ASYMPTOTIC PROPERTIES OF A TWO SAMPLE RANDOMIZED TEST FOR PARTIALLY DEPENDENT DATA

By

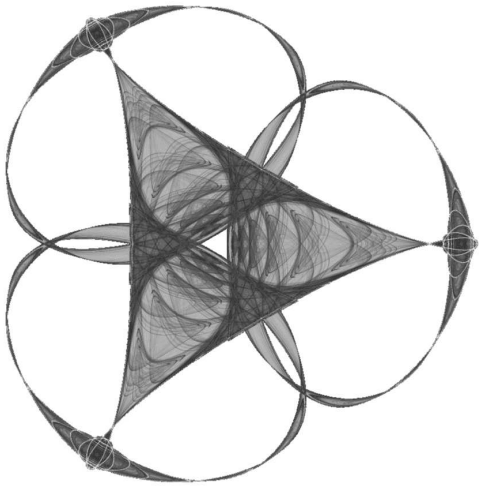
Grzegorz A. Rempala

and

Stephen W. Looney

IMA Preprint Series # 1963

(February 2004)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Asymptotic Properties of a Two Sample Randomized Test for Partially Dependent Data

Grzegorz A. Rempala
Institute for Mathematics and Its Applications
University of Minnesota
and
Department of Mathematics
University of Louisville
and
Stephen W. Looney
Department of Biostatistics
Louisiana State University

revised draft copy - do not cite
February 2004

Abstract

We are concerned with an issue of asymptotic validity of a non-parametric randomization test for the two sample location problem under the assumption of partially dependent observations, in which case the validity of the usual permutation t-test breaks down. We show that a certain modification of the permutation group used in the randomization procedure yields an unconditional asymptotically valid test in the sense that its probability of Type I error tends to the nominal level with increasing sample sizes. We show that this unconditional test is equivalent to the one based on a linear combination of two- and one-sample t-statistics and enjoys some optimal power properties. Finally, we present an example of application of the test in a medical study on functional status assessment at the end of life.

Key Words and Phrases: Consistency, hypothesis test, randomization test, two sample problem, relative efficiency.

Running Head: Randomized Test for Partially Dependent Data

1 Introduction

The construction of a randomized test was first described in the celebrated works of Fisher (1935) and Pitman (1937) as a powerful tool to yield exact nonparametric tests, i.e., those having their level equal exactly to the nominal level. Since then the research on randomized tests has been quite extensive, resulting in both finite sample and asymptotic results. For instance, for finite samples Lehmann and Stein (1949) have shown that in several situations including the classical two sample location problem, randomization tests enjoy certain optimality properties. On the other hand, Hoeffding (1952) has shown that in many cases the asymptotic power of randomization tests is equal to that of the standard optimal procedures based on parametric models. More recently, Romano (1990) has shown that some randomization tests, including, e.g., permutation t-tests in the two sample location problem, are asymptotically valid (i.e., their probability of Type I error tends to the nominal level with increasing sample sizes) even when the assumptions leading to their construction are violated.

The general idea of a randomization test can be summarized as follows (in notation borrowed from Hoeffding 1952). Let $\mathbf{G}$ be a finite group of M transformations of a sample space $\mathbb{X}$ into itself. Based on the data X taking values in $\mathbb{X}$ we desire to test the null hypothesis H_0 that the underlying probability law P which generates X belongs to a certain family of distributions Ω_0. We assume that the null hypothesis implies that the distribution of X is invariant under the transformations in $\mathbf{G}$. That is, if P belongs to Ω_0 then for every g in $\mathbf{G}$, gX and X have the same distribution. Let $T(X)$ be a test statistic for testing H_0. For every x in $\mathbb{X}$ let

$$T^{(1)}(x) \leq T^{(2)}(x) \leq \ldots \leq T^{(M)}(x) \tag{1}$$

be the ordered values of $T(gx)$ as g varies in $\mathbf{G}$. Given a nominal level α ($0 < \alpha < 1$), let k be

defined by

$$k = M - [M\alpha] \tag{2}$$

where $[M\alpha]$ denotes the largest integer less than or equal to $M\alpha$. Let also

$$a(x) = \frac{M\alpha - M^{+}(x)}{M^{0}(x)}$$

where $M^{+}(x)$ and $M^{0}(x)$ denote the number of values of $T^{(i)}$ for $i = 1, \ldots, M$ that are greater than and equal to $T^{(k)}$, respectively. It is well known (see e.g. Hoeffding 1952) that if we define the randomization test function $\phi(x)$ to be equal to one, zero, or $a(x)$ according to whether $T(x) > T^{(k)}(x)$, $T(x) < T^{(k)}(x)$, or $T(x) = T^{(k)}(x)$, then, under the assumption that the distribution P is invariant under all g in $\mathbf{G}$, we have $E_P[\phi(x)] = \alpha$, that is, ϕ has the exact finite sample level α.

As an example of the above construction let us consider the randomized test in the two sample problem. Let $X_1 \ldots, X_n$ be a sample of n independent observations from a distribution F_X and let $Y_1 \ldots, Y_m$ be a sample of m independent observations from a distribution F_Y. In this case, $X = (X_1, \ldots, X_n, Y_1 \ldots, Y_m)$. Accordingly, we put $N = n + m$ and for every $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, define $gx \in \mathbb{R}^N$ by $(x_{\pi(1)}, \ldots, x_{\pi(N)})$, where $(\pi(1), \ldots, \pi(N))$ is a permutation of $(1, \ldots, N)$. Let $\mathbf{G}_N$ be a collection of all such g so that $M = N!$. Under the null hypothesis $H_0 : F_X = G_Y = F$, clearly gX and X have the same distribution for any g in $\mathbf{G}_N$ and the above described randomization construction of ϕ applies, yielding exact α-level test for any test statistic $T(X)$. We now specialize to the case when the test statistic is

$$T_{(n)}(X) = \sqrt{n}\,(\overline{X} - \overline{Y}). \tag{3}$$

This test statistic is appropriate, for instance, under further assumption that the observed data comes from the same location family, i.e., $F_X(x) = F_Y(x - \Delta) = F(x)$ in which case the null hypothesis H_0 is reduced to $H_0 : \Delta = 0$. The latter is often referred to, in medical and other research studies, as the hypothesis of *the lack of treatment effect or lack of treatment difference* and plays a fundamental role, for instance, in comparison experiments when the observed values of the X's and the Y's represent the data on the subjects undergoing one of the two competing treatments.

The application of the randomization principle to (3) yields the usual randomized (or permutation) t-test in the classical two sample problem, asymptotically equivalent to the unconditional, pooled-variance, two-sample t-test. The beauty of the randomization approach, pointed out by Fisher (1935), lies in the fact that for the above construction we do not assume that F is normal or even that its variance $\sigma^2(F)$ is finite.

In general, if the assumption $F_X = F_Y$ is violated the randomization test based on (3) is no longer of α-level. However, as pointed out by Romano (1990), in some circumstances such tests may be still *asymptotically valid* in the sense that under the appropriate H_0 (possibly no longer ensuring the invariance of the test statistic) the level of the test satisfies

$$E_F[\phi_n(X)] \to \alpha, \tag{4}$$

as $n, m \to \infty$ with $n/N \to \lambda \in (0, 1)$.

The purpose of the current article is to establish the asymptotic properties of the randomization procedure for testing $H_0 : \Delta = 0$ based on the statistic (3) in the presence of dependence structure in the data, in which case the invariance assumption in the model described above is violated. More precisely, we are considering a situation when the data may consist of one sub-sample in which the observations from both treatments are independent of each other, and another sub-sample which consists of paired observations taken under both treatments. We

also allow the marginals of the paired observations to differ (under the null hypothesis) from the remainder of the sample, i.e., the unpaired observations.

