdot K} = \bar{t}_{\cdot 0}$), $|Z_K| \rightarrow \infty$ as $M_K \rightarrow \infty$. We thus can approximate the variance of $\tilde{p}_u$ for this set of permutations:

$$\widehat{\mathrm{var}}\left(\bar{Y} \mid \mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_B\right) = \sum_{K=1}^{B} \frac{\Phi(Z_K)\{1 - \Phi(Z_K)\}}{B^2}$$

The overall variance estimate for our *p* value $\bar{Y}$ is

$$\begin{aligned} \mathrm{var}(\bar{Y}) &= E\left\{\mathrm{var}\left(\bar{Y} \mid \mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_B\right)\right\} + \mathrm{var}\left\{E\left(\bar{Y} \mid \mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_B\right)\right\} \\ &\approx E\left[\sum_{K=1}^{B} \frac{\Phi(Z_K)\{1 - \Phi(Z_K)\}}{B^2}\right] + \mathrm{var}\left(\frac{1}{B} \sum_{K=1}^{B} y_K\right) \\ &\approx \sum_{K=1}^{B} \frac{\Phi(Z_K)\{1 - \Phi(Z_K)\}}{B^2} + \frac{\bar{Y}(1 - \bar{Y})}{B} \end{aligned} \qquad (10)$$

The second line approximation is because we approximate $P(Y_K = 1)$ by $\Phi(Z_K)$. The third line approximation is because we replace the expectation of a random variable with a realization of that random variable and we use $\bar{Y}$ to estimate the exhaustive WCR *p* value $\bar{y}$. Note that with the 6-tens algorithm, most $|Z_K|$s are large so $\bar{Y}(1 - \bar{Y})/B$ is the dominant term of Eq. (10).

REFERENCES

Anderson M, Legendre P (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

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

Cai J, Shen Y (2000). Permutation tests for comparing marginal survival functions with clustered failure time data. Statistics in Medicine 19:2963–2973. [PubMed: 11042626]

Datta S, Satten G (2005). Rank tests for clustered data. Journal of the American Statistical Association 100:908–915.

Datta S, Satten G (2008). A signed-rank test for for clustered data. Biometrics Association 64:501–507.

Di Mascio M, Markowitz M, Louie M, Hurley A, Hogan C, Simon V, Follmann D, Ho DD, Perelson AS (2004). Dynamics of intermittent viremia during highly active antiretroviral therapy in patients who initiate therapy during chronic versus acute and early human immunodeficiency virus type 1 infection. Journal of Virology 78(19):10566–10573. [PubMed: 15367623]

Domanski M, Sloas M, Follmann DA, Scallise PP, Tucker EE, Egan D, Pizzo P (1995). Effect of zidovudine and didanosine treatment on cardiac function in human immunodeficiency virus infected children. Journal of Pediatrics 127:137–146. [PubMed: 7608800]

Edgington ES (1995). Randomization Test. 3rd ed. New York: Marcel Dekker.

Fay MP, Follmann D (2002). Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. American Statistician 56:63–70.

Fay MP, Graubard BI (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 57(4):1198–1206. [PubMed: 11764261]

Fay MP, Shih JH (1998). Permutation tests using estimated distribution functions. Journal of the American Statistical Association 93:387–396.

Follmann DA, Proschan MA, Leifer E (2003). Multiple outputation: Inference for complex multivariate data by averaging analyses from univariate data. Biometrics 59:420–429. [PubMed: 12926727]

Follmann DA, Proschan MA (1999). A simple permutation-type method for testing circular uniformity with correlated angular measurements. Biometrics 55:782–791. [PubMed: 11315007]

Freedman D, Lane D (1983). A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics 1:292–298.

Gail M, Mark S, Carroll R, Green S, Pee D (1996). On design considerations and randomization-based inference for community intervention trials. Statistics in Medicine 15:1069–1092. [PubMed: 8804140]

Hoffman EB, Sen PK, Weinberg C (2001). Within-cluster resampling. Biometrika 88:1121–1134.

Kennedy PE, Cade BS (1996). Randomization tests for multiple regression. Communications in Statistics: Simulation and Computation 25:923–936.

Liang K-Y, Zeger S (1986). Longitudinal analysis using generalized linear models. Biometrika 73:13–22.

Manly BFJ (1997). Randomization and Monte Carlo Methods in Biology. London: Chapman and Hall.

Mardia KV (1972). Statistics of Directional Data. New York: Academic Press.

