Robust Permutation Tests for Matched-Pairs Designs

William J. Welch; Leopoldo G. Gutierrez

Journal of the American Statistical Association, Volume 83, Issue 402 (Jun., 1988), 450-455.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198806%2983%3A402%3C450%3ARPTFMD%3E2.0.CO%3B2-N

Robust Permutation Tests for Matched-Pairs Designs

WILLIAM J. WELCH and LEOPOLDO G. GUTIERREZ*

A branch-and-bound algorithm is described for finding the permutation (randomization) *P* value in matched-pairs designs without enumeration of the entire reference distribution. It is not restricted to test statistics that are linear in functions of the observations, and permutation tests based on trimmed means are investigated. We apply the algorithm to six examples, demonstrating that the use of a moderately trimmed, instead of an untrimmed, mean can sometimes lead to substantially smaller *P* values and shorter confidence intervals. Confidence intervals are obtained by trial-and-error inversion of the *P* value. Permutation tests arise in randomization inference, though they can be applied to nonrandomized studies. Under the randomization model, permutation tests are exact, giving the correct probability of a Type I error, without distributional assumptions. The observed test statistic is compared with the reference set of test statistics that would occur under all possible randomizations. Thus inference is based on the *known* randomization distribution. This robustness of validity, however, does not necessarily carry over to robustness of efficiency. The mean pair difference, the test statistic often suggested for permutation analysis of matched-pairs designs, is well known to lack robustness to outliers and long-tailed distributions. With a trimmed mean, however, robustness of efficiency also appears possible. Trimming two observations from each tail performs well, relative to no trimming, in the six examples studied. This stretegy reduces a one-sided *P* value of .380 (no trimming) to .028 in a 14-pair example comparing fault rates on telephone equipment. The width of the 95% confidence interval is similarly reduced, by 42%. In the largest example, a cloud-seeding experiment with 37 pairs, the width of the 95% confidence interval for the effect of seeding is reduced by 24%. Thus gains in efficiency large enough to be of practical interest appear possible.

KEY WORDS: Branch and bound; Randomization; Randomization test; Rerandomization; Trimmed mean.

1. INTRODUCTION

Permutation tests were first described by Fisher [1966, chap. 3 (1st ed. 1935)]. In the context of randomized experiments, they are also known as randomization or rerandomization tests. Even if normality assumptions are correct, permutation tests can be as powerful as standard parametric tests (e.g., see Hoeffding 1952; Lehmann and Stein 1949), yet no distributional assumptions are necessary. The only assumption, treatment-unit additivity (Kempthorne 1955), is questionable for some experiments but unnecessary under the null hypothesis of no treatment effects.

Why, then, are permutation tests not widely applied? Basu (1980) argued that the randomization distribution is irrelevant for inference, but even among statisticians with no philosophical objections, the technique is not widespread. We believe there are two fundamental problems.

The first is computational difficulty. For the matched-pairs design considered in this article, the reference set has 2^n test statistics for an experiment with n pairs. Dwass (1957) suggested sampling the randomization distribution, but to avoid different investigators obtaining different analyses, many test statistics must be sampled. Pagano and Tritchler (1983) and Tritchler (1984) introduced pathbreaking algorithms for testing and for constructing confidence intervals. These methods have running times that are polynomial in n, but they are restricted to test statistics that are linear in the observations or in functions of the observations (such as ranks). John and Robinson (1983) gave an algorithm for one- and two-sample problems. Gabriel and Hall (1983) described methods based on pivotal statistics that make testing, confidence intervals, and (even more formidable) power computations feasible (see also Hall 1985). The main disadvantage of these recent developments is that they are restricted to simple—linear or pivotal—test statistics.

The second problem is the limited practical advantage of permutation tests as currently advocated. The much more convenient parametric t test is usually a good approximation to the permutation test for matched-pairs designs if the test statistic is the mean difference, a common choice. [Pitman (1937) and B. L. Welch (1937) gave results for randomized, complete-block designs in general.]

The randomization argument does not depend on the particular test statistic employed. Winsorized means were considered by Lambert (1985) for the two-sample problem. W. J. Welch (1987) showed that rerandomizing the median in matched-pairs designs offers protection against outliers and is computationally very straightforward. In this article we use trimmed means, aiming to combine much of the median's robustness with the mean's effiency for approximately Gaussian distributions. Moreover, the branch-and-bound algorithm we describe enables the exact *P* value to be calculated without explicit enumeration of the entire reference distribution.

