Journal of the Korean Statistical Society (2004), 33: 1, pp 79–97

A PERMUTATION APPROACH TO THE BEHRENS-FISHER PROBLEM

MICHAEL A. PROSCHAN[1] AND DEAN A. FOLLMANN[2]

ABSTRACT

We propose a permutation approach to the classic Behrens-Fisher problem of comparing two means in the presence of unequal variances. It is motivated by the observation that a paired test is valid whether or not the variances are equal. Rather than using a single arbitrary pairing of the data, we average over all possible pairings. We do this in both a parametric and nonparametric setting. When the sample sizes are equal, the parametric version is equivalent to referral of the unpaired t-statistic to a t-table with half the usual degrees of freedom. The derivation provides an interesting representation of the unpaired t-statistic in terms of all possible pairwise t-statistics. The nonparametric version uses the same idea of considering all different pairings of data from the two groups, but applies it to a permutation test setting. Each pairing gives rise to a permutation distribution obtained by relabeling treatment and control within pairs. The totality of different mean differences across all possible pairings and relabelings forms the null distribution upon which the p-value is based. The conservatism of this procedure diminishes as the disparity in variances increases, disappearing completely when the ratio of the smaller to larger variance approaches 0. The nonparametric procedure behaves increasingly like a paired t-test as the sample sizes increase.

AMS 2000 subject classifications. Primary 62Fxx, 62Gxx; Secondary 60C05.

Keywords. Unequal variances, permutation test, t-test, nonparametric methods.

Received November 2002; accepted October 2003.

[1]Office of Biostatistics Research, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, Bethesda, Maryland 20892-7938, U.S.A.

[2]National Institute of Allergy and Infectious Diseases, 6700 A Rockledge Drive, Bethesda, Maryland 20892, U.S.A.

1. INTRODUCTION

The classic Behrens-Fisher problem (Fisher, 1935) involves comparing two treatment means in the presence of possibly unequal variances. This is important in practice because treatments may increase variability. This happened in the Dietary Approaches to Stop Hypertension (DASH) study (Appel *et al.*, 1997) comparing two diets to a control diet with respect to change in diastolic blood pressure; the two diets produced greater blood pressure reductions than the control diet, but the variances were also greater. The usual, pooled variance t-statistic is uniformly most powerful among unbiased tests when the variances are equal, but it has problems in the setting of unequal variances. Let $\{X_i\}_{i=1}^{n_X}$ and $\{Y_i\}_{i=1}^{n_Y}$ be random samples from $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$, respectively, and we wish to test whether $\mu_X = \mu_Y$.

The usual pooled estimate of variance is close to $(n_X s_X^2 + n_Y s_Y^2)/(n_X + n_Y)$, where s_X^2 and s_Y^2 are sample variances, so the squared denominator of the usual t-statistic is close to $\{(n_X s_X^2 + n_Y s_Y^2)/(n_X + n_Y)\}(1/n_X + 1/n_Y) = s_X^2/n_Y + s_Y^2/n_X$. In other words, it is as if we used a large sample normal approximation, but mistakenly reversed the sample sizes in the formula. It is well-recognized that the type I error rate is inflated, even asymptotically, if the larger sample comes from the population with the smaller variance. What is not so well recognized is that the type I error rate can be inflated even when the sample sizes are equal. For example, suppose that $n_1 = n_2 = 5$ and σ_X^2 is extremely tiny compared to σ_Y^2. The pooled variance will essentially be s_Y^2, but we will use 8 degrees of freedom instead of 4. A two-tailed test at $\alpha = 0.05$ rejects if $|t_8| > 2.306$, which will happen with probability close to $Pr(|t_4| > 2.306) = 0.082$.

Several proposed solutions to the Behrens-Fisher problem involve the d-statistic

$$d = \frac{\overline{Y} - \overline{X}}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}. \tag{1.1}$$

When $n_X = n_Y$, the d-statistic is the usual, pooled variance t-statistic. When the population variances are unequal, d does not have a t-distribution. Nonetheless, one could naively refer d to a t-distribution with $M = n_X + n_Y - 2$ degrees of freedom. A more conservative approach uses $m = \min(n_X - 1, n_Y - 1)$ degrees of freedom. Mickey and Brown (1966) proved that the true distribution of d lies somewhere between a t with df, m and a t with df, M (see also Scheffe, 1970). Welch (1949) gave an approximate number, r, of df to use with the t-table; his formula for r always yields a value between m and M. Fisher (1935)'s critical

value for d was derived using the controversial method of fiducial probability, and is therefore not generally accepted.

Barnard (1982, 1984) considered different ratios of standard deviations, $\rho = \sigma_Y/\sigma_X$. For each ρ, one could construct a valid t-test. If the p-value remains small for reasonable values of ρ, one can be confident of a treatment difference. Sprott and Farewell (1993) also discuss this approach.

We take a very different tack to the Behrens-Fisher problem by considering paired statistics. Section 2 discusses the parametric setting. We consider the equal sample size case, $n_X = n_Y = n$. The idea is that although the unpaired t-test assumes that $\sigma_X^2 = \sigma_Y^2$, the paired t-test does not. We could do a paired t-test on $(X_1, Y_1), \ldots, (X_n, Y_n)$. But there is equal justification for any other pairing of the data. That is, we could just as easily have paired $(X_1, Y_{\pi_1}), \ldots, (X_n, Y_{\pi_n})$ for any permutation π of the integers $\{1, 2, \ldots, n\}$. Of course the numerator of the t-statistic remains $\overline{Y} - \overline{X}$ for any π. The sample variance of the n differences changes with π; we show in Section 2 that the average squared denominator across all possible pairings is the squared denominator of the d statistic. If we had used the variance of differences from a single pairing, the resulting test statistic would have had a t-distribution with $n - 1$ degrees of freedom, so it stands to reason that when we use a better estimate of variance and refer to a t_{n-1} distribution, the procedure should be conservative. We prove this and show that the conservatism dissipates as the disparity in variances increases; it becomes an exact procedure when the ratio of the smaller to larger variance tends to 0. Under this circumstance our procedure essentially reduces to a one sample problem for which a t_{n-1} null distribution is exactly right.

The idea of considering all possible pairings and combining the corresponding one-sample paired difference statistics is applied in a nonparametric setting in Section 3. A permutation test is used in lieu of a t-test of the paired differences. We show in Section 3 that, like its parametric counterpart, the nonparametric procedure is conservative, but the degree of conservatism disappears as the ratio of the smaller to larger variance tends to 0.

