Rerandomizing the Median in Matched-Pairs Designs

William J. Welch

Biometrika, Vol. 74, No. 3. (Sep., 1987), pp. 609-614.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198709%2974%3A3%3C609%3ARTMIMD%3E2.0.CO%3B2-B

Biometrika is currently published by Biometrika Trust.

Biometrika (1987), **74**, 3, *pp.* 609-14
Printed in Great Britain

Rerandomizing the median in matched-pairs designs

BY WILLIAM J. WELCH
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

SUMMARY

When a rerandomization or permutation test is applied to a matched-pairs design, the observed mean difference is typically compared with the null distribution of means that would have occurred under all possible randomizations. It is shown that rerandomizing the median instead of the mean transforms a formidable computational problem into one amenable to a hand calculator. Moreover, the robustness derived from rerandomizing the median rather than the mean is illustrated with an example where outliers were anticipated. Computational methods for constructing confidence intervals are also presented, and extensions to trimmed means are discussed.

Some key words: Matched-pairs design; Median; Permutation test; Randomization; Rerandomization test.

1. INTRODUCTION

Fisher (1966 and earlier editions, Ch. 3) showed that the physical act of randomization can provide a basis for statistical inference. The only assumption involved is treatment-unit additivity (Kempthorne, 1955). In the case of the randomized matched-pairs design this model simplifies to

$$d_i = \Delta \pm u_i \quad (i = 1, \ldots, n), \tag{1}$$

where d_i is the observed difference between treatments A and B for pair i, Δ is a constant treatment effect, and u_i is the absolute difference in the experimental-unit effects for pair i. 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 respectively. 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. A significance test follows by comparing an observed test statistic, T_{obs}, with the reference set of the 2^n statistics $T(\pm|d_1|, \ldots, \pm|d_n|)$. Typically, T is the sample mean, and we shall refer to the rerandomized-means test. These tests are also called randomization or permutation tests.

For large n, complete enumeration of the reference set is impractical, even with a computer. Dwass (1957) suggested sampling the reference set. More recently, Pagano & Tritchler (1983) described a path-breaking method of inverting the characteristic function for which computational effort increases only as a polynomial in n. Their algorithm is restricted to statistics linear in functions of the observations, however, and extensions are not obvious. These methods make rerandomization inference feasible, but there is little practical advantage over, say, the t test. Pitman (1937) and B. L. Welch (1937) showed that the rerandomization test is often adequately approximated by methods based on the normal distribution.

The rerandomized-means test has no distributional assumptions, thus ensuring robustness of validity, but there is no guarantee of robustness of efficiency. Mosteller & Tukey

(1977, Ch. 1) discuss these two types of robustness. Rosenberger & Gasko (1983), for example, show that the sample mean is a poor estimator of location for long-tailed distributions and advocate the use of medians or trimmed means. An example in § 4 illustrates that the rerandomized-means test lacks robustness of efficiency to an outlier.

Problems of computation and robustness are both ameliorated by rerandomizing the median rather than the mean. The P value of the rerandomized-medians test for testing H_0: $\Delta = 0$ against H_a: $\Delta > 0$, the only case considered here, is the proportion of the 2^n medians $\tilde{d} = \text{med}\,(\pm|d_1|, \ldots, \pm|d_n|)$ greater than or equal to the observed median $\tilde{d}_{\text{obs}}$. Tests of H_0: $\Delta = \Delta_0$ are converted to H_0: $\Delta = 0$ by subtracting Δ_0 from all pair differences. Similarly, to convert H_a: $\Delta < 0$ or H_a: $\Delta \neq 0$ to H_a: $\Delta > 0$ replace $\tilde{d}_{\text{obs}}$ by $-\tilde{d}_{\text{obs}}$ or $|\tilde{d}_{\text{obs}}|$ respectively, and double the P value for the two-sided test.

In § 2 we develop bounds on the P value from rerandomizing trimmed means in general. In the case of the median, these bounds also lead to a method of calculating the exact P value suitable for hand calculation if n is moderate. These methods would have been particularly advantageous before electronic computers, yet, to the author's knowledge, they were not developed. Confidence intervals, admittedly rather more tedious, follow by inverting significance tests; a systematic method is outlined in § 3. Two examples are discussed in § 4; one demonstrates the robustness of the rerandomized-medians test to outliers. Finally, § 5 discusses extensions of these ideas to calculating the exact P value for trimmed means in general.

