Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test

Paul R. ROSENBAUM

Wilcoxon's signed rank test is often inverted, using Walsh averages, to yield exact, distribution free, randomization based confidence intervals for a constant treatment effect or center of symmetry; however, many treatment effects are not constant, so that this interval is not applicable. This article proposes a new way to invert the signed rank test, again using Walsh averages, to produce an exact, distribution free, randomization based confidence interval describing treatment effects that are not constant. The procedure is simple to apply, comparable to the signed rank test itself. Also, the procedure permits a sensitivity analysis in observational studies that estimate treatment effects in the absence of randomization. The method is illustrated using an observational study of the frequency of micronuclei in the cells of alcoholics and matched controls.

1. WALSH AVERAGES AND THEIR USES

1.1 Measuring Treatment Effects That Vary

If a treatment effect is not constant, not the same for everyone, then one cannot describe it by describing the effect for one person. The smallest increase in complexity is to describe the effect for two people. One might reasonably judge the treatment to have a favorable result for two people if either both benefited in comparison to untreated controls, or if one benefited much more than the other was harmed.

Consider a randomized experiment in I matched pairs of one treated subject and one control. In the I treated-minus-control differences in response, D_i, $i = 1, \ldots, I$, the treated subject in one pair may have a much better, much higher response than the control, while in another pair, the difference may be small or negative. For two pairs, say i and k, the Walsh average, $(D_i + D_k)/2$, is positive if the more favorable result in the two pairs, $\max\{D_i, D_k\}$, was sufficiently positive to offset the less favorable result, $\min\{D_i, D_k\}$, so that, at least, the average of the two results is positive. For brevity, a positive Walsh average will sometimes be called an *offset*; however, this is just an abbreviation. In a randomized experiment, a Walsh average, $(D_i + D_k)/2$, may be positive by *chance*—by the flip of a fair coin in assigning treatment or control—or the Walsh average may be positive because of effects *caused* by the treatment. Can we draw exact, distribution free, randomization inferences about the number of Walsh averages that were positive because of effects caused by the treatment? Can we infer the number of offsets attributable to treatment?

It is a familiar fact (e.g., Lehmann 1998, sec. 3.2) that the number of positive Walsh averages, $i \leq k$, equals Wilcoxon's signed rank statistic. This offers the hope, realized in Section 3, that the desired inferences can be obtained by inverting the randomization distribution of the signed rank statistic. First, some context and history. [Note: Pratt and Gibbons (1981, sec. 3.5, pp. 158–160) considered an alternative definition of the signed rank statistic as the number of positive Walsh averages with $i < k$, that is, excluding the case of $i = k$. They concluded that there is little basis to choose between the two definitions, and then return to the traditional definition, with $i \leq k$. Notice that, if one adopts the alternative definition, excluding $i = k$, then the sign of the difference D_i with the smallest $|D_i|$ does not affect the value of the statistic; this minor oddity argues in favor of the traditional definition, which attaches a small weight to this smallest absolute difference.]

1.2 Traditional Uses of the Probability of a Positive Walsh Average

In theoretical discussions of Wilcoxon's signed rank test in an infinite population model, the probability of a positive Walsh average plays a central role because this probability is the population quantity that is estimated when the signed rank statistic is expressed approximately as one of Hoeffding's (1948) U-statistics; for example, Lehmann (1998, p. 368) or Randles and Wolfe (1979, p. 64). In particular, the probability of a positive Walsh average shows up in: (i) a general definition of the null and alternative hypotheses tested by the signed rank test, (ii) expressions for the asymptotic power and relative efficiency of the test, and (iii) Noether's (1987) convenient method for sample size calculations.

Despite having a simple, practical interpretation in terms of favorable results offsetting less favorable ones, positive Walsh averages are little used in practice. One reason is that the behavior of the signed rank statistic under the null hypothesis of no effect is much simpler than its non-null behavior when the treatment has nonconstant effects. The null distribution of the signed rank statistic is a known distribution, does not depend on the common underlying continuous symmetric distribution of the differences, and is widely tabulated; this is part of the convenience and attraction of the signed rank test. Moreover, this exact null distribution is often inverted to obtain confidence intervals for a constant treatment effect; see Section 2.3 for review of this standard procedure. However, if the treatment effect is not constant, and one estimates the probability of a positive

Paul R. Rosenbaum is Robert G. Putzel Professor, Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 (E-mail: rosenbaum@stat.wharton.upenn.edu). This work was supported by grant SES-0004205 from the Methodology, Measurement, and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation.

DOI: 10.1198/0003130031405

Walsh average using a U-statistic, then the null distribution is no longer the relevant distribution, and the non-null distribution of the estimator does depend upon the unknown distribution of the responses, so its distribution is not known; for instance, this is true of the variance of the estimator (Randles and Wolfe 1979, p. 66; Pratt and Gibbons 1981, sec. 3.5). Obviously, one can work around this, at least in sufficiently large samples, where the unknown features of the distribution of the U-statistic can be consistently estimated; however, the exact distribution, the fully distribution free property, some of the ease of use and elegant simplicity of the signed rank test are lost. In contrast, Section 3 develops exact, distribution free inferences about effects attributable to treatment by inverting the known null distribution of the signed rank test, and the procedure requires no additional calculations beyond those required to perform the test itself.