The study of this phenomenon was originally motivated by the data from the study on symptom management among hospice patients in the last days of life (Hermann and Looney 2001). In the study the Karnofsky Performance Status (KPS) scale (Karnofsky and Burchenal 1949) was used to assess the functional status of the study subjects. However, for various reasons, the Karnofsky assessment was not performed for each of the participating patients every day, resulting, e.g., in consecutive days samples being only partially paired. We discuss the KPS data in some detail in Section 5.

As noted in Looney and Jones (2003), there are also many other examples of studies involving the two-sample problem in which the data consist of a combination of dependent and independent observations (e.g, Dimery et al. 1987, Nurnberger et al. 1982, Steere et al. 1985). The approach for handling the data in each of these studies has been to ignore the fact that dependent data were present, and to analyze all of the data using the two-sample pooled variance t-statistic. This can lead to biased estimation of the variance of the difference in treatment means and can seriously affect the performance of the statistical test in terms of control of Type I error rate and power. Therefore, it would be helpful to have a method of analysis that makes use of all of the available data, maintains the Type I error at the nominal level, and has optimal or near-optimal power. Under many circumstances the meta-analysis type methods, like e.g., Fisher's method of combining multiple p-values could also be used, however, as pointed out, in Guerra et al. (1999) multiple simulation studies seem to indicate that these types of procedures would be perhaps better viewed as a preliminary step towards the goal of analyzing the pooled raw data.

In this context it is perhaps of interest to investigate to what extend the procedures based on Fisher's randomization principle in the two sample location problem are robust against the

violation of the underlying assumption of symmetry and independence between samples. In particular, it is of interest to see if for partially dependent data the relation (4) holds true. As it turns out (see next section) it typically does not, except for the trivial cases when the dependence structure vanishes asymptotically. In general, the application of a slightly different group of transformations than $\mathbf{G}_N$ leads under $H_0 : \Delta = 0$ to an asymptotically valid test when the randomization procedure is based on a certain modified version of the statistic (3). This modified test procedure is seen to have optimal power properties in the Gaussian family as well as to be asymptotically equivalent to the test based on a linear combination of two- and one- sample t-statistics. However, it is also seen that the efficacy of the randomization test may or may not exceed the efficacy of the asymptotic (unconditional) test based on (3). This depends upon the proportion of dependent data in the combined sample, as well as the strength of the dependence as measured by the correlation coefficient.

The paper is organized as follows. In Section 2 below we provide a setup for the general problem of testing for the lack of treatment effect with partially dependent data and show that the randomization test based on (3) and $\mathbf{G}_N$ does not yield an asymptotically valid permutation test. We also suggest a test statistic and a permutation group, for which (4) holds and obtain the expression for the asymptotic power of the resulting test. In Section 3 we show that our derived test is asymptotically equivalent to a test based on a linear combination of one- and two- sample t-statistics, which is most powerful unbiased in a certain family of Gaussian distributions. In Section 4 we compare the asymptotic relative efficiency of our test and the (biased) unconditional test based on the statistic (3). In Section 5 we present an example of the application of the proposed test to partially dependent data in the study of assessment of functional status at the end of life (Hermann and Looney 2001) as well as make some further comparisons between our randomized test, its large sample approximations, and their competitors. Section 6 contains the summary and offers some conclusions. The proofs of the theorems are straightforward and

along with some auxiliary technical results are provided in the Appendix.

In the sequel, $\widehat{R}_n$ shall denote the randomization distribution which is the empirical distribution of the M values $T(gx)$ as g varies in $\mathbf{G}$. Since its exact computation may be difficult for large M, an approximation of $\widehat{R}_n$ is often employed based on sampling $g_1, \ldots, g_l$ without replacement from $\mathbf{G}$. $\widehat{R}_n$ is then approximated by $\widetilde{R}_n$, the empirical distribution of these l values. It is readily seen (cf. e.g., Romano 1989) that the results concerning $\widehat{R}_n$ contained herein also apply to $\widetilde{R}_n$. It should be also clear that the results extend quite easily to the rank-tests settings as well as have direct implications for the confidence intervals by the usual way of inversion of the appropriate critical regions.

2 Randomization test with partially dependent observations

It was shown by Romano (1990) that the randomization test based on (3) is asymptotically valid provided that the distribution functions share a common mean, i.e., $\mu(F_X) = \mu(F_Y)$ and either share a common variance, i.e., $\sigma^2(F_X) = \sigma^2(F_Y)$ or the sample sizes are asymptotically equal. Note that these are the usual assumptions yielding the asymptotic validity of the two sample t-test. To appreciate why (4) holds for (3), consider the following. Denote by $\widehat{R}_n(x, \mathbf{G}_N)$ the randomization distribution based on (1), (note that the ordering is unaffected by the normalizing constant $\sqrt{n}$) and denote by $J_n(x, F)$ the actual, unconditional distribution of (3) under H_0 then (see Romano 1990) we have as $n \to \infty$

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_N) - J_n(x, F)| \to 0 \qquad \text{wp1} \tag{5}$$

(here and in the sequel wp1 stands for "with probability one"); as well as

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_N) - \Phi(x/\sigma(F))| \to 0 \qquad \text{wp1}, \tag{6}$$

which in turn implies that the critical value $r_n = T^{(k_n)}(X)$ of the randomized test (recall that k_n is defined by (2)) satisfies

$$r_n \to \sigma(F) z_{1-\alpha} \qquad \text{wp1.} \tag{7}$$

The above implies (4). Here $\Phi(\cdot)$ denotes the standard normal distribution function and $z_{1-\alpha}$ denotes its upper α-th quantile, i.e., $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. In view of the usual (unconditional) central limit theorem for the statistic (3) it is immediate that the relations (7) and (5) are both consequences of (6), i.e., the central limit theorem for the randomized statistic.

We consider now the following modification of the classical two sample problem described in the previous section. Assume that for fixed positive integer $k \leq \min(n, m)$ the observations $(X_1 \ldots, X_{n-k})$ and $(Y_1, \ldots, Y_{m-k})$ are, as before, arriving independently from two distribution functions F_X and F_Y which are only assumed to share a common variance, i.e., $\sigma^2(F_X) = \sigma^2(F_Y) = \sigma^2$. Suppose that in addition to the X's and the Y's we also have a sample of independent observations $(Z_1, \ldots, Z_k)$ from the bivariate distribution F_Z, with $Z = (Z^{(1)}, Z^{(2)})$, whose marginals $F_Z^{(1)}$ and $F_Z^{(2)}$ share a common variance, i.e., $\sigma^2(F_Z^{(1)}) = \sigma^2(F_Z^{(2)}) = \tau^2$, and possibly $\sigma^2 \neq \tau^2$. Finally, assume that the means of F_X and $F_Z^{(1)}$, coincide, i.e., $\mu(F_X) = \mu(F_Z^{(1)}) = \xi$, say, and that a similar relation holds true for the means of F_Y and $F_Z^{(2)}$, i.e., $\mu(F_Y) = \mu(F_Z^{(2)}) = \eta$. Note that the assumptions on the equality of the variances are essential here, since otherwise (4) is, in general, not true even for $k = 0$ (cf., e.g., Romano 1990). In this setting, with the help of the randomization procedure, one is interested in testing a general two-part hypothesis

$$H_0' : F_X = F_Y \quad \text{and} \quad F_Z^{(1)} = F_Z^{(2)}. \tag{8}$$

or a slightly more restrictive one (but perhaps also more natural)

$$H_0'' : H_0' \text{ and } (Z^{(1)}, Z^{(2)}) \stackrel{d}{=} (Z^{(2)}, Z^{(1)}) \tag{9}$$

where the last equality is in distribution.