Rosner B, Glynn R, Lee M-L (2003). Incorporation of clustering effects for the Wilcoxon Rank Sum Test: A large-sample approach. Biometrics 59:1089–1098. [PubMed: 14969489]

Sen PK (1985). Permutational central limit theorem. In: Kotz Balakrishnan, Read Vidakovic, Johnson, eds. Encyclopedia of Statistical Sciences. 2nd ed. pp. 6069–6073.

Williamson J, Datta S, Satten G (2003). Marginal analyses of clustered data when cluster size is informative. Biometrics 59:36–42. [PubMed: 12762439]

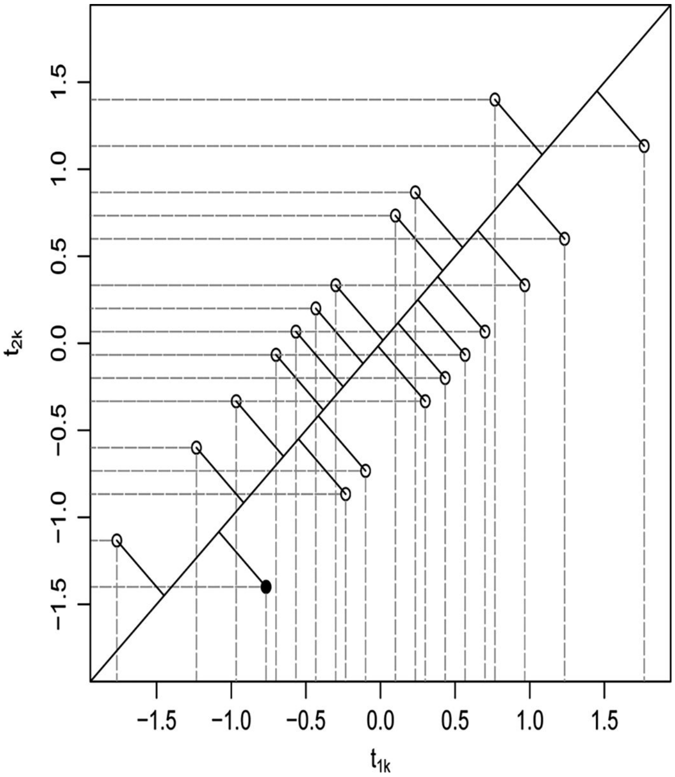

Figure 1.

Geometric representation of permutation tests for the data of section 3.1, where everyone provides one measurement except the last person in group 0 who gives 2. Each (x, y) dot corresponds to the pair of test statistics for one of the 20 possible permutations. The x-value (y-value) is based on the first (second) measurement from this last person, the permutation distribution is the horizontal (vertical) projection, and the one-sided p value is 4/20 (1/20). Exhaustive WCR permutation is equivalent to projecting the pair of test statistics onto the 45-degree line. Here, the WCR permutation one-sided p value is 2/20.

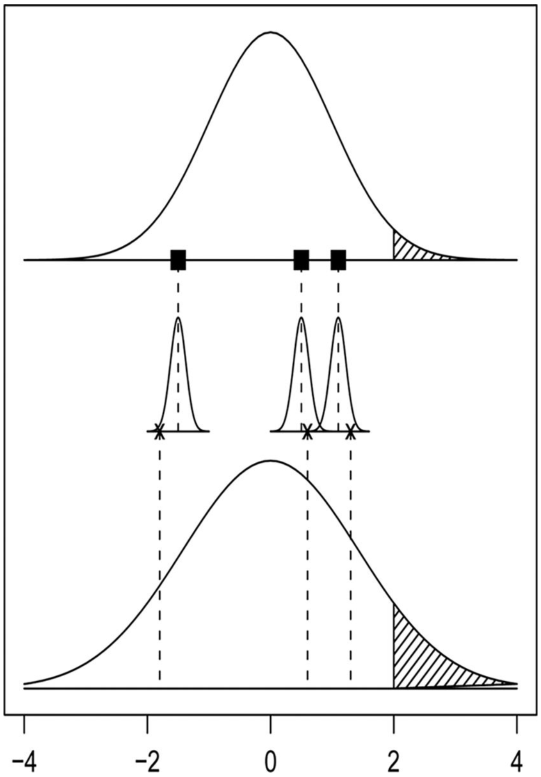

Figure 2.

Illustration of the sources of variability in the averaged permutation distribution for the difference in means statistic. The top panel shows the approximate Gaussian (0, τ^2) distribution of the $\bar{t}_k$s, i.e., exhaustive WCR permutation distribution. Conditional on $\bar{t}_k$, $\bar{T}_{\cdot k}$ has an approximate Gaussian $\left(\bar{t}_k, \sigma^2/M\right)$ distribution. Thus unconditionally, the Monte Carlo WCR permutation distribution (bottom panel) is more spread out than the exhaustive permutation distribution. The shaded area of the bottom panel is larger than for the top panel, illustrating that the Monte Carlo *p* value is likely to be larger than the exhaustive *p* value.