2. ALL POSSIBLE PAIRINGS AND t-STATISTICS

As noted above, scrambling the Y data does not change the mean difference, but does change the sample variance $s_{D_\pi}^2$ of the differences $\{D_{\pi_i}\}_{i=1}^n = \{Y_{\pi_i} - X_i\}_{i=1}^n$. It seems reasonable to estimate the variance of the difference by the

average variance across all scramblings,

$$\bar{s}_D^2 = \frac{1}{n!}\sum_{\pi} s_{D_\pi}^2.$$

To see what results, consider the data $\underline{x}$ and $\underline{y}$ fixed, and randomly select an index I from $\{1, 2, \ldots, n\}$ and a permutation π. Then X_I and Y_{π_I} are random independent draws from $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, respectively; $\{(n-1)/n\}s_X^2$ and $\{(n-1)/n\}s_Y^2$ are the variances of the distributions of those random draws, and $\{(n-1)/n\}(s_X^2 + s_Y^2)$ is the variance of the distribution of $Y_{\pi_I} - X_I$ given $\underline{x}$, $\underline{y}$. Furthermore, $\bar{s}_D^2$ may be viewed as a conditional expectation:

$$\frac{n-1}{n}\bar{s}_D^2 = E\{\mathrm{Var}(Y_{\pi_I} - X_I | \underline{x}, \underline{y}, \pi) | \underline{x}, \underline{y}\}.$$

Putting these facts together, we obtain

$$\begin{aligned}
\frac{n-1}{n}(s_X^2 + s_Y^2) &= \mathrm{Var}(D_{\pi_I} \mid \underline{x}, \underline{y}) \\
&= E\{\mathrm{Var}(D_{\pi_I} \mid \underline{x}, \underline{y}, \pi) \mid \underline{x}, \underline{y}\} + \mathrm{Var}\{E(D_{\pi_I} \mid \underline{x}, \underline{y}, \pi) \mid \underline{x}, \underline{y}\} \\
&= \frac{n-1}{n}\bar{s}_D^2 + \mathrm{Var}(\bar{y} - \bar{x} \mid \underline{x}, \underline{y}) \\
&= \frac{n-1}{n}\bar{s}_D^2
\end{aligned}$$

Thus, $(\bar{s}_D^2/n)^{1/2} = \{(s_X^2 + s_Y^2)/n\}^{1/2}$. That is,

RESULT 1. *The numerator and squared denominator of the unpaired t-statistic are averages, over all possible pairings, of the numerators and squared denominators of paired t-statistics.*

The all pairwise comparisons approach calls for referral of the unpaired t-statistic to a t-distribution with df, $n-1$ — the df associated with a paired t-statistic. Mickey and Brown (1966) prove that it is always conservative to refer d to $t_{\min(m,n)-1}$. Here is an easier proof when $m = n$.

$$\begin{aligned}
P_{H_0}\left(\frac{\overline{Y} - \overline{X}}{\sqrt{\sigma_D^2/n}\sqrt{\bar{s}_D^2/\sigma_D^2}} > t_{n-1,\alpha}\right) &= E\left\{1 - \Phi\left(t_{n-1,\alpha}\sqrt{\frac{\bar{s}_D^2}{\sigma_D^2}}\right)\right\} \\
&= E\left\{1 - \Phi\left(t_{n-1,\alpha}\sqrt{\frac{E(s_{D_\pi}^2 \mid \underline{x}, \underline{y})}{\sigma_D^2}}\right)\right\} \\
&\leq E\left[E\left\{1 - \Phi\left(t_{n-1,\alpha}\sqrt{\frac{s_{D_\pi}^2}{\sigma_D^2}}\right) \,\middle|\, \underline{x}, \underline{y}\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= E\left\{1-\Phi\left(t_{n-1,\alpha}\sqrt{\frac{s^2_{D_\pi}}{\sigma^2_D}}\right)\right\} \\
&= E\left[E\left\{1-\Phi\left(t_{n-1,\alpha}\sqrt{\frac{s^2_{D_\pi}}{\sigma^2_D}}\right)\,\middle|\,\pi\right\}\right] \\
&= \alpha.
\end{aligned}
$$

The second step is from Result 1, and the third step is from Jensen's inequality because $f(x) = 1 - \Phi(tx)$ is convex on $x > 0$ for positive t. Although the procedure is conservative when the variances are equal, it becomes exact if the ratio of population variances tends to 0. This is illustrated by Table 1, which compares the rejection rates when the standard, two-sample t-statistic is referred to a t-distribution with the usual $2(n-1)$ versus $(n-1)$ degrees of freedom. To simulate under the alternative hypothesis, we chose μ_Y to yield approximately 80% power using the conservative procedure of referring the unpaired t-statistic to a t-distribution with $n-1$ degrees of freedom. The respective type I error rates are 0.077 and 0.050 for a sample size of 6/arm when one standard deviation is 10 times the other. Because the ratio of population variances is rarely known in advance, it is prudent to use a method that protects the type I error rate under all circumstances.

Table 1 also illustrates the loss in power from using a T_{n-1} null distribution. When $\sigma_Y/\sigma_X = 1$, T_{2n-2} is a valid null distribution. The power disparity between T_{2n-2} and T_{n-1} is 0.88 versus 0.80 for $n = 6$, but essentially disappears for $n = 24$. When $\sigma_Y/\sigma_X = 10$ it is not fair to compare powers because use of a T_{2n-2} null distribution does not control the type I error rate.

3. The Scramble-Relabel (S-R) Procedure

A less parametric version is to use a permutation test instead of a t-test on the paired differences. As before, the motivation is that it would be valid but arbitrary to pair the data once and use an appropriate test statistic, so we consider all possible pairings instead. We first scramble the treatment data: $y_{\pi_1}, \ldots, y_{\pi_n}$, and then switch the treatment and control labels of between 0 and n of the pairs $(x_1, y_{\pi_1}), \ldots, (x_n, y_{\pi_n})$. This constitutes a single scramble-relabel. We compute the average treatment minus control difference for each scrambling and relabeling. This produces $n!2^n$ average differences, not all unique. The *scramble-relabel* (S-R) p-value for a one-tailed test is the proportion of the $n!2^n$ scramble-relabel

TABLE 1

$\Delta = \mu_Y - \mu_X$	σ_Y/σ_X	n	T_{2n-2}	T_{n-1}
0	1	6	0.054	0.029
		12	0.053	0.041
		24	0.048	0.043
0	10	6	0.077	0.050
		12	0.060	0.049
		24	0.057	0.049
$\neq 0$	1	6	0.876	0.795
		12	0.830	0.799
		24	0.811	0.798
$\neq 0$	10	6	0.870	0.800
		12	0.830	0.799
		24	0.811	0.798

NOTE : *Rejection rates for referral of the standard, two-sample t-statistic to a t-distribution with* $2(n-1)$ *and* $(n-1)$ *degrees of freedom. Results are based on simulation of 10,000 data sets.*

mean differences that lie at least as far in the tail as the mean difference of the original data.