2. CALCULATING P VALUES

The results for the median follow as special cases of rerandomizing trimmed means. In general, let t observations be trimmed from each tail so that T is a mean of the $c = n - 2t$ central observations. For any T in the reference set, a value $-T$ exists by reversing all signs attached to $|d_1|, \ldots, |d_n|$. Thus, $P > \frac{1}{2}$ if and only if $T_{\text{obs}} \leqslant 0$, a case we ignore hereafter.

Let $|d_{\langle 1\rangle}| \geqslant \ldots \geqslant |d_{\langle n\rangle}|$ be the ordered absolute differences, and define

$$k_s = \max\left(j: |d_{\langle j-s+1\rangle}| + \ldots + |d_{\langle j\rangle}| \geqslant cT_{\text{obs}} - \sum_{i=1}^{c-s} a_i\right) \tag{2}$$

for $s = c, \ldots, 1$, where

$$a_i = \begin{cases} |d_{\langle k_c+i\rangle}| & (k_c + i \leqslant n), \\ -|d_{\langle k_c+1-i\rangle}| & \text{otherwise.} \end{cases}$$

Note that k_c must be calculated first. For $s \leqslant c-1$ if no j satisfies (2) we write $k_s = 0$. In the case of the median with odd n, k_1 is simply the number of d_i's such that $|d_i| \geqslant \tilde{d}_{\text{obs}}$, and sorting the absolute differences is unnecessary. Definition (2) leads to the following result, proved in the Appendix.

LEMMA 1. *The P value obtained by rerandomizing a trimmed mean T with $T_{\text{obs}} > 0$ satisfies*

$$b(k_c, t+c) \leqslant P \leqslant b(k_c, t+c) + (\tfrac{1}{2})^{k_c} \sum_{s=1}^{c-1} \sum_{j=t+1}^{t+s} \binom{k_s - s + 1}{j} \binom{k_c - k_s + s - 1}{t+s-j},$$

where $b(k, m)$ is the binomial probability of $m, \ldots, k$ successes in k trials when pr (*success*) = $\frac{1}{2}$.

For the median, Lemma 1 gives

$$P = b(k_1, \tfrac{1}{2}n + \tfrac{1}{2}) \tag{3}$$

if n is odd, and

$$b(k_2, \tfrac{1}{2}n+1) \leqslant P \leqslant b(k_2, \tfrac{1}{2}n+1) + (\tfrac{1}{2})^{k_2}\binom{k_1}{\frac{1}{2}n} \tag{4}$$

if n is even.

The author's experience is that Lemma 1 typically gives excellent results for small c. Furthermore, if the exact P value is required, Lemma 1 reduces the number of test statistics in the reference set requiring evaluation.

The remainder of this section is concerned with calculating the exact P value for the median with even n. From (4) the unresolved medians are those with $\frac{1}{2}n$ and 0 positive signs allocated to $|d_{\langle 1\rangle}|, \ldots, |d_{\langle k_1\rangle}|$ and $|d_{\langle k_1+1\rangle}|, \ldots, |d_{\langle k_2\rangle}|$ respectively. If $k_2 = n$, these medians are $\tilde{d}_i = \frac{1}{2}(|d_{\langle i\rangle}| - |d_{\langle n\rangle}|)$ for $i = \frac{1}{2}n, \ldots, k_1$, and, from the definition of k_1, the exact P value equals the upper bound in (4). If $k_2 < n$, however, the unresolved medians are

$$\tilde{d}_i = \tfrac{1}{2}(|d_{\langle i\rangle}| + \max_{j=k_2+1,\ldots,n} s_j|d_{\langle j\rangle}|) \quad (i = \tfrac{1}{2}n, \ldots, k_1),$$

where $s_j = \pm 1$ is the sign allocated to $|d_{\langle j\rangle}|$. Thus,

$$P = b(k_2, \tfrac{1}{2}n+1) + (\tfrac{1}{2})^{k_2} \sum_{i=\frac{1}{2}n}^{k_1} f_i \binom{i-1}{\frac{1}{2}n-1},$$

where

$$f_i = 1 - (\tfrac{1}{2})^{n-k_2} \sum_{j=k_2+1}^{n} t_{ij},$$

t_{ij} is the number of sign allocations $s_j = \pm 1$ with $s_j|d_{\langle j\rangle}| < u_i$, and $u_i = 2\tilde{d}_{\text{obs}} - |d_{\langle i\rangle}|$.