2. REVIEW: SIGNED RANK TEST WITH RANDOMIZATION

2.1 Notation: Treatment Assignment; Treatment Effect

There are I matched pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one of whom received the treatment, signified by $Z_{ij} = 1$, the other received the control, signified by $Z_{ij} = 0$, so that $Z_{i1} + Z_{i2} = 1$ for each i. In a randomized experiment, the treatment is assigned at random in each pair, $\Pr(Z_{ij} = 1) = 1/2$ for each i, j, with mutually independent assignments in distinct pairs. Each subject has two potential responses, (r_{Tij}, r_{Cij}), where r_{Tij} is observed if the jth subject in pair i receives the treatment, $Z_{ij} = 1$, or else r_{Cij} is observed if this subject receives the control, $Z_{ij} = 0$; see Neyman (1923) and Rubin (1974). It simplifies the discussion to assume there are no "ties" of any kind among the responses or their differences.

The effect of the treatment on the jth person in pair i is $\tau_{ij} = r_{Tij} - r_{Cij}$. Because any one person (i, j) either receives the treatment, $Z_{ij} = 1$, or not, $Z_{ij} = 0$, either r_{Tij} or r_{Cij} is observed but never both, so we never observe the effects τ_{ij} caused by the treatment, and must instead use the observed data to draw inferences about them. Write $\boldsymbol{\tau} = (\tau_{11}, \tau_{12}, \ldots, \tau_{I2})$ for the $2I$-dimensional vector of effects.

Using $Z_{i2} = 1 - Z_{i1}$, the observed treated-minus-control difference in responses in pair i is then

$$\begin{aligned} D_i &= Z_{i1}(r_{Ti1} - r_{Ci2}) + Z_{i2}\,(r_{Ti2} - r_{Ci1}) && (1)\\ &= (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2}) + Z_{i1}\,\tau_{i1} + Z_{i2}\,\tau_{i2} && (2)\\ &= (2Z_{i1} - 1)\,d_i + \Delta_i, && (3)\end{aligned}$$

where $d_i = r_{Ci1} - r_{Ci2}$ and $\Delta_i = Z_{i1}\,\tau_{i1} + Z_{i2}\,\tau_{i2}$. In Wilcoxon's signed rank test, the $|D_i|$ are ranked from 1 to I, and the test statistic, T, is the sum of these ranks for pairs i with positive differences, $D_i > 0$. An alternative expression for T uses the $I(I+1)/2$ Walsh averages, $D_i + D_k/2$, of the differences for two pairs, $i \leq k$; specifically, T is the number of positive Walsh averages.

2.2 Review: Randomization Test of No Effect

The same null distribution of T may be derived from several starting points (Lehmann 1998, secs. 3.3 and 4.2); in particular, Lehmann showed that T is a randomization test, as will now be reviewed. In a randomization inference, as developed by Fisher (1935), stochastic properties of the inference are derived solely from the random assignment of treatments, Z_{ij}. In this way, randomization in an experiment forms the "reasoned basis for inference," in Fisher's phrase, in that the only distributions used in the inference are those created by random treatment assignment. In Fisher's approach, quantities that depend upon the treatment assignment, Z_{ij}, such as the treated-minus-control difference D_i in (1), are random variables, but quantities that do not involve Z_{ij}, such as the potential responses, (r_{Tij}, r_{Cij}), are fixed features of the finite population of $2I$ subjects.

The null hypothesis of no treatment effect asserts that the response of each subject, i, j, is unchanged by receiving the treatment rather than the control, that is, $\mathrm{H}_0 : r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$, $j = 1, 2$. Under the null hypothesis, the treated-minus-control difference D_i equals

$$\begin{aligned} \widetilde{D}_i &= Z_{i1}\,(r_{Ci1} - r_{Ci2}) + Z_{i2}\,(r_{Ci2} - r_{Ci1})\\ &= (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2}) = (2Z_{i1} - 1)\,d_i \qquad (4)\end{aligned}$$

that is, $\pm(r_{Ci1} - r_{Ci2}) = \pm d_i$, depending upon Z_{i1}, and the rank q_i of $|D_i|$ equals the rank of $\left|\widetilde{D}_i\right| = |r_{Ci1} - r_{Ci2}| = |d_i|$, which is fixed, not varying with Z_{ij}. Notice that in a randomized experiment with $\Pr(Z_{i1} = 1) = \Pr(Z_{i1} = 0) = 1/2$, the difference in potential control responses $\widetilde{D}_i$ is $\pm\left|\widetilde{D}_i\right|$ with equal probabilities, so the $\widetilde{D}_i$'s are symmetrically distributed about zero. It follows from this that, under the null hypothesis of no effect in a randomized experiment, Wilcoxon's signed rank statistic T has the distribution of the sum of I independent random variables taking the values i or 0 each with probability $1/2$, $i = 1, \ldots, I$. Throughout the article, write c_α for the upper α critical value of this standard, exact, null distribution for the signed rank test, so that when the null hypothesis is true in a randomized experiment, $\Pr(T \geq c_\alpha) = \alpha$; then, a test which rejects the null hypothesis of no effect when $T \geq c_\alpha$ has exactly level α, is a distribution free test, and is a randomization test. (Generally, to avoid repetition, I will speak of rejecting for large T, but of course, with minor adjustments, one can reject for small T or for both large and small T.)

