A CLASS OF PERMUTATION TESTS OF BIVARIATE INTERCHANGEABILITY

Michael D. Ernst
Division of Biostatistics
Department of Statistics
University of Florida
Gainesville, FL 32610-0212

William R. Schucany
Department of Statistical Science
Southern Methodist University
Dallas, TX 75275-0332

SMU / DS / TR - 288
January 1998

A Class of Permutation Tests of Bivariate Interchangeability

Michael D. Ernst
Division of Biostatistics
Department of Statistics
University of Florida
P. O. Box 100212
Gainesville, FL 32610-0212
mernst@stat.ufl.edu

William R. Schucany
Department of Statistical Science
Southern Methodist University
P. O. Box 750332
Dallas, TX 75275-0332
schucany@mail.smu.edu

Michael D. Ernst is Visiting Research Assistant Professor, Division of Biostatistics, Department of Statistics, University of Florida, P. O. Box 100212, Gainesville, FL 32610-0212 (E-mail: mernst@stat.ufl.edu); and William R. Schucany is Professor, Department of Statistical Science, Southern Methodist University, P. O. Box 750332, Dallas, TX 75275-0332 (E-mail: schucany@mail.smu.edu). This work contains results from the first author's Ph.D. dissertation, written under the direction of the second author at Southern Methodist University. The authors wish to thank Volker Wenzel and Paul Kubilis for supplying the dataset used in the example.

ABSTRACT

To simultaneously detect differences in marginal locations and/or scales in bivariate data, a set of permutation tests that are both exact and distribution-free are proposed. The tests take advantage of the fact that only under the null hypothesis of equal means and variances are the pairwise differences symmetrically distributed about zero and uncorrelated with the pairwise sums. Two statistics for detecting the marginal location and scale differences are combined in a quadratic form. A permutation distribution for this quadratic form follows from considering all 2^n conditionally equally likely sign changes on the differences. Several methods of estimating the covariance matrix of the quadratic form are examined including conditional and unconditional (plug-in) approaches. These new tests are compared with the standard tests in the literature and are found, through simulation for several families of bivariate distributions, to perform quite favorably. This paper also brings to light the overlooked likelihood ratio test for equal means and variances in the bivariate normal and shows its relationship to more recent approaches, including those presented here.

KEY WORDS: Bioequivalence; Bivariate Symmetry; Conditional Test; Location-Scale Test; Pitman-Morgan Test; Randomization.

1. INTRODUCTION

When the same experimental units are used in both the treatment and control groups the result is often highly correlated paired observations. Situations like this naturally involve the null hypothesis that Bell and Haller (1969) termed bivariate symmetry, meaning that the group labels are in fact arbitrary. Perhaps a more descriptive term, used by Sen (1967), is *bivariate interchangeability*. Several tests have been proposed over the years that are designed to detect certain alternatives to the hypothesis of bivariate interchangeability. Hollander (1971) proposed a nonparametric test to detect general alternatives. Recently, Hilton and Gee (1997a,b) have given an efficient algorithm that makes it more reasonable to get the exact distribution of Hollander's test. Many approaches, including the present one, concentrate on the alternative hypothesis of unequal location and/or scale parameters in the marginal distributions. For bivariate normal data, tests of this type have been proposed by Hsu (1940), Bell and Haller (1969), and Bradley and Blackwood (1989). Sen (1967) and Kepner and Randles (1984) proposed some rank-based conditionally distribution-free location/scale tests.

Section 2 introduces a new class of permutation tests designed to detect differences in marginal locations and/or scales. Estimation of nuisance parameters is discussed in Section 3. Various approaches are illustrated in Section 4 for a test-retest dataset. The results of a simulation study, designed to compare the new tests to some of those previously proposed, are presented in Section 5. On a historical note, we point out how the test by Hsu (1940) has been overlooked in the literature. We derive an interesting relationship between his test, the test proposed by Bradley and Blackwood (1989), and one proposed here.

2. TEST PROCEDURE

2.1 Bivariate Interchangeability

The bivariate random variable $\mathbf{X} = (X_1, X_2)'$ is said to possess bivariate interchangeability if its cumulative distribution function (cdf) $F(\,\cdot\,,\,\cdot\,)$ satisfies

$$F(x_1, x_2) = F(x_2, x_1) \tag{1}$$

for every pair of real numbers (x_1, x_2). An important consequence of bivariate interchangeability is that X_1 and X_2 have the same marginal distributions, denoted by $X_1 \overset{d}{=} X_2$.

The approach taken here is to parameterize the problem in terms of location and scale parameters. Bivariate interchangeability is required to hold for $F(\,\cdot\,,\,\cdot\,)$ under a null hypothesis, $\mathcal{H}_0$. While $F(\,\cdot\,,\,\cdot\,)$ does not need to satisfy (1) under the alternative hypothesis, the specific alternatives to bivariate interchangeability being tested are those of location and scale differences.

To introduce location and scale parameters, suppose that $\mathbf{X}$ has finite second moments and cdf $F\left(\frac{x_1-\mu_1}{\sigma_1}, \frac{x_2-\mu_2}{\sigma_2}\right)$. Without loss of generality, also assume that

$$\iint\limits_{\mathbb{R}^2} x_i \, dF(x_1, x_2) = 0 \quad \text{and} \quad \iint\limits_{\mathbb{R}^2} x_i^2 \, dF(x_1, x_2) = 1$$

for $i = 1$ and 2. Then, the expectation and variance of $\mathbf{X}$ is $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\mathbb{V}[\mathbf{X}] = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, respectively, where $|\rho| \leq 1$. The null hypothesis of interest is that $\mathbf{X}$ possesses bivariate interchangeability. This can be stated as

$$\mathcal{H}_0\colon \mu_1 = \mu_2,\ \sigma_1^2 = \sigma_2^2, \text{ and } F(\,\cdot\,,\,\cdot\,) \text{ satisfies (1).}$$

The alternative hypothesis is

$$\mathcal{H}_1\colon \mu_1 \neq \mu_2 \text{ or } \sigma_1^2 \neq \sigma_2^2,$$

that the marginals of $\mathbf{X}$ differ in either location or scale (or both). It is important to note that under $\mathcal{H}_0$, $X_1 \overset{d}{=} X_2$, yet X_1 and X_2 may still be (and typically are) correlated.

2.2 Test Statistics

Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample of size n from $F\left(\frac{x_1-\mu_1}{\sigma_1}, \frac{x_2-\mu_2}{\sigma_2}\right)$ where $\mathbf{X}_i = (X_{1i}, X_{2i})'$. Two test statistics, one designed to detect location differences and the other to detect scale differences, can be used to measure the sample's departure from $\mathcal{H}_0$. It is useful to transform the paired data into their pairwise differences and sums,

$$\mathbf{Y}_i = (X_{1i} - X_{2i}, X_{1i} + X_{2i})' = (D_i, S_i)',$$

for $i = 1, \ldots, n$. Clearly, a suitable function of the differences (such as $\overline{D}$) can be used to detect differences in marginal location since $\mathbb{E}[D_i] = \mu_1 - \mu_2$.

Pitman (1939) and Morgan (1939) observed for the bivariate normal distribution that the resulting covariance is

$$\begin{aligned}
\mathbb{C}[D_i, S_i] &= \mathbb{C}[X_{1i} - X_{2i}, X_{1i} + X_{2i}] \\
&= \mathbb{C}[X_{1i}, X_{1i}] + \mathbb{C}[X_{1i}, X_{2i}] - \mathbb{C}[X_{2i}, X_{1i}] - \mathbb{C}[X_{2i}, X_{2i}] \\
&= \sigma_1^2 - \sigma_2^2.
\end{aligned}$$

Hence, the marginal variances are equal if and only if the differences and sums are uncorrelated. In fact, this is true for any bivariate distribution with finite second moments. Therefore, the difference in scale can be judged by measuring the covariance between the differences and sums. With this in mind, we define the statistic $\mathbf{U} = (U_1, U_2)'$, where $U_1 = \overline{D}$ and $U_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)\left(S_i - \overline{S}\right)$. This pair can be used to simultaneously detect differences in marginal location or scale.

2.3 Permutation Distribution

The null hypothesis of interchangeability implies that a permutation distribution for $\mathbf{U}$ can be constructed from all 2^n interchanges of X_{1i} and X_{2i}, $i = 1, \ldots, n$. This is equivalent to considering all 2^n possible signs on the differences. Recalculating $\mathbf{U}$ for each of the conditionally equally likely 2^n sign changes of the differences will yield a permutation distribution of 2^n points in $\mathbb{R}^2$. This permutation distribution is illustrated in Figure 1 for the dataset consisting of 10 pairs of observations displayed in Figure 2(a). This permutation distribution contains $2^{10} = 1024$ points. Figure 2(b) displays the differences and sums from the data in Figure 2(a) that are used to calculate U_1 and U_2.

The testing principle here is that $\mathcal{H}_0$ is to be rejected if $\mathbf{U}$ is "extreme" in this permutation distribution. Under $\mathcal{H}_0$ $\mathbb{E}[\mathbf{U}] = \mathbf{0}$, hence the farther $\mathbf{U}$ is from the origin compared to the other points in the permutation distribution, the more "extreme" it is. The distance of $\mathbf{U}$ from the origin will be measured by an estimated Mahalanobis distance given by the quadratic form $E = \mathbf{U}'\widehat{\mathbf{\Gamma}}^{-1}\mathbf{U}$, where $\widehat{\mathbf{\Gamma}}$ is an estimate of $\mathbf{\Gamma} = \mathbb{V}[\mathbf{U}]$, the 2×2 covariance matrix of $\mathbf{U}$.

The following notation will prove helpful in constructing a reference distribution for E. Define $\mathfrak{b}_i^k$ to be the i^{th} digit (from the right) in the unique binary representation of the integer k. The binary expansion of an integer index between 0 and $2^n - 1$ consists of n digits, which can be used to identify the sign of each D_i. This representation is useful for identifying each of the 2^n permutations of the data.

Now, let

$$\mathbf{Y} = [\,\mathbf{Y}_1 \quad \mathbf{Y}_2 \quad \cdots \quad \mathbf{Y}_n\,] = \begin{bmatrix} D_1 & D_2 & \cdots & D_n \\ S_1 & S_2 & \cdots & S_n \end{bmatrix}$$

be the $2 \times n$ data matrix of observed differences and sums. In a similar fashion, define the k^{th} permutation of the data as

$$\mathbf{Y}_{\langle k \rangle} = \begin{bmatrix} (-1)^{\mathfrak{b}_1^k} D_1 & (-1)^{\mathfrak{b}_2^k} D_2 & \cdots & (-1)^{\mathfrak{b}_n^k} D_n \\ S_1 & S_2 & \cdots & S_n \end{bmatrix}$$

for $k = 0, 1, \ldots, 2^n - 1$. Notice that $\mathbf{Y}_{\langle 0 \rangle}$ ($\mathfrak{b}_i^0 = 0, i = 1, \ldots, n$) is the observed value of the random variable $\mathbf{Y}$. Then $\mathcal{Y} = \{\mathbf{Y}_{\langle k \rangle} \,|\, k = 0, 1, \ldots, 2^n - 1\}$ forms the set of all equally likely data matrices conditional on the observed data matrix $\mathbf{Y}_{\langle 0\rangle}$. It should be noted that the sums remain unchanged for every permutation of the data.

Next define $\mathbf{U}_{\langle k\rangle}$ to be the statistic $\mathbf{U}$ calculated from $\mathbf{Y}_{\langle k\rangle}$, the k^{th} permutation of the observed data. This can be written as

$$\mathbf{U}_{\langle k\rangle} = \begin{bmatrix} U_{1\langle k\rangle} \\ U_{2\langle k\rangle} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(-1)^{b_i^k} D_i \\ \frac{1}{n-1}\sum_{i=1}^{n}(-1)^{b_i^k} D_i\left(S_i - \overline{S}\right) \end{bmatrix}.$$

Notice that $U_{2\langle k\rangle}$, the sample covariance between the permuted differences and sums, can be written without using the mean of the permuted differences. Now let $\mathcal{U} = \{\mathbf{U}_{\langle k\rangle} \,|\, k = 0, 1, \ldots, 2^n - 1\}$ be the set of 2^n conditionally equally likely outcomes of $\mathbf{U} = (U_1, U_2)'$. The 2^{10} points in Figure 1 form such a permutation set of values of $\mathbf{U}$. Again, $\mathbf{U}_{\langle 0\rangle}$ is the observed value of the random variable $\mathbf{U}$.