However, as discussed before, in many practical situations the above non-parametric hypothesis would be replaced by the hypothesis of the lack of treatment effect (this is the case, in particular, in all of the medical literature examples cited in Section 1). Henceforth, we thus assume that based on the data contained in the combined samples $(X_1, \ldots, X_{n-k}, Z_1^{(1)}, \ldots, Z_k^{(1)})$ and $(Y_1, \ldots, Y_{m-k}, Z_1^{(2)}, \ldots, Z_k^{(2)})$ it is desired to test the null hypothesis

$$H_0 : \Delta = \xi - \eta = 0. \tag{10}$$

which is equivalent to (8) under the additional assumption that

$$F_X(x) = F_Y(x - \Delta) \quad \text{and} \quad F_Z^{(1)}(x) = F_Z^{(2)}(x - \Delta) \tag{11}$$

Consider first the question of asymptotic validity of the randomized test based on (3). For convenience, denote $Z_i^{(1)} = X_{n-k+i}$ for $i = 1, \ldots, n-k$ and $Z_i^{(2)} = Y_{m-k+i}$ for $i = 1, \ldots, m-k$. Retaining the previously used notation with this modification, we have the following result. Note that we are not assuming that the model (11) holds true.

Theorem 1. *Suppose that H_0 given by* (10) *holds true and that the distribution functions $F_X, F_Y, F_Z^{(1)}, F_Z^{(2)}$ all have finite absolute moment of order greater then two. Assume that $1 \leq k \leq \min(m, n)$ and suppose that as $m, n, k \to \infty$ we have $n/N \to \lambda \in (0, 1)$ and $k/n \to \delta \in [0, \min(1, \frac{1-\lambda}{\lambda})]$. Then*

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_N) - \Phi(x/\sigma_1)| \to 0 \qquad wp1$$

where $\sigma_1^2 = (1-\lambda)^{-1}\left[(1-2\delta\lambda)\,\sigma^2 + 2\delta\lambda\,\tau^2\right]$. *Therefore,* $r_n \to \sigma_1\, z_{1-\alpha}$ *wp*1.

Under the conditions of Theorem 1 it is fairly routine to check that Lindeberg's central limit theorem implies that the actual (unconditional) distribution of the test statistic $T_{(n)}$ given by (3) is asymptotically Gaussian with mean zero and variance

$$\nu_p^2 = (1-\lambda)^{-2}[(1-\lambda-\delta(1-2\lambda+2\lambda^2))\,\sigma^2 + \delta\,(1-2\lambda(1+\rho)(1-\lambda))\,\tau^2], \tag{12}$$

where ρ is the correlation coefficient between the components of Z. According to the theorem, however, the randomization distribution is asymptotically Gaussian with mean zero and variance σ_1^2. Therefore, in general, for the randomization test based on $T_{(n)}$ and $\mathbf{G}_N$ with dependent samples, the relation (4) will hold if and only if $\sigma_1^2 = \nu_p^2$. For arbitrary σ^2 and τ^2 this condition is seen to hold if $\delta = 0$ or $\lambda = 1/2$ and $\rho = 0$, i.e., when the dependence structure vanishes asymptotically, but typically not otherwise.

The reason for the failure of the randomization test is the presence of the dependence structure which introduces heterogeneity (even when $\sigma^2 = \tau^2$) of the components of the test statistic (3) but is not accounted for by the randomization procedure. In other words, the permutation test based on G_N treats all the points of the sample in the same way, disregarding the fact that parts of the sample are dependent. In order to remedy this deficiency we need to consider a different group of transformations, so as to preserve both the dependence structure of the data and the symmetry of the general null hypothesis (9) (or (8)). The idea is to split the sample into dependent and independent components and consider a group of transformations which shall act separately on each part. This is equivalent to considering the product groups of transformations under which the null hypothesis (9) remains invariant.

To this end, consider $z = (z^{(1)}, z^{(2)}) \in \mathbf{R}^2$, and let transformations h_0 and h_1 be defined as $h_0 z = (z^{(1)}, z^{(2)})$ (identity) and $h_1 z = (z^{(2)}, z^{(1)})$ (coordinates swap). For $u \in \mathbf{R}^{N-2k}$ and $z_i \in \mathbf{R}^2$

$(i = 1, \ldots, k)$ let $x = (u, z_1 \ldots, z_k) \in \mathbf{R}^N$ and let $\mathbf{G}_{[N-2k,k]}$ be a collection of all transformations of the form

$$gx = \tilde{g}u \times h_{j_1} z_1 \times \cdots \times h_{j_k} z_k \tag{13}$$

where $\tilde{g}$ is a transformation belonging to $\mathbf{G}_{N-2k}$, the permutation group of the vector $(1, \ldots, N-2k)$ described before, and j_i equals zero or one for $i = 1, \ldots, k$. Here $M = 2^k(N-2k)!$ and the general construction of a randomization test based on (13) applies as described before. We shall take $(U_1 \ldots, U_{N-2k})$ to be the values of the combined sample $(X_1 \ldots, X_{n-k}, Y_1 \ldots, Y_{m-k})$, whereas the Z's shall be the values of the bivariate data sample $(Z_1, \ldots, Z_k)$. s It is easily seen that under group (13) the statistic (3) is no longer centered at zero, which is the case under $\mathbf{G}_N$. The consideration of the randomization distribution of (3) under transformations (13) suggests the following adjustment (bias correction) of the original statistic. For another justification of this adjustment see also the next section. Define

$$W_{(n)} = \sqrt{n}\left(\overline{X} - \overline{Y} - \left(\frac{1}{n} - \frac{1}{m}\right)\sum_{i=1}^{k}(\overline{Z_i} - \overline{U})\right), \tag{14}$$

where $\overline{Z_i} = (Z_i^{(1)} + Z_i^{(2)})/2$ for $i = 1, \ldots, k$ and $\overline{U} = \sum_{i=1}^{N-2k} U_i/(N-2k)$. Note that an expression for $W_{(n)}$ alternative to (14) is

$$W_{(n)} = \sqrt{n}\left(\frac{1}{n} + \frac{1}{m}\right)\left[\sum_{i=1}^{n-k}(X_i - \overline{U}) + \sum_{i=1}^{k}(Z_i^{(1)} - Z_i^{(2)})/2\right]. \tag{15}$$

It is clearly seen from (14) that the statistic $W_{(n)}$ is equal to $T_{(n)}$ if $m = n$ or if the bivariate component vanishes ($k = 0$). Let us also note that the correction factor in (14) above is invariant under the group $\mathbf{G}_{[N-2k,k]}$ and thus under this group the permutation test based on $W_{(n)}$ is

equivalent to the one based on $T_{(n)}$. We now show that this randomization test is asymptotically valid. For the sake of convenience in the sequel we shall take our test statistic to be $W_{(n)}$.

Denote by $J_n(\cdot, F_X, F_Y, F_Z)$ the actual (unconditional) distribution of $W_{(n)}$ and otherwise retain the notation of Theorem 1 and the preceding discussions except replace the statistic (3) with (14) and the group $\mathbf{G}_N$ with $\mathbf{G}_{[N-2k,k]}$. With this convention we have the following

Theorem 2. *Let*

$$\sigma_2^2 = \frac{1}{(1-\lambda)^2}\left[\frac{(1-\delta)(1-\delta\lambda-\lambda)}{1-2\delta\lambda}\,\sigma^2 + \delta\,\frac{(1-\rho)}{2}\tau^2\right]$$

be the asymptotic variance of the statistic $W_{(n)}$ defined in (14)*. Then, under the assumptions of Theorem 1,*

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_{[N-2k,k]}) - J_n(x, F_X, F_Y, F_Z)| \to 0 \qquad wp1, \tag{16}$$

as well as

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_{[N-2k,k]}) - \Phi(x/\sigma_2)| \to 0 \qquad wp1, \tag{17}$$

which implies $r_n \to \sigma_2\, z_{1-\alpha}$ wp1 and thus also (4)*.*

Note that in view of the preceding discussion the above theorem indicates, in particular, that the statistic $T_{(n)}$ given by (3) yields also an asymptotically valid randomization test under $\mathbf{G}_{[N-2k,k]}$.

The proof of Theorem 2 relies on the straightforward application of the usual limit theorems for the sample mean in sampling with and without replacement and is detailed in the appendix. The same method of argument also allows one to investigate the consistency of the test on any arbitrary alternative hypothesis

$$H_1 : \Delta > 0, \tag{18}$$

as well as its asymptotic power.