2.3 Review: Confidence Intervals for Constant Effect

In a randomized experiment, if the unobservable treatment effects are constant, $r_{Tij} - r_{Cij} = \zeta$, the same for all $2I$ subjects i, j, then the observable treated-minus-control differences, $D_i = \widetilde{D}_i + \zeta$, are symmetrically distributed about ζ. In particular, if the effect is constant, then the effect in pair i is $\Delta_i = Z_{i1}\,\tau_{i1} + Z_{i2}\,\tau_{i2} = Z_{i1}\,\zeta + Z_{i2}\,\zeta = \zeta$ for every i, so variation among the Δ_i's expresses departures from constant effect. With a constant effect, ζ, Wilcoxon's exact test may be inverted to give an exact confidence interval for ζ. Specifically, if $r_{Tij} - r_{Cij} = \zeta$, then $D_i - \zeta = \widetilde{D}_i$, and one can test any hypothesis $\mathrm{H}_0 : \zeta = \zeta_0$ by applying the signed rank test to $D_i - \zeta_0$; then the confidence interval is the set of hypotheses ζ_0 not rejected by the test. The endpoints of the confidence interval turn out to be two of the Walsh averages; see, for instance, Pratt and Gibbons (1981, sec. 3.4), Lehmann (1998, sec. 4.5), or Hollander and Wolfe (1999, sec. 3.3). If the D_i do not have a symmetric distribution, then this confidence interval is not valid. Can Wilcoxon's exact test be inverted when treatment effects $r_{Tij} - r_{Cij}$ vary from person to person and the differences D_i are not symmetrically distributed?

2.4 Review: Attributable Effects

Fisher's (1935) theory of randomization inference in experiments has two desirable consequences: (i) it formally justifies and encourages randomized experimentation whenever ethical and feasible, and (ii) it forces nonrandomized or observational studies to acknowledge, and then measure as part of the quantitative findings, sources of uncertainty that would not have been present had randomization been used.

Attributable effects are a device that expands substantially the circumstances in which randomization tests can be inverted to yield confidence intervals for magnitudes of effect. An attributable effect, say A, is a quantity that describes how treated subjects would have responded differently had they received the control. An attributable effect is typically an unobservable random variable rather than a fixed parameter; nonetheless, exact inference about A is often possible using randomization as the sole basis for inference, that is, by inverting a randomization test of no effect. For instance, when inverting Fisher's exact test for a 2×2 table, the attributable effect A is the number of successes caused by the treatment. Rosenbaum (2001, 2002b) discussed confidence intervals for attributable effects that invert Fisher's exact test for a 2×2 table, certain median and quantile tests, Wilcoxon's rank sum test, McNemar's test for paired binary responses, and the Mantel–Haenszel test for matched binary responses with multiple controls.

With a certain natural definition of the attributable effect, A, Wilcoxon's signed rank test may also be inverted to yield confidence statements without the assumption that the effect is constant. This is developed for experiments in Section 3 and for observational studies in Section 5, with examples in Section 4 and Section 5.2. Section 6 extends the discussion to situations in which the difference, D_i, does not reflect the effect of a treatment on one person in pair i, but rather an effect on the pair itself.

3. RANDOMIZATION INFERENCE WITH NONCONSTANT EFFECTS

3.1 Testing General Hypotheses About Treatment Effects

A general hypothesis about the effect of the treatment specifies each of the $2I$ effects, $\mathrm{H}_0 : r_{Tij} - r_{Cij} = \tau_{0ij}$ for $i = 1, \ldots, I$, $j = 1, 2$ or $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ with $\boldsymbol{\tau}_0 = (\tau_{011}, \tau_{012}, \ldots, \tau_{0I2})$. In practice, it is tedious to work with hypotheses that specify $2I$ parameters; however, in principle, there is nothing special about testing such hypotheses. Later on, a pivotal argument will permit us to bypass the specification of $\boldsymbol{\tau}_0$. Write $\Delta_{0i} = Z_{i1}\,\tau_{0i1} + Z_{i2}\,\tau_{0i2}$ for the hypothesized effect in pair i.

The testing of a general hypothesis is similar to the testing of hypotheses about a constant effect. Essentially, one subtracts the hypothesized effects, $\boldsymbol{\tau}_0$, and checks whether, after subtraction, the null hypothesis of no effect appears to hold. For the hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, write $\widetilde{D}_{\boldsymbol{\tau}_0,i}$ for the adjusted difference $D_i - Z_{i1}\tau_{0i1} - Z_{i2}\tau_{0i2} = D_i - \Delta_{0i}$ and write $T_{\boldsymbol{\tau}_0}$ for the signed rank statistic computed from the adjusted differences, $\widetilde{D}_{\boldsymbol{\tau}_0,i}$, so that, of course, $T_{\boldsymbol{\tau}_0}$ is the number of pairs $i \leq k$ such that $\widetilde{D}_{\boldsymbol{\tau}_0,i} + \widetilde{D}_{\boldsymbol{\tau}_0,k} > 0$. If the null hypothesis, $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, were true, then $\widetilde{D}_{\boldsymbol{\tau}_0,i}$ would equal $\widetilde{D}_i$ and $T_{\boldsymbol{\tau}_0}$ would equal the signed rank statistic computed from the $\widetilde{D}_i$, that is, $T_{\boldsymbol{\tau}_0}$ would be the number of of pairs $i \leq k$ such that $\widetilde{D}_i + \widetilde{D}_k > 0$. Now the $\widetilde{D}_i$'s satisfy the null hypothesis of no effect—randomization simply flips the signs of the $\widetilde{D}_i$ without changing their absolute magnitudes—so if $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ is true, then $\Pr\left(T_{\boldsymbol{\tau}_0} \geq c_\alpha\right) = \alpha$.

3.2 Effects Attributable to Treatment

Suppose that a particular Walsh average is positive, say $\left(D_i + D_k\right)/2 > 0$. Did the treatment cause this to happen? Or would it have happened anyway even if all four subjects in these two pairs had received the control? The Walsh average for i and k is *an offset attributable to treatment* if the Walsh average is positive, $D_i + D_k > 0$, but it would not have been positive had all four subjects received the control, $\widetilde{D}_i + \widetilde{D}_k \leq 0$. How many Walsh averages are positive because of effects of the treatment? Write A for this unknown number, and note that A depends on the treatment assignments, Z_{ij}, so A is a random variable.