2. THE PERMUTATION TEST

For randomized, matched-pairs designs the model of treatment-unit additivity can be written as

$$d_i = \Delta \pm u_i, \qquad i = 1, \ldots, n, \tag{1}$$

where d_i is the observed difference between treatments A and B for pair i, Δ is a constant treatment effect, and u_i

* William J. Welch is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Leopoldo G. Gutierrez is a graduate student in the Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, British Columbia V6T 1Y8, Canada. Welch's research was funded by the Natural Sciences and Engineering Research Council of Canada and National Science Foundation workshop on efficient data collection. The authors are grateful for comments from the associate editor and referees, leading to an improved presentation.

Journal of the American Statistical Association
June 1988, Vol. 83, No. 402, Theory and Methods

is the absolute difference in experimental-unit effects for pair i. Randomization permutes the treatment labels within each pair, and we have $+u_i$ or $-u_i$ if the randomization happens to assign treatment A to the unit with the higher or lower unit effect.

Hence under H_0: $\Delta = 0$, the n observed differences would have been $\pm d_1, \ldots, \pm d_n$, or equivalently $\pm|d_1|, \ldots, \pm|d_n|$, for the 2^n possible randomizations. We shall consider testing H_0: $\Delta = 0$ against H_a: $\Delta > 0$ throughout. The P value for this test is the proportion of the 2^n test statistics $T(\pm|d_1|, \ldots, \pm|d_n|)$ greater than or equal to the observed statistic T_{obs}. In this article, T will be a trimmed sum with n_t (possibly zero) observations trimmed from each tail. The trimmed mean and trimmed sum are equivalent statistics, but the sum avoids division. The alternative H_a: $\Delta < 0$ is handled by reversing the sign of T_{obs}. Similarly, for H_a: $\Delta \neq 0$, because of the symmetry present, T_{obs} is replaced by $|T_{\text{obs}}|$ and the resulting P value is doubled. Tests of H_0: $\Delta = \Delta_0$ are converted to H_0: $\Delta = 0$ by subtracting Δ_0 from all observed differences.

This article considers the case of unrestricted randomization. If the units within pairs can be meaningfully distinguished—for example, older–younger—then a restricted randomization with AB and BA each occurring $n/2$ times may be more appropriate.

3. THE ALGORITHM

3.1 Branch and Bound

The branch-and-bound algorithm extends Fisher's (1966, chap. 3) idea of partitioning the reference set of test statistics according to the number of positive signs attached to $|d_1|, \ldots, |d_n|$. Branch and bound is typically applied to optimization problems, but the idea carries over to the P-value counting problem. Our extension also allows trimming.

The algorithm recursively generates a tree structure. Nodes of the tree correspond to subsets of test statistics in the randomization reference set defined by one or more constraints. If the absolute differences are ordered $|d_{\langle 1\rangle}| \geq \cdots \geq |d_{\langle n\rangle}|$ and $s_{\langle i\rangle}$ takes the value 1 or 0 to indicate whether $|d_{\langle i\rangle}|$ is assigned a positive or a negative sign ($i = 1, \ldots, n$), then these constraints are of the form

$$s_{\min} \leq \sum_{i=f}^{f+e-1} s_{\langle i\rangle} \leq s_{\max}. \tag{2}$$

A constraint can therefore be written $(f, e, s_{\min}, s_{\max})$, where $s_{\min}$ and $s_{\max}$ bound the number of positive signs attached to $|d_{\langle f\rangle}|, \ldots, |d_{\langle f+e-1\rangle}|$, f indexes the first ordered absolute difference with a sign affected by the constraint, and e is the number of signs affected or the extent of the constraint. It is also convenient to keep track of the trimming associated with each constraint: the t_{L} largest and t_{S} smallest signed differences from $|d_{\langle f\rangle}|, \ldots, |d_{\langle f+e-1\rangle}|$ are trimmed.

Let C be one or more constaints (2), and let $\tau(C)$ denote the subset of test statistics allowed by C. The algorithm commences at the root node defined by

$$C_{\text{root}}: \quad 0 \leq \sum_{i=1}^{n} s_{\langle i\rangle} \leq n, \quad t_{\text{L}} = n_t, \quad t_{\text{S}} = n_t,$$

for which $\tau(C_{\text{root}})$ is the entire reference set. If $n_e(C)$ is a function returning the number of test statistics T in $\tau(C)$ at least as extreme as T_{obs}, then the P value is $2^{-n}n_e(C_{\text{root}})$.