Theorem 3. *Under the assumptions of Theorem 1 the α-level randomization test for testing* (10) *vs* (18) *based on $W_{(n)}$ and $\mathbf{G}_{[N-2k,k]}$ is consistent. Moreover, its asymptotic power on Pitman's alternatives $\Delta_n = \Delta/\sqrt{n}$ is given by*

$$\beta(\Delta) = \Phi\left(\frac{\Delta'}{\sigma_2} - z_{1-\alpha}\right),$$

where

$$\Delta' = \frac{\Delta}{1-\lambda}\left[\frac{(1-\delta)(1-\delta\lambda-\lambda)}{1-2\delta\lambda} + \frac{\delta}{2}\right] = \frac{\Delta}{1-\lambda}\left[\frac{1-\lambda-\delta/2}{1-2\delta\lambda}\right].$$

The results of Theorems 1-3 can be extended in obvious manner to the statistics of the form $|T(X)|$ and the two-sided tests.

It is perhaps worth noticing at this point that since the exact randomization distributions are often approximated by the empirical distributions based on sampling without replacement from the set of all possible group actions on the observations, there is an obvious similarity of the randomization procedure with the bootstrap method (see Efron and Tibshirani 1993, for example). In our setting, the bootstrap distribution is also asymptotically normal, and the results contained Theorems 1-3 have, generally speaking, their analogues for the bootstrap as well. For a more detailed comparison, see the discussion in Romano (1989).

3 The Gaussian family

It has been shown by Fisher (1935) that the permutation test in the classical two sample location problem can be approximated by the corresponding t-test. In this section we shall consider the approximation of the randomization test discussed in Section 2 by certain uniformly most powerful (UMP) unbiased tests related to the t-test. In order to do so we shall consider the special case when F_X, F_Y, F_Z are all Gaussian distributions. Namely, throughout this section

we shall assume that F_X, F_Y are independent Gaussian distributions with common variance σ^2 and respective means ξ and η and that F_Z is a bivariate Gaussian distribution with the variance vector (τ^2,τ^2), correlation coefficient ρ, and the mean vector (ξ, η). The joint distribution of the combined sample $\mathbb{X} = (X_1 \dots, X_{n-k}, Y_1 \dots, Y_{m-k}, Z_1, \dots, Z_k)$ constitutes thus an exponential family

$$dP_{\underline{\theta},\underline{\nu}}(x) = C(\underline{\theta}, \underline{\nu}) \exp[\theta_1 V_1(x) + \theta_2 V_2(x) + \sum_{i=1}^{5} \nu_i Q_i(x)]\, dx \tag{19}$$

with vectors of parameters $\underline{\theta} = (\theta_1, \theta_2)$ and $\underline{\nu} = (\nu_1, \nu_2, \nu_3, \nu_4, \nu_5)$, where

$$\theta_1 = \frac{\xi - \eta}{\sigma^2}, \quad \theta_2 = \frac{\xi - \eta}{\tau^2(1-\rho)},$$

and

$$\nu_1 = \frac{n\xi + m\eta}{\sigma^2}, \quad \nu_2 = \frac{\xi + \eta}{(n+m)\,\tau^2}, \quad \nu_3 = -\frac{1}{2\,\sigma^2}, \quad \nu_4 = -\frac{1}{2\,\tau^2\,(1-\rho^2)}, \quad \nu_5 = \frac{\rho}{\tau^2\,(1-\rho^2)}.$$

Denote $\widetilde{X} = \sum_{i=1}^{n-k} X_i/(n-k)$ and $\widetilde{Y} = \sum_{i=1}^{m-k} Y_i/(m-k)$. The corresponding set of sufficient statistics for $(\underline{\theta}, \underline{\nu})$ is $(V_1, V_2, Q_1, Q_2, Q_3, Q_4, Q_5)$, where

$$V_1 = \frac{\widetilde{X} - \widetilde{Y}}{\frac{1}{n-k} + \frac{1}{m-k}}, \qquad V_2 = \sum_{i=1}^{k} (Z_i^{(1)} - Z_i^{(2)})/2, \tag{20}$$

and

$$Q_1 = \frac{(n-k)\widetilde{X} + (m-k)\widetilde{Y}}{n+m-2k}, \quad Q_2 = \sum_{i=1}^{k} (Z_i^{(1)} + Z_i^{(2)})/2, \quad Q_3 = \sum_{i=1}^{n-k} X_i^2 + \sum_{i=1}^{m-k} Y_i^2,$$
$$Q_4 = \sum_{i=1}^{k} \left(\left[Z_i^{(1)} \right]^2 + \left[Z_i^{(2)} \right]^2 \right), \qquad Q_5 = \sum_{i=1}^{k} Z_i^{(1)} Z_i^{(2)}.$$

In general, the problem of testing (10) in this setting is equivalent to the problem of testing the hypothesis that $\underline{\theta} = (\theta_1, \theta_2) = (0, 0)$ and is seen to be a version of the Behrens-Fisher problem, in which case no single uniformly best test over a reasonable class of alternatives exists. However,

under the simplifying assumption that $\theta_2 = A\theta_1$, where $A \neq 0$ is assumed to be a known constant, the problem of finding the UMP test for testing (10) vs (18) can be solved in the family (19), if we restrict our attention to unbiased tests only. In our setting, this is typically done by the well known principle of conditioning on Q, i.e., the part of the set of sufficient statistics associated with the vector of nuisance parameters $\underline{\nu}$.

Without losing generality (since we can always rescale the Z_i's) we shall assume in the sequel that $A = 1$, which is equivalent to requiring that

$$\sigma^2 = (1 - \rho)\,\tau^2. \tag{21}$$

Under this proviso it follows from the general theory (see e.g., Lehmann 1997, Chapter 5) that the UMP unbiased test for testing (10) vs (18) is based on the function of the statistic

$$V_{(n)} = V_1 + V_2$$

where V_1 and V_2 are given by (20). Note that in the notation of the previous section we have

$$V_1 = \frac{\widetilde{X} - \widetilde{Y}}{\frac{1}{n-k} + \frac{1}{m-k}} = \sum_{i=1}^{n-k} (X_i - \overline{U})$$

and in view of (15) this gives the following relation between the statistics $W_{(n)}$ and $V_{(n)}$

$$W_{(n)} = \sqrt{n}\,\left(\frac{1}{n} + \frac{1}{m}\right)\,V_{(n)}.$$

Let us denote

$$S_p^2 = \frac{\sum_{i=1}^{k} \frac{1}{2}\left(Z_i^{(1)} - Z_i^{(2)}\right)^2 + \sum_{i=1}^{N-2k} \left(U_i - \overline{U}\right)^2}{N - k - 1}.$$

Invoking Basu's theorem we see that in the family (19) under our assumption (21) the statistic

$$t_n = \frac{V_{(n)}}{S_p \left[\frac{2nm-km-kn}{2(N-2k)}\right]^{1/2}} = \frac{W_{(n)}}{S_p \left[(1+\frac{n}{m})(\frac{1}{n}+\frac{1}{m})\,\frac{2nm-km-kn}{2(N-2k)}\right]^{1/2}} = \frac{W_{(n)}}{S_{(n,2)}}$$

is independent of Q under the null hypothesis (10) and proviso (21), as well as is increasing in $V_{(n)}$ for fixed Q. Using the standard argument based on conditioning (see, e.g, Theorem 1 in Chapter 5 of Lehmann 1997) and the fact that, under the null hypothesis, t_n follows Student's t-distribution with $N-k-1$ degrees of freedom, we have

Proposition 1. *The UMP unbiased α-level test for testing* (10) *vs* (18) *in the family* (19) *under condition* (21) *rejects when $t_n > t_{1-\alpha}(N-k-1)$, i.e., when*

$$W_{(n)} > S_{(n,2)} t_{1-\alpha}(N-k-1) \tag{22}$$

where $t_{1-\alpha}(N-k-1)$ is the upper α-quantile of a t-distribution with $N-k-1$ degrees of freedom.