Section 4 gives a detailed example using a subset of data from an actual trial. We illustrate the test using only the first 3 data points in the control and treatment arms, $x = (-1.43, 1.54, 5.89)$ and $y = (-6.75, -0.75, 3.86)$. The six different scramblings of the $\underline{y}$ vector are enumerated on the vertical axis of Figure 1. For each scrambling there are 2^3 ways to relabel treatments, and hence 8 average difference dots in each row. The original average difference, -3.213, shown in dark, is common to each scrambling, as is its reflection about the origin, $+3.213$. These two support points receive greater weight than the others as illustrated by the S-R probability mass function along the horizontal axis. The one-tailed S-R p-value for this example is $14/48 = 0.29$ because 14 of the 48 points on the graph are -3.213 or smaller. The two-tailed p-value is 0.58.

Calculation of the S-R p-value is less computationally intensive than it seems because there are simpler equivalent formulations. For example, for each π, one could use existing software to compute the p-value for a permutation test of whether the mean of $Y_{\pi_1} - X_1, \ldots, Y_{\pi_n} - X_n$ is 0; the average over π of these p-values is the S-R p-value. This representation suggests that the S-R procedure ought to be conservative. After all, for each scrambling π, the permutation p-value p_π is valid in that $Pr(p_\pi \leq \alpha) \leq \alpha$; the average p-value across π has the same mean and no greater variance than p_π, yet we refer it to the same critical

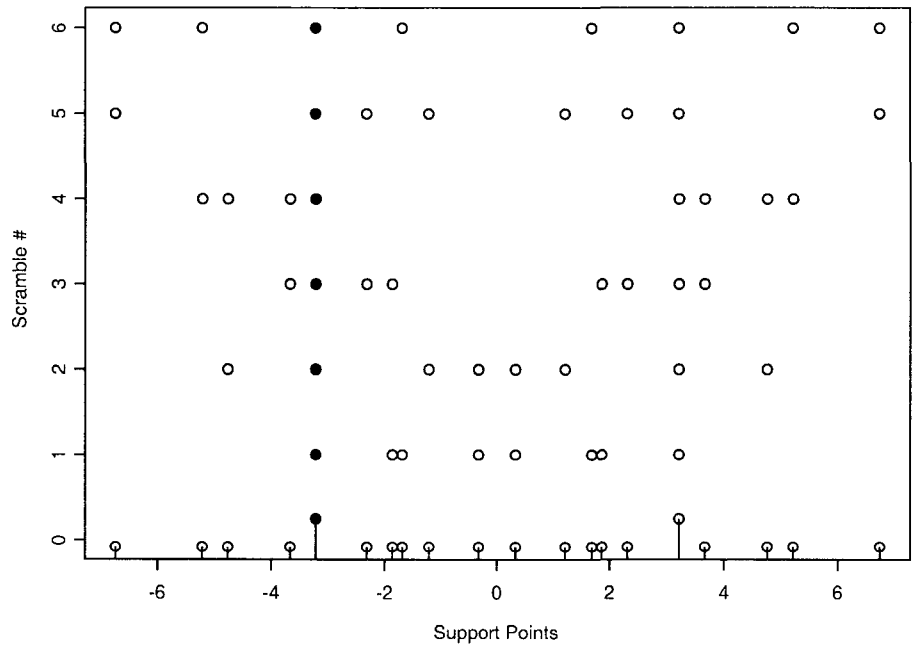

FIGURE 1 *The S-R p-value illustrated for the data* $\underline{x} = (-1.43, 1.54, 5.89)$, $\underline{y} = (-6.75, -0.75, 3.86)$. *For each scrambling enumerated on the vertical axis, the* 8 *values of the mean difference, corresponding to different treatment relabelings within pairs, are shown horizontally. The original mean difference,* $\bar{y} - \bar{x} = -3.213$, *is shown in dark circles and is always a possible value for any scrambling. The one-tailed, S-R p-value is the proportion of the* 48 *circles that lie to the left or include* -3.213; $p = 14/48 = 0.29$. *The two-tailed p-value is* 0.58.

value, *i.e.*, α.

Another way to view the S-R p-value is as follows. Each scrambling and relabeling corresponds to taking a subset of x values and interchanging them with an equal sized subset of y values. Consider the probability that exactly k interchanges occur, and that they occur among specified sets, A and B, of x and y values. There are $\binom{n}{k}$ possible y values that could have matched up, in some order, with the x values in A. Thus, the probability that A and B match up, in some order, is $1/\binom{n}{k}$. Given that they match up, the probability that they, and only they, will be interchanged is $(1/2)^n$. Thus, the probability associated with the interchange of a specific set of k x and k y values is $\{\binom{n}{k}2^n\}^{-1}$. Given this specific interchange, the probability of a mean difference at least as extreme as

TABLE 2

0-switches	*1−interchanges*			*2−interchanges*	*3−interchanges*
−1.43	−1.43	**3.86**	**−0.75**	**−0.75**	-
1.54	**3.86**	1.54	1.54	**3.86**	-
5.89	5.89	5.89	5.89	5.89	-

NOTE : *The $\underline{x}$ vectors following the S-R procedure for which the mean difference is −3.213 or less, by number of interchanges. Original data is $\underline{x} = (-1.43, 1.54, 5.89), \underline{y} = (-6.75, -0.75, 3.86)$. Original x data not bolded, original y data bolded.*

$\bar{y} - \bar{x}$ is simply the indicator that this happens. Thus, the S-R probability of a mean difference at least as extreme as $\bar{y} - \bar{x}$ is simply

$$\sum_{k=0}^{n} \binom{n}{k}^{-1} \frac{m_k}{2^n}, \tag{3.1}$$

where m_k is the number of k-interchange mean differences that are at least as extreme as $\bar{y} - \bar{x}$.