Table 1

Schematic representation of exhaustive WCR permutation

Cluster	Covariate	Permutation	Outcome						
1	$Z_1 = 1$	Z_3	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
2	$Z_2 = 1$	Z_1	X_{21}						
3	$Z_3 = 1$	Z_6	X_{31}	X_{32}	X_{33}	X_{34}			
4	$Z_4 = 0$	Z_4	X_{41}	X_{42}	X_{43}				
5	$Z_5 = 0$	Z_2	X_{41}	X_{52}	X_{53}	X_{54}	X_{55}	X_{56}	
6	$Z_6 = 0$	Z_5	X_{71}	X_{62}					
$\boldsymbol{j} = (3, 1, 2, 3, 6, 1)$									
$\boldsymbol{Z}\ (\boldsymbol{\pi}) = (Z_3, Z_1, Z_6, Z_4, Z_2, Z_5)$									
$\boldsymbol{X}(\boldsymbol{j}) = (X_{13}, X_{21}, X_{32}, X_{43}, X_{56}, X_{71})$									
$t\{\boldsymbol{X}(\boldsymbol{j}), \boldsymbol{Z}(\boldsymbol{\pi})\} = (1/3)\sum_{i=1}^{3} Z_i(\boldsymbol{\pi})X_i(\boldsymbol{j}) - (1/3)\sum_{i=1}^{3}\{1 - Z_i(\boldsymbol{\pi})\}X_i(\boldsymbol{j})$									

Note. Within each cluster an outcome is randomly selected and denoted by a box. The indices of the randomly selected outcomes from each cluster are given by $\boldsymbol{j}$. A permutation of the covariates $\boldsymbol{\pi}$ is given by $\mathbf{Z}(\boldsymbol{\pi})$. A WCR is given by $\boldsymbol{X}(\boldsymbol{j})$. A difference in means test statistic for permutation $\boldsymbol{\pi}$ and WCR $\boldsymbol{j}$ is provided in the bottom row.

Table 2

Matrix of test statistics for exhaustive WCR permutation (WCRP)

Within-cluster resample (*j*)	**Permutation (*k*)**				**Original data**
	$\boldsymbol{\pi}_{\mathbf{1}}$	$\boldsymbol{\pi}_{\mathbf{2}}$	**…**	$\boldsymbol{\pi}_{\boldsymbol{b}}$	$\boldsymbol{\pi}_{\mathbf{0}}$
$\boldsymbol{j}_1$	t_{11}	t_{12}	…	t_{1b}	t_{10}
$\boldsymbol{j}_2$	t_{21}	t_{22}	…	t_{2b}	t_{20}
.	.	.		.	.
.	.	.		.	.
.	.	.	t_{jk}	.	.
.	.	.		.	.
$\boldsymbol{j}_m$	t_{m1}	t_{m2}	…	t_{mb}	t_{m0}
Average	$\bar{t}_{.1}$	$\bar{t}_{.2}$	…	$\bar{t}_{.b}$	$\bar{t}_{.0}$

Note. For each of m within cluster resamples, the test statistic is calculated over the same b permutations. The final row provides the single permutation distribution for the averaged test statistic. The "permutation" $\boldsymbol{\pi}_0$ denotes the original unpermuted setting.

Table 3

Simulated performance of the 6-tens algorithm

σ	$\bar{t}_{\cdot 0}$	**max(M_K)**	$\bar{y}$	**%$\lvert Z_K \rvert$s < 6**	**bias**	**MSE**	$\widehat{\text{var}}[\bar{Y}]$	$S^2(\bar{Y})$
1	1	100	.144	.415	4.2×10^{-3}	3.5×10^{-5}	1.9×10^{-5}	1.8×10^{-5}
1	1	10000	.161	.040	-5.0×10^{-5}	1.9×10^{-6}	1.7×10^{-6}	1.6×10^{-6}
1	1	1000000	.156	.003	-2.6×10^{-5}	2.6×10^{-8}	6.2×10^{-8}	2.5×10^{-8}
1	2	100	.020	.133	2.5×10^{-3}	1.1×10^{-5}	4.6×10^{-6}	4.8×10^{-6}
1	2	10000	.019	.006	6.6×10^{-4}	9.3×10^{-7}	4.0×10^{-7}	4.9×10^{-7}
1	2	1000000	.017	.000	0	0	2.6×10^{-14}	0
.1	1	100	.174	.040	-1.5×10^{-3}	3.4×10^{-6}	1.5×10^{-6}	1.2×10^{-6}
.1	1	10000	.175	.004	-1.7×10^{-6}	9.3×10^{-9}	5.2×10^{-8}	9.3×10^{-9}
.1	1	1000000	.158	.0	0	0	4.7×10^{-15}	0
.1	2	100	.022	.009	-1.9×10^{-4}	3.8×10^{-7}	3.7×10^{-7}	3.4×10^{-7}
.1	2	10000	.023	0	0	0	1.1×10^{-15}	1.2×10^{-35}
.1	2	1000000	.020	0	0	0	1.8×10^{-15}	0