Since the statistic $S_{(n,2)}$ remains invariant under any transformation belonging to $\mathbf{G}_{[N-2k,k]}$, as well as $S_{(n,2)} \to \sigma_2$ with probability one, it follows immediately from the above proposition and Theorems 2 and 3 of Section 2 that the critical region of the randomization test based on $W_{(n)}$ is (with probability one) asymptotically equivalent to (22). Let us formulate this as

Theorem 4. *Under the assumptions of Theorem 1 and with the condition* (21) *satisfied, the α-level randomization test for testing* (10) *vs* (18) *based on $W_{(n)}$ and $\mathbf{G}_{[N-2k,k]}$ is asymptotically equivalent to the UMP unbiased test* (22) *i.e.,*

$$|r_n - S_{(n,2)} t_{1-\alpha}(N-k-1)| \to 0 \qquad wp1.$$

Moreover, the asymptotic power of the test (22) *on Pitman's alternatives $\Delta_n = \Delta/\sqrt{n}$ is also given by*

$\beta(\Delta)$ of Theorem 3.

It is perhaps of interest to also note that although the optimality result given above is concerned with asymptotic power only, it is in fact not difficult to extend it to the finite sample size in much the same way as in the case of the classical two-sample randomization test based on the statistic (3). Namely, only a slight modification of the standard arguments (see, for instance, Lehmann 1997, Chapter 5, Lemma 3 and Chapter 6, Lemma 4) is required in order to show that the test rejecting for large values of $W_{(n)}$ is the most powerful among the α-level randomized tests based on the finite group of transformations (13) for the normal shift alternatives in the family (19) under the constraint (21). One can also show that it is unbiased in the family of α-level randomization tests based on (13) in the one-parameter location family of all absolutely continuous distributions satisfying (11) in the problem of testing (10) vs (18).

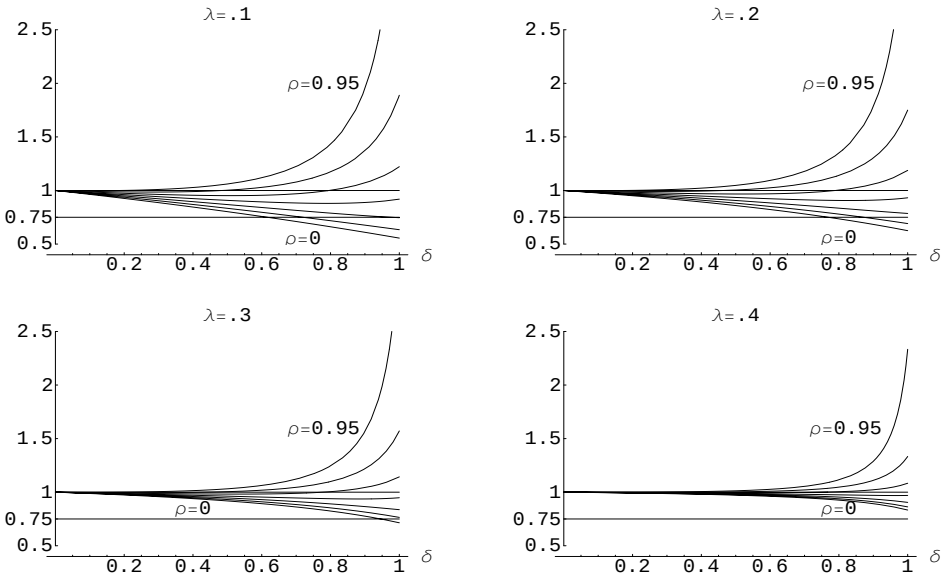
4 Efficacy and ARE vs the means-difference statistic

In this section we shall consider the asymptotic efficiency of our randomization test based on $W_{(n)}$ versus the unconditional test based on the statistic (3). Even though the randomization test based on (3) is not asymptotically valid under $\mathbf{G}_N$ (see Section 2) we still may wish to perform an unconditional test based on the ratio of (3) and a consistent estimator of the variance ν_p^2, where ν_p^2 is defined by (12). Let us note that regardless of condition (21) being satisfied or not, the unconditional test based on (3) is generally biased, but its bias is typically not too large and vanishes asymptotically (see, e.g., Sheffé 1970).

Let $B = \tau^2/\sigma^2$. By Theorem 3 the squared efficacy of the one-sided permutation test based on $W_{(n)}$ as a function of the variance ratio B and the parameters λ, δ, and ρ is

$$c_W^2(B,\delta,\lambda,\rho) = \frac{(1-\lambda-\delta/2)^2}{\sigma^2\,(1-2\delta\lambda)\,\{1-\lambda-\delta[1-\delta\lambda-B(1-2\delta\lambda)(1-\rho)/2]\}}. \tag{23}$$

Figure 1: $ARE(B, \delta, \lambda, \rho)$ for different values of ρ and λ with $B = 1$



Similarly, with the help of formula (12) we obtain that the squared efficacy of the one-sided test based on the statistic (3) is

$$c_T^2(B, \delta, \lambda, \rho) = \frac{(1-\lambda)^2}{\sigma^2\{1-\lambda-2\delta\rho\, B(1-\lambda)\,\lambda - \delta(1-B)\,[1-2\,(1-\lambda)\lambda]\}}. \tag{24}$$

Note that in the special case corresponding to (21) these formulae simplify to

$$c_W^2(1/(1-\rho), \cdot) = \frac{1-\lambda-\delta/2}{\sigma^2\,(1-2\lambda\delta)}$$

and

$$c_T^2(1/(1-\rho), \cdot) = \frac{(1-\lambda)^2}{\sigma^2(1-\lambda+\delta(1-2\lambda)^2\rho/(1-\rho))}.$$

Using (23) and (24) we may compare the asymptotic relative efficiency (ARE) of the one-

sided test based on $W_{(n)}$ versus the one-sided test based on the statistic $T_{(n)}$ given by (3). ARE, as a function of the parameters B, λ, δ, and ρ, is given by the ratio of (23) and (24)

$$ARE(B, \delta, \lambda, \rho) = c_W^2(B, \delta, \lambda, \rho)/c_T^2(B, \delta, \lambda, \rho). \tag{25}$$

Depending on the particular values of the parameters, ARE may be greater then or less then one, indicating that neither test is asymptotically uniformly better. Note that ARE=1 whenever $\lambda = 1/2$ or $\delta = 0$ as the test statistics coincide in these two cases. Based on the inspection of the efficacy formulae (23) and (24) we can clearly see that in general $c_W \geq c_T$ for large positive values of ρ and δ, with the inequality reversed when ρ is negative or has small to moderate positive values, or δ is small. However, whereas there is no upper bound for ARE, its lower bound seems to be at least .5 and for λ values between 0.3 and 0.6 or so, it seems to be at least .75, except for values of δ very close to one. In Figure 1 we present the plots of ARE for $B = 1$ with $0.1 \leq \lambda \leq 0.4$ (for $\lambda > 0.5$ use the symmetry of the problem with respect to X and Y samples) as well as ρ varying by the increments of 0.15 between 0 and 0.95.

5 Numerical examples

In this section we would like to offer some numerical examples illustrating the issues of true versus nominal-level coverage under different testing schemes. We begin with a small sample simulation study comparing the nominal and empirical tests levels for some of the relevant test statistics discussed previously.