We illustrate this method of calculating the p-value using the data of Figure 1, where $\bar{y} - \bar{x} = -3.213$. The $\underline{x}$ vectors following S-R that give rise to a mean difference less than or equal to −3.213 are given in Table 2.

From Table 2 we see that the respective numbers of 0−, 1−, 2− and 3− interchange differences that are −3.213 or less are $m_0 = 1$, $m_1 = 3$, $m_2 = 1$ and $m_3 = 0$. Thus the S-R p-value using (3.1) can be calculated as $1/(1 \cdot 2^3) + 3/(3 \cdot 2^3) + 1/(3 \cdot 2^3) + 0/(1 \cdot 2^3) = 7/24$.

It is interesting to contrast the S-R procedure with the usual two-sample permutation (UP) test. The UP procedure can be described as permuting all $2n$ data points and pretending that the 1^{st} n of these are from the control group. Each permutation results in a mean difference and the $\binom{2n}{n}$ mean differences provide the null reference distribution. The UP p-value is the proportion of these mean differences that are more extreme than the original $\bar{y} - \bar{x}$. The UP p-value can also be expressed in a fashion similar to (3.1):

$$\sum_{k=0}^{n} \binom{2n}{n}^{-1} m_k. \tag{3.2}$$

The UP p-value for the data of Figure 1 is $1/\binom{6}{3} + 3/\binom{6}{3} + 1/\binom{6}{3} + 0/\binom{6}{3} = 5/20$.

The UP and S-R procedures actually test different null hypotheses. The UP procedure tests the *strong null* that the distribution of X is the same as

the distribution of Y; in particular, this null requires that $\sigma_X^2 = \sigma_Y^2$. Given a particular scramble, the relabeling procedure provides a valid test of the *weak null* that the distribution of $X - Y$ is the same as that of $Y - X$, so $\sigma_X^2 \neq \sigma_Y^2$ is allowed. The S-R permutation p-value is an average, over all possible pairings, of p-values of tests of the *weak null*. Here is the intuitive reasoning that the S-R test is more conservative than the UP test. The $\binom{2n}{n}$ support points for the UP and S-R procedures are identical, but the weights are different. Under UP, the weights are the same for all support points. Under S-R, greater weight is given to support points "closer" to $\bar{y} - \bar{x}$ in a certain sense. The weight is $1/2^n$ for $\bar{y} - \bar{x}$, $1/(2^n n)$ for any support point obtained by switching exactly one (x, y) pair, $2/\{2^n n(n-1)\}$ for any support point obtained by switching exactly two (x, y) pairs, *etc.* The closer a point is to $\bar{y} - \bar{x}$ in terms of this 'switching metric', the greater the weight it has. The support point $\bar{y} - \bar{x}$ is always counted in the p-value, and it receives much greater weight under S-R than UP (see Figure 2 for an example where $n = 6$). Thus, the S-R procedure tends to produce a larger p-value than the UP procedure. Because the UP procedure is valid under equal variances, the S-R procedure should be conservative under equal variances.

Now consider the opposite extreme, namely when the variances of X and Y are markedly different. In this case, the UP procedure is anti-conservative, while the S-R procedure retains control of the type I error rate. To see this, take σ_X fixed and let $\sigma_Y \to \infty$. We can imagine $\underline{X}$ as independent draws from $N(0, \sigma_X^2)$ and $\underline{Y}$ as $\sigma_Y \underline{Z}$, where $\underline{Z}$ are independent draws from $N(0, 1)$. The S-R p-value for $(\underline{X}, \underline{Y})$ is identical to that of $(\underline{X}/\sigma_Y, \underline{Y}/\sigma_Y) = (\underline{X}/\sigma_Y, \underline{Z})$. It is

$$\frac{1}{n!}\sum_{\pi} Pr\left(\frac{1}{n}\sum_{i=1}^{n}(-1)^{U_i}\left|z_{\pi_i} - \frac{x_i}{\sigma_Y}\right| \geq \bar{z} - \frac{\bar{x}}{\sigma_Y}\,\middle|\, \underline{x}, \underline{z}\right). \tag{3.3}$$

where the U_i are independent Bernoulli(1/2) random variables. It is tempting to conclude straightaway that as $\sigma_Y \to \infty$, (3.3) tends to

$$\frac{1}{n!}\sum_{\pi} Pr\left(\frac{1}{n}\sum_{i=1}^{n}(-1)^{U_i}|z_{\pi_i}| \geq \bar{z}\,\middle|\, \underline{z}\right) = Pr\left(\sum_{i=1}^{n}(-1)^{U_i}|z_i| \geq \bar{z}\,\middle|\, \underline{z}\right). \tag{3.4}$$

Some care is needed because of the jump points of (3.3) and (3.4) at $\bar{z} - \bar{x}/\sigma_Y$ and $\bar{z}$, respectively. Let $A = \{u : (1/n)\sum |z_{\pi_i} - x_i/\sigma_Y|(-1)^{u_i} \geq \bar{z} - \bar{x}/\sigma_Y\}$ and $B = \{u : (1/n)\sum |z_{\pi_i}|(-1)^{u_i} \geq \bar{z}\}$. Except for a set of $\underline{z}$ with probability 0, there is exactly one vector $\underline{u}$ such that $\sum |z_{\pi_i}|(-1)^{u_i} = \bar{z}$. Outside this exceptional set, $A \subset B$ for σ_Y sufficiently large. Thus, $Pr(A) \leq Pr(B)$. To see that $Pr(B) \leq$