As motivation, consider two simple, extreme situations. If the treatment has no effect, $r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$, $j = 1, 2$, then none of the positive Walsh averages are attributable to effects of the treatment, $A = 0$, because $D_i + D_k = \widetilde{D}_i + \widetilde{D}_k$. If every one of the $2I$ effects, $r_{Tij} - r_{Cij}$, is sufficient large (formally, if all $2I$ effects $r_{Tij} - r_{Cij} \to \infty$), then every one of the $I\left(I+1\right)/2$ choices of two pairs $i \leq k$ exhibits a positive Walsh average, $D_i + D_k > 0$, and in expectation, half of these are attributable to the treatment, as $\widetilde{D}_i + \widetilde{D}_k \leq 0$, and the other half were expected anyway by chance, as $\widetilde{D}_i + \widetilde{D}_k > 0$.

The magnitude of the *number* A reflects, in part, the sample size I, and for some purposes, this is not desirable. Instead, in a randomized experiment, multiplying by a constant gives $4A/\left\{I\left(I+1\right)\right\}$, which: (i) is exactly 0 under the null hypothesis of no effect, (ii) tends to 1 as all $2I$ effects increase, $r_{Tij} - r_{Cij} \to \infty$ and $I \to \infty$, (iii) and is, for instance, $1/2$ if the treatment has increased the number of positive Walsh averages by 50% above the number expected by chance.

3.3 The Attributable Effect is a Pivot

Each general hypothesis, $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, implies a certain number of Walsh averages that are positive because of effects of the treatment. For any hypothesis, $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, we may use the hypothesis and the observed data to calculate $\widetilde{D}_{\boldsymbol{\tau}_0,i} = D_i - Z_{i1}\tau_{0i1} - Z_{i2}\tau_{0i2} = D_i - \Delta_{0i}$, and from this, the number, say $A_{\boldsymbol{\tau}_0}$, of pairs, $i \leq k$, with $D_i + D_k > 0$ and $\widetilde{D}_{\boldsymbol{\tau}_0,i} + \widetilde{D}_{\boldsymbol{\tau}_0,k} \leq 0$. There is a simple, useful relationship between (i) the conventional signed rank statistic T for testing $\mathrm{H}_0 : \boldsymbol{\tau} = 0$ computed from the observed differences D_i, (ii) the signed rank statistic for testing $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ computed from the adjusted differences, $\widetilde{D}_{\boldsymbol{\tau}_0,i}$, and (iii) the number $A_{\boldsymbol{\tau}_0}$ of Walsh averages that are positive because of effects of the treatment under the hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$. The relationship is summarized in the Proposition 1. If the hypothesis, $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, were true, then $\widetilde{D}_{\boldsymbol{\tau}_0,i} = \widetilde{D}_i$ and $A_{\boldsymbol{\tau}_0} = A$.

Proposition 1. For any hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$,

$$A_{\boldsymbol{\tau}_0} \geq T - T_{\boldsymbol{\tau}_0}, \tag{5}$$

and if the hypothesis is true in a randomized experiment, then $A = A_{\boldsymbol{\tau}_0}$ and

$$\Pr\left(A \geq T - c_\alpha + 1\right) \geq 1 - \alpha. \tag{6}$$

These inequalities are sharp in the following sense. If, in the hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, all the effects are non-negative, $\tau_{0ij} \geq 0$, for $i = 1 = 1, \ldots, I$, $j = 1, 2$, then the inequalities become equalities; specifically, $A_{\boldsymbol{\tau}_0} = T - T_{\boldsymbol{\tau}_0}$, and if the hypothesis is true in a randomized experiment, then $A = A_{\boldsymbol{\tau}_0}$ and $\Pr\left(A \geq T - c_\alpha + 1\right) = 1 - \alpha$.

Proof: For any finite set $\mathcal{S}$, write $|\mathcal{S}|$ for the number of elements of $\mathcal{S}$. Let

$$\mathcal{A} = \left\{(i,k)\,,\; i \leq k : D_i + D_k > 0\right\}, \quad \text{so } T = |\mathcal{A}|,$$

$$\mathcal{B} = \left\{(i,k)\,,\; i \leq k : \widetilde{D}_{\boldsymbol{\tau}_0,i} + \widetilde{D}_{\boldsymbol{\tau}_0,k} > 0\right\}, \quad \text{so } T_{\boldsymbol{\tau}_0} = |\mathcal{B}|,$$

and

$$\mathcal{C} = \left\{(i,k)\,,\; i \leq k : D_i + D_k > 0, \right.$$
$$\left. \widetilde{D}_{\boldsymbol{\tau}_0,i} + \widetilde{D}_{\boldsymbol{\tau}_0,k} \leq 0\right\}, \quad \text{so } A_{\boldsymbol{\tau}_0} = |\mathcal{C}|.$$