In our algorithm, the function $n_e(C)$ generates k new sets of constraints $C_1, \ldots, C_k$ that partition the test statistics $\tau(C)$ of the current, *ancestor* node into sets $\tau(C_1), \ldots, \tau(C_k)$ associated with k *descendant* nodes. The dependence of k on C is omitted. For each of $C_1, \ldots, C_k$ easily computed lower and upper bounds, $T_{\min}$ and $T_{\max}$, on the test statistics in $\tau(C_i)$ are then employed to try to avoid enumerating $\tau(C_i)$. If $T_{\min} \geq T_{\text{obs}}$, then all test statistics T in $\tau(C_i)$ must satisfy $T \geq T_{\text{obs}}$, and the P value is incremented. Similarly, if $T_{\max} < T_{\text{obs}}$ then $T < T_{\text{obs}}$ for all T in $\tau(C_i)$. Only if $T_{\min} < T_{\text{obs}} \leq T_{\max}$ do we enumerate $\tau(C_i)$ by recursively calling $n_e(C_i)$.

Thus the branch-and-bound algorithm branches to subproblems by partitioning the test statistics and uses bounds to avoid enumeration of some subproblems. If the bounds fail, the procedure calls itself recursively.

Partitioning an ancestor commences by choosing one of its c constraints $(f, e, s_{\min}, s_{\max})$. This constraint, acting on $|d_{\langle i\rangle}|$ ($i = f, \ldots, f + e - 1$), is replaced by a pair of new constraints acting on $|d_{\langle i\rangle}|$ [$i = f, \ldots, f + e^{(1)} - 1$] and $|d_{\langle i\rangle}|$ [$i = f + e^{(1)}, \ldots, f + e - 1$], respectively. The value of $e^{(1)}$ is common to all descendants. Thus each descendant copies $c - 1$ ancestor constraints unchanged and replaces the selected constraint by two new constraints $[f^{(j)}, e^{(j)}, s_{\min}^{(j)}, s_{\max}^{(j)}]$ ($j = 1, 2$), where $f^{(1)} = f$, $f^{(2)} = f + e^{(1)}$, and $e^{(2)} = e - e^{(1)}$. Heuristics for choosing both the constraint to be replaced and $e^{(1)}$ will be outlined later.

The descendants differ in the values of $s_{\min}^{(j)}$ and $s_{\max}^{(j)}$ ($j = 1, 2$) for the new constraints. Computation of the bounds $T_{\min}$ and $T_{\max}$ is considerably simplified if $s_{\min}^{(1)} = s_{\max}^{(1)}$ for all descendants. As the replaced ancestor constraint assigned $s_{\min}, \ldots, s_{\max}$ positive signs and the two new constraints have extents $e^{(1)}$ and $e^{(2)}$, one descendant must be created for each value of $s_{\min}^{(1)}$ in the range $s_{\min}^{(1)} = \max[s_{\min} - e^{(2)}, 0], \ldots, \min[e^{(1)}, s_{\max}]$. Remaining positive signs are allocated to the second new constraint: $s_{\min}^{(2)} = \max[s_{\min} - s_{\min}^{(1)}, 0]$ and $s_{\max}^{(2)} = \min[s_{\max} - s_{\min}^{(1)}, e^{(2)}]$. For each descendant it is also convenient for bound computation if the replaced constraint's t_{L} and t_{S} are divided so that the new constraint j has $t_{\text{L}}^{(j)}$ and $t_{\text{S}}^{(j)}$ trimmed signed differences ($j = 1, 2$); rules are given in the Appendix.

The foregoing partitioning rules lead to independent constraints, facilitating bound computations. For a node with c constraints, $T_{\max} = \sum_{j=1}^{c} T_{\max}^{(j)}$ is a sum of contributions from each constraint. Calculation of $T_{\max}^{(j)}$ is described in the Appendix. $T_{\min}$ is given by an analogous sum.