In the next section, four methods of estimating $\mathbf{\Gamma}$ will be considered. Estimates of $\mathbf{\Gamma}$ are denoted by $\widehat{\mathbf{\Gamma}}_M$, where M identifies the particular estimation method. For a specific estimate of $\mathbf{\Gamma}$ the distance from a point in $\mathcal{U}$ to the origin is determined by $E_{M\langle k\rangle} = \mathbf{U}'_{\langle k\rangle}\widehat{\mathbf{\Gamma}}_M^{-1}\mathbf{U}_{\langle k\rangle}$, $k = 0, 1, \ldots, 2^n - 1$. The conditional permutation distribution of E_M is given by $\mathcal{E} = \{E_{M\langle k\rangle} \,|\, k = 0, 1, \ldots, 2^n - 1\}$. Again, $E_{M\langle 0\rangle}$ is the observed value of the random variable E_M.

As noted earlier, values of $\mathbf{U}$ far from the origin are evidence against $\mathcal{H}_0$, therefore large values of E_M are evidence in favor of $\mathcal{H}_1$. Conditional on the observed data, the elements of $\mathcal{Y}$, $\mathcal{U}$, and $\mathcal{E}$ are equally likely under $\mathcal{H}_0$ (Kepner and Randles 1982). Hence for each $k = 0, 1, \ldots, 2^n - 1$, $\mathbb{P}_{\mathcal{E}}[E_M = E_{M\langle k\rangle} \,|\, \mathcal{H}_0] = 2^{-n}$, where the probability is taken with respect to $\mathcal{E}$. A conditionally distribution-free test of $\mathcal{H}_0$ with exact nominal level $\alpha = m/2^n$ is obtained by rejecting $\mathcal{H}_0$ when $E_{M\langle 0\rangle}$ is one of the m largest elements in $\mathcal{E}$.

Another interpretation of this test can be seen by reconsidering Figure 1 which graphically depicts $\mathcal{U}$. Of the 2^n points in $\mathcal{U}$, if $\mathbf{U}_{\langle 0 \rangle}$, the observed value of $\mathbf{U}$, is one of the m furthest points from the origin (measured in Mahalanobis distance) then $\mathcal{H}_0$ is rejected.

3. ESTIMATING $\boldsymbol{\Gamma}$

The test proposed in the previous section depends on $\boldsymbol{\Gamma}$, the covariance matrix of $\mathbf{U}$. Since in practice $\boldsymbol{\Gamma}$ will not be known, it is necessary to estimate it from the data. In this section, four methods of estimating $\boldsymbol{\Gamma}$ are considered along with some properties of the associated test statistics E_M. The proofs to all the theorems appear in Appendix A.

3.1 A Conditional Estimate of $\boldsymbol{\Gamma}$

Conditional on the observed data, $\mathcal{U}$ provides a distribution from which an estimate of $\boldsymbol{\Gamma}$ can be calculated directly. Note that $\mathbb{E}_{\mathcal{U}}[\mathbf{U}] = \mathbf{0}$ [Lemma A.1(a)] and define the conditional estimate of $\boldsymbol{\Gamma}$ by

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_C = \mathbb{V}_{\mathcal{U}}[\mathbf{U}] &= \frac{1}{2^n}\sum_{k=0}^{2^n-1}\left(\mathbf{U}_{\langle k \rangle} - \mathbb{E}_{\mathcal{U}}[\mathbf{U}]\right)\left(\mathbf{U}_{\langle k \rangle} - \mathbb{E}_{\mathcal{U}}[\mathbf{U}]\right)' \\ &= \frac{1}{2^n}\sum_{k=0}^{2^n-1}\mathbf{U}_{\langle k \rangle}\mathbf{U}'_{\langle k \rangle}.\end{aligned} \tag{2}$$

By considering $\mathcal{U}$ to be the conditional population from which $\mathbf{U}$ is drawn, $\widehat{\boldsymbol{\Gamma}}_C$ is the conditional population covariance matrix of $\mathbf{U}$.

The definition in (2) is a computationally cumbersome method of calculating $\widehat{\boldsymbol{\Gamma}}_C$. This requires calculating a 2×2 covariance matrix from 2^n points. The following theorem gives an alternate and more efficient way to calculate $\widehat{\boldsymbol{\Gamma}}_C$.

Theorem 3.1. In (2), $\widehat{\mathbf{\Gamma}}_C$ can be written as

$$\widehat{\mathbf{\Gamma}}_C = \begin{bmatrix} \frac{1}{n^2}\sum_{i=1}^{n} D_i^2 & \frac{1}{n(n-1)}\sum_{i=1}^{n} D_i^2\left(S_i - \overline{S}\right) \\ \frac{1}{n(n-1)}\sum_{i=1}^{n} D_i^2\left(S_i - \overline{S}\right) & \frac{1}{(n-1)^2}\sum_{i=1}^{n} D_i^2\left(S_i - \overline{S}\right)^2 \end{bmatrix}. \tag{3}$$

This simplifies the calculation of $\widehat{\mathbf{\Gamma}}_C$ greatly in that the covariance matrix can be calculated from the original n points rather than the full permutation distribution of 2^n points.

3.2 Unconditional Estimates of $\mathbf{\Gamma}$

Another approach to estimating $\mathbf{\Gamma}$ is to express it in terms of the moments of the differences and sums and then estimate these from the data. The following theorem gives $\mathbf{\Gamma}$ in these terms.

Theorem 3.2. If $\mathbf{\Gamma} = \mathbb{V}[\mathbf{U}] = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$ and $\mathbb{V}[\mathbf{Y}_i] = \begin{bmatrix} \sigma_D^2 & \sigma_{DS} \\ \sigma_{DS} & \sigma_S^2 \end{bmatrix}$, then $\gamma_{11} = \sigma_D^2/n$, $\gamma_{12} = \gamma_{21} = \delta_{21}/n$, and

$$\gamma_{22} = \frac{(n-1)\delta_{22} - (n-2)\sigma_{DS}^2 + \sigma_D^2\sigma_S^2}{n(n-1)},$$

provided $\delta_{22} < \infty$, where $\delta_{2a} = \mathbb{E}\left[(D_i - \mu_D)^2 (S_i - \mu_S)^a\right]$.

Now an unconditional estimate of $\mathbf{\Gamma}$ can be obtained by replacing the unknown parameters in $\mathbf{\Gamma}$ by their appropriate sample estimates. This estimate is unconditional in the sense that $\mathbf{\Gamma}$ is derived without dependence on the conditional permutation distribution. This "plug-in" method will result in a different unconditional estimate of $\mathbf{\Gamma}$ for each $\mathbf{Y}_{\langle k \rangle} \in \mathcal{Y}$, call it $\widehat{\mathbf{\Gamma}}_{P\langle k \rangle}$. More formally, define the following estimates based on $\mathbf{Y}_{\langle k \rangle}$:

$$
\begin{aligned}
\widehat{\mu}_{D\langle k\rangle} &= \frac{1}{n}\sum_{i=1}^{n}(-1)^{b_i^k} D_i = U_{1\langle k\rangle}, \\
\widehat{\sigma}^2_{D\langle k\rangle} &= \frac{1}{n-1}\sum_{i=1}^{n}\left[(-1)^{b_i^k} D_i - \widehat{\mu}_{D\langle k\rangle}\right]^2, \\
\widehat{\sigma}^2_S &= \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2, \\
\widehat{\sigma}_{DS\langle k\rangle} &= \frac{1}{n-1}\sum_{i=1}^{n}\left[(-1)^{b_i^k} D_i - \widehat{\mu}_{D\langle k\rangle}\right]\left(S_i - \overline{S}\right) = U_{2\langle k\rangle}, \\
\widehat{\delta}_{2a\langle k\rangle} &= \frac{1}{n}\sum_{i=1}^{n}\left[(-1)^{b_i^k} D_i - \widehat{\mu}_{D\langle k\rangle}\right]^2\left(S_i - \overline{S}\right)^a, a = 1, 2.
\end{aligned}
\tag{4}
$$

Then the "plug-in" estimate of $\mathbf{\Gamma}$ is given by

$$
\widehat{\mathbf{\Gamma}}_{P\langle k\rangle} = \frac{1}{n}\begin{bmatrix} \widehat{\sigma}^2_{D\langle k\rangle} & \widehat{\delta}_{21\langle k\rangle} \\ \widehat{\delta}_{21\langle k\rangle} & \widehat{\delta}_{22\langle k\rangle} - \left[(n-2)\widehat{\sigma}^2_{DS\langle k\rangle} - \widehat{\sigma}^2_{D\langle k\rangle}\widehat{\sigma}^2_S\right]/(n-1) \end{bmatrix}. \tag{5}
$$

Thus the elements of $\mathcal{E}$ are $E_{P\langle k\rangle} = \mathbf{U}'_{\langle k\rangle}\widehat{\mathbf{\Gamma}}^{-1}_{P\langle k\rangle}\mathbf{U}_{\langle k\rangle}$, $k = 0, 1, \ldots, 2^n - 1$.

An advantage of the conditional estimate of $\mathbf{\Gamma}$ in the previous section is its invariance over $\mathcal{Y}$. That is, if any element of $\mathcal{Y}$ were the observed value of the data, the resulting permutation distribution, and thus $\widehat{\mathbf{\Gamma}}_C$, would be unchanged. Hence $\widehat{\mathbf{\Gamma}}_C$ only needs to be calculated once rather than 2^n times as $\widehat{\mathbf{\Gamma}}_P$ does. This has obvious computational advantages.

A similar invariant version of $\widehat{\mathbf{\Gamma}}_P$ can be obtained by more strongly imposing the null hypothesis in the estimation of $\mathbf{\Gamma}$. A common testing principle is to use the information in the null hypothesis to estimate nuisance parameters. For example, when testing whether a binomial proportion p is equal to a hypothesized value p_0, it is not uncommon to use p_0 rather than the observed proportion $\widehat{p}$ in the estimation of $\mathbb{V}[\widehat{p}]$.

Under the null hypothesis, $\mu_D = \sigma_{DS} = 0$. If these null values are used in the estimation of the parameters in $\mathbf{\Gamma}$, then the estimates in (4) become

$$
\begin{aligned}
\widehat{\mu}_D &= 0,\\
\widehat{\sigma}_D^2 &= \frac{1}{n}\sum_{i=1}^{n} D_i^2,\\
\widehat{\sigma}_S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i-\overline{S}\right)^2,\\
\widehat{\sigma}_{DS} &= 0,\\
\widehat{\delta}_{2a} &= \frac{1}{n}\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)^a, a=1,2.
\end{aligned}
\tag{6}
$$

These estimates are invariant over $\mathcal{Y}$, hence, the invariant "plug-in" estimate of $\mathbf{\Gamma}$ is given by

$$
\widehat{\mathbf{\Gamma}}_I = \frac{1}{n}\begin{bmatrix} \widehat{\sigma}_D^2 & \widehat{\delta}_{21} \\ \widehat{\delta}_{21} & \widehat{\delta}_{22}+\widehat{\sigma}_D^2\widehat{\sigma}_S^2/(n-1) \end{bmatrix}
\tag{7}
$$

and the elements of $\mathcal{E}$ are $E_{I\langle k\rangle} = \mathbf{U}'_{\langle k\rangle}\widehat{\mathbf{\Gamma}}_I^{-1}\mathbf{U}_{\langle k\rangle}$, $k = 0, 1, \ldots, 2^n - 1$.

If the underlying distribution from which the data are generated is bivariate normal, then so too are the differences and sums. Since the differences and sums are uncorrelated under $\mathcal{H}_0$, bivariate normality implies that they are in fact independent. Therefore, under the null hypothesis and bivariate normality

$$
\begin{aligned}
\delta_{21} &= \mathbb{E}\big[(D_i-\mu_D)^2(S_i-\mu_S)|\,\mathcal{H}_0\big]\\
&= \mathbb{E}\big[(D_i-\mu_D)^2|\,\mathcal{H}_0\big]\mathbb{E}[(S_i-\mu_S)|\,\mathcal{H}_0] = 0
\end{aligned}
$$

and

$$
\begin{aligned}
\delta_{22} &= \mathbb{E}\big[(D_i-\mu_D)^2(S_i-\mu_S)^2|\,\mathcal{H}_0\big]\\
&= \mathbb{E}\big[(D_i-\mu_D)^2|\,\mathcal{H}_0\big]\mathbb{E}\big[(S_i-\mu_S)^2|\,\mathcal{H}_0\big] = \sigma_D^2\sigma_S^2.
\end{aligned}
$$

Thus, $\mathbf{\Gamma}$ simplifies to

$$
\mathbf{\Gamma}_N = \begin{bmatrix} \sigma_D^2/n & 0 \\ 0 & \sigma_D^2\sigma_S^2/(n-1) \end{bmatrix},
$$

which can be estimated by

$$\widehat{\boldsymbol{\Gamma}}_N = \begin{bmatrix} \widehat{\sigma}_D^2/n & 0 \\ 0 & \widehat{\sigma}_D^2\widehat{\sigma}_S^2/(n-1) \end{bmatrix}, \tag{8}$$

and the elements of $\mathcal{E}$ are $E_{N\langle k\rangle} = \mathbf{U}'_{\langle k\rangle}\widehat{\boldsymbol{\Gamma}}_N^{-1}\mathbf{U}_{\langle k\rangle}$, $k = 0, 1, \ldots, 2^n - 1$.