Now $\mathcal{C}$ is the set difference, $\mathcal{C} = \mathcal{A} - \mathcal{B} = \mathcal{A} \cap \overline{\mathcal{B}}$, so $|\mathcal{C}| \geq |\mathcal{A}| - |\mathcal{B}|$, proving (5). If $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ is true, then $A = A_{\boldsymbol{\tau}_0}$ and $T_{\boldsymbol{\tau}_0}$ has the null distribution of Wilcoxon's signed rank statistic, so that $\Pr\left(T_{\boldsymbol{\tau}_0} \geq c_\alpha\right) = \alpha$, and hence $\Pr\left(T_{\boldsymbol{\tau}_0} \leq c_\alpha - 1\right) = 1 - \alpha = \Pr\left(-T_{\boldsymbol{\tau}_0} \geq -c_\alpha + 1\right)$; therefore, using (5), it follows that $\Pr\left(A \geq T - c_\alpha + 1\right) \geq 1 - \alpha$, proving (6). By definition, $\widetilde{D}_{\boldsymbol{\tau}_0,i} = D_i - Z_{i1}\tau_{0i1} - Z_{i2}\tau_{0i2}$, so if $\tau_{0ij} \geq 0$, for $i = 1 = 1, \ldots, I, j = 1, 2$, then $D_i \geq \widetilde{D}_{\boldsymbol{\tau}_0,i}$, for every i, so that $\mathcal{B} \subseteq \mathcal{A}$, whereupon $|\mathcal{C}| = |\mathcal{A} - \mathcal{B}| = |\mathcal{A}| - |\mathcal{B}|$, that is, $A_{\boldsymbol{\tau}_0} = T - T_{\boldsymbol{\tau}_0}$, proving the cases of equality.

Table 1. Frequency of Micronuclei Harboring Whole Chromosomes in 20 Alcoholics and 20 Controls Matched for Age and Gender. Source: Maffei et al. (2000).

Treated			Control			C+ MN
Gender	Age	C+ MN	Gender	Age	C+ MN	difference
F	36	3.1	F	37	2.7	0.4
F	40	3.1	F	41	2.9	0.2
F	42	5.4	F	42	2.3	3.1
F	52	5.4	F	49	3.4	2.0
F	60	16.6	F	58	4.1	12.5
F	62	20.3	F	61	5.4	14.9
F	62	9.2	F	62	7.8	1.4
M	36	9.6	M	33	2.5	7.1
M	36	3.2	M	35	2.0	1.2
M	37	4.5	M	35	2.2	2.3
M	40	16.1	M	36	2.2	13.9
M	43	6.2	M	41	2.4	3.8
M	47	4.3	M	43	4.0	0.3
M	47	7.8	M	45	2.3	5.5
M	50	5.3	M	48	4.6	0.7
M	54	8.4	M	50	3.0	5.4
M	54	11.2	M	58	3.5	7.7
M	60	9.9	M	59	4.0	5.9
M	61	9.3	M	60	3.6	5.7
M	62	5.9	M	60	4.0	1.9

Notice that (6) asserts that the chance that $A \geq T - c_\alpha + 1$ is at least $1 - \alpha$ no matter what hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ happens to be the one true hypothesis. Moreover, because the inequality (6) is sharp, although we are $1 - \alpha$ confident that $A \geq T - c_\alpha + 1$, we are not more than $1 - \alpha$ confident, because for some possible hypotheses $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$, namely those with non-negative effects, the confidence rate is exactly $1 - \alpha$. In short, (6) says the signed rank statistic T combined with its α critical value c_α immediately translates into a $1 - \alpha$ confidence statement about the magnitude of the treatment effect.

3.4 Confidence Procedure

In a randomized experiment, the procedure is very simple. Calculate Wilcoxon's signed rank statistic T from the treated-minus-control differences D_i in the usual way, and from the standard tables for this statistic, determine c_α so that, under the null hypothesis of no treatment effect, $\Pr\left(T \geq c_\alpha\right) = \alpha$. Then describe the magnitude of the effect by asserting with confidence $1 - \alpha$ that the number A of Walsh averages that are positive because of effects of the treatment satisfies $A \geq T - c_\alpha + 1$. This assertion will be true in at least $100\,(1 - \alpha)\,\%$ of randomized experiments, by virtue of (6). The procedure just described is illustrated in Section 4, and then Section 5 discusses inference in observational studies where treatments are not randomly assigned.

4. EXAMPLE: GENETIC DAMAGE FROM ALCOHOL

Maffei et al. (2000) examined possible genetic damage caused by alcohol consumption by comparing alcoholics to controls; their data, matched for age and gender, appear in Table 1. The alcoholics consumed $>$ 120 grams per day of pure alcohol, whereas the controls consumed between 8 and 13 grams per day. The measure of genetic damage in this table is the frequency of micronuclei harboring whole chromosomes per 1,000 binucleated cells (C+ MN), where higher values indicate greater damage. Notice that a few alcoholics had exceptionally high frequencies of micronuclei containing whole chromosomes, and the distribution of differences D_i appears to have a long right hand tail. The greatest genetic damage is in this right tail, which is too thick for a mean, but is largely ignored by a median.

In a randomized experiment with $I = 20$ pairs, if $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ is true, then $\Pr\left(T_{\boldsymbol{\tau}_0} \geq 150\right) = 0.0487$, so any hypothesis $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ yielding $T_{\boldsymbol{\tau}_0} \geq 150$ would be rejected at the 0.05 level, but any hypothesis yielding $T_{\boldsymbol{\tau}_0} \leq 149$ would not be rejected at this level. The signed rank statistic of $T = 210$ signifies that in all $210 = I\,(I + 1)/2$ comparisons, the Walsh average is positive. If this were a randomized experiment, which of course it is not, then we would be 95% confident that at least $T - c_{0.05} + 1 = 210 - 150 + 1 = 61$ of these Walsh averages were positive because of effects of the treatment. In a randomized experiment, if the treatment had no effect, $\boldsymbol{\tau} = \mathbf{0}$, so that for every i, k, $D_i + D_k = \widetilde{D}_i + \widetilde{D}_k$, one expects half of the comparisons, here $105 = I\,(I + 1)/4$, to produce a positive Walsh average by chance, that is, as a consequence of the random assignment. So the observed number of positive Walsh averages, namely $T = 210$, is $100\% = (210 - 105)\,/105$ greater than expected by chance, and in an experiment we would be 95% confident that at least $58\% = 61/105$ of this excess was caused by the treatment, and not by chance fluctuations.