5.1 Simulation study with small sample sizes

In order to compare the true versus nominal coverage for various testing procedures discussed in the paper we have performed a simulation study following the setup of Section 3. Namely,

Table 1: Empirical estimates of the nominal level of 0.05

test based on	$T^{(p)}_{(k)}$ (paired)	$T_{(n-k)}(-2k)$ (unpaired)	Combined p-value	$T_{(n)}$ $\mathbf{G}_N$	$W_{(n)}$ $\mathbf{G}_{[N-2k,k]}$	$W_{(n)}$ (z-approx.)	$W_{(n)}$ (t-approx.)
value	0.246	0.056	0.274	0.033	0.055	0.075	0.048
lower 95%CI	0.227	0.046	0.254	0.025	0.045	0.063	0.039
upper 95%CI	0.264	0.066	0.294	0.040	0.064	0.086	0.057

with the help of statistical software R (see, e.g., Ihaka and Gentleman 1996) we have simulated the data vector $\mathbb{X} = (X_1 \dots, X_{n-k}, Y_1 \dots, Y_{m-k}, Z_1, \dots, Z_k)$ with $m = n = 7$ and $k = 3$ multiple times ($l = 3000$) taking F_X, F_Y to be independent Gaussian distributions with common unit variance ($\sigma^2 = 1$) and respective means equal zero, ($\xi = \eta = 0$) as well as F_Z to be a bivariate Gaussian distribution with the unit variance vector ($\tau^2 = 1$), and correlation coefficient ρ=.6. Since in this example we are primarily concerned about the test size i.e., Type I error, we have not insisted on (21) being satisfied.

Several different test statistics were considered and the empirical levels of the corresponding tests were compared with the nominal level of $\alpha = .05$. The overall comparison is presented in Table 1 above where for each testing procedure considered we give the corresponding Monte-Carol estimate of its nominal level along with the 95% confidence interval based on $l = 3000$ simulated datasets. The estimate of the nominal level is based on the count of p-values that have fallen below α. The results given in the first two columns of the table correspond to the separate permutation tests performed on paired and unpaired data only. The p-values for the statistic $T^{(p)}_{(k)}$ were computed via the randomization procedure based on the normalized mean of the differences of paired observation and the permutation group of all g's of the form

$$g(z_1, \dots, z_k) = h_{j_1} z_1 \times \cdots \times h_{j_k} z_k \tag{26}$$

where h_{j_i} are as in (13). On the other hand, the p-values for the test based on $T_{(n-k)}(-2k)$ given

by (3), with the sample sizes $n-k=4$ and $m-k=4$ were computed based on the permutation group $\mathbf{G}_{N-2k}$. The p-values obtained from the two tests were then combined using Fisher's method (Fisher 1958, section 21.1, pp 99–101) to yield an estimate of the nominal level for the combined testing procedure. The numerical value of that estimate is given in column three. In the next two columns we have presented the estimates for the respective tests based on the statistics $T_{(n)}$ and $W_{(n)}$ and the randomization procedures for the complete data set ($N=14$) under the corresponding groups $\mathbf{G}_N$ and $\mathbf{G}_{[N-2k,k]}$. Finally, in the last two columns we have presented the estimates of the nominal levels for the tests based on relating $W_{(n)}$ to the quantiles of the normal and Welch t-distributions (see also next example).

As can be clearly seen from the comparison in Table 1 the randomized tests based on $T_{(n-k)}(-2k)$, and $W_{(n)}$ both control the Type I error at the nominal level and so does the test based on Welch t-approximation to $W_{(n)}$. Indeed, on the basis of the discussion in Section 1 it follows that the first two are both of exact level α in the Gaussian family (and the third one is "almost" so). On the other hand, the tests based on permutation distribution of $T_{(n)}$ under the group G_N as well as based on $W_{(n)}$ and normal approximation are seen not to control the Type I error at the nominal level. Finally, we point out that the test based on the combined p-values also doesn't control the Type I error at the correct level, albeit for a somewhat different reason. Indeed, note that in our current setup the number of elements g of the form (26) is $2^k=8$, and thus the permutation distribution of $T_{(k)}^{(p)}$ is not rich enough in the sense that its quantiles are multiples of $1/8$ only. This deficiency could be, of course, rectified by considering the randomized test, however, the phenomenon illustrates the point of Guerra et al. (1999) that a deficiency of one of the components (say, due to insufficient sample size) may very severely effect the combined testing procedure.

5.2 Karnofsky Performance Status data

We illustrate the performance of the competing test statistics for moderate sample sizes, using the data from the research study on the Karnofsky Performance Status (KPS) scale (Karnofsky and Burchenal 1949) which has motivated our present research. As already outlined in the introduction, in a study of symptom management among hospice patients in the last seven days of life (Hermann and Looney 2001), KPS scale was used to assess the functional status of the study subjects. However, the Karnofsky assessment was not performed for each of the 100 patients participating in the study on each of their last seven days of life. For example, the KPS was obtained for $m = 32$ patients only on their last day of life and for $n = 37$ patients on the day before they died. Of these, $k = 9$ patients were observed on both days, $m - k = 23$ were observed only on their last day of life, and $n - k = 28$ were observed only on the day before they died. The data are presented in Table 2.

Table 2: Hospice patients KPS data

Next-to-last day KPS only (X)
10, 20, 25, 30, 20, 30, 15, 20, 30, 15, 15, 20, 10, 25, 30, 20, 20, 30, 25, 30, 20, 20, 10, 25, 20, 10, 20, 20
Last day KPS only (Y)
15, 25, 30, 20, 10, 20, 10, 30, 10, 10, 10, 25, 15, 20, 20, 20, 20, 10, 10, 10, 20, 30, 10
Next-to-last and last day KPS (Z)
(20,10), (30,20), (25,10), (20,20), (25,20), (10,10), (15,15), (20,20), (30,30).

We shall consider KPS data and the problem of testing the null hypothesis of equality of mean KPS rating on the last day and next-to-last day of life, in order to illustrate the performance of the large sample approximations to the randomization test based on $W_{(n)}$ and the permutation group $\mathbf{G}_{[N-2k,k]}$. We shall compare the results with that obtained by using the means-difference statistic $T_{(n)}$ given by (3) as well as the statistic $T_{(n)}(-k)$, computed after deleting one part of the dependent sample component (in which case (3) and (14) coincide). Note that

the unconditional test based on the studentized $T_{(n)}(-k)$ is asymptotically equivalent to the randomization test based on the group of transformations $\mathbf{G}_{(N-k)}$ as described in Section 1.

Let us first consider the statistics $W_{(n)}$ and $T_{(n)}$. Note that for the data in Table 1 we have $\hat{\rho} = .52$ and $\hat{B} = 1$ (using the usual estimates of τ^2 and σ^2) as well as $\hat{\lambda} = 37/69$ and $\hat{\delta} = 9/37$. Hence from (25)

$$ARE(\hat{B}, \hat{\delta}, \hat{\lambda}, \hat{\rho}) = 1 \tag{27}$$

(up to three significant digits at least) and thus the tests based on the statistics $W_{(n)}$ and $T_{(n)}$ are equally efficient asymptotically. In order to compare the bounds of the critical regions for both tests, the two-sided, symmetric confidence intervals with the asymptotic coverage of 95% have been calculated via normal and Welch t-approximations. The latter has been employed (instead of the pooled-variance t-approximation of Proposition 1) in view of the fact that the condition (21) is readily not satisfied for the KPS data. Additionally, the exact confidence interval for the (two-sided) randomization test based on the statistic $W_{(n)}$ and the group of transformations $\mathbf{G}_{[N-2k,k]}$ were computed using sampling without replacement with the sample size $l = 5000$, as described in Section 1. The numerical results are presented in Table 2.

Table 3: Randomization tests and their approximations for KPS data

Statistic	Value	Method	95% CI	Length	P-value
$W_{(n)}$	24.536	normal approximation	(5.774, 43.298)	37.524	0.010
		Welch t-approximation	(5.377, 43.695)	38.317	0.013
		randomization wrt $\mathbf{G}_{[N-2k,k]}$	(4.761, 43.755)	38.993	0.009
$T_{(n)}$	24.634	normal approximation	(6.700, 42.568)	35.868	0.007
		Welch t-approximation	(6.370, 42.898)	36.528	0.009
$T_{(n)}(-k)$	20.434	normal approximation	(2.829 38.038)	35.209	0.023
		Welch t-approximation	(2.454 38.413)	35.959	0.027
		randomization wrt $\mathbf{G}_N$	(1.890 40.867)	38.978	0.036

Let us note that in view of (27) and since the computed values of $W_{(n)}$ and $T_{(n)}$ are seen to be very close, one would expect the critical region bounds of the randomization test based on $W_{(n)}$ and $\mathbf{G}_{[N-2k,k]}$ to be also indicative of the true bounds of the critical region in the unconditional test based on $T_{(n)}$.