3. Constructing confidence intervals

Confidence intervals based on rerandomizing the median may be derived by the usual method of test inversion: A $(1-\alpha)$ lower limit for the treatment difference Δ is the set of values Δ_0 where a test of H_0: $\Delta = \Delta_0$ against H_a: $\Delta > \Delta_0$ yields $P > \alpha$. As the method of P-value computation in § 2 does not distinguish one-sided P values where $P > \frac{1}{2}$, we assume $\alpha \in (0, \frac{1}{2}]$. Kempthorne & Doerfler's (1969) monotonicity argument is readily adapted to show that this procedure does lead to an interval. Upper limits, and hence two-sided intervals, are analogous.

A finite, conservative limit exists if and only if $P \leqslant \alpha$ for some Δ_0. Let $d_{(1)} \leqslant \ldots \leqslant d_{(n)}$ be the ordered differences. For any $\Delta_0 < d_{(1)}$ all differences are positive after subtracting Δ_0, and it can be shown that the P value is minimized. We assume this minimum P value does not exceed α.

In the simpler case of odd n, the P value (3) is determined by k_1 if $\Delta_0 < \tilde{d}_{\text{obs}}$. Furthermore, k_1 and hence P are step functions of Δ_0 with steps possible only at

$$\Delta_{0i} = \tfrac{1}{2}(d_{(i)} + \tilde{d}_{\text{obs}}) \quad (i = 1, \ldots, m-1), \tag{5}$$

where $m = \frac{1}{2}(n+1)$, the solutions of $|d_{(i)} - \Delta_{0i}| = \tilde{d}_{\text{obs}} - \Delta_{0i}$. Therefore, a conservative lower limit is the smallest Δ_{0i} in (5) for which $P > \alpha$; if no Δ_{0i} leads to $P > \alpha$ then the limit is

$\tilde{d}_{\text{obs}}$. If there are no ties amongst $d_{(1)}, \ldots, d_{(m)}$ then $k_1 = m$ for $\Delta_0 < \Delta_{01}$ and k_1 jumps to $m+i$ at Δ_{0i} for $i = 1, \ldots, m-1$.

Search by bisection is recommended for even n, though special structure can be exploited. The left bracket, L, of the bisection search is initialized to a value less than $d_{(1)}$, and the right bracket, R, is set to $\tilde{d}_{\text{obs}}$. The P value is again a step function, with steps possible at Δ_0 satisfying

$$\pm(d_i - \Delta_0) \pm (d_j - \Delta_0) = 2(\tilde{d}_{\text{obs}} - \Delta_0) \quad (1 \leqslant i < j \leqslant n). \tag{6}$$

Therefore, if the observed differences are recorded as integers, say, the confidence limit must be an integer multiple of $\frac{1}{4}$, and successive midpoints of the bracketed interval are rounded to the nearest $\frac{1}{4}$. Either L or R is replaced by the midpoint to maintain $P \leqslant \alpha$ at $\Delta_0 = L$ and $P > \alpha$ at $\Delta_0 = R$. When $R - L = \frac{1}{4}$, R is a conservative lower limit with actual confidence $(1 - P_L)$, where P_L is the P value at $\Delta_0 = L$.

4. EXAMPLES

The first example, taken from a programme of experiments that motivated this research, illustrates calculation of the exact P value for the median with even n. The experiment tested a method of reducing faults on telephone lines. Fourteen matched pairs of areas were available. Experience suggested that outliers could be anticipated: accidental damage of large cables would generate large batches of fault reports. In this experiment, then, the rerandomized-medians test was expected to perform well a priori relative to tests based on criteria less robust than the median.