Of course, Table 1 describes an observational study, not an experiment, so the alcoholics and their matched controls might differ systematically with respect to relevant unobserved covariates. This possibility is examined in Section 5.2.

5. INFERENCE IN OBSERVATIONAL STUDIES

5.1 Sensitivity to Hidden Bias Without Randomization

In an observational or nonrandomized study of treatment effects, treated subjects are often matched to controls on the basis of observed pretreatment covariates, for instance, age and gender in Table 1. Even within such ostensibly similar matched pairs, subjects may differ in terms of covariates not observed, so they may differ in their chances of receiving treatment, and so $\Pr(Z_{ij} = 1)$ may not equal $1/2$, and the randomization inference in Sections 2 and 3 may be unwarranted and incorrect. In this context, a sensitivity analysis asks how far the distribution of treatment assignments must depart from random assignment to materially alter the study's conclusions. The sensitivity analysis converts a qualitative concern about unobserved covariates into a quantitative appraisal: Would small hidden biases alter the conclusions? Or would it take very large biases to do so? In the first sensitivity analysis, Cornfield et al. (1959) concluded that to explain away the observed association between heavy smoking and lung cancer, an unobserved binary covariate would need to be a near perfect predictor of lung cancer and nine times more common among heavy smokers than nonsmokers, so this association is highly insensitive to hidden bias.

In a similar, though more general, approach to sensitivity analysis, prior to matching, the treatment assignments Z_{ij} are assumed mutually independent with unknown probabilities, such that two subjects with the same values of the observed covariates may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$; then matched treated-control pairs are formed, so the relevant distribution is conditional given $Z_{i1} + Z_{i2} = 1$; see Rosenbaum (1987, 2002a, sec. 4). Here, Γ is an unknown sensitivity parameter that is varied to display the degree of sensitivity of the inference. From this, one obtains the bounds,

$$\frac{1}{\Gamma} \leq \frac{\Pr(Z_{i1} = 1, Z_{i2} = 0)}{\Pr(Z_{i1} = 0, Z_{i2} = 1)} \leq \Gamma, \qquad (7)$$

with mutually independent assignments in distinct pairs, which yields bounds on the distribution of the signed rank statistic. Specifically, under the null hypothesis of no treatment effect, for each value of $\Gamma \geq 1$, there are two random variables with known distributions, $\overline{T}_\Gamma$ and $\overline{\overline{T}}_\Gamma$, that bound the unknown distribution of the signed rank statistic, T, for all treatment assignment distributions satisfying (7); that is, $\Pr\left(\overline{\overline{T}}_\Gamma \geq c\right) \geq \Pr(T \geq c) \geq \Pr\left(\overline{T}_\Gamma \geq c\right)$. For instance, $\overline{\overline{T}}_\Gamma$ is the sum of I independent random variables, $i = 1, \ldots, I$, taking values i or 0, with probabilities $\Gamma/1 + \Gamma$ or $1/1 + \Gamma$. Both bounds equal the randomization distribution when $\Gamma = 1$. For fixed $\Gamma \geq 1$, find $\overline{\overline{c}}_{\Gamma,\alpha}$ so that $\Pr\left(\overline{\overline{T}}_\Gamma \geq \overline{\overline{c}}_{\Gamma,\alpha}\right) = \alpha$; see the Appendix for details.

Proposition 2. Under the sensitivity model (7), for each fixed $\Gamma \geq 1$, we may assert with confidence $1 - \alpha$ that $A \geq T - \overline{\overline{c}}_{\Gamma,\alpha} + 1$.

Proof: If $\mathrm{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ is true, then $A = A_{\boldsymbol{\tau}_0} \geq T - T_{\boldsymbol{\tau}_0}$ by Proposition 1. Moreover, because

$$\alpha = \Pr\left(\overline{\overline{T}}_\Gamma \geq \overline{\overline{c}}_{\Gamma,\alpha}\right) \geq \Pr\left(T_{\boldsymbol{\tau}_0} \geq \overline{\overline{c}}_{\Gamma,\alpha}\right),$$

it follows that

$$1 - \alpha \leq \Pr\left(T_{\boldsymbol{\tau}_0} \leq \overline{\overline{c}}_{\Gamma,\alpha} - 1\right) \leq \Pr\left(A \geq T - \overline{\overline{c}}_{\Gamma,\alpha} + 1\right).$$

5.2 Sensitivity Analysis in the Example

The sensitivity analysis of Section 5.1 will now be illustrated using the example of Section 4. With a moderately large hidden bias, that is, with an unobserved covariate associated with a $\Gamma = 2$ fold increase in the odds of alcohol abuse, compute $\Pr\left(\overline{\overline{T}}_2 \geq 181\right) = 0.048$, so we would be slightly more than 95% confident that at least $A \geq 210 - 181 + 1 = 30$ Walsh averages were positive because of effects of the treatment. Alternatively, as in Section 3.2, for $\Gamma = 2$, we would be slightly more than 95% confident that the number of Walsh averages that are positive because of effects of the treatment is at least $4A/\{I(I+1)\} \geq (4 \times 30)/\{20(20+1)\} = 29\%$ greater than the number expected to be positive by chance in a randomized experiment. The sensitivity analysis is given in Table 2. For $\Gamma = 8$, the null hypothesis of no treatment effect is plausible, the upper bound on the significance level being $\Pr\left(\overline{\overline{T}}_8 \geq 210\right) = .095$. By comparison with other observational studies (see Rosenbaum 2002b, sec. 4), this study is quite insensitive even to fairly large biases: an unobserved covariate associated with a four-fold increase in the odds of alcoholism would still leave us 95% confident that at least 9 positive Walsh averages were attributable to alcoholism.