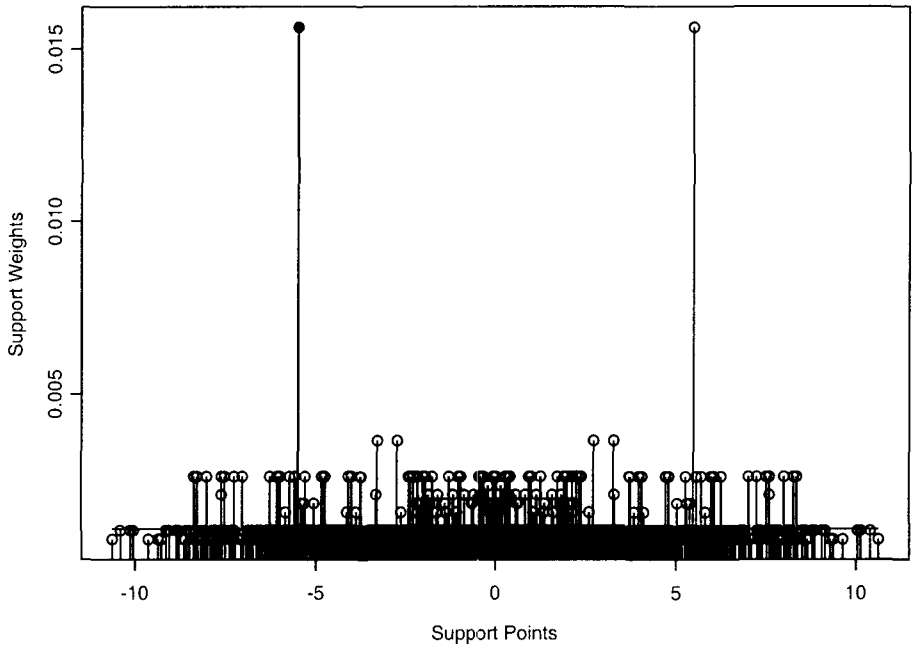

FIGURE 2 *Probability mass functions for the strong and weak null test statistics applied to the 6 selected patients in each of the Control and Fruits & Vegetables arms of the Dietary Approaches to Stop Hypertension (DASH) study. The tall spikes illustrate the large probabilities associated with the sample mean difference, −5.49, and its mirror reflection, +5.49, under the weak null hypothesis. The horizontal line represents equal probabilities across all support points under the strong null hypothesis.*

$Pr(A)$, Note that $Pr(A) = 1/2^n + Pr(A')$, $Pr(B) = 1/2^n + Pr(B')$, where A' is obtained from A by replacing $\geq$ with $>$, and similarly for B'. Clearly, $B' \subset A'$ for σ_Y sufficiently large, so $Pr(B) = 1/2^n + Pr(B') \leq 1/2^n + Pr(A') = Pr(A)$.

In summary, when the null hypothesis is true, the S-R procedure is conservative when the variances are equal and just right when one variance dwarfs the other, because in that case the p-value is like that from a 1-sample permutation test in which the true mean is 0.

We conducted a simulation study to compare the peformance of the S-R procedure to a simple paired difference permutation (PP) test. The PP test is equivalent to the S-R procedure with a single scramble and theoretically controls the type I error rate for fixed $\underline{x}, \underline{y}$ under the weak null. Intuitively, the S-R

procedure should perform better. Each simulated trial consisted of n X's and Y's from normal distributions with respective means $\mu_X = 0$ and μ_Y and respective standard deviations $\sigma_X = 1$ and σ_Y. The value of μ_Y was the same as in Table 1. The standard deviation σ_Y was chosen to be either equal to or 10 times as large as σ_X. For each simulated data set, we approximated the PP procedure by randomly relabeling the original pairs of X's and Y's 199 times. We approximated the S-R procedure by randomly selecting 199 permutations π and for each of these scramblings, we randomly relabeled the pairs $(x_1, y_{\pi 1}), \ldots, (x_n, y_{\pi n})$. The approximate PP and S-R p-values were used to reject or not for each data set, and the proportion of rejections among the 10,000 generated data sets was recorded.

The results are shown in Table 3. Interestingly, under the null, the S-R procedure is less likely to reject than the PP procedure when $\sigma_Y/\sigma_X = 1$, especially for small n. Nevertheless, under the alternatve, the S-R procedure always has better power than the paired permutation test.

TABLE 3

$\Delta = \mu_Y - \mu_X$	σ_Y/σ_X	n	*Paired Permutation*	*Scramble-Relabel*
0	1	6	0.042	0.015
		12	0.051	0.039
		24	0.048	0.042
0	10	6	0.043	0.039
		12	0.048	0.049
		24	0.050	0.050
$\neq 0$	1	6	0.628	0.658
		12	0.788	0.796
		24	0.788	0.797
$\neq 0$	10	6	0.640	0.669
		12	0.788	0.796
		24	0.788	0.797

NOTE : *Simulated rejection rates for nonparametric tests of equality of two means. The PP and S-R methods provide a permutation distribution for the mean difference. For each line, 10,000 data sets are simulated. The PP null distribution is approximated by relabeling the original pairs 199 times. The S-R null distribution is approximated by scrambling the* Y*'s 199 times. For each scrambling, 199 relabelings are done.*

4. EXAMPLE

The Dietary Approaches to Stop Hypertension (DASH) trial (Appel *et al.*, 1997) compared the change in diastolic blood pressure from baseline to end of study in each of three dietary patterns. The Control pattern was similar to what many Americans eat. The Fruits & Vegetables dietary pattern contained, aptly enough, more fruits and vegetables, while the Combination dietary pattern emphasized both fruits and vegetables and lowfat dairy products. Because it is easier to demonstrate and contrast permutation tests with small data sets, we restrict attention to data from only 6 participants each from the Control and Fruits & Vegetables diets.

Table 4 shows the diastolic blood pressure changes from baseline to end of study and the means and standard deviations for the six participants selected from the Control and Fruits & Vegetables diets. Of concern is the fact that the sample standard deviation is appreciably larger in the Fruits & Vegetables arm relative to the Control arm. In fact, the F-test comparing the two variances has a p-value of about 0.03.

One question of interest might be: Does diet affect blood pressure change? The relevant null hypothesis is that the distribution of blood pressure changes is the same in both arms. The standard two-sample permutation test is a valid test of this null hypothesis. An alternative question is: Does diet affect the mean blood pressure change? This is a weaker null hypothesis that includes the situation in which the mean changes are equal but the variances are different. The standard two-sample permutation test is not appropriate, so the S-R procedure provides a better test of this null hypothesis

Figure 2 compares the probability mass functions associated with the tests of the strong and weak null hypotheses. The support points are the same. The horizontal line indicates constant probability over those support points for the test of the strong null, whereas the test of the weak null gives much larger probability to the observed mean difference and its mirror image about the origin.

TABLE 4

Control	−1.43	1.54	5.89	4.49	−2.49	−2.23	$\bar{x} =$ 0.97	$s = 3.60$
F & V	−6.75	−0.75	3.86	−11.89	−17.80	6.14	$\bar{y} = -4.53$	$s = 9.31$

NOTE : *Diastolic blood pressure changes from baseline to end of study from 6 participants in each of the Control and Fruits & Vegetables arms of the Dietary Approaches to Stop Hypertension trial.*

For this example, the strong and weak null tests give p-values of 0.11 and 0.13, respectively.