Even if the underlying distribution is not normal, the elements of $\mathcal{E}$ are conditionally equally likely since they are functions of the elements of $\mathcal{U}$ and the mapping from $\mathcal{U}$ to $\mathcal{E}$ (through $\widehat{\boldsymbol{\Gamma}}_N$) does not depend on the permutation distribution. Hence, regardless of the underlying distribution, a test based on E_N is conditionally distribution-free.

Historical Note. The likelihood ratio test of $\mathcal{H}_0$ under bivariate normality was given by Hsu (1940). Even though it was published in *The Annals of Mathematical Statistics* and is such a fundamental problem, it has never been referenced in the bivariate symmetry literature and appears to be essentially unrecognized. The only prominent publication that we found referencing it is Kotz, Johnson, and Read (1985, p. 740). Bradley and Blackwood (1989) proposed a test of $\mathcal{H}_0$ assuming bivariate normality that consists of regressing the differences on the sums and testing parameters of that regression model. Hsu's likelihood ratio test (LRT) was apparently unknown to Bradley and Blackwood, and in Appendix B we see that these two tests are equivalent. In addition, it is shown that these two statistics are monotone functions of E_N. Hence, E_N is a randomization version of the bivariate normal LRT.

3.3 Rejection Region

Figure 1 is an example graphical depiction of $\mathcal{U}$, the permutation distribution of $\mathbf{U}$. Each point in the scatterplot represents a calculated value of U_1 and U_2 from one of the 2^n permutations of the observed data. The rejection region for each test is determined by the m points corresponding to the largest values of $E_{M\langle k\rangle} = \mathbf{U}'_{\langle k\rangle}\widehat{\boldsymbol{\Gamma}}_M^{-1}\mathbf{U}_{\langle k\rangle}$, $k = 0, 1, \ldots, 2^n - 1$, where $\alpha = m/2^n$ is the nominal level of the test. When $M = C$, I, or N, $\widehat{\boldsymbol{\Gamma}}_M$ is the same for all permutations of the data ($k = 0, 1, \ldots, 2^n - 1$), hence the distance from the origin to each point in $\mathcal{U}$ is being measured using the same metric. Therefore, constant values of E_M define equidistant points. Since constant values of a quadratic form define an ellipse, an elliptical contour can be drawn with constant distance equal to the $100(1-\alpha)^{\text{th}}$ percentile of the $E_{M\langle k\rangle}$ distribution. Points outside this ellipse will correspond to the m largest values of $E_{M\langle k\rangle}$ and constitute the rejection region for the test.

Different estimates of $\boldsymbol{\Gamma}$ will lead to different ellipses, defining the appropriate rejection regions. The three ellipses for E_C, E_I, and E_N are shown in Figure 3 for the permutation distribution in Figure 1. These ellipses use $\alpha = 102/2^{10} = .0996$.

Since $\widehat{\boldsymbol{\Gamma}}_P$, the unconditional plug-in estimator, changes for each permutation of the data, an ellipse will not define the rejection region. But in Figure 3 the rejection region for E_P is depicted in another way; the points corresponding to the $m = 102$ largest values of $E_{P\langle k\rangle}$ are marked with a $+$.

This graphical representation allows us to see similarities and differences in the behavior of the different tests. Notice that the ellipse for E_N has axes that are parallel to the (U_1, U_2) axes. This is due to the diagonality of $\widehat{\boldsymbol{\Gamma}}_N$. Also, the ellipses for E_C and E_I are very similar, almost overlapping. This similarity comes from the algebraic similarity of $\widehat{\boldsymbol{\Gamma}}_C$ and $\widehat{\boldsymbol{\Gamma}}_I$, which can be seen by comparing (3) with (6) and (7).

The permutation distribution in Figure 3 is generated by one particular permutation of the data, namely, the permutation of the data that was observed (in Figure 2). This observed value of $\mathbf{U} = (-.76, 3.86)'$ is marked with a $\square$. When this point falls outside one of the ellipses, then it is in the rejection region for the corresponding test. For each of the three ellipses, the legend shows the critical value for E_M that the ellipse represents, followed by the observed value of E_M for each test. Points outside the ellipse have larger values of E_M than the critical value. If the $\square$ is overlayed with a $+$, then it is in the rejection region for E_P. For the permutation distribution in Figure 3, E_N would reject $\mathcal{H}_0$ while the other tests would not.

3.4 Some Properties of E_M

To calculate the test statistic E_M, $\widehat{\mathbf{\Gamma}}$ must be invertable. This condition and the asymptotic distribution of E_M are addressed in the following theorem. For all four methods the estimates are non-negative definite. The four corresponding tests all have the same limiting chi-squared distribution.

Theorem 3.3. For $M = C, P, I$, and N,

(a) $|\widehat{\mathbf{\Gamma}}_M| \geq 0$,

(b) when $\mathcal{H}_0$ is true, $E_M \xrightarrow{\mathcal{D}} \chi_2^2$ as $n \longrightarrow \infty$.

The validity of (b) for E_N also requires the underlying distribution to be normal.

As the sample size increases, enumerating the entire permutation distribution becomes prohibitive. An alternative is to take a random sample from the permutation distribution. This was first proposed by Dwass (1957). The test remains exact and conditionally distribution-free, the only penalty being a loss of efficiency. The previous theorem suggests that the asymptotic distribution of E_M may be used as an alternative to enumerating the permutation distribution, but relying on the asymptotic χ_2^2 distribution is only an approximation that has no guarantees for small and moderate sample sizes. Sampling from the permutation distribution is still an exact, distribution-free procedure.

An interesting feature of the asymptotic distribution of E_C is found from the expectation of E_C over the permutation distribution,

$$\mathbb{E}_{\mathcal{E}}[E_C] = \frac{1}{2^n} \sum_{k=0}^{2^n-1} \mathbf{U}'_{\langle k \rangle} \widehat{\mathbf{\Gamma}}_C^{-1} \mathbf{U}_{\langle k \rangle}.$$

If we let $\widehat{\gamma}_C^{(i,j)}$ be the ij^{th} element of $\widehat{\mathbf{\Gamma}}_C$ (see the notation in Appendix A), then by explicitly inverting $\widehat{\boldsymbol{\Gamma}}_C$ and writing out the quadratic form, we see that

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}[E_C] &= \frac{1}{2^n}\sum_{k=0}^{2^n-1}\frac{U_{1\langle k\rangle}^2\widehat{\gamma}_C^{(2,2)} + U_{2\langle k\rangle}^2\widehat{\gamma}_C^{(1,1)} - 2U_{1\langle k\rangle}U_{2\langle k\rangle}\widehat{\gamma}_C^{(1,2)}}{\widehat{\gamma}_C^{(1,1)}\widehat{\gamma}_C^{(2,2)} - \widehat{\gamma}_C^{(1,2)}\widehat{\gamma}_C^{(1,2)}} \\ &= \frac{\widehat{\gamma}_C^{(1,1)}\widehat{\gamma}_C^{(2,2)} + \widehat{\gamma}_C^{(2,2)}\widehat{\gamma}_C^{(1,1)} - 2\,\widehat{\gamma}_C^{(1,2)}\widehat{\gamma}_C^{(1,2)}}{\widehat{\gamma}_C^{(1,1)}\widehat{\gamma}_C^{(2,2)} - \widehat{\gamma}_C^{(1,2)}\widehat{\gamma}_C^{(1,2)}} = 2.\end{aligned}$$

That is, regardless of the original data, the conditional expectation of E_C is 2 for every n, matching the mean of the limiting chi-squared distribution.

4. AN EXAMPLE

We will demonstrate the application of these tests with an example dataset. A study by Wenzel, Lehmkuhl, Kubilis, Idris, and Pichlmayr (1997) evaluated the cardiopulmonary resuscitation (CPR) skills retention of 113 medical students. The students were given a CPR skills test following a two-hour basic life support class and again four to seven months later. There were several pertinent variables, but for this example we will concentrate on the rate of mouth-to-mouth ventilation (breaths per minute). The two-minute test was performed on a CPR manikin that automatically transferred the performance data to a computer.

Figure 4 displays a scatterplot of the baseline versus retest results. Because of the discrete nature of the test scores there are many ties, hence some jittering (uniform noise) was added to allow overlapping points to be seen. Figure 5 shows boxplots of the baseline and retest scores with an individual's results for each test connected by a line. Again, because of ties these lines are jittered. The dashed lines indicate recommended limits (10-12) for the ventilation rate. While it appears that as a group the students' ventilation rate was slightly higher than recommended at baseline, this seemed to worsen for the retest. It appears that the group became slightly faster, but also more variable, due to a dramatic increase in some individuals' ventilation rate.

For this example, the permutation distribution contains 2^{113} ($> 10^{34}$) points. Since enumerating that many points is impractical, we took a random sample of 1999 points from the permutation distribution and did the four permutation tests at the $\alpha = .05$ significance level. This is illustrated graphically in Figure 6. The observed value of $\mathbf{U} = (-2.11, \ -21.42)'$ is marked by a $\square$ (in the lower left corner) and is in the rejection region for all four tests. We can estimate a p-value for the test by noting the rank of the observed value of E_M among its permutation distribution. Graphically for E_C, E_I, and E_N, this is equivalent to using the observed value of $\mathbf{U}$ as the critical value to draw the ellipse and then counting how many points in the permutation distribution fall outside the ellipse. In this case, for all four tests, the estimated p-value is $1/2000 = .0005$.

Since the observed value of $\mathbf{U}$ in Figure 6 appears extreme for both the U_1 and U_2 components it seems that the difference between the baseline test and the retest is a combination of a location difference (as reported by Wenzel et al. 1997) and a scale difference. Therefore, the students' ventilation rates at retest months later were faster and more sporadic than immediately after their CPR instruction.

5. EMPIRICAL POWER STUDY

5.1 Description

We designed a simulation study to compare the empirical powers of the proposed permutation tests among themselves and with several other tests that have been previously proposed. Besides the four statistics introduced in Section 3 (E_C, E_P, E_I, E_N), the simulation included two conditionally distribution-free tests due to Kepner and Randles (1984) and Sen (1967) as well as the normal likelihood ratio test of Hsu (1940). The tests of Kepner and Randles (1984) and Sen (1967), denoted K and S, respectively, are rank-like tests that each combine a location statistic and a scale statistic in a quadratic form.

These tests use the same permutations of the data as E_M to form their reference distributions.

Sen's test uses the overall ranking of the $2n$ paired values and combines the Wilcoxon rank-sum statistic with Mood's statistic for scale. The test of Kepner and Randles combines the Wilcoxon signed-rank statistic on the differences with Kendall's tau calculated from the paired differences and sums. In our simulation, both quadratic forms used the conditional estimates of the covariance matrix between the location and scale statistics. Kepner and Randles (1984) derived unconditional estimates of the covariance matrices for both K and S, but their simulation revealed that these estimates could occasionally be singular. The seventh competitor is the F for Hsu's (1940) normal LRT.

There are five different families of bivariate distributions in our study. These are the normal, t, generalized Laplace, Cook-Johnson, and the lognormal. The multivariate generalized Laplace distribution is a family of elliptically contoured distributions that includes the multivariate Laplace, normal, and uniform distributions. In the bivariate case, its density function for uncorrelated components is

$$f(x_1, x_2; \lambda) = \frac{\lambda}{2\pi\Gamma(2/\lambda)} e^{-(x_1^2+x_2^2)^{\lambda/2}},$$

where $\lambda > 0$ is a shape parameter. A value of $\lambda = 5$ provides a fairly light-tailed bivariate elliptically contoured distribution.

The Cook-Johnson distribution, which is a type of multivariate uniform distribution, was introduced by Cook and Johnson (1981). In its bivariate form, the density function is

$$f(u_1, u_2; \alpha) = \frac{\alpha+1}{\alpha}(u_1 u_2)^{(-1/\alpha)-1}\left(u_1^{-1/\alpha} + u_2^{-1/\alpha} - 1\right)^{-(\alpha+2)},$$

for $\alpha > 0$ and $0 < u_i \leq 1, i = 1, 2$. Since the marginals are uniform on $(0, 1)$, a variety of marginal distributions can be obtained by applying the appropriate inverse probability integral transformation. Normal marginals were used in our simulation. The bivariate t

distribution used in the simulation has four degrees of freedom and the marginal shape parameters for the lognormal distribution are $\sigma_1 = \sigma_2 = 1$.