Table 2. Sensitivity Analysis of Confidence Intervals for the Number of Offsets Attributable to Treatment.

Γ	$\Pr\left(\overline{\overline{T}}_\Gamma \geq \overline{\overline{c}}_{\Gamma,\alpha}\right) = \alpha$	$A \geq T - \overline{\overline{c}}_{\Gamma,\alpha} + 1$
1	$\Pr\left(\overline{\overline{T}}_1 \geq 150\right) = .049$	$A \geq 61$
2	$\Pr\left(\overline{\overline{T}}_2 \geq 181\right) = .048$	$A \geq 30$
4	$\Pr\left(\overline{\overline{T}}_4 \geq 202\right) = .044$	$A \geq 9$
6	$\Pr\left(\overline{\overline{T}}_6 \geq 210\right) = .046$	$A \geq 1$

6. DIFFERENCES WITHOUT INDIVIDUAL EFFECTS

The discussion so far has presumed that there are two individuals in a matched pair, one treated, one control, and that the treatment has an effect on the treated subject in the pair but no effect on the control. This is the traditional way that treatment effects are described in randomization inference, and in most contexts it is quite natural. However, one can dispense with individual effects and focus entirely on the differences, D_i. That is, one can think about the pair i as a single unit and consider the effect Δ_i on the pair, rather than the separate effects on the individuals in the pair. This section presents some motivation and a few details.

The notation for causal effects in Section 2.1 assumes that the treatment has an effect on the person who receives it, but has no effect on the untreated person in the same pair. Cox (1958, sec. 2.4) refers to this as "no interference between units." In the example in Section 4, this is to be expected: the two matched people are unrelated, and the consumption of alcohol by one person is not likely to damage the genes of the other. However, situations do arise in which there can be interference within but not between pairs, and in this case the notation in Section 2.1 is not adequate, but the method proposed is still applicable, although it now describes differential effects rather than absolute effects.

An interesting example of interference within but not between pairs arises in psychiatric genetics (Neiderhiser 2001). Parents transmit to their offspring a distribution of genotypes that is exchangeable within sibships. Using this fact in an appropriate way, together with measures of disease outcomes and direct measures of genotype markers, yields a relatively persuasive test that a marker is linked to a gene contributing to the disease (Spielman and Ewens 1998). For instance, one might compare sibling pairs in terms of a measure of disease outcome where the parents transmitted the marker to one sibling but not to the other. Most genes affect their owner, but do not affect other people—that is, there is no interference between units—however, this is not necessarily true for genes that affect behavior. A gene that affects behavior may be "evocative ... For example, a child who has a difficult temperament may be more likely to elicit aggressive responses from family members ..." (Neiderhiser 2001, p. s15). In other words, a gene that affects behavior may not only affect the sibling who received the gene, but may indirectly affect to some degree the control sibling as well—that is, there may be interference between siblings in the same family, with no interference between children in different families. With such interference, a comparison of siblings describes not the absolute effect of a gene, but rather the differential effect, that is, the difference in effect between being the owner of the gene and the sibling of the owner of the gene. The differential effect may be larger or smaller than the absolute effect; for example, aggressive behavior by one sibling might either promote or inhibit such behavior by the other. However, if the differential effect is of interest, it may be studied using the method of this article.

With this motivation, consider the simplest case, namely a randomized experiment with interference within but not between pairs i. Imagine, first, randomly assigning treatment or control labels within a pair, but withholding the treatment. In this imagined situation, the labels are without effect and they simply determine which response is subtracted from the other; hence, the difference, $\widetilde{D}_i$, within pair i would take values $\pm |d_i|$ with equal probabilities $1/2$, and the signed rank statistic computed from the $\widetilde{D}_i$ would have its usual, null randomization distribution. Of course, the treatment is actually given to the treated subject in each pair, possibly affecting both subjects, resulting in a difference of $D_i = \widetilde{D}_i + \Delta_i$, where Δ_i takes one of two values with equal probabilities, namely $\Delta_i = \delta_{i1}$ if the treatment is given to the first subject or $\Delta_i = \delta_{i2}$ if the treatment is given to the second. Although the δ's are interpreted as differential effects on pair i, rather than effects on one person ij in pair i, inference about the δ's is essentially the same as inference about the individual effects τ's. The number A of pairs, (i,k), $i \leq k$, with $D_i + D_k > 0$ but $\widetilde{D}_i + \widetilde{D}_k \leq 0$ is the number of Walsh averages that are positive because of differential effects of the treatment, and so on, and confidence statements are constructed as before. In other words, shifting attention from the effect of the treatment on one person, τ_{ij}, to the effect of the treatment on the matched pair difference, Δ_i, might under some circumstances alter the interpretation of results, but it does not alter the technical details of the statistical argument.