Our partitioning heuristics aim for small values of $T_{\max} - T_{\min}$ for the descendant nodes. Any descendant with $T_{\max} = T_{\min}$ must have $T_{\min} \geq T_{\text{obs}}$ or $T_{\max} < T_{\text{obs}}$. Conversely, large values of $T_{\max} - T_{\min}$ often lead to $T_{\min} < T_{\text{obs}} \leq T_{\max}$ and further partitioning.

We replace the weakest of the c ancestor constraints: the one with the largest contribution to the ancestor $T_{\max} - T_{\min}$. Let $(f, e, s_{\min}, s_{\max})$ denote the chosen constraint. To make $T_{\max}^{(j)} - T_{\min}^{(j)}$ $(j = 1, 2)$ small for both new constraints, $e^{(1)}$ should partition $|d_{\langle f\rangle}|, \ldots, |d_{\langle f+e-1\rangle}|$ into two homogeneous sets. This is easiest to see in the case of the first new constraint, where $s_{\min}^{(1)} = s_{\max}^{(1)}$. If the $|d_{\langle i\rangle}|$ are identical for $i = f, \ldots, f + e^{(1)} - 1$, then $T_{\min}^{(1)} = T_{\max}^{(1)}$. Any $|d_{\langle i\rangle}|$'s that are trimmed when calculating both $T_{\min}$ and $T_{\max}$ contributions for the ancestor constraint are ignored: Let i_1 and i_2 be the first and last i in $i = f, \ldots, f + e - 1$ such that $|d_{\langle i\rangle}|$ is not trimmed. Then if $i_1 < i_2$, we find the i for which $|d_{\langle i\rangle}| - |d_{\langle i+1\rangle}|$ is maximized over $i = i_1, \ldots, i_2 - 1$ and put $e^{(1)} = i + 1 - f$ to separate $|d_{\langle i\rangle}|$ and $|d_{\langle i+1\rangle}|$. If $i_1 = i_2$, however, we arbitrarily partition to the left of i_1 and choose $e^{(1)} = \max(i_1 - f, 1)$.

Finally, it is sometimes possible to tighten a constraint $(f, e, s_{\min}, s_{\max})$ with $s_{\min} < s_{\max}$. The procedure applies to the root node, as outlined by Fisher (1966, chap. 3), and to the second new constraint formed when partitioning (the first new constraint always has $s_{\min} = s_{\max}$). When partitioning, the tightening algorithm is only invoked if the new node cannot be bounded immediately. Consider assigning exactly $s_{\max}$ positive signs to $|d_{\langle i\rangle}|$ $(i = f, \ldots, f + e - 1)$. The revised contribution to the node $T_{\min}$ and hence $T_{\min}$ itself are calculated. If $T_{\min} \geq T_{\text{obs}}$, then all test statistics with exactly $s_{\max}$ positive signs assigned to $|d_{\langle i\rangle}|$ $(i = f, \ldots, f + e - 1)$ are known to contribute to the P value, and $S_{\max}$ may be decremented by 1. The procedure continues iteratively, decrementing $s_{\max}$ while $T_{\min} \geq T_{\text{obs}}$. Analogous iterations then commence, incrementing $s_{\min}$ while $T_{\max} < T_{\text{obs}}$.

Readers interested in the detailed implementation of the algorithm may obtain a C source listing from the first author.

3.2 Example Branch-and-Bound Computations

Ryan, Joiner, and Ryan (1985, pp. 101–104) outlined an experiment to compare two materials for the soles of boys' shoes. The differences (multiplied by 10) in measured wear for 10 boys are 8, 6, 3, −1, 11, −2, 3, 5, 5, and 3.

Assuming that the two materials were randomized to the boys' left and right shoes (the details are not given), we shall test H_0: $\Delta = 0$ against H_a: $\Delta > 0$. With two observations trimmed from each tail, $T_{\text{obs}} = 6 + 3 + 3 + 5 + 5 + 3 = 25$. (Other levels of trimming are explored in Sec. 4.) The ordered absolute differences $|d_{\langle i\rangle}|$ are 11, 8, 6, 5, 5, 3, 3, 3, 2, and 1. Note that different observations will be trimmed when calculating the 1,024 trimmed sums $T(\pm 11, \pm 8, \pm 6, \pm 5, \pm 5, \pm 3, \pm 3, \pm 3, \pm 2, \pm 1)$ in the reference set.