It is readily seen from the comparison in Table 3 that the asymptotic tests based on $W_{(n)}$ and $T_{(n)}$ seem to be underestimating the probability of Type I error at the nominal 5% level, with the t-approximations performing slightly better than the corresponding normal ones. Judging by the respective p-values, the approximations seem to be in more agreement at the empirical level due to the strong evidence in the data against the null hypothesis.

In addition to the comparison of the confidence intervals based on the statistics $T_{(n)}$ and $W_{(n)}$, in Table 3 we have also calculated the confidence intervals for the exact and approximate tests based on the statistic $T_{(n)}(-k)$ which is a version of $T_{(n)}$ (and also $W_{(n)}$) calculated for the reduced set of $N - k = 60$ data points obtained after removing all of the data values for $Z^{(1)}$ but retaining the values for $Z^{(2)}$. As before, the critical region bounds for the exact randomization test based on the group $\mathbf{G}_N$ are approximated by the bounds of the distribution of $l = 5000$ samples without replacement from the permutation distribution of $T_{(n)}(-k)$. Due to the removal of some data points the confidence bounds have noticeably shifted for $T_{(n)}(-k)$. However, the length of the corresponding approximate confidence intervals compared to those computed under $T_{(n)}$ has changed only slightly. Note that the length of the confidence interval for both randomization test statistics remains almost the same, indicating that both tests guard against Type I error at the correct level.

Overall, the numerical comparison in Table 3 seems to indicate that for the KPS data the approximate, unconditional tests based on $W_{(n)}$, $T_{(n)}$, and $T_{(n)}(-k)$ are all underestimating the true levels, with tests based on $W_{(n)}$ statistic being slightly less inaccurate then their competitors. This is consistent with the findings from the small sample simulation study discussed above.

6 Summary and conclusions

The results of Sections 2 and 3 indicate that, in general, in the two sample problem with partially dependent data the usual α-level randomization test based on the groups of transformations $\mathbf{G}_N$ or $\mathbf{G}_{[N-2k,k]}$ and the statistic (3) is not asymptotically valid in the sense of the relation (4). The consideration of the invariance properties of the general non-parametric null hypothesis (8) and (9) leads to the adjustment of the test statistic (3) and thus to the statistic $W_{(n)}$ given by (14). Under the group of transformations $\mathbf{G}_{[N-2k,k]}$ the randomization test which rejects for large values of $W_{(n)}$ is of exact level α under the hypothesis (9) and is seen to be equivalent to the test based on (3). However, unlike the latter, the former is also of asymptotic level α under the hypothesis (10) with the assumption of finite variances. It is further demonstrated in Section 3 that the asymptotic power of this randomization test is optimal among all of the unbiased tests for testing (10) vs (18) in the Gaussian family (20) under the condition (21). The result with obvious modifications extends to the case when the proportion of the population variances for dependent and independent data is a known fixed constant and to the two-sided test setting. The asymptotic relative efficiency of the randomization test based on the statistic $W_{(n)}$ versus the unconditional one based on $T_{(n)}$ was shown to exceed unity for high values of the correlation coefficient and the sufficiently high sample proportion of dependent observations. Using the numerical examples based on simulated data it was shown that the corrected permutation test may have an advantage over the testing procedure based on combined p-values even for small sample sizes. For moderate sample sizes it was also argued via an example taken from the analysis of the KPS data that the approximate unconditional tests based on $W_{(n)}$ seem to be more robust then their competitors against deflating Type I error probabilities as compared to the asymptotic nominal levels.

References

Dimery I.W, Nishioka K, Grossie B, Ota D.M. et al. (1987). Polyamine metabolism in carcinoma of the oral cavity compared with adjacent and normal oral mucosa. *The American Journal of Surgery*, 154, 429–433.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, New York: Chapman and Hall.

Fisher, R.A. (1935). *The Design of Experiments*, London: Oliver & Boyd.

Fisher, R.A. (1958). Statistical Methods for Research Workers (13th edition, revised), Hafner Publishing Company Inc., New York

Fligner, M.A. and Policello, G.E. (1981). Robust rank procedures for the Behrens-Fisher problem. *J. Amer. Statist. Assoc.*, 76(373), 162–168.

Guerra, R., Etzel, C.J., Goldstein, D.R., and Sain, S.R. (1999). Meta-analysis by combining P-values: simulated linkage studies, *Genetic Epidemiology, (Suppl. 1)*:S605-S609

Hermann C. and Looney S. (2001). The effectiveness of symptom management in hospice patients during the last seven days of life. *Journal of Hospice and Palliative Care*, 3, 88–96.

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

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299-314.

Karnofsky D.A, Burchenal J.H. *The clinical evaluation of chemotherapeutic agents in cancer.* In C.M. McLeod (Ed.), Evaluation of Chemotherapeutic Agents (pp. 191-205). New York: Columbia Press, 1949.

Lee, A.J. (1992). *U-statistics. Theory and practice,* Dekker: New York.

Lehmann, E. (1999). *Elements of large-sample theory*, Springer: New York.

Lehmann, E. (1997). *Testing statistical hypotheses*, Reprint of the 1986 second edition, Springer: New York.

Lehmann, E. and Stein, C. (1949). On the theory of some nonparametric hypotheses. *Ann. Math. Statistics* 20, 28–45.

Looney S.W. and Jones P.W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, in press.

Nurnberger J, Jimerson D. C., Allen J. R., Simmons S., Gershon E. (1982). Red cell ouabain-sensitive Na+-K+- adenosine triphosphatase: A state marker in affective disorder inversely related to plasma cortisol. *Biological Psychiatry 1982,* 17(9), 981-992.

Pitman, E. J. G. (1937). Significance Tests Which May Be Applied to Samples From Any Populations I, *Journal of the Royal Statistical Society Supplement*, 4, 119–130.

Romano, J. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, 17(1), 141–159.

Romano, J. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.*, 85(411), 686–692.

Steere A. C., Green J, Schoen R. T., Taylor E, et al. (1985). Successful parenteral penicillin therapy of established Lyme arthritis. *The New England Journal of Medicine,* 312(14,: 869–874.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *J. Amer. Statist. Assoc.*, 65, 1501–1508.

Appendix. Proofs of the theorems

Let $N = m + n$ and let $R_1 \ldots, R_N$ be the ordered values (not the ranks) of the combined sample $(X_1 \ldots, X_n, Y_1 \ldots, Y_m)$ where $X_{n-k+i} = Z_i^{(1)}$ for $i = 1, \ldots, n-k$ and $Y_{m-k+i} = Z_i^{(2)}$ for $i =$

$1, \ldots, m-k$. Additionally, let $\overline{R} = \sum_{i=1}^{N} R_i/N$ and $S_N^2 = \sum_{i=1}^{N}(R_i - \overline{R})^2/N$.