5. UNEQUAL SAMPLE SIZES

Thus far we have assumed equal sample sizes. For many applications, sample sizes are very close to equal. We use a hypothetical example to illustrate how to proceed when the sample sizes are almost equal. Suppose that the sample sizes are $n_X = 18$ and $n_Y = 20$. First discard two Y observations at random and obtain the S-R p-value comparing the 18 X values to the 18 Y values. Repeat many times the procedure of randomly discarding two Y observations and obtaining the S-R p-value on the resulting equally sized samples, and then average the p-values.

There are some settings in which the sample sizes may differ markedly. For example, in a multi-armed clinical trial there may be interest in combining two or more active arms and comparing to the control. The sample size in the combined arm is a multiple of the sample size in the control. When $n_Y = kn_X$, there is an alternative that is preferable to the procedure outlined above. Randomly divide the Y observations into n_X groups of size k. Let $\underline{\boldsymbol{G}}_i$ and $\overline{Y}(\underline{\boldsymbol{G}}_i)$ be the set of indices and sample mean Y value, respectively, for group i, $i = 1, \ldots, n_X$. Obtain an S-R p-value applied to $(X_1, \overline{Y}(\underline{\boldsymbol{G}}_1)), \ldots, (X_{n_X}, \overline{Y}(\underline{\boldsymbol{G}}_{n_X}))$. Repeat many times this procedure of randomly forming n_X groups of Y values and computing an S-R p-value applied to $(X_1, \overline{Y}(\underline{\boldsymbol{G}}_1)), \ldots, (X_{n_X}, \overline{Y}(\underline{\boldsymbol{G}}_{n_X}))$. Average the resulting p-values. We call this the grouped scramble-relabel procedure (GS-R).

Table 5 evaluates the rejection probabilities of the GS-R procedure when one arm is twice the size of the other. Rejection rates are contrasted with those of the grouped paired permutation procedure (GPP) in which the Ys are grouped into n_X pairs just one time and a permutation test is applied to paired differences between X values and Y averages.

Here is a parametric variant motivated by Result 1 for the case $n_Y = kn_X$. For notational ease, we use the letters m and n for n_X and n_Y, respectively. Let $\underline{\boldsymbol{G}} = (\underline{\boldsymbol{G}}_1, \ldots, \underline{\boldsymbol{G}}_m)$. Randomly scramble $\overline{Y}(\underline{\boldsymbol{G}}_1), \ldots, \overline{Y}(\underline{\boldsymbol{G}}_m)$ and compute the numerator and squared denominator of the paired t-statistic comparing $(X_1, \overline{Y}(G_{\pi_1}))$, $\ldots, (X_m, \overline{Y}(G_{\pi_m}))$. Average these numerators and squared denominators over all possible scramblings π. The average numerator will clearly be $\bar{y}_n - \bar{x}_m$. By Result 1, the average squared denominator is $(s_X^2 + s_{\overline{Y}}^2(\underline{\boldsymbol{G}}))/m$, where $s_{\overline{Y}}^2(\underline{\boldsymbol{G}})$ is the sample variance of $(\overline{Y}(\underline{\boldsymbol{G}}_1), \ldots, \overline{Y}(\underline{\boldsymbol{G}}_m))$. We can write $s_{\overline{Y}}^2(\underline{\boldsymbol{G}})$ as $\{m/(m-1)\}\mathrm{Var}(\overline{Y}(I, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}, \underline{\boldsymbol{G}})$, where $\overline{Y}(I, G)$ is randomly selected from $(\overline{Y}(\underline{\boldsymbol{G}}_1), \ldots,$

TABLE 5

$\Delta = \mu_Y - \mu_X$	σ_Y/σ_X	n_X	*Paired Permutation*	*Scramble-Relabel*
0	1	6	0.043	0.015
		12	0.051	0.037
		24	0.050	0.045
0	0.1	6	0.046	0.043
		12	0.050	0.049
		24	0.052	0.050
0	10	6	0.045	0.016
		12	0.051	0.036
		24	0.048	0.043
$\neq 0$	1	6	0.640	0.687
		12	0.780	0.788
		24	0.797	0.803
$\neq 0$	0.1	6	0.660	0.696
		12	0.791	0.802
		24	0.795	0.804
$\neq 0$	10	6	0.645	0.689
		12	0.775	0.787
		24	0.789	0.795

NOTE : *Simulated rejection rates for nonparametric tests of equality of two means when* $n_Y = 2n_X$. *The GPP and GS-R methods provide a permutation distribution for the mean difference. For each line, 10,000 data sets are simulated. The GPP null distribution is approximated by 1) averaging random pairs of* Y*s, randomly matching these* n_X *averages with the* X*s and then 2) randomly relabeling these synthetic pairs 199 times. The GS-R null distribution is approximated by repeating part 1) above 199 times and for each synthetic pairing, 199 relabelings are done.*

$\overline{Y}(\underline{\boldsymbol{G}}_m)$). Now average over all possible $\underline{\boldsymbol{G}}$. We obtain $\{m/(m-1)\}E\{\mathrm{Var}(\overline{Y}(I, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}\}$. Because $E\{\mathrm{Var}(\overline{Y}(I, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}\} + \mathrm{Var}\{E(\overline{Y}(I, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}\} = \mathrm{Var}(\overline{Y}(I, \underline{\boldsymbol{G}}) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}})$ and $\mathrm{Var}\{E(\overline{Y}(I, G) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}, \underline{\boldsymbol{G}})\} = \mathrm{Var}\{\bar{y}_n - \bar{x}_m \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}) = 0$, the squared denominator averaged over groupings and permutations is $\mathrm{Var}(\overline{Y}(G, I) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}})$. But this may be viewed as the variance of the mean of a sample of k items drawn without replacement from a population of n items. Thus, $\mathrm{Var}(\overline{Y}(G, I) \mid \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}) = \{1 - (k-1)/(n-1)\}\hat{\sigma}^2_{Y\cdot}/k$, where $\hat{\sigma}^2_{Y\cdot} = (n-1)s^2_{Y\cdot}/n$ and $s^2_{Y\cdot}$ is the sample variance of Y. Substituting $n = km$ and simplifying yields an average squared denominator of $(s^2_{X\cdot} + s^2_{Y\cdot}/k)/m = s^2_{X\cdot}/m + s^2_{Y\cdot}/n$. We summarize as follows.