These five distributions provide a wide range of bivariate distributions. Besides the bivariate normal distribution, we include a heavily skewed distribution (lognormal), a non-elliptically contoured distribution with normal marginals (Cook-Johnson), and two elliptically contoured distributions: one with heavier tails than the normal (t) and one with lighter tails (generalized Laplace).

The simulation considered four small to moderate sample sizes: $n = 8, 12, 16, 20$, and three different correlations: $\rho = 0, .5, .8$. For the six distribution-free tests, the entire permutation distribution was enumerated for $n = 8$. For the larger sample sizes, a random sample of 499 elements was drawn from the permutation distribution. The same sample was used for all six tests. A nominal significance level of $\alpha = .05$ was used throughout (except for $n = 8$, where $\alpha = 12/2^8 = .046875$).

The marginal mean and variance of one variable, X_1, is fixed at $(\mu_1, \sigma_1^2) = (0, 1)$ in all cases while the mean and variance of X_2 varies. The different values of μ_2 are $(\mu_{20}, \mu_{21}, \mu_{22}) = (0, .25, .5)$ and the values of σ_2^2 are $(\sigma_{20}^2, \sigma_{21}^2, \sigma_{22}^2) = (1, 2, 3)$. Therefore, of the nine combinations, one value of (μ_2, σ_2^2) satisfies $\mathcal{H}_0$ and eight imply $\mathcal{H}_1$. The empirical power of each test is the percentage of the 10,000 pseudo-random samples for which the test rejected $\mathcal{H}_0$. All pseudo-uniform variates on $(0, 1)$ were obtained using the `ran2` random number generator from Press, Teukolsky, Vetterling, and Flannery (1992, p. 282). A more detailed description of the simulation is available from the authors.

5.2 Results

Some of the results of the power study are summarized in Tables 1-3. The observed relative efficiencies of the tests are similar at all four sample sizes. Therefore, to simplify comparisons, only the results for $n = 20$ are reported. Each table contains the empirical power for each of the seven tests (based on the same 10,000 repeats) independently at every one of the $3 \times 3 \times 3 = 27$ combinations of ρ, μ_2, and σ_2^2. The standard errors of the empirical powers range from .2% when $\mathcal{H}_0$ is true to a maximum of .5% when the true power is 50%.

Table 1 presents the results for the bivariate normal distribution. The likelihood ratio F test is clearly superior when there is some variance difference in the marginals, while E_N comes in second. When there is only a mean difference in the marginals, E_C and E_N have slightly higher power than the F test. The differences among E_C, E_P, and E_I are small or insignificant, not consistently favoring any one of the three. This agreement is suggested by the example critical regions displayed in Figure 3. The pattern holds for the other four distributions, thus the remaining tables contain only E_C to save space. The full set of all tables may be obtained from the authors.

Table 2 displays some results for the bivariate Cook-Johnson distribution with normal marginals. When $\rho = 0$, the marginals are uncorrelated and the Cook-Johnson distribution is exactly a bivariate normal distribution. The results for that special case are consistent with the results in Table 1, so we do not present them here.

For $\rho = .5$, E_N is the clear winner among the distribution-free tests when there is some difference in both means and variances. The other E_M share second place, with K in most cases. When there is only a difference in marginal means, E_C does best. While there is no theoretical reason for F to achieve the nominal level of $\alpha = .05$, it appears to do so here. Indeed it is significantly more powerful than the other tests except when there is only a mean difference in the marginals.

For $\rho = .8$, E_N is again the convincing winner when there is a difference in both means and variances, with the other permutation tests forging ahead when there is only a difference in the marginal means. When there is any difference at all in marginal variances,

K makes a strong second place showing, although it is less efficient at the smaller sample sizes where E_P makes a stronger showing.

Some of the results for the remaining three distributions are displayed in Table 3. These are given as ratios of empirical powers relative to E_C. We know that E_M, K, and S are exact 5% level tests and the simulation outcomes under the null hypothesis in Tables 1 and 2 are consistent with this fact. The patterns in Table 3 are easier to see if we suppress the 27 redundant estimates of 1.00. For the generalized Laplace distribution in part (a) the clear winner among the tests achieving the nominal level is E_N, while the other variants of E_M claim second place. One interesting feature of here is that the F test is quite powerful in some cases despite being conservative. Similar to the bivariate normal distribution, the F test does best when there is some variance difference, but loses by a wider margin to E_N when there is only a mean difference.

Part (b) of Table 3 displays the results for the bivariate t distribution. When there is some difference in variances E_N tends to be the winner, while E_C does well when there is only a mean difference. Although K is a strong second place, this is less pronounced in smaller samples where E_C tends to do better. The level of the F test is more than twice the nominal level.

The results for the bivariate lognormal distribution are given in part (c) of Table 3. The clear winner in most cases is S, although E_C and E_N tend to win on the diagonal of each of the three 3×3 sub-tables. Running a strong second is K, although this is a slightly better showing than in the smaller sample sizes. In reality, none of the tests do very well when there is both a small mean and small variance difference. The F test has even more serious difficulties with its level more than eight times the nominal level.

6. CONCLUDING REMARKS

The work here provides some new methods for testing the hypothesis of bivariate interchangeability with a location and scale parameterization. The graphical representation of the permutation distribution is offered as an additional way to interpret the results of the test and to provide more insight into the individual contributions of the location and scale components to the outcome of the test. We have recognized the early work of Hsu (1940) for the bivariate normal and have shown the equivalence of his LRT to that proposed by Bradley and Blackwood (1989).

An example illustrated the new tests in detecting location and scale differences in bivariate data. It also demonstrated the utility of sampling from a permutation distribution when full enumeration is impractical.

In our empirical power study, the proposed tests do quite well under all of the distributions except the heavily skewed lognormal distribution. In fact, E_N does especially well in the elliptically contoured distributions, namely the normal, t, and generalized Laplace. This should not be surprising for the normal distribution since $\mathbf{\Gamma}_N$ was derived under bivariate normality. It turns out that $\mathbf{\Gamma}$ is diagonal for every distribution in the broad class of elliptically contoured distributions, not just the normal distribution (Anderson 1993, eq. 3.18). This explains the success of E_N in the t and generalized Laplace distributions as well.

APPENDIX A: PROOFS OF THEOREMS

Proof of Theorem 3.1

Lemma A.1. For any real numbers $a_1, a_2, \ldots, a_n, c_1, c_2, \ldots, c_n$,

(a) $\sum_{k=0}^{2^n-1} \sum_{i=1}^{n} a_i(-1)^{\mathfrak{b}_i^k} = 0,$

(b) $\sum_{k=0}^{2^n-1} \left[\sum_{i=1}^{n} a_i(-1)^{\mathfrak{b}_i^k}\right] \left[\sum_{j=1}^{n} c_j(-1)^{\mathfrak{b}_j^k}\right] = 2^n \sum_{i=1}^{n} a_i c_i.$

Proof. The key is to understand the patterned behavior of $\mathfrak{b}_i^k$. For illustration, consider Table A.1 which contains values of $\mathfrak{b}_i^k$ for $i \in \{1, 2, 3, 4\}$ and $k \in \{0, 1, \ldots, 2^4 - 1\}$. Each column gives the binary representation of k. Noting that $\mathfrak{b}_i^k = \omega(\lfloor k/2^{i-1} \rfloor)$, where $\lfloor x \rfloor$ is the integer part of x and ω is the indicator function

$$\omega(t) = \begin{cases} 0 & \text{for } t \text{ even} \\ 1 & \text{for } t \text{ odd,} \end{cases}$$

consider the behavior of $\mathfrak{b}_i^k$ for fixed i as k goes from 0 to $2^n - 1$. That is, consider a fixed row in Table A.1. It can be observed that incrementing k by 2^{i-1} will change the value of $\mathfrak{b}_i^k$ (from zero to one, or one to zero). The result is that as k goes from 0 to $2^n - 1$, $\mathfrak{b}_i^k$ exhibits an alternating pattern of 2^{i-1} zeros followed by 2^{i-1} ones. This pattern is repeated 2^{n-i} times. This patterned behavior of $\mathfrak{b}_i^k$ leads to the following result.

Result. Let $i, j \in \{1, 2, \ldots, n\}, i \neq j$, be fixed. Then as k goes from 0 to $2^n - 1$,

$$\mathfrak{b}_i^k = \begin{cases} 0 & 2^{n-1} \text{ times} \\ 1 & 2^{n-1} \text{ times} \end{cases}$$

and

$$(\mathfrak{b}_i^k, \mathfrak{b}_j^k) = \begin{cases} (0,0) & 2^{n-2} \text{ times} \\ (0,1) & 2^{n-2} \text{ times} \\ (1,0) & 2^{n-2} \text{ times} \\ (1,1) & 2^{n-2} \text{ times.} \end{cases}$$

With this result, part (a) of the lemma can be written as

$$\sum_{k=0}^{2^n-1}\sum_{i=1}^{n} a_i(-1)^{\mathfrak{b}_i^k} = \sum_{i=1}^{n}\left[a_i\sum_{k=0}^{2^n-1}(-1)^{\mathfrak{b}_i^k}\right] = \sum_{i=1}^{n} a_i\left[2^{n-1}(-1)^0 + 2^{n-1}(-1)^1\right]$$
$$= \sum_{i=1}^{n} a_i\left(2^{n-1} - 2^{n-1}\right) = 0,$$

and part (b) can be written as

$$\sum_{k=0}^{2^n-1}\left[\sum_{i=1}^{n} a_i(-1)^{\mathfrak{b}_i^k}\right]\left[\sum_{j=1}^{n} c_j(-1)^{\mathfrak{b}_j^k}\right] = \sum_{k=0}^{2^n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i c_j(-1)^{\mathfrak{b}_i^k+\mathfrak{b}_j^k}$$
$$= \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} a_i c_j\left[\sum_{k=0}^{2^n-1}(-1)^{\mathfrak{b}_i^k+\mathfrak{b}_j^k}\right] + \sum_{k=0}^{2^n-1}\sum_{i=1}^{n} a_i c_i(-1)^{2\mathfrak{b}_i^k}$$
$$= \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} a_i c_j\left[2^{n-2}(-1)^0 + 2\cdot 2^{n-2}(-1)^1 + 2^{n-2}(-1)^2\right] + 2^n\sum_{i=1}^{n} a_i c_i$$
$$= \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} a_i c_j\left(2^{n-2} - 2^{n-1} + 2^{n-2}\right) + 2^n\sum_{i=1}^{n} a_i c_i = 2^n\sum_{i=1}^{n} a_i c_i.$$

We can now prove the theorem. The quantity $\widehat{\mathbf{\Gamma}}_C$ is the population covariance matrix of the elements of $\mathcal{U}$. If we let the ij^{th} element of $\widehat{\mathbf{\Gamma}}_C$ be $\widehat{\gamma}_C^{(i,j)}$, then we can calculate each of the elements directly as

$$\widehat{\gamma}_C^{(1,1)} = \frac{1}{2^n}\sum_{k=0}^{2^n-1} U_{1\langle k\rangle}^2 = \frac{1}{2^n}\sum_{k=0}^{2^n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(-1)^{\mathfrak{b}_i^k}D_i\right]\left[\frac{1}{n}\sum_{j=1}^{n}(-1)^{\mathfrak{b}_j^k}D_j\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} D_i^2,$$
$$\widehat{\gamma}_C^{(1,2)} = \widehat{\gamma}_C^{(2,1)} = \frac{1}{2^n}\sum_{k=0}^{2^n-1} U_{1\langle k\rangle}U_{2\langle k\rangle}$$
$$= \frac{1}{2^n}\sum_{k=0}^{2^n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(-1)^{\mathfrak{b}_i^k}D_i\right]\left[\frac{1}{n-1}\sum_{j=1}^{n}(-1)^{\mathfrak{b}_j^k}D_j\left(S_j - \overline{S}\right)\right]$$
$$= \frac{1}{n(n-1)}\sum_{i=1}^{n} D_i^2\left(S_i - \overline{S}\right),$$

and

$$\begin{aligned}
\widehat{\gamma}_C^{(2,2)} &= \frac{1}{2^n}\sum_{k=0}^{2^n-1} U_{2\langle k\rangle}^2 \\
&= \frac{1}{2^n}\sum_{k=0}^{2^n-1}\left[\frac{1}{n-1}\sum_{i=1}^{n}(-1)^{b_i^k} D_i\left(S_i-\overline{S}\right)\right]\left[\frac{1}{n-1}\sum_{j=1}^{n}(-1)^{b_j^k} D_j\left(S_j-\overline{S}\right)\right] \\
&= \frac{1}{(n-1)^2}\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)^2,
\end{aligned}$$

where the last equality in each case comes from an application of part (b) of the lemma.