Table 1 shows the observed test and control fault rates r_T and r_C. Plotting the pair difference against the pair average suggests that the model of treatment-unit additivity (1) appears plausible on the reciprocal scale. The pairs are ordered in Table 1 so that the absolute values of $d = 1/r_T - 1/r_C$ are nonincreasing. The observed median is 101·5; from (2), $k_2 = 10$ and $k_1 = 8$, and (4) gives $0{\cdot}0547 \leqslant P \leqslant 0{\cdot}0625$ for H_0: $\Delta = 0$ against the alternative H_a: $\Delta > 0$ of a test-method improvement.

Here the bounds are adequate for practical purposes, but, for illustration, Table 2 sets out the method described in § 2 for finding the exact P value in a format convenient for hand calculation. Thus the lower bound needs to be incremented by 0·0060 to give an exact P value of 0·0607.

Table 1. *Fault rates on telephone lines in* 14 *pairs of areas*

	Pair													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Test rate r_T	676	206	230	256	280	433	337	466	497	512	794	428	452	512
Control rate r_C	88	570	605	617	653	2913	924	286	1098	982	2346	321	615	519
$(1/r_T - 1/r_C) \times 10^5$	−988	310	269	229	204	197	189	−135	110	93	83	−78	59	3

Table 2. *Calculating the exact P value for the telephone-faults experiment*

			$\pm\lvert d_{(i)}\rvert < u_i$						
i	$\lvert d_{(i)}\rvert$	u_i	83	78	59	3	f_i	$\left(\frac{1}{2}\right)^{10}\binom{i-1}{6}$	Increment
7	189	14	−	−	−	±	0·875	0·00098	0·0009
8	135	68	−	−	±	±	0·750	0·00684	0·0051

Table 3. *P values for the telephone-faults experiment*

	All data ($n = 14$)	Excluding pair 1 ($n = 13$)
Rerandomized-medians test	0·0607	0·0352
Rerandomized-means test	0·3796	0·0052
t test	0·3292	0·0038

The P values obtained from the rerandomized-medians, rerandomized-means, and t tests are compared in Table 3. Pair 1, with the lowest control rate and the second-largest test-method rate, is arguably an outlier, and P values are also computed with this pair omitted. The rerandomized-means and t tests show extreme sensitivity to exclusion of pair 1.

To obtain say a 0·95 two-sided confidence interval for $\Delta = 1/r_T - 1/r_C$ we perform the rerandomized-medians test of H_0: $\Delta = \Delta_0$ against H_a: $\Delta \neq \Delta_0$ and find that $P > 0{\cdot}05$ for $-21{\cdot}0 \leqslant \Delta_0 \leqslant 189{\cdot}5$, whereas $P \leqslant 0{\cdot}05$ outside this range. The bisection method in § 3 required 25 P-value computations to find the two limits.

The confidence intervals corresponding to some other tests are: $-146{\cdot}2 \leqslant \Delta \leqslant 183{\cdot}8$ for the rerandomized-means test; $-146{\cdot}9 \leqslant \Delta \leqslant 224{\cdot}8$ for the t test; $-21{\cdot}0 \leqslant \Delta \leqslant 196{\cdot}5$ for the Wilcoxon (1945) test; and $-78{\cdot}0 \leqslant \Delta \leqslant 229{\cdot}0$ for the sign test. Thus, the rerandomized-means and t tests give similarly wide intervals for this example. The similarity of the Wilcoxon and rerandomized-medians confidence intervals is not unusual. The Wilcoxon endpoints are Walsh averages $(d_i + d_j)/2$, as are the possible limits (5) and some of the solutions to (6).

Finally, we should caution against the application of multiple tests to a set of data. The rerandomized-medians test was chosen a priori; a biased P value would result from choosing the best of several tests.