RESULT 2. *When* $n = km$, *the numerator and squared denominator of the d-statistic (1.1) are averages over all groupings* $\underline{\boldsymbol{G}}$ *and permutations* π *of the nu-*

TABLE 6

$\Delta = \mu_Y - \mu_X$	σ_Y/σ_X	n_X	T_{3n_X-2}	T_{n_X-1}
0	1	6	0.059	0.025
		12	0.056	0.038
		24	0.054	0.045
0	0.1	6	0.091	0.054
		12	0.067	0.049
		24	0.060	0.051
0	10	6	0.060	0.026
		12	0.053	0.037
		24	0.051	0.043
$\neq 0$	1	6	0.886	0.791
		12	0.838	0.794
		24	0.825	0.805
$\neq 0$	0.1	6	0.893	0.802
		12	0.843	0.801
		24	0.823	0.803
$\neq 0$	10	6	0.890	0.793
		12	0.833	0.792
		24	0.818	0.797

NOTE : *Rejection rates for referral of the d-statistic to a t-distribution with* $3n - 2$ *or* $(n - 1)$ *degrees of freedom, when* $n_Y = 2n_X$. *Results are based on simulation of 10,000 data sets.*

merator and squared denominator of the paired t-statistic applied to $(X_1, \overline{Y}(\underline{G}_1))$, $\ldots, (X_m, \overline{Y}(\underline{G}_m))$.

Thus, the parametric procedure refers the d-statistic to a t-distribution with $m-1$ degrees of freedom. As noted earlier, Mickey and Brown (1966) have proven that this is always conservative. Table 6 is an extension of Table 1 to the case in which the sample size in one arm is twice that of the other.

6. DISCUSSION

The pooled variance t-statistic has remained the most popular choice for comparing two means despite the fact that it can inflate the type I error rate when the population variances differ. Had the data been paired, the sample variance of paired differences would have taken account of differences in population variances automatically. We applied the simple idea of considering every possible pairing and essentially averaging over them, in both a parametric and nonparametric setting.

Loosely speaking, we showed that in the parametric setting the usual, unpaired t-statistic can be viewed in terms of averages of paired t-statistics. Therefore, the principle of averaging over all possible pairings calls for referral of the unpaired t-statistic to a t-table with half the usual degrees of freedom. The procedure is conservative, but the degree of conservatism diminishes as the disparity in variances increases.

The all-possible-pairings principle applied to a permutation test instead of the t-test leads to the scramble-relabel procedure. Simulation results show that the S-R procedure is conservative when the population variances are equal, and becomes progressively less so as the disparity between them increases. Simulation also shows that the S-R procedure has better power than a paired permutation test. The S-R procedure becomes equivalent to the paired t-test as the sample size increases. The S-R method is a valuable tool for robust testing.

ACKNOWLEDGEMENTS

We thank Stephen Senn for his careful review and constructive comments, including calling our attention to relevant literature. We also thank the referees whose comments led to a more complete exposition of the unequal sample size case and a clearer manuscript.

APPENDIX : ASYMPTOTIC DISTRIBUTION OF S-R TEST STATISTIC

It is helpful to write the scramble-relabel distribution function in a more formal way. Let U_i be independent Bernoulli random variables with parameter 1/2. For fixed $\underline{x} = (x_1, \ldots, x_n)$, $\underline{y} = (y_1, \ldots, y_n)$, and permutation π, define $W(\pi_i) = W_{\underline{x},\underline{y}}(\pi_i) = (-1)^{U_i}|y_{\pi_i} - x_i|$, $i = 1, \ldots, n$. Let $S_n^\pi = \sum_i W(\pi_i)$. The scramble-relabel distribution function is defined by

$$F_n(s \mid \underline{x}, \underline{y}) = \frac{1}{n!}\sum_\pi F_n^\pi(s) = \frac{1}{n!}\sum_\pi Pr(S_n^\pi \le s \mid \underline{x}, \underline{y}, \pi).$$

THEOREM. *If $x_1, \ldots, x_n, \ldots$ and $y_1, \ldots, y_n, \ldots$ are realized values of random samples from distributions with equal means, unequal variances, and finite fourth moments, then*

$$Pr\left\{x_1, \ldots, x_n, \ldots; y_1, \ldots, y_n, \ldots : F_n\left(s(n\sigma_D^2)^{1/2} \mid \underline{x}, \underline{y}\right) \to \Phi(s)\right\} = 1.$$

PROOF. Without loss we take the common mean to be 0. Let $G_n^\pi(s)$ be the scramble-relabel distribution function of the standardized sum, $S_n^\pi/(v_n^\pi)^{1/2}$.

where $v_n^\pi = \sum_{i=1}^n (y_{\pi_i} - x_i)^2$ is the conditional variance of S_n^π given $\underline{x}$, $\underline{y}$, π. Then

$$\begin{aligned}
&\left| F_n\left(s(n\sigma_D^2)^{1/2} \mid \underline{x}, \underline{y}\right) - \Phi(s) \right| \\
&= \left| \frac{1}{n!} \sum_\pi \left\{ G_n^\pi \left(s\sqrt{\frac{n\sigma_D^2}{v_n^\pi}} \right) - \Phi(s) \right\} \right| \\
&\le \frac{1}{n!} \sum_\pi \left\{ \left| G_n^\pi \left(s\sqrt{\frac{n\sigma_D^2}{v_n^\pi}} \right) - \Phi\left(s\sqrt{\frac{n\sigma_D^2}{v_n^\pi}} \right) \right| \right\} \\
&\quad + \left| \frac{1}{n!} \sum_\pi \left\{ \Phi\left(s\sqrt{\frac{n\sigma_D^2}{v_n^\pi}} \right) - \Phi(s) \right\} \right| \qquad \text{(A.1)}
\end{aligned}$$

We will show that each component of (A.1) tends to 0 almost surely.