Note. Each line represents a summary of 10,000 estimates ($\bar{y}$) of a single WCR permutation *p* value $\bar{y}$ based on a specific set of 1000 permutations (support points). Bias is the average of the 10,000 differences $\bar{Y} - \bar{y}$. MSE is the average of the 10,000 squared differences $(\bar{Y} - \bar{y})^2$.$S^2(\bar{Y})$ is the sample variance of the $\bar{Y}$s.

Table 4

Proportion of rejections for different tests of equality of two distributions

m_is	**μ**	**ρ**	**SWCRP**	**EWCRP**	**DS**	***n*/2**
Rectangular	0	0	0.0510	0.0505	0.0538	15
Rectangular	0	.5	0.0504	0.0495	0.0529	15
Rectangular	0	1	0.0519	0.0519	0.0545	15
Rectangular	a	0	0.3080	0.8180	0.8164	15
Rectangular	a	.5	0.5466	0.7341	0.7451	15
Rectangular	a	1	0.7299	0.7299	0.7260	15
Triangular	0	0	0.0504	0.0508	0.0540	15
Triangular	0	.5	0.0498	0.0503	0.0532	15
Triangular	0	1	0.0505	0.0505	0.0526	15
Triangular	a	0	0.3045	0.7874	0.7719	15
Triangular	a	.5	0.5456	0.7278	0.7338	15
Triangular	a	1	0.7272	0.7272	0.7165	15
Informative	0	0	0.0503	0.0503	0.0539	15
Informative	0	.5	0.0501	0.0504	0.0375	15
Informative	0	1	0.0499	0.0499	0.0180	15
Informative	a	0	0.5819	0.8061	0.7924	15
Informative	a	.5	0.6471	0.7438	0.6858	15
Informative	a	1	0.7307	0.7307	0.5751	15
Rectangular	0	0	0.0472	0.0487	0.0615	5
Rectangular	0	.5	0.0486	0.0478	0.0621	5
Rectangular	0	1	0.0478	0.0478	0.0471	5
Rectangular	a	0	0.2736	0.7547	0.7943	5
Rectangular	a	.5	0.6651	0.6651	0.6512	5
Rectangular	a	1	0.4932	0.6723	0.7313	5
Triangular	0	0	0.2744	0.4847	0.5148	5
Triangular	0	.5	0.4900	0.6008	0.6474	5
Triangular	0	1	0.6667	0.6667	0.6797	5
Triangular	a	0	0.0468	0.0463	0.0624	5
Triangular	a	.5	0.0487	0.0465	0.0607	5
Triangular	a	1	0.0480	0.0480	0.0558	5
Informative	0	0	0.0484	0.0483	0.0653	5
Informative	0	.5	0.0471	0.0481	0.0436	5
Informative	0	1	0.0484	0.0484	0.0187	5
Informative	a	0	0.5223	0.7567	0.7518	5
Informative	a	.5	0.5838	0.6813	0.6572	5
Informative	a	1	0.6653	0.6653	0.5437	5

Note. Within clusters, data are multivariate normal with correlation ρ. The mean cluster difference between groups is μ and varies for different scenarios. Cluster sizes and tests are described in the text.

Table 5

Descriptive statistics for the study of factors on the fractional shortening of the left ventricle in children with AIDS (Domanski et al., 1995)

	Σm_i = 486 measurements			
Variable	**Mean**	**Minimum**	**Maximum**	**Mean of *n* cluster averages**
Fractional shortening	0.34	0.05	0.59	0.34
CD4 cells	455.6	0	3421	544.8
Percent CD4 cells	16.1	0	63	18.1
I(AZT)	0.53	0	1	0.53
I(vertical)	0.51	0	1	0.60
Age	7.2	.33	17.11	6.4