Proof of Theorem 3.2

To derive $\mathbf{\Gamma}$, notice that U_1 and U_2 are bivariate one-sample U-statistics of degrees 1 and 2, respectively. The symmetric kernel for U_1 is $h^{(1)}(\mathbf{Y}_i) = D_i$ so that $U_1 = \binom{n}{1}^{-1}\sum_{i=1}^{n} h^{(1)}(\mathbf{Y}_i) = \frac{1}{n}\sum_{i=1}^{n} D_i = \overline{D}$. The symmetric kernel for U_2 is $h^{(2)}(\mathbf{Y}_i, \mathbf{Y}_j) = \frac{1}{2}(D_i - D_j)(S_i - S_j)$ so that

$$\begin{aligned}
U_2 &= \binom{n}{2}^{-1}\sum\sum_{1\le i<j\le n} h^{(2)}(\mathbf{Y}_i,\mathbf{Y}_j) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(D_i-D_j)(S_i-S_j) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(D_iS_i - D_iS_j) = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} D_iS_i - n^2\overline{D}\,\overline{S}\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i-\overline{D}\right)\left(S_i-\overline{S}\right).
\end{aligned}$$

Let $\mathbf{y}_i = (d_i, s_i)'$ be an arbitrary fixed vector. These are involved in standard decompositions of U-statistics. Using the notation of Randles and Wolfe (1979), define the conditional expectations

$$\begin{aligned}
h_1^{(1)}(\mathbf{y}_i) &= \mathbb{E}\left[h^{(1)}(\mathbf{Y}_i)\middle|\, \mathbf{Y}_i = \mathbf{y}_i\right] = d_i, \\
h_1^{(2)}(\mathbf{y}_i) &= \mathbb{E}\left[h^{(2)}(\mathbf{Y}_i,\mathbf{Y}_j)\middle|\, \mathbf{Y}_i = \mathbf{y}_i\right] = \mathbb{E}\left[\frac{1}{2}(d_i - D_j)(s_i - S_j)\right] \\
&= \frac{1}{2}\left[(d_i-\mu_D)(s_i-\mu_S) + \sigma_{DS}\right],
\end{aligned}$$

and

$$h_2^{(2)}(\mathbf{y}_i,\mathbf{y}_j) = \mathbb{E}\big[h^{(2)}(\mathbf{Y}_i,\mathbf{Y}_j)\big|\,\mathbf{Y}_i=\mathbf{y}_i,\mathbf{Y}_j=\mathbf{y}_j\big] = \frac{1}{2}(d_i-d_j)(s_i-s_j).$$

Introducing the centered differences and sums $D_i^* = D_i - \mu_D$ and $S_i^* = S_i - \mu_S$, define the quantities

$$\begin{aligned}
\zeta_1^{(1)} &= \mathbb{V}\Big[h_1^{(1)}(\mathbf{Y}_i)\Big] = \sigma_D^2,\\
\zeta_1^{(2)} &= \mathbb{V}\Big[h_1^{(2)}(\mathbf{Y}_i)\Big] = \frac{1}{4}\mathbb{V}[(D_i-\mu_D)(S_i-\mu_S)]\\
&= \frac{1}{4}\{\mathbb{E}[D_i^{*2}S_i^{*2}] - \mathbb{E}[D_i^*S_i^*]^2\} = \frac{1}{4}\big(\delta_{22}-\sigma_{DS}^2\big),\\
\zeta_2^{(2)} &= \mathbb{V}\Big[h_2^{(2)}(\mathbf{Y}_i,\mathbf{Y}_j)\Big] = \frac{1}{4}\mathbb{V}[(D_i-D_j)(S_i-S_j)]\\
&= \frac{1}{4}\mathbb{V}\big[\big(D_i^*-D_j^*\big)\big(S_i^*-S_j^*\big)\big]\\
&= \frac{1}{4}\mathbb{V}\big[D_i^*S_i^* - D_i^*S_j^* - D_j^*S_i^* + D_j^*S_j^*\big]\\
&= \frac{1}{4}\Big\{\mathbb{E}\Big[\big(D_i^*S_i^* - D_i^*S_j^* - D_j^*S_i^* + D_j^*S_j^*\big)^2\Big]\\
&\qquad - \mathbb{E}\big[D_i^*S_i^* - D_i^*S_j^* - D_j^*S_i^* + D_j^*S_j^*\big]^2\Big\}\\
&= \frac{1}{4}\{\mathbb{E}\big[D_i^{*2}S_i^{*2} - 2D_i^{*2}S_i^*S_j^* - 2D_i^*D_j^*S_i^{*2} + 4D_i^*S_i^*D_j^*S_j^* + D_i^{*2}S_j^{*2}\\
&\qquad - 2D_i^*D_j^*S_j^{*2} + D_j^{*2}S_i^{*2} - 2D_j^{*2}S_i^*S_j^* + D_j^{*2}S_j^{*2}\big] - (2\sigma_{DS})^2\}\\
&= \frac{1}{4}\big(2\delta_{22} + 2\sigma_D^2\sigma_S^2 + 4\sigma_{DS}^2 - 4\sigma_{DS}^2\big) = \frac{1}{2}\big(\delta_{22}+\sigma_D^2\sigma_S^2\big),
\end{aligned}$$

and

$$\begin{aligned}
\zeta_1^{(1,2)} &= \mathbb{C}\Big[h_1^{(1)}(\mathbf{Y}_i), h_1^{(2)}(\mathbf{Y}_i)\Big] = \mathbb{C}\Big[D_i, \frac{1}{2}(D_i^*S_i^* + \sigma_{DS})\Big]\\
&= \frac{1}{2}\mathbb{C}[D_i^*, D_i^*S_i^*] = \frac{1}{2}\{\mathbb{E}[D_i^{*2}S_i^*] - \mathbb{E}[D_i^*]\mathbb{E}[D_i^*S_i^*]\}\\
&= \frac{1}{2}\delta_{21}.
\end{aligned}$$

Lee (1990) gives the variance of a U-statistic (Theorem 3, p. 12) and the covariance between two U-statistics (Theorem 2, p. 17) in terms of these ζs. Using these theorems, the elements of $\boldsymbol{\Gamma}$ can be found to be

$$\gamma_{11} = \mathbb{V}[U_1] = \binom{n}{1}^{-1}\binom{1}{1}\binom{n-1}{1-1}\zeta_1^{(1)} = \frac{\sigma_D^2}{n},$$

$$\begin{aligned}
\gamma_{22} = \mathbb{V}[U_2] &= \binom{n}{2}^{-1}\sum_{c=1}^{2}\binom{2}{c}\binom{n-2}{2-c}\zeta_c^{(2)} \\
&= \frac{2}{n(n-1)}\left[2(n-2)\frac{1}{4}(\delta_{22}-\sigma_{DS}^2)+\frac{1}{2}(\delta_{22}+\sigma_D^2\sigma_S^2)\right] \\
&= \frac{1}{n(n-1)}\left[(n-1)\delta_{22}-(n-2)\sigma_{DS}^2+\sigma_D^2\sigma_S^2\right],
\end{aligned}$$

and

$$\gamma_{12} = \mathbb{C}[U_1, U_2] = \binom{n}{1}^{-1}\binom{2}{1}\binom{n-2}{1-1}\zeta_1^{(1,2)} = \frac{\delta_{21}}{n}.$$

Proof of Theorem 3.3

(a) Each of the four methods of estimating $\mathbf{\Gamma}$ will be considered separately. Considering $M = N$ first, it is clear from (8) that $|\widehat{\mathbf{\Gamma}}_N| = \frac{1}{n(n-1)}\widehat{\sigma}_D^4\widehat{\sigma}_S^2$, which is clearly nonnegative.

Next for $M = C$, (3) and an application of the Cauchy-Schwarz inequality implies that

$$|\widehat{\mathbf{\Gamma}}_C| = \frac{1}{n^2(n-1)^2}\left\{\left[\sum_{i=1}^{n} D_i^2\right]\left[\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)^2\right] - \left[\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)\right]^2\right\} \geq 0.$$

The proof for $\widehat{\mathbf{\Gamma}}_I$ follows easily from that for $\widehat{\mathbf{\Gamma}}_C$. Simply notice from (7) that

$$\begin{aligned}
|\widehat{\mathbf{\Gamma}}_I| &= \frac{1}{n^2}\left\{\widehat{\sigma}_D^2\left[\widehat{\delta}_{22}+\widehat{\sigma}_D^2\widehat{\sigma}_S^2/(n-1)\right]-\widehat{\delta}_{21}^2\right\} \geq \frac{1}{n^2}\left(\widehat{\sigma}_D^2\widehat{\delta}_{22}-\widehat{\delta}_{21}^2\right) \\
&= \frac{1}{n^4}\left\{\left[\sum_{i=1}^{n} D_i^2\right]\left[\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)^2\right] - \left[\sum_{i=1}^{n} D_i^2\left(S_i-\overline{S}\right)\right]^2\right\} \\
&= \frac{(n-1)^2}{n^2}|\widehat{\mathbf{\Gamma}}_C| \geq 0.
\end{aligned}$$

The proof for $\widehat{\boldsymbol{\Gamma}}_{P\langle k\rangle}$ is a bit more complicated. The method of proof is valid for all 2^n permutations of the data, therefore we concentrate on $\widehat{\boldsymbol{\Gamma}}_{P\langle k\rangle}$ calculated from the observed data, namely $\widehat{\boldsymbol{\Gamma}}_{P\langle 0\rangle}$.

Begin by defining $A_i = \left(D_i - \overline{D}\right)\left(S_i - \overline{S}\right), i = 1, \ldots, n$. Recalling the notation used in (4), notice that $\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)\left(S_i - \overline{S}\right) = \frac{n}{n-1}\widehat{\sigma}_{DS\langle 0\rangle}$. Define $\widehat{\boldsymbol{\Theta}}$ to be the maximum likelihood estimate of the covariance matrix of the D_is and A_is. Then the elements of $\widehat{\boldsymbol{\Theta}}$ are

$$
\begin{aligned}
\widehat{\theta}_{11} &= \frac{1}{n}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2 = \frac{n}{n-1}\widehat{\sigma}^2_{D\langle 0\rangle}, \\
\widehat{\theta}_{12} &= \frac{1}{n}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)\left(A_i - \overline{A}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)A_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2\left(S_i - \overline{S}\right) = \widehat{\delta}_{21\langle 0\rangle},
\end{aligned}
$$

and

$$
\widehat{\theta}_{22} = \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}A_i^2 - \overline{A}^2 = \widehat{\delta}_{22\langle 0\rangle} - \frac{n^2}{(n-1)^2}\widehat{\sigma}^2_{DS\langle 0\rangle}.
$$

Now, by applying the Cauchy-Schwarz inequality as before, we see that

$$
\begin{aligned}
\left|\widehat{\boldsymbol{\Theta}}\right| &= \widehat{\theta}_{11}\widehat{\theta}_{22} - \widehat{\theta}^2_{12} \\
&= \frac{1}{n^2}\left\{\left[\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2\right]\left[\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2\right] - \left[\sum_{i=1}^{n}\left(D_i - \overline{D}\right)\left(A_i - \overline{A}\right)\right]^2\right\} \\
&\geq 0.
\end{aligned}
$$