Confidence-interval calculations are much simpler when n is odd. Consider the following 15 differences in heights of cross and self-fertilised plants used by Fisher (1966, Ch. 3) to introduce the rerandomization argument: −67, −48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75. Following the method of § 3, we have, for $\Delta_0 < -21{\cdot}5$, $k_1 = 8$, $P = b(k_1, 8) = 0{\cdot}0039$; for $\Delta_0 = -21{\cdot}5$, $k_1 = 9$, $P = 0{\cdot}0195$ and, for $\Delta_0 - 12$, $k_1 = 10$, $P = 0{\cdot}0547$, for the lower limit. A conservative 0·95 two-sided interval for the treatment difference is $-12 \leqslant \Delta \leqslant 42$, with actual confidence $1 - 2 \times 0{\cdot}0195 = 0{\cdot}961$.

5. Discussion

The rerandomized-medians test is readily computed and robust. Unlike the Wilcoxon (1945) signed-ranks test, no modification is needed for ties. In many situations the median is also easier to communicate and interpret than a sum of ranks.

One disadvantage of the proposed method is the limited number of achievable P values or confidence levels when n is odd. A remedy might be to use the mean of the three central order statistics, for which Lemma 1 often gives adequate results. For example, with $c = 3$, Lemma 1 gives $0{\cdot}0327 \leqslant P \leqslant 0{\cdot}0347$ for the plant-height data of § 4, and only a few further trimmed means need to be calculated to establish $P = 0{\cdot}0344$. In general, less-extreme trimming than the median might retain much of the median's robustness and ease of computation, while nearly equalling the mean's performance when pair differences are approximately Gaussian. The rerandomized-means calculations in § 4 were performed using an implicit-enumeration, backtracking algorithm. The algorithm is adaptable to trimmed means and will be reported elsewhere.

Much of the research in rerandomization inference has concentrated on test statistics suggested by classical parametric techniques. The rerandomization argument applies equally well to robust statistics, though, and this is where the greatest benefits appear possible.

The methods forwarded here may also find application in observational studies involving matched pairs if we assume exchangeability of the observations within a pair under the null hypothesis.

Hand calculation has been stressed: P values and confidence intervals are trivial to compute for the median with odd n. With even n the author finds that the exact P value may be calculated in about 15 minutes if n is less than 25, but confidence intervals are fairly tedious. Source code for rerandomized-medians P values and confidence intervals, written in C, is available from the author.

ACKNOWLEDGMENTS

This research was supported by the National Science and Engineering Research Council of Canada and a Workshop on efficient data collection funded by the National Science Foundation. The author is also grateful for the referees' constructive suggestions.

APPENDIX

Proof of Lemma 1

Suppose $t+s$ positive signs are allocated to $|d_{\langle 1\rangle}|, \ldots, |d_{\langle k_c\rangle}|$. From the definition of k_c in (2), if $s \geqslant c$ then $T \geqslant T_{\text{obs}}$ and the lower bound follows. If $s \leqslant 0$ then $T < T_{\text{obs}}$. In the remaining cases, T is the mean of s positively-signed terms from $|d_{\langle 1\rangle}|, \ldots, |d_{\langle k_c\rangle}|$ and $c-s$ other terms. The sum of the other terms is bounded above by $\Sigma\, a_i$ in (2). From the definition of k_s in (2), $T < T_{\text{obs}}$ if t or fewer positive signs are allocated to $|d_{\langle 1\rangle}|, \ldots, |d_{\langle k_s-s+1\rangle}|$, and the upper bound follows.

REFERENCES

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Statist.* **28**, 181–7.

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

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Am. Statist. Assoc.* **50**, 946–67.

KEMPTHORNE, O. & DOERFLER, T. E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika* **56**, 231–48.

MOSTELLER, F. & TUKEY, J. W. (1977). *Data Analysis and Regression.* Reading, Mass: Addison-Wesley.

PAGANO, M. & TRITCHLER, D. (1983). On obtaining permutation distributions in polynomial time. *J. Am. Statist. Assoc.* **78**, 435–40.

PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika* **29**, 322–35.

ROSENBERGER, J. L. & GASKO, M. (1983). Comparing location estimators: trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis*, Ed. D. C. Hoaglin, F. Mosteller and J. W. Tukey, pp. 297–338. New York: Wiley.

WELCH, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika* **29**, 21–52.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80–3.

[*Received August* 1985. *Revised August* 1986]