Lemma 1. *Under the assumptions of Theorem 1* $S_N^2 \to (1-2\delta\lambda)\,\sigma^2 + 2\delta\lambda\,\tau^2$ *wp1.*

Proof. Denote $\widetilde{X} = \sum_{i=1}^{n-k} X_i/(n-k)$, $\widetilde{Y} = \sum_{i=1}^{m-k} Y_i/(m-k)$, and $\overline{Z^{(j)}} = \sum_{i=1}^{k} Z_i^{(j)}/k$ for $j = 1, 2$. The result follows in view of the decomposition

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(R_i - \overline{R})^2 = {} & \frac{n-k}{N}\sum_{i=1}^{n-k}(X_i - \widetilde{X})^2 + \frac{m-k}{N}\sum_{i=1}^{m-k}(Y_i - \widetilde{Y})^2 + \frac{k}{N}\sum_{i=1}^{k}\left(Z_i^{(1)} - \overline{Z^{(1)}}\right)^2 \\ & + \frac{k}{N}\sum_{i=1}^{k}\left(Z_i^{(2)} - \overline{Z^{(2)}}\right)^2 + \frac{n-k}{N}\,\widetilde{X}^2 + \frac{m-k}{N}\,\widetilde{Y}^2 + \frac{k}{N}\,\overline{Z^{(1)}}^2 + \frac{k}{N}\,\overline{Z^{(2)}}^2 \\ & - \overline{R}^2\end{aligned}$$

and the strong law of large numbers (SLLN) for U-statistics (see, e.g., Lee 1990 p.122). □

Proof of Theorem 1. To describe a permutation distribution conditional on the $R_i's$, let $(X_1^* \ldots, X_n^*)$ be a sample of size n chosen without replacement from the values $R_1 \ldots, R_N$. Note that $\overline{X} - \overline{Y} = (1/n + 1/m)\sum_{i=1}^{n}(X_i - \overline{R})$. By the usual limit theorem for finite population sampling (see, e.g., Lehmann 1999 p.116) we have that for almost all sample sequences $X_1 \ldots, X_n, Y_1 \ldots, Y_m$ the statistic

$$\frac{\sqrt{n}\,(\overline{X^*} - \overline{R})}{\sqrt{\frac{m}{N-1}\,S_N^2}}$$

is asymptotically Gaussian with mean zero and unit variance, provided that

$$\max_{1 \le i \le N} \frac{(R_i - \overline{R})^2}{N\,S_N^2} \to 0 \qquad \text{wp1},$$

for which, by Lemma 1 and SLLN for $\overline{R}$, it suffices that

$$\max_{1 \le i \le N} \frac{R_i^2}{N} \to 0 \qquad \text{wp1}.$$

To show the latest, take $\varepsilon > 0$ so that the absolute moment of order $(2+\varepsilon)$ exists for the distribution functions $F_X, F_Y, F_Z^{(1)}, F_Z^{(2)}$ and note

$$\left(\max_{1\le i\le N}\frac{R_i^2}{N}\right)^{2+\varepsilon} = \frac{\left(\max_{1\le i\le N}|R_i|^{2+\varepsilon}\right)^2}{N^{2+\varepsilon}} \le \frac{\left(\sum_{i=1}^N |R_i|^{2+\varepsilon}\right)^2}{N^2}\,\frac{1}{N^\varepsilon} \to 0 \qquad \text{wp1.} \quad \square$$

Before we prove the statements contained in theorems 2 and 3 we need the following

Lemma 2. *Let $\{d_i\}$ for $i = 1, 2, 3\ldots$ be a sequence of real numbers such that $\sum_{i=1}^k d_i^2/k \to \gamma^2 > 0$ and let ε_i $(i = 1, 2, 3, \ldots)$ be iid random variables such that $P(\varepsilon_i = -1) = P(\varepsilon_i = 1) = 1/2$. Set $\mathcal{S}_k = k^{-1/2}\sum_{i=1}^k \varepsilon_i\, d_{ik}$. Then $\mathcal{S}_k$ is asymptotically normal with mean zero and variance γ^2.*

Proof. The statement is proved by verifying Lindeberg's condition (see e.g., Romano 1990). $\square$

Proof of Theorems 2 and 3. Let $\Delta \ge 0$ be given. As in Section 2 denote by $(U_1 \ldots, U_{N-2k})$ the combined sample of the unpaired data $(X_1 \ldots, X_{n-k}, Y_1 \ldots, Y_{m-k})$. According to (15)

$$W_{(n)} = \sqrt{n}\left(\frac{1}{n}+\frac{1}{m}\right)\sum_{i=1}^{n-k}(X_i - \overline{U}) + \sqrt{n}\left(\frac{1}{n}+\frac{1}{m}\right)\sum_{i=1}^{k}(Z_i^{(1)} - Z_i^{(2)})/2 = \sqrt{n}\,(W_{(n)}^{(1)} + W_{(n)}^{(2)}).$$

Due to the fact that $\mathbf{G}_{[N-2k,k]}$ is a product group acting separately on the $U's$ and the $Z's$ (which are assumed to be independent) it is enough to consider separately the permutation distributions of the statistics $\sqrt{n}\,W_{(n)}^{(i)}$, for $i = 1, 2$, conditionally on the $U's$ and the $Z's$. Treating the $Z's$ as fixed, the permutation distribution of $\sqrt{n}\,W_{(n)}^{(2)}$ is the distribution function of $\sqrt{n/k}\,(1/n + 1/m)\,\mathcal{S}_k$ with $d_i = |Z_i^{(1)} - Z_i^{(2)}|/2$. By Lemma 2 and SLLN we have that this distribution for almost all sequences $(Z_1, \ldots Z_k)$ is asymptotically Gaussian with mean zero and variance

$$\gamma_U^2 = \frac{(1-\rho)\delta\,\tau^2}{2(1-\lambda)^2}.$$

Note that this quantity does not depend on Δ. Next, set

$$\sigma_\Delta^2 = \sigma^2 + \lambda(1-\lambda)\Delta^2$$

It is readily seen that the argument similar to that used in the proof of Theorem 1 shows that for almost all sample sequences $X_1 \dots, X_{n-k}, Y_1 \dots, Y_{m-k}$ the permutation distribution of $\sqrt{n}\, W_{(n)}^{(2)}$ under (10) or (18) is also asymptotically Gaussian with mean zero and variance

$$\gamma_Z^2 = \frac{(1-\delta)(1-\delta\lambda-\lambda)\,\sigma_\Delta^2}{(1-2\delta\lambda)\,(1-\lambda)^2}.$$

Thus

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_{[N-2k,k]}) - \Phi(x/\gamma_R)| \to 0 \qquad wp1, \tag{28}$$

where $\gamma_R^2 = \gamma_U^2 + \gamma_Z^2$. In view of the above, we see that under the null hypothesis $\Delta = 0$ or under the local alternatives $\Delta_n = \Delta/\sqrt{n}$

$$\sup_x |\widehat{R}_n(x, \mathbf{G}_{[N-2k,k]}) - \Phi(x/\sigma_2)| \to 0 \qquad wp1, \tag{29}$$

and, in particular, (17) holds. Now, let $J_n(\cdot, F_X, F_Y, F_Z, \Delta)$ be the distribution function of the statistic

$$W_{(n)}(\Delta) = W_{(n)} - \sqrt{n}\,\Delta\,\frac{2nm-kn-km}{2\,(n+m-2k)}\left(\frac{1}{n}+\frac{1}{m}\right).$$

By Lindeberg's central limit theorem it follows that

$$\sup_x |J_n(x, F_X, F_Y, F_Z, \Delta) - \Phi(x/\sigma_2)| \to 0. \tag{30}$$

In particular, under the null hypothesis $\Delta = 0$ or under the local alternatives $\Delta_n = \Delta/\sqrt{n}$, the relations (29) and (30) together imply that the critical region of the randomization test satisfies $r_n \to \sigma_2 \, z_{1-\alpha}$ wp1. Hence the remaining assertions of Theorem 2 as well as the asymptotic power formula of Theorem 3 follow immediately. In order to show the last remaining assertion of Theorem 3, i.e., the consistency of the test, fix arbitrary $\Delta > 0$ and note that (28) implies that $r_n \to \gamma_R \, z_{1-\alpha}$ with probability one, but by (30) the actual value of the test statistic $W_{(n)}$ tends to infinity. □

Proof of Theorem 4. The result follows immediately from (30) and the fact that under the null hypothesis $\Delta = 0$ or under the local alternatives $\Delta_n = \Delta/\sqrt{n}$ we have $S_{(n,2)} \to \sigma_2$ with probability one as well as $t_{1-\alpha}(N-k-1) \to z_{1-\alpha}$. □