But we can also write

$$
\begin{aligned}
|\widehat{\boldsymbol{\Theta}}| = \widehat{\theta}_{11}\widehat{\theta}_{22} - \widehat{\theta}_{12}^2 &= \frac{n}{n-1}\widehat{\sigma}^2_{D\langle 0\rangle}\left[\widehat{\delta}_{22\langle 0\rangle} - \frac{n^2}{(n-1)^2}\widehat{\sigma}^2_{DS\langle 0\rangle}\right] - \widehat{\delta}^2_{21\langle 0\rangle} \\
&= \widehat{\sigma}^2_{D\langle 0\rangle}\left[\frac{n}{n-1}\widehat{\delta}_{22\langle 0\rangle} - \frac{n^3}{(n-1)^3}\widehat{\sigma}^2_{DS\langle 0\rangle}\right] - \widehat{\delta}^2_{21\langle 0\rangle} \\
&\leq \widehat{\sigma}^2_{D\langle 0\rangle}\left[\frac{n}{n-1}\widehat{\delta}_{22\langle 0\rangle} - \frac{n-2}{n-1}\widehat{\sigma}^2_{DS\langle 0\rangle}\right] - \widehat{\delta}^2_{21\langle 0\rangle},
\end{aligned} \tag{A.1}
$$

where the inequality results from the fact that $\frac{n^3}{(n-1)^3} \geq \frac{n-2}{n-1}$. Another application of the Cauchy-Schwarz inequality shows that $\frac{1}{n-1}\left(\widehat{\sigma}^2_{D\langle 0\rangle}\widehat{\sigma}^2_S - \widehat{\delta}_{22\langle 0\rangle}\right) \geq 0$. Adding this quantity inside the brackets in (A.1) and comparing the result to (5) leads to the required string of inequalities

$$
0 \leq |\widehat{\boldsymbol{\Theta}}| \leq \widehat{\sigma}^2_{D\langle 0\rangle}\left[\widehat{\delta}_{22\langle 0\rangle} - \frac{n-2}{n-1}\widehat{\sigma}^2_{DS\langle 0\rangle} + \frac{\widehat{\sigma}^2_{D\langle 0\rangle}\widehat{\sigma}^2_S}{n-1}\right] - \widehat{\delta}^2_{21\langle 0\rangle} = n^2|\widehat{\boldsymbol{\Gamma}}_{P\langle 0\rangle}|.
$$

(b) Recall that in the proof of Theorem 3.2 it was noted that U_1 and U_2 are bivariate one-sample U-statistics of degrees 1 and 2, respectively. Then it follows from the work there and Theorem 3.6.9 of Randles and Wolfe (1979, p. 107) concerning the asymptotic normality of U-statistics that under $\mathcal{H}_0$

$$
\sqrt{n}\mathbf{U} \xrightarrow{\mathcal{D}} N_2(\mathbf{0}, \boldsymbol{\Delta}_0),
$$

as $n \longrightarrow \infty$, where $\boldsymbol{\Delta}_0 = \begin{bmatrix} \sigma_D^2 & \delta_{21} \\ \delta_{21} & \delta_{22} \end{bmatrix}$.

For $M = C$, P, and I, it is easily seen from Sections 2.2.2 and 2.2.3 of Serfling (1980, pp. 68-69) that all the sample estimates in $\widehat{\boldsymbol{\Gamma}}_M$ converge in probability to their population counterparts and hence $n\widehat{\boldsymbol{\Gamma}}_M \xrightarrow{\mathcal{P}} \boldsymbol{\Delta}_0$ as $n \longrightarrow \infty$. Furthermore, $\frac{1}{n}\widehat{\boldsymbol{\Gamma}}_M^{-1} \xrightarrow{\mathcal{P}} \boldsymbol{\Delta}_0^{-1}$. An application of Theorem 3.4.8 of Sen and Singer (1993, pp. 137-138) implies that

$$
E_M = (\sqrt{n}\mathbf{U})'\left(\frac{1}{n}\widehat{\boldsymbol{\Gamma}}_M^{-1}\right)(\sqrt{n}\mathbf{U}) \xrightarrow{\mathcal{D}} \chi_2^2,
$$

as $n \longrightarrow \infty$.

It only remains to be shown that under $\mathcal{H}_0$ and bivariate normality, $E_N \xrightarrow{\mathcal{D}} \chi_2^2$ as $n \longrightarrow \infty$. In Appendix B it is established that $\lambda = (1 - E_N/n)^{n/2}$, where λ is the likelihood ratio statistic under bivariate normality. Since the number of constraints imposed by $\mathcal{H}_0$ is two ($\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$), then Wilk's likelihood ratio statistic, $-2\ln\lambda$, is asymptotically χ_2^2. But

$$\begin{aligned}-2\ln\lambda &= -2\ln\left[\left(1 - \frac{E_N}{n}\right)^{n/2}\right] = -n\ln\left(1 - \frac{E_N}{n}\right) = -n\sum_{i=1}^{\infty}\frac{1}{i}\left(-\frac{E_N}{n}\right)^i \\ &= E_N + O_p\left(n^{-1}\right),\end{aligned}$$

and therefore $E_N \xrightarrow{\mathcal{D}} \chi_2^2$.

APPENDIX B: RELATIONSHIPS BETWEEN TESTS ASSUMING BIVARIATE NORMALITY

This section establishes the equivalence of the two tests for equal means and variances in bivariate normal data proposed by Hsu (1940) and Bradley and Blackwood (1989). In addition, it is shown that these two test statistics are monotonic functions of E_N.

We will proceed by showing the relationship between E_N and each of the two test statistics. This will allow us to then show the equivalence of the two normal theory tests. Before proceeding, it can be easily seen from (8) that

$$E_N = \mathbf{U}'\widehat{\mathbf{\Gamma}}_N^{-1}\mathbf{U} = \frac{nU_1^2}{\widehat{\sigma}_D^2} + \frac{(n-1)U_2^2}{\widehat{\sigma}_D^2\widehat{\sigma}_S^2},$$

where $\widehat{\sigma}_D^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2$ and $\widehat{\sigma}_S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2$.

Hsu (1940) gives a monotone function of the likelihood ratio, λ, and derives its distribution. More specifically, after correcting a typographical error in his equation (9), he considers

$$L_5 = \lambda^{2/n} = \frac{4s_1^2 s_2^2(1-r_{12}^2)}{\left[s_1^2+s_2^2+\frac{1}{2}(\overline{x}_1-\overline{x}_2)^2\right]^2 - \left[2s_1s_2r_{12}-\frac{1}{2}(\overline{x}_1-\overline{x}_2)^2\right]^2}, \tag{B.1}$$

where $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2, r_{12})$ are the MLEs of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. To facilitate the change to our notation, notice that

$$\begin{aligned}
s_1^2+s_2^2-2r_{12}s_1s_2 &= \frac{1}{n}\sum_{i=1}^n\left(D_i-\overline{D}\right)^2 = \frac{1}{n}\sum_{i=1}^n D_i^2-\overline{D}^2 = \widehat{\sigma}_D^2-U_1^2,\\
s_1^2+s_2^2+2r_{12}s_1s_2 &= \frac{1}{n}\sum_{i=1}^n\left(S_i-\overline{S}\right)^2 = \frac{n-1}{n}\widehat{\sigma}_S^2,
\end{aligned}$$

and

$$\begin{aligned}
s_1^2-s_2^2 &= \frac{1}{n}\sum_{i=1}^n\left[\left(X_{1i}-\overline{X}_1\right)^2-\left(X_{2i}-\overline{X}_2\right)^2\right]\\
&= \frac{1}{n}\sum_{i=1}^n\left(X_{1i}-\overline{X}_1-X_{2i}+\overline{X}_2\right)\left(X_{1i}-\overline{X}_1+X_{2i}-\overline{X}_2\right)\\
&= \frac{1}{n}\sum_{i=1}^n\left(D_i-\overline{D}\right)\left(S_i-\overline{S}\right) = \frac{n-1}{n}U_2.
\end{aligned}$$

Now, by expanding the numerator and denominator of (B.1) and adding and subtracting $(s_1^2+s_2^2)^2$ in the numerator, we get

$$\begin{aligned}
L_5 &= \frac{(s_1^2+s_2^2)^2-4s_1^2s_2^2r_{12}^2-(s_1^2+s_2^2)^2+4s_1^2s_2^2}{(s_1^2+s_2^2)^2+\frac{1}{4}U_1^2+U_1^2(s_1^2+s_2^2)-4s_1^2s_2^2r_{12}^2-\frac{1}{4}U_1^2+2s_1s_2r_{12}U_1^2}\\
&= \frac{\left[(s_1^2+s_2^2)^2-4s_1^2s_2^2r_{12}^2\right]-\left[(s_1^2+s_2^2)^2-4s_1^2s_2^2\right]}{(s_1^2+s_2^2)^2-4s_1^2s_2^2r_{12}^2+U_1^2(s_1^2+s_2^2+2s_1s_2r_{12})}\\
&= \frac{(s_1^2+s_2^2+2s_1s_2r_{12})(s_1^2+s_2^2-2s_1s_2r_{12})-(s_1^2-s_2^2)^2}{(s_1^2+s_2^2+2s_1s_2r_{12})(s_1^2+s_2^2-2s_1s_2r_{12})+U_1^2(s_1^2+s_2^2+2s_1s_2r_{12})}\\
&= \frac{\frac{n-1}{n}\widehat{\sigma}_S^2\left(\widehat{\sigma}_D^2-U_1^2\right)-\left(\frac{n-1}{n}U_2\right)^2}{\frac{n-1}{n}\widehat{\sigma}_S^2\left(\widehat{\sigma}_D^2-U_1^2\right)+\frac{n-1}{n}U_1^2\widehat{\sigma}_S^2} = \frac{\widehat{\sigma}_S^2\left(\widehat{\sigma}_D^2-U_1^2\right)-\frac{n-1}{n}U_2^2}{\widehat{\sigma}_D^2\widehat{\sigma}_S^2}\\
&= 1-\frac{U_1^2}{\widehat{\sigma}_D^2}-\frac{(n-1)U_2^2}{n\widehat{\sigma}_D^2\widehat{\sigma}_S^2} = 1-\frac{1}{n}\left[\frac{nU_1^2}{\widehat{\sigma}_D^2}+\frac{(n-1)U_2^2}{\widehat{\sigma}_D^2\widehat{\sigma}_S^2}\right] = 1-\frac{E_N}{n}.
\end{aligned}$$

Therefore, L_5 is a monotone function of E_N.

The test proposed by Bradley and Blackwood (1989) comes from regressing the differences on the sums. They show that testing $\mathcal{H}_0'\colon \beta_0 = \beta_1 = 0$, where β_0 and β_1 are the intercept and slope coefficients, respectively, is equivalent to testing $\mathcal{H}_0$. The statistic for testing $\mathcal{H}_0'$ has an F distribution with 2 and $n-2$ degrees of freedom and is given by

$$F = \frac{\left(\sum_{i=1}^{n} D_i^2 - \mathrm{SSE}\right)\Big/2}{\mathrm{SSE}/(n-2)},$$

where $\mathrm{SSE} = \sum_{i=1}^{n}\left(D_i - \widehat{\beta}_0 - \widehat{\beta}_1 S_i\right)^2$ is the error sum of squares, $\widehat{\beta}_0 = \overline{D} - \widehat{\beta}_1\overline{S}$, and $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(D_i-\overline{D})(S_i-\overline{S})}{\sum_{i=1}^{n}(S_i-\overline{S})^2} = U_2/\widehat{\sigma}_S^2$.

Now, we can write SSE as

$$\begin{aligned}
\mathrm{SSE} &= \sum_{i=1}^{n}\left[D_i - (\overline{D} - \widehat{\beta}_1\overline{S}) - \widehat{\beta}_1 S_i\right]^2 = \sum_{i=1}^{n}\left[D_i - \overline{D} - \widehat{\beta}_1(S_i - \overline{S})\right]^2 \\
&= \sum_{i=1}^{n}\left[\left(D_i - \overline{D}\right)^2 + \widehat{\beta}_1^2\left(S_i - \overline{S}\right)^2 - 2\widehat{\beta}_1\left(S_i - \overline{S}\right)\left(D_i - \overline{D}\right)\right] \\
&= \sum_{i=1}^{n} D_i^2 - n\overline{D}^2 + \widehat{\beta}_1^2(n-1)\widehat{\sigma}_S^2 - 2\widehat{\beta}_1(n-1)U_2 \\
&= n\widehat{\sigma}_D^2 - nU_1^2 - \frac{(n-1)U_2^2}{\widehat{\sigma}_S^2} = \widehat{\sigma}_D^2(n - E_N).
\end{aligned}$$

Therefore, the F statistic is

$$F = \frac{n-2}{2}\left[\frac{n\widehat{\sigma}_D^2 - \widehat{\sigma}_D^2(n-E_N)}{\widehat{\sigma}_D^2(n-E_N)}\right] = \frac{n-2}{2}\left(\frac{E_N}{n-E_N}\right), \tag{B.2}$$

which is also a monotone function of E_N.

Now, substituting $E_N = n(1 - L_5)$ into (B.2), we see that

$$F = \frac{n-2}{2}\left(\frac{1-L_5}{L_5}\right).$$

Hsu (1940) showed that L_5 has a beta distribution with parameters $\frac{1}{2}(n-2)$ and 1.

Applying the relationship between the beta and F distributions to L_5, we see that

$$\frac{n-2}{2}\left(\frac{1-L_5}{L_5}\right) \sim F_{2,n-2}.$$

Therefore, the LRT of Hsu (1940) and the F test given by Bradley and Blackwood (1989) are equivalent tests.