To see that the first component tends to 0, apply the Berry-Esseen Lemma:

$$\begin{aligned}
&\left| G_n^\pi \left\{ s\left(\frac{n\sigma_D^2}{v_n^\pi}\right)^{1/2} \right\} - \Phi\left\{ s\left(\frac{n\sigma_D^2}{v_n^\pi}\right)^{1/2} \right\} \right| \\
&\le 6 \sum_{i=1}^n \frac{E\left(|W_{\pi_i}|^3 \mid \underline{x}, \underline{y}, \pi \right)}{(v_n^\pi)^{3/2}} \\
&= (v_n^\pi)^{-3/2} \sum_{i=1}^n |y_{\pi_i} - x_i|^3.
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{v_n^\pi}{n} &\ge \sum_{i=1}^n \frac{y_{\pi_i}^2 + x_i^2}{n} - 2\sqrt{\sum_{i=1}^n \frac{y_{\pi_i}^2}{n} \sum_{i=1}^n \frac{x_i^2}{n}} \\
&= \sum \frac{y_i^2}{n} + \sum \frac{x_i^2}{n} - 2\sqrt{\sum_{i=1}^n \frac{y_i^2}{n} \sum_{i=1}^n \frac{x_i^2}{n}} \\
&= h(\underline{x}, \underline{y}) \to (\sigma_Y - \sigma_X)^2 > 0. \qquad \text{(A.2)}
\end{aligned}$$

This, in conjunction with the elementary inequality

$$\sum |y_{\pi_i} - x_i|^3 \le 2^3 \sum (|y_{\pi_i}|^3 + |x_i|^3) = 8 \sum (|y_i^3| + |x_i|^3),$$

shows that the first term of (A.1) is bounded above by $48 \sum (|y_i|^3 + |x_i|^3)\, n^{-3/2} \times \{h(\underline{x}, \underline{y})\}^{-3/2}$. This tends to 0 almost surely by the strong law of large numbers applied to $|X|^3$ and $|Y^3|$.

To see that the second term of (A.1) tends to 0 almost surely, let $T_n^\pi = v_n^\pi/(n\sigma_D^2)$ and expand the function $f_s(T_n^\pi) = \Phi\{s/(T_n^\pi)^{1/2}\} - \Phi(s)$ in a Taylor series about 1:

$$f_s(T_n^\pi) = f_s'(1)(T_n^\pi - 1) + f_s''(\eta_n^\pi)(T_n^\pi - 1)^2/2,$$

where η_n^π is between 1 and T_n^π. The linear part $(1/n!)\sum_\pi f_s'(1)(T_n^\pi - 1) = f_s'(1)\{(n-1)(s_X^2 + s_Y^2)/(n\sigma_D^2) - 1\}$ tends to 0 almost surely, so we need show only that the remainder term $(1/n!)\sum_\pi f''(\eta_n^\pi)(T_n^\pi - 1)^2$ does so as well.

By the elmentary inequality (A.2), T_n^π can be no smaller than $h(\underline{x}, \underline{y})/\sigma_D^2$, so η_n^π can be no smaller than $L = L(\underline{x}, \underline{y}) = \min\{1, h(\underline{x}, \underline{y})/\sigma_D^2\}$. It follows that

$$\frac{1}{n!}\sum_\pi f''(\eta_n^\pi)(T_n^\pi - 1)^2 \le B\sum_\pi (T_n^\pi - 1)^2,$$

where $B = B(\underline{x}, \underline{y}) = \sup_{a \ge L} f''(a)$. But $h(\underline{x}, \underline{y})/\sigma_D^2 \to (\sigma_Y - \sigma_X)^2/\sigma_D^2 > 0$ as $n \to \infty$, so for n sufficiently large, $B \le \sup_{a > \min\{1, (1/2)(\sigma_Y - \sigma_X)^2/\sigma_D^2\}} f''(a) < \infty$ because $f''(x)$ is bounded on any interval $[b, \infty)$, $b > 0$. Thus, the proof will be complete if we establish that $(1/n!)\sum_\pi (T_n^\pi - 1)^2 \to 0$ almost surely.

Routine but tedious algebra involving interchanging the order of summation reveals that

$$\begin{aligned}
& \frac{1}{n!}\sum_\pi (T_n^\pi - 1)^2 \\
= \; & \frac{1}{n^2\sigma_D^4}\left[\left(\sum x_i^2 + \sum y_i^2\right)^2 - \frac{4}{n}\sum x_i \sum y_i \left(\sum x_i^2 + \sum y_i^2\right)\right. \\
& \left. + \frac{4}{n}\sum x_i^2 \sum y_i^2 + \frac{4}{n(n-1)}\left\{\left(\sum x_i\right)^2 - \sum x_i^2\right\}\left\{\left(\sum y_i\right)^2 - \sum y_i^2\right\}\right] \\
& - 2\frac{(n-1)(s_X^2 + s_Y^2)}{n\sigma_D^2} + 1 \\
\longrightarrow \; & \frac{(\sigma_X^2 + \sigma_Y^2)^2}{\sigma_D^4} - 0 + 0 + 0 - 2 + 1 = 0
\end{aligned}$$

This completes the proof. □

REFERENCES

APPEL, L. J., MOORE, T. J., OBARZANEK, E., VOLLMER, W. M., SVETKEY, L. P., SACKS, F. M., BRAY, G. A., VOGT, T. M., CUTLER, J. A., WINDHAUSER, M. M., LIN, P. AND KARANJA, N. (1997). "A clinical trial of the effects of dietary patterns on blood pressure", *The New England Journal of Medicine*. **336**, 1117-1124.

BARNARD, G. A. (1982). "A new approach to the Behrens-Fisher problem". *Utilitas Mathematica*, **21B**, 261–271.

BARNARD, G. A. (1984). "Comparing the means of two independent samples", *Journal of the Royal Statistical Society*, **C33**, 266–271.

COCHRAN, W. G. (1953). *Sampling Techniques*, Wiley, New York.

FISHER, R. A. (1935). "The fiducial argument in statistical inference", *Annals of Eugenics (London)*, **6**, 391–398.

MICKEY, M. R. AND BROWN, M. B. (1966). "Bounds on the distribution functions of the Behrens-Fisher statistic", *The Annals of Mathematical Statistics*, **37**, 639 642.

SCHEFFE, H. (1970). "Practical solutions of the Behrens-Fisher problem", *Journal of the American Statistical Association*, **65**, 1501-1508.

SPROTT, D. A. AND FAREWELL, V. T. (1993). "The difference between two normal means", *The American Statistician*, **47**, 126–128.

WELCH, B. L. (1949). "Further note on Mrs. Aspin's tables and on certain approximations to the tabled function", *Biometrika*, **36**, 293 296.