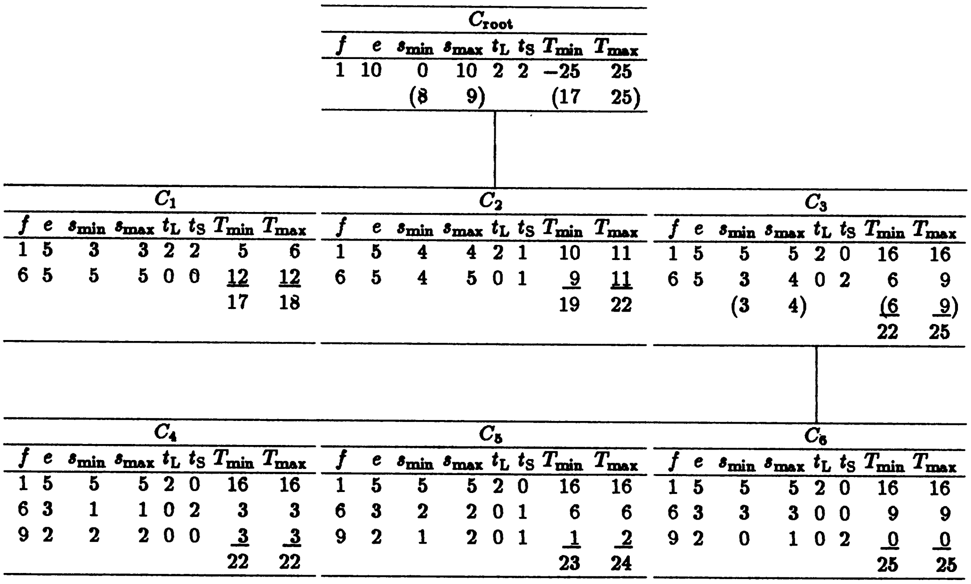
Figure 1 shows the tree generated by the algorithm. The root-node constraint is defined by $f = 1$, $e = 10$, $s_{\min} = 0$, and $s_{\max} = 10$; but tightening leads to $s_{\min} = 8$ and $s_{\max} = 9$, as shown in parentheses in Figure 1. The revised $T_{\min}$ and $T_{\max}$ are also shown in parentheses.

To partition the revised root node, the single constraint is replaced by pairs of new constraints to form descendants $\tau(C_1), \ldots, \tau(C_3)$. Only C_3 is unbounded. Tightening is also ineffective for C_3, the second constraint is replaced, and descendants C_4, C_5, and C_6 are generated. All of these descendants are bounded, and the algorithm terminates.

Contributions to the P value occur when C_{root} is tightened by reducing $s_{\max}$ and when node C_6 is bounded by $T_{\min} \geq T_{\text{obs}}$. From e, $s_{\min}$, and $s_{\max}$, the one-sided P value is $\{\binom{10}{10} + \binom{5}{5}\binom{3}{3}[\binom{2}{0} + \binom{2}{1}]\}/2^{10} = .0039$.

3.3 Differences Equal to Zero

If $|d_{\langle i\rangle}| = 0$, then the sign attached to $|d_{\langle i\rangle}|$ is immaterial. Thus differences exactly equal to 0 may be discarded, n is adjusted, and some computational simplification occurs.



C_{root}

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	10	0	10	2	2	−25	25
		(8	9)			(17	25)

C_1

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	3	3	2	2	5	6
6	5	5	5	0	0	12	12
						17	18

C_2

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	4	4	2	1	10	11
6	5	4	5	0	1	9	11
						19	22

C_3

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	5	5	2	0	16	16
6	5	3	4	0	2	6	9
		(3	4)			(6	9)
						22	25

C_4

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	5	5	2	0	16	16
6	3	1	1	0	2	3	3
9	2	2	2	0	0	3	3
						22	22

C_5

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	5	5	2	0	16	16
6	3	2	2	0	1	6	6
9	2	1	2	0	1	1	2
						23	24

C_6

f	e	$s_{\min}$	$s_{\max}$	t_L	t_S	$T_{\min}$	$T_{\max}$
1	5	5	5	2	0	16	16
6	3	3	3	0	0	9	9
9	2	0	1	0	2	0	0
						25	25

Figure 1. Example of a Branch-and-Bound Tree.