REFERENCES

Anderson, T. W. (1993). "Nonnormal Multivariate Distributions: Inference Based on Elliptically Contoured Distributions," in *Multivariate Analysis: Future Directions*, ed. C. R. Rao, Amsterdam: North-Holland, pp. 1-24.

Bell, C. B. and Haller, H. S. (1969). "Bivariate Symmetry Tests: Parametric and Nonparametric," *The Annals of Mathematical Statistics*, **40**, 259-269.

Bradley, E. L. and Blackwood, L. G. (1989). "Comparing Paired Data: A Simultaneous Test for Means and Variances," *The American Statistician*, **43**, 234-235.

Cook, R. D. and Johnson, M. E. (1981). "A Family of Distributions for Modelling Non-elliptically Symmetric Multivariate Data," *Journal of the Royal Statistical Society*, Ser. B, **43**, 210-218.

Dwass, M. (1957). "Modified Randomization Tests for Nonparametric Hypotheses," *The Annals of Mathematical Statistics*, **28**, 181-187.

Hilton, J. F. and Gee, L. (1997a). "The size and power of the exact bivariate symmetry test," *Computational Statistics and Data Analysis*, **26**, 53-69.

——— (1997b). "An Exact Hollander Bivariate Symmetry Test," *Applied Statistics*, **46**, 533-540.

Hollander, M. (1971). "A Nonparametric Test for Bivariate Symmetry," *Biometrika*, **58**, 203-212.

Hsu, C. T. (1940). "On Samples From a Normal Bivariate Population," *The Annals of Mathematical Statistics*, **11**, 410-426.

Kepner, J. L. and Randles, R. H. (1982). "Detecting Unequal Marginal Scales in a Bivariate Population," *Journal of the American Statistical Association*, **77**, 475-482.

Kepner, J. L. and Randles, R. H. (1984). "Comparison of Tests for Bivariate Symmetry Versus Location and/or Scale Alternatives," *Communications in Statistics – Theory and Methods*, **13**, 915-930.

Kotz, S., Johnson, N. L., and Read, C. B. (eds.) (1985). *Encyclopedia of Statistical Science* (Vol. 6), New York: John Wiley.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*, New York: Marcel Dekker.

Morgan, W. A. (1939). "A Test for the Significance of the Difference Between the Two Variances in a Sample from a Normal Bivariate Population," *Biometrika*, **31**, 13-19.

Pitman, E. J. G. (1939). "A Note on Normal Correlation," *Biometrika*, **31**, 9-12.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C* (2nd ed.), Cambridge: Cambridge University Press.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to The Theory of Nonparametric Statistics*, New York: John Wiley.

Sen, P. K. (1967). "Nonparametric Tests for Multivariate Interchangeability. Part 1: Problems of Location and Scale in Bivariate Distributions," *Sankhyā*, Ser. A, **29**, 351-372.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*, New York: Chapman & Hall.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

Wenzel, V., Lehmkuhl, P., Kubilis, P. S., Idris, A. H., Pichlmayr, I. (1997). "Poor correlation of mouth-to-mouth ventilation skills after basic life support training and 6 months later," *Resuscitation*, **35**, 129-134.

Table 1. Empirical power (in percent) based on 10,000 samples of size $n = 20$ from the bivariate normal distribution.

		$\rho = 0$			$\rho = .5$			$\rho = .8$		
		σ_{20}^2	σ_{21}^2	σ_{22}^2	σ_{20}^2	σ_{21}^2	σ_{22}^2	σ_{20}^2	σ_{21}^2	σ_{22}^2
E_C	μ_{20}	4.8	16.8	36.1	4.8	21.5	45.7	5.0	37.8	70.6
	μ_{21}	9.2	19.8	38.9	14.0	26.6	49.1	29.4	47.4	73.3
	μ_{22}	24.4	30.0	44.9	45.0	45.3	57.8	83.9	77.1	82.5
E_P	μ_{20}	4.8	17.3	38.4	4.6	22.5	47.8	5.1	40.1	74.2
	μ_{21}	9.2	21.4	41.6	13.6	29.0	53.2	28.0	52.6	79.2
	μ_{22}	23.8	32.5	49.5	43.5	49.4	64.9	81.9	82.1	89.4
E_I	μ_{20}	4.8	17.6	38.5	4.8	22.7	48.3	5.0	40.2	73.4
	μ_{21}	9.1	20.5	40.7	13.7	27.7	51.6	28.7	49.0	76.1
	μ_{22}	23.8	30.4	46.9	44.0	45.7	59.7	83.2	77.6	84.2
E_N	μ_{20}	4.6	19.6	46.1	4.7	25.3	56.8	5.1	47.8	85.6
	μ_{21}	9.5	23.8	48.0	14.8	32.8	62.1	30.7	59.6	88.7
	μ_{22}	25.4	35.6	56.8	46.2	53.7	72.7	83.9	86.7	95.3
K	μ_{20}	4.6	17.1	38.4	4.8	21.6	47.0	5.0	39.2	75.2
	μ_{21}	8.6	19.9	40.0	13.5	27.2	51.4	27.5	48.9	78.3
	μ_{22}	23.2	29.8	46.6	43.0	44.9	61.2	81.7	77.6	87.3
S	μ_{20}	5.1	16.6	36.5	4.6	20.7	44.5	5.0	35.9	69.4
	μ_{21}	8.4	20.0	38.9	12.9	25.8	48.7	24.0	45.1	73.6
	μ_{22}	22.5	29.0	45.6	39.6	42.8	58.5	75.5	73.5	84.4
F	μ_{20}	4.9	23.5	54.1	4.7	30.4	65.7	5.1	56.8	91.6
	μ_{21}	9.1	27.4	56.2	14.2	37.4	70.4	28.9	66.6	93.7
	μ_{22}	24.2	38.3	64.0	44.0	57.1	79.8	83.1	89.7	97.3

Table 2. Empirical power (in percent) based on 10,000 samples of size $n = 20$ from the Cook-Johnson distribution with normal marginals.

		$\rho = .5$			$\rho = .8$		
		σ^2_{20}	σ^2_{21}	σ^2_{22}	σ^2_{20}	σ^2_{21}	σ^2_{22}
E_C	μ_{20}	4.9	27.1	53.0	4.8	47.1	73.5
	μ_{21}	16.5	24.9	50.0	47.1	35.5	66.4
	μ_{22}	51.3	37.2	53.0	93.8	61.0	69.0
E_N	μ_{20}	5.0	26.2	59.6	5.0	43.0	80.5
	μ_{21}	15.7	33.0	61.5	26.7	50.8	81.3
	μ_{22}	46.9	50.3	70.0	79.2	75.7	88.5
K	μ_{20}	4.9	26.5	55.0	4.8	50.1	79.3
	μ_{21}	15.8	25.3	53.2	44.4	43.0	76.5
	μ_{22}	49.7	37.9	57.3	93.1	64.3	79.9
S	μ_{20}	5.2	24.5	50.5	4.8	45.4	75.4
	μ_{21}	14.3	24.8	49.8	37.0	38.7	70.5
	μ_{22}	43.7	36.8	54.1	87.7	62.9	75.6
F	μ_{20}	5.0	30.6	68.2	7.6	54.9	89.7
	μ_{21}	13.9	37.7	70.3	29.6	61.7	89.2
	μ_{22}	44.3	54.2	77.0	82.2	81.9	94.0

Table 3. Ratios of empirical powers relative to E_C based on 10,000 samples of size $n = 20$ from the (a) generalized Laplace, (b) Student t, and (c) lognormal distributions.

		$\rho = 0$			$\rho = .5$			$\rho = .8$		
		σ^2_{20}	σ^2_{21}	σ^2_{22}	σ^2_{20}	σ^2_{21}	σ^2_{22}	σ^2_{20}	σ^2_{21}	σ^2_{22}
					(a) generalized Laplace					
E_N	μ_{20}		1.01	1.07		1.00	1.08		1.07	1.07
	μ_{21}	1.21	1.09	1.09	1.20	1.14	1.14	1.18	1.20	1.10
	μ_{22}	1.22	1.20	1.16	1.14	1.23	1.18	1.03	1.15	1.11
K	μ_{20}		.93	.92		.90	.93		.91	.97
	μ_{21}	.99	.90	.91	.96	.93	.96	.95	.97	.99
	μ_{22}	.94	.95	.95	.94	.95	.97	.95	.97	1.01
S	μ_{20}		.87	.86		.86	.86		.82	.90
	μ_{21}	.96	.91	.86	.90	.89	.89	.83	.89	.93
	μ_{22}	.94	.94	.91	.86	.93	.94	.88	.95	.98
	μ_{20}	.71	1.00	1.13	.75	1.01	1.15	.72	1.12	1.11
F	μ_{21}	.89	1.06	1.14	.93	1.13	1.20	1.00	1.26	1.13
	μ_{22}	.99	1.17	1.22	1.01	1.24	1.24	1.00	1.19	1.13
					(b) Student t with 4 df					
E_N	μ_{20}		1.53	1.66		1.54	1.65		1.66	1.62
	μ_{21}	.84	1.21	1.43	.83	1.13	1.41	.87	1.12	1.32
	μ_{22}	.85	.98	1.18	.91	.97	1.10	.98	.99	1.04
K	μ_{20}		1.33	1.45		1.34	1.46		1.45	1.48
	μ_{21}	.94	1.13	1.29	.95	1.09	1.26	.98	1.10	1.27
	μ_{22}	.98	1.03	1.13	1.00	1.02	1.09	1.00	1.01	1.04
S	μ_{20}		1.30	1.40		1.32	1.40		1.33	1.38
	μ_{21}	.91	1.10	1.25	.87	1.04	1.23	.91	1.03	1.20
	μ_{22}	.95	1.01	1.11	.97	.99	1.06	.99	.99	1.03
F	μ_{20}	2.48			2.49			2.38		
					(c) lognormal					
E_N	μ_{20}		.56	.57		.53	.63		.63	.92
	μ_{21}	.26	1.20	1.78	.32	1.08	2.15	.51	.99	2.71
	μ_{22}	.34	.38	.94	.59	.46	.78	.85	.72	.79
K	μ_{20}		2.07	1.98		2.02	1.66		1.71	1.43
	μ_{21}	1.32	1.06	3.32	1.25	.93	3.95	1.08	.87	4.09
	μ_{22}	1.11	1.09	.85	1.03	1.10	.83	1.00	1.01	.88
S	μ_{20}		2.66	2.28		2.20	1.72		1.79	1.45
	μ_{21}	1.60	1.12	4.34	1.27	1.05	4.46	1.08	.97	4.39
	μ_{22}	1.15	1.27	.93	1.03	1.09	.88	1.00	1.01	.93
F	μ_{20}	8.78			8.37			8.16		

Table A.1. Values of $\mathfrak{b}_i^k$ for $i = 1, 2, 3, 4$.

		k															
i	2^{i-1}	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
2	2	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
3	4	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
4	8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

Figure Captions

Figure 1. Permutation distribution of **U** for the data in Figure 2(a).

Figure 2. Scatterplots of (a) the random sample used to generate Figure 1 and (b) the same data transformed to differences and sums.

Figure 3. Permutation distribution of **U** with the rejection regions for E_C, E_P, E_I, and E_N depicted. The observed value of **U** marked with a $\square$.

Figure 4. Mouth-to-mouth ventilation rate (breaths per minute) of 113 medical students after a CPR skills course (baseline) and four to seven months later (retest).

Figure 5. Mouth-to-mouth ventilation rate (breaths per minute) of 113 medical students after a CPR skills course (baseline) and four to seven months later (retest). Solid lines indicate individual profiles. Dashed lines indicate recommended limits.

Figure 6. A random sample of 1999 points from the permutation distribution of **U** with the rejection regions for E_C, E_P, E_I, and E_N depicted. The observed value of **U** is marked with a $\square$.