APPENDIX: COMPUTING THE EXACT SENSITIVITY BOUNDS

In moderate to large samples, I, there is a very simple Normal approximation to the sensitivity bounds in Section 5 for the signed rank statistic using, for example, $E\left(\overline{\overline{T}}_{\Gamma}\right) = \lambda I\left(I+1\right)/2$ and $\operatorname{var}\left(\overline{\overline{T}}_{\Gamma}\right) = \lambda\left(1-\lambda\right)I\left(I+1\right)\left(2I+1\right)/6$ where $\lambda = \Gamma/\left(\Gamma+1\right)$; see Rosenbaum (1987, 2002b, Section 4.3.3). For relatively small I, exact calculations are possible, as discussed in this appendix. Specifically, let $\pi_I\left(m,k\right)$ be the number of subsets of m elements $\{q_1, q_2, \ldots, q_m\}$ of $\{1, 2, \ldots, I\}$ which sum to $k = \sum_{j=1}^{m} q_j$. Hájek, Šidák, and Sen (1999, sec. 5.3.1, thm. 2) gave a recurrence formula for $\pi_I\left(k\right) = \sum_{m=1}^{I} \pi_I\left(m,k\right)$ and show that the null, randomization distribution of the signed rank statistic is $\Pr\left(T = k\right) = \pi_I\left(k\right)/2^I$. A similar approach works for the sensitivity bounds. Specifically, $\pi_I\left(m,k\right) = \pi_{I-1}\left(m,k\right) + \pi_{I-1}\left(m-1, k-I\right)$, with initial conditions $\pi_I(m,k) = 0$ for $k < 0$ or $m < 0$, $\pi_0(0,0) = 1$, $\pi_0(m,k) = 0$ otherwise. Under the sensitivity analysis model in Rosenbaum (1987), the distribution of the random variable $\overline{\overline{T}}_{\Gamma}$ that provides the upper bound is:

$$\Pr\left(\overline{\overline{T}}_{\Gamma} = k\right) = \frac{\sum_{m=0}^{I} \Gamma^m \, \pi_I\left(m,k\right)}{\left(1+\Gamma\right)^I}. \tag{A.1}$$

S-PLUS CODE FOR EXACT SENSITIVITY ANALYSIS CALCULATIONS

Standard tables suffice in a randomized experiment. For a sensitivity analysis in an observational study, the S-plus code `exact(gamma, I, c)` computes $\Pr\left(\overline{\overline{T}}_{\Gamma} \geq c\right)$ for I matched pairs using (A.1). Use of the code is illustrated by computing the central column in Table 2. The subfunction `exactPI(I, m, k)` computes $\pi_I\left(m,k\right)$ and is called by `exact(gamma, I, c)`. The S-Plus code is useful for small I; for large I, the Normal approximation may be used.

```
>  exact(1,20,150)
[1] 0.0486536
»  exact(2,20,181)
[1] 0.0480461
> exact(4,20,202)
[1] 0.04395513
> exact(6,20,210)
```

```
[1] 0.04582096
>
> exact
function(gamma, I, c)
{
# Computes the upper bound on Prob(T >=c) for %
# Wilcoxon's signed rank statistic T in
# a sample of size I
    prob <- 0
    p <- gamma/(1 + gamma)
    max <- (I * (I + 1))/2
    mmin  <- 1 + sum(c > cumsum(I:1))
    for(m in mmin:I)
        for(k in c:max)
            prob <- prob + (p^m) * ((1 - p)^(I - m))
            * exactPI(I, m, k)
        prob
}
>  exactPI
function(I, m, k)
{
# Finds the number of subsets of  {1,...,I} with m
# elements totalling k
   if(k <  0 | m < 0) 0 else if(I == 0)  {
        if(k == 0  & m == 0)
             1
        else 0
   }
   else if(((I * (I + 1))/2) - (((I - m)
   * (I + 1 - m))/2) <  k)
        0
   else exactPI(I - 1, m, k) + exactPI(I - 1,
   m - 1, k - I)
}
```

[Received August 2002. Revised December 2002.]

REFERENCES

Cornfield, J., Haenszel, W., Hammond, E., et al. (1959), "Smoking and Lung Cancer," *Journal of the National Cancer Institute*, 22, 173–203.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

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

Hájek, J., Šidák, Z., and Sen, P. K. (1999), *Theory of Rank Tests* (2nd ed.), New York: Academic.

Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

Hollander, M., and Wolfe, D. A. (1999), *Nonparametric Statistical Methods* (2nd ed.), New York: Wiley.

Lehmann, E. L. (1998), *Nonparametrics*, Saddle River, NJ: Prentice Hall.

Maffei, F., Fimognari, C., Castelli, E., Stefanini, G., Forti, G., and Hrelia, P. (2000), "Increased Cytogenetic Damage Detected by FISH Analysis on Micronuclei in Peripheral Lymphocytes from Alcoholics," *Mutagenesis*, 15, 517–523.

Neiderhiser, J. M. (2001), "Understanding the Roles of Genome and Enviroment: Methods in Genetic Epidemiology," *British Journal of Psychiatry*, 178, s12–s17.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles," *Roczniki Nauk Roiniczych,* Tom X, pp. 1–51; reprinted in English in *Statistical Science*, 1990, 5, 463–480.

Noether, G. E. (1987), "Sample Size Determination for Some Common Nonparametric Tests," *Journal of the American Statistical Association,* 82, 645–647.

Pratt, J. W., and Gibbons, J. D. (1981), *Concepts of Nonparametric Theory*, New York: Springer.

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

Rosenbaum, P. R. (1987), "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.

——— (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231.

——— (2002a), "Attributing Effects to Treatment in Matched Observational Studies," *Journal of the American Statistical Association,* 97, 183–192.

——— (2002b), *Observational Studies*, New York: Springer-Verlag.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Spielman, R. S., and Ewens, W. J. (1998), "A Sibship Test for Linkage in the Presence of Association," *American Journal of Human Genetics*, 62, 450–458.