One must take care, however, to determine whether the discarded zeros are trimmed. For example, suppose $n = 10$ differences include two that are exactly 0 and $n_t = 1$. Let n_+ denote the number of positives attached to the undiscarded $|d_{\langle 1\rangle}|, \ldots, |d_{\langle 8\rangle}|$. If $n_+ = 0$, we trim $t_S = 1$, but $t_L = 0$ of the signed $|d_{\langle 1\rangle}|, \ldots, |d_{\langle 8\rangle}|$; the largest signed difference is trimmed from one of the discarded zeros. Similarly, if $n_+ = 8$ then $t_S = 0$. Thus we need to start from three root nodes defined by

$$0 \le \sum_{i=1}^{8} s_{(i)} \le 0, \quad t_L = 0, \quad t_S = 1,$$

$$1 \le \sum_{i=1}^{8} s_{(i)} \le 7, \quad t_L = 1, \quad t_S = 1,$$

and

$$8 \le \sum_{i=1}^{8} s_{(i)} \le 8, \quad t_L = 1, \quad t_S = 0.$$

4. SOME EXAMPLES

We ran the branch-and-bound algorithm for the six data sets briefly described in Table 1. In each case we varied n_t, the number of observations trimmed from each tail, from 0 to $n/2 - 1$ (n even) or to $(n - 1)/2$ (n odd).

Table 2 gives the P values. Clearly, choice of test statistic can be important: For the faults data, where $n = 14$, the P value ranges from .024 ($n_t = 4$) to .380 ($n_t = 0$). The sample mean with no trimming lacks robustness to the single outlier present in this data set. The smallest P values are also achieved with moderate trimming for the marijuana data ($n = 9$, $n_t = 1$), the shoes data ($n = 10$, $n_t = 2$), the plants data ($n = 15$, $n_t = 2$), and the rain data ($n = 37$, $n_t = 2$); but the comparisons with no trimming are less striking.

Because a permutation test is exact for the model of treatment-unit additivity, Equation (1), under the null hypothesis, small P values suggest greater power under the alternative. Moreover, as will be seen, tests with smaller P values appear (not surprisingly) to be associated with narrower confidence intervals for the treatment effect. It would definitely not be valid, however, to choose a test statistic minimizing the P value a posteriori. For an exact test, the test statistic must be chosen a priori.

There is, of course, an enormous literature on trimmed means and other robust estimators of location. Rosenberger and Gasko (1983), for example, recommended trimming 25% of the observations from each tail, or less if very-heavy-tailed distributions are excluded. For the small-to-moderate samples of Table 2 (where a permutation is feasible), the simple rule of trimming two observations from each tail would have given a P value close to the minimum.

The computer time to calculate these P values tended to increase with n but decrease with the amount of trimming. Requirements were trivial for the five smallest data sets where $n = 9$, 10, 14, 15, and 25. The grapefuit data with 25 pairs have a reference set of about 34 million test statistics. Even in the worst case, however, the branch-and-bound algorithm called n_e only 555 times and took about three seconds on a Data General MV/1000 minicomputer.

The rain data with 37 pairs were much more expensive to analyze. The worst case ($n_t = 0$) ran for about 98 minutes on a Data General MV/1000, calling n_e about 8 million times; but trimming nine observations from each tail (about 25% trimming), for example, reduced the running time to about 30 seconds.

P values for the rain data are all fairly high: For all trimming levels, there is little evidence to reject H_0: $\Delta = 0$ against H_a: $\Delta \neq 0$. Confidence-interval comparisons are more interesting, though. A $100(1 - \alpha)\%$ two-sided confidence interval for Δ can be constructed from the set of values of Δ_0 with a P value for H_0: $\Delta = \Delta_0$ against H_a: $\Delta \neq \Delta_0$ exceeding α. Trial and error with the branch-and-bound algorithm yielded 95% confidence intervals of -78 to 30 for $n_t = 0$ and -64 to 18 for $n_t = 2$. Thus trimming two observations from each tail reduced the width of the confidence interval by 24%. The corresponding comparison for the faults data showed a 42% reduction in confidence-interval length.