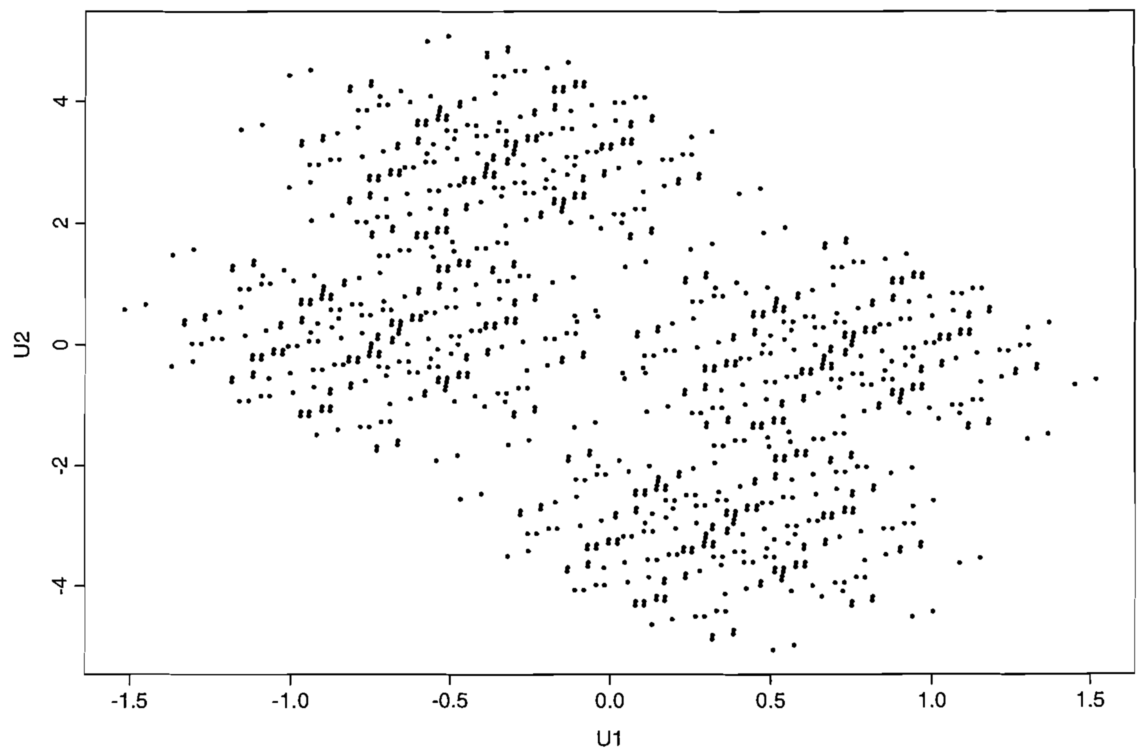

Figure 1. Permutation distribution of **U** for the data in Figure 2(a).

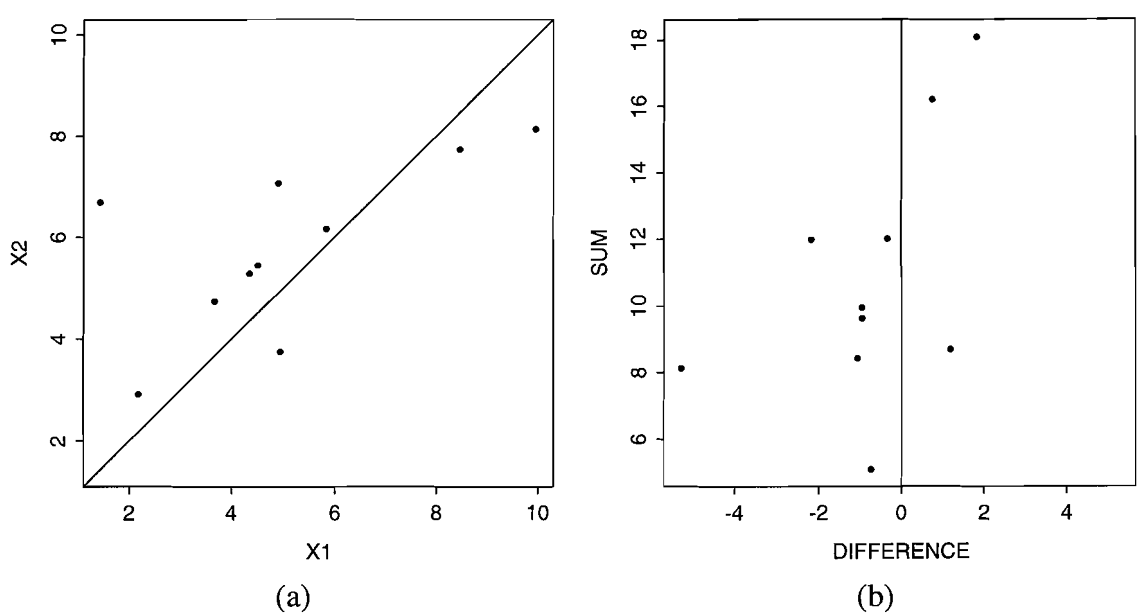

Figure 2. Scatterplots of (a) the random sample used to generate Figure 1 and (b) the same data transformed to differences and sums.

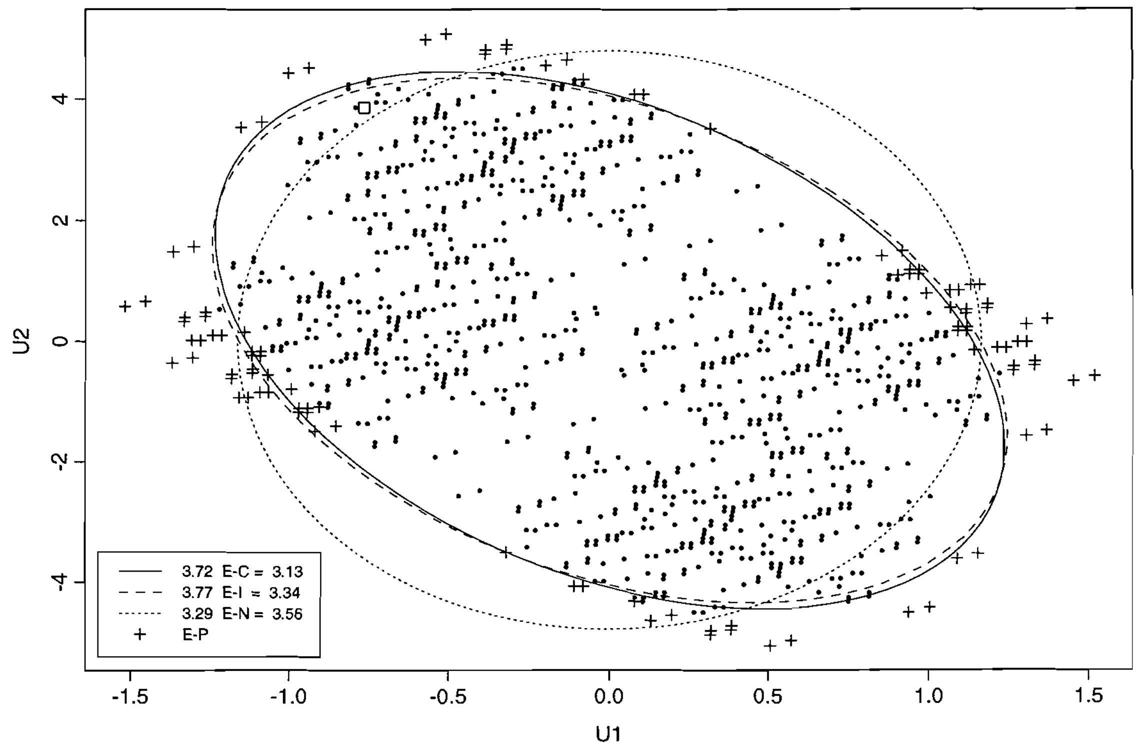

Figure 3. Permutation distribution of **U** with the rejection regions for E_C, E_P, E_I, and E_N depicted. The observed value of **U** marked with a □.

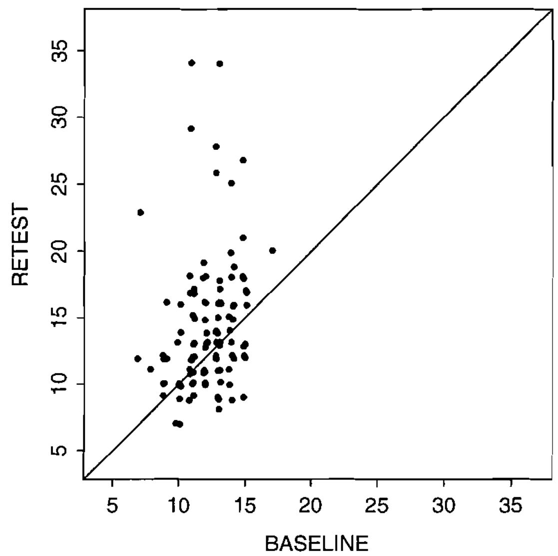

Figure 4. Mouth-to-mouth ventilation rate (breaths per minute) of 113 medical students after a CPR skills course (baseline) and four to seven months later (retest).

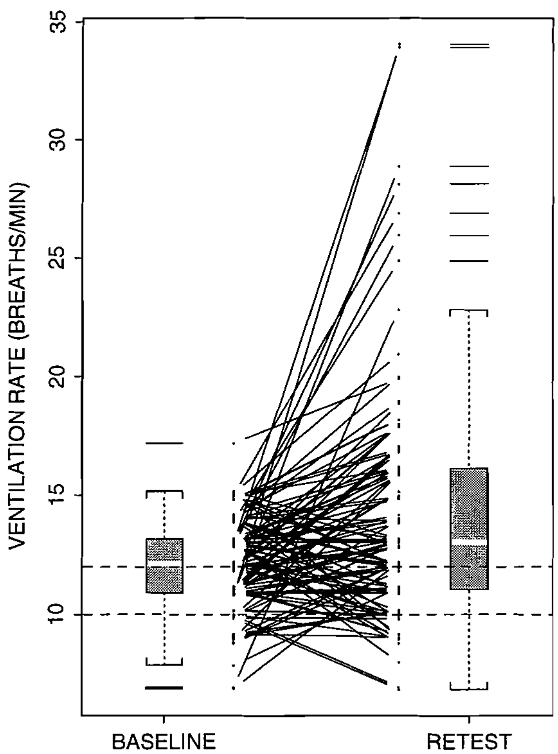

Figure 5. Mouth-to-mouth ventilation rate (breaths per minute) of 113 medical students after a CPR skills course (baseline) and four to seven months later (retest). Solid lines indicate individual profiles. Dashed lines indicate recommended limits.

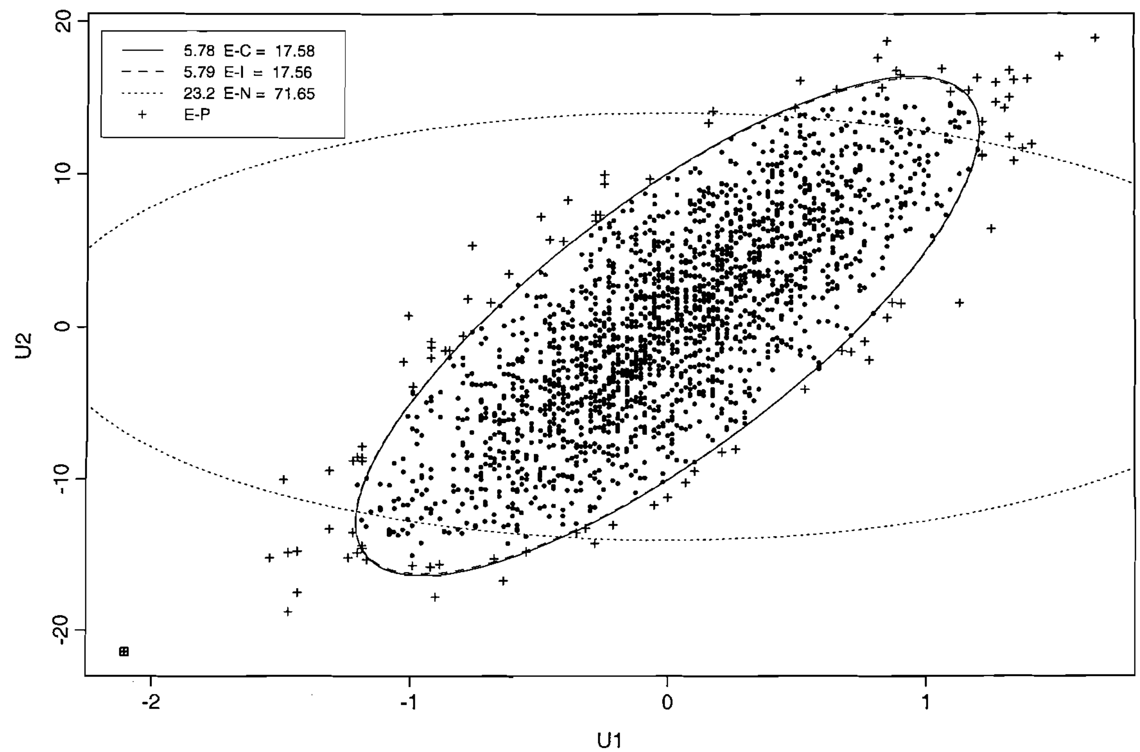

Figure 6. A random sample of 1999 points from the permutation distribution of **U** with the rejection regions for E_C, E_P, E_I, and E_N depicted. The observed value of **U** is marked with a □.