5. SUMMARY AND DISCUSSION

The branch-and-bound algorithm makes the matched-pairs permutation test computationally trivial for up to about 25 pairs. Larger experiments seem rare, but we successfully enumerated one example with 37 pairs, which has a reference set of about 1.4×10^{11} test statistics. Use

Table 1. Example Data Sets

Name	*Reference*	*n*	*Comparing*
Marijuana	Weil, Zinberg, and Nelsen (1968)	9	Changes in performance on the Digit Symbolization Test for naive subjects smoking high-marijuana and placebo cigarettes for 15 minutes
Shoes	Ryan et al. (1985, pp. 101–104)	10	Wear on two materials for the soles of boys' shoes
Faults	Welch (1987)	14	Reciprocals of fault rates for test and control telephone equipment
Plants	Fisher (1966, chap. 3)	15	Heights of self- and cross-fertilized *Zea mays* plants
Grapefruit	Croxton, Cowden, and Klein (1967)	25	Percentages of solids in shaded and exposed grapefruit halves
Rain	Battan (1966)	37	Rainfalls for seeded and nonseeded days in the Second Arizona Cloud-Seeding Experiment

NOTE: Possible defects in the randomization of the grapefruit data, noted by Preece (1982), are ignored.

Table 2. Permutation P Values When n_t Observations Are Trimmed From Each Tail

	Data set					
n_t	*Marijuana (n = 9)*	*Shoes (n = 10)*	*Faults (n = 14)*	*Plants (n = 15)*	*Grapefruit (n = 25)*	*Rain (n = 37)*
t test	.050	.0085	.329	.025	.0021	.38
0	.047	.0137	.380	.026	.0021	.39
1	.043	.0137	.031	.021	.0025	.32
2	.055	.0078	.028	.012	.0028	.28
3	.074	.0313	.026	.019	.0028	.30
4	.109	.0547	.024	.028	.0027	.34
5			.031	.035	.0031	.36
6			.061	.034	.0030	.39
7				.055	.0050	.38
8					.0047	.38
9					.0041	.38
10					.0034	.37
11					.0051	.37
12					.0048	.36
13						.40
14						.43
15						.41
16						.36
17						.37
18						.38

NOTE: *P* values are two-tailed for the shoes and rain data, one-tailed otherwise.

of a trimmed mean causes no difficulty for the algorithm; indeed, computational effort decreases with the amount of trimming.

Six data sets were analyzed in Section 4. Histograms of the observed differences suggest long-tailed distributions or outliers in most cases. The permutation test with no trimming, however, agreed well with the *t* test, which assumes normality. If the model of unit-treatment additivity (1) holds, the permutation test has the correct probability of a Type I error. The close agreement of the *t* test in this study, therefore, adds to the evidence that the *t* test is robust.

As Miller (1986) pointed out, though, the *t* test's robustness of *validity* does not extend to robustness of *efficiency*. The same comment applies to permutation tests based on the mean. Smaller *P* values and shorter confidence intervals were obtained with a trimmed mean in most of the six examples. We should again caution against choosing the best of several test statistics. The a priori strategy of trimming two observations from each tail would have done well in the examples investigated. Clearly, there is potential for further work here.

One of the greatest strengths of the randomization argument is that it is independent of the test statistic chosen. In principle, therefore, inference is no more complicated for trimmed means or other robust statistics (computational difficulties may arise, however). The matched-pairs design is just the simplest experimental plan; the advantages of permutation tests and randomization inference undoubtedly have wider application.

APPENDIX: SOME COMPUTATIONAL DETAILS

The rules for allocating a replaced constraint's t_L and t_S so that new constraint j trims $t_L^{(j)}$ and $t_S^{(j)}$ signed differences ($j = 1, 2$) are

$$t_L^{(1)} = \min[t_L, s_{\min}^{(1)}] + \max[t_L - s_{\min}^{(1)} - e^{(2)}, 0],$$

$$t_L^{(2)} = t_L - t_L^{(1)},$$

$$t_S^{(1)} = \min[t_S, e^{(1)} - s_{\min}^{(1)}] + \max\{t_S - [e^{(1)} - s_{\min}^{(1)}] - e^{(2)}, 0\},$$

and

$$t_S^{(2)} = t_S - t_S^{(1)}.$$

The t_L largest signed differences are trimmed from, first, the $s_{\min}^{(1)}$ positively signed $|d_{\langle i\rangle}|$ of the first constraint; second, if $t_L - s_{\min}^{(1)} > 0$, from the second constraint; and finally, if $t_L - s_{\min}^{(1)} > e^{(2)}$, from the negatively signed $|d_{\langle i\rangle}|$ of the first constraint. Dividing t_S is analogous. Note that $e^{(1)} - s_{\min}^{(1)}$ is the number of negative signs allocated to $|d_{\langle i\rangle}|$ $[i = f^{(1)}, \ldots, f^{(1)} + e^{(1)} - 1]$ by the first constraint.

The contribution $T_{\max}^{(j)}$ to $T_{\max}$ from a constraint $[f^{(j)}, e^{(j)}, s_{\min}^{(j)}, s_{\max}^{(j)}]$ is given by

$$T_{\max}^{(j)} = \sum_{i=a}^{b} |d_{\langle i\rangle}| - \sum_{i=u}^{v} |d_{\langle i\rangle}|, \qquad \text{(A.1)}$$

where

$$a = f^{(j)} + t_L^{(j)},$$

$$b = f^{(j)} + s_{\max}^{(j)} - 1 - \max\{t_S^{(j)} - [e^{(j)} - s_{\max}^{(j)}], 0\},$$

$$u = f^{(j)} + s_{\max}^{(j)} + t_S^{(j)},$$

and

$$v = f^{(j)} + e^{(j)} - 1 - \max[t_L^{(j)} - s_{\max}^{(j)}, 0].$$

The calculation of $T_{\min}^{(j)}$ is analogous. The partial sums in (A.1) can be looked up in a table.

[*Received June 1986. Revised July 1987.*]

REFERENCES

Basu, D. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test" (with comments), *Journal of the American Statistical Association*, 75, 575–595.

Battan, L. J. (1966), "Silver-Iodide Seeding and Rainfall From Convective Clouds," *Journal of Applied Meteorology*, 5, 669–683.

Croxton, F. E., Cowden, D. J., and Klein, S. (1967), *Applied General Statistics* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

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

Fisher, R. A. (1966), *The Design of Experiments* (8th ed.), Edinburgh: Oliver & Boyd.

Gabriel, K. R., and Hall, W. J. (1983), "Rerandomization Inference on Regression and Shift Effects: Computationally Feasible Methods," *Journal of the American Statistical Association*, 78, 827–836.

Hall, W. J. (1985), "Confidence Intervals, by Rerandomization, for Additive and Multiplicative Effects," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 60–69.

Hoeffding, W. (1952), "The Large-Sample Power of Tests Based on Permutations of Observations," *Annals of Mathematical Statistics*, 23, 169–192.

John, R. D., and Robinson, J. (1983), "Significance Levels and Confidence Intervals for Permutation Tests," *Journal of Statistical Computation and Simulation*, 16, 161–173.

Kempthorne, O. (1955), "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.

Lambert, D. (1985), "Robust Two-Sample Permutation Tests," *The Annals of Statistics*, 13, 606–625.

Lehmann, E. L., and Stein, C. (1949), "On the Theory of Some Non-Parametric Hypotheses," *Annals of Mathematical Statistics*, 20, 28–45.

Miller, R. G., Jr. (1986), *Beyond ANOVA, Basics of Applied Statistics*, New York: John Wiley.

Pagano, M., and Tritchler, D. (1983), "On Obtaining Permutation Distributions in Polynomial Time," *Journal of the American Statistical Association*, 78, 435–440.

Pitman, E. J. G. (1937), "Significance Tests Which May Be Applied to Samples From Any Populations, III: The Analysis of Variance Test," *Biometrika*, 29, 322–335.

Preece, D. A. (1982), "*t* Is for Trouble (and Textbooks): A Critique of Some Examples of the Paired-Samples *t*-Test," *The Statistician*, 31, 169–195.

Rosenberger, J. L., and Gasko, M. (1983), "Comparing Location Estimators: Trimmed Means, Medians, and Trimean," in *Understanding Robust and Exploratory Data Analysis*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: John Wiley, pp. 297–334.

Ryan, B. F., Joiner, B. L., and Ryan, T. A., Jr. (1985), *Minitab Handbook* (2nd ed.), Boston: Duxbury Press.

Tritchler, D. (1984), "On Inverting Permutation Tests," *Journal of the American Statistical Association*, 79, 200–207.

Weil, A. T., Zinberg, N. E., and Nelsen, J. M. (1968), "Clinical and Psychological Effects of Marihuana in Man," *Science*, 162, 1234–1242.

Welch, B. L. (1937), "On the *z*-Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52.

Welch, W. J. (1987), "Rerandomizing the Median in Matched-Pairs Designs," *Biometrika*, 74, 609–614.