Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies

Author(s): Paul R. Rosenbaum

Source: *Biometrika*, Mar., 1987, Vol. 74, No. 1 (Mar., 1987), pp. 13-26

Published by: Oxford University Press on behalf of Biometrika Trust

Stable URL: https://www.jstor.org/stable/2336017

Biometrika (1987), **74**, 1, *pp.* 13–26
Printed in Great Britain

Sensitivity analysis for certain permutation inferences in matched observational studies

BY PAUL R. ROSENBAUM
Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

SUMMARY

In observational studies, treatments are not randomly assigned to experimental units, so that randomization tests and their associated interval estimates are not generally applicable. In an effort to compensate for the lack of randomization, treated and control units are often matched on the basis of observed covariates; however, the possibility remains of bias due to residual imbalances in unobserved covariates. A general though simple method is proposed for displaying the sensitivity of permutation inferences to a range of assumptions about unobserved covariates in matched observational studies. The sensitivity analysis is applicable to Wilcoxon's signed rank test, to the McNemar–Cox test for paired binary responses, and to some matching problems with a variable number of controls.

Some key words: Logit model; Observational study; Permutation test; Probability inequalities; Propensity score; Reflection group; Sensitivity analysis.

1. OBSERVATIONAL STUDIES: REVIEW OF DEFINITIONS AND NOTATION

1·1. *Observational studies*

Randomized experiments and observational studies share a common purpose, namely inference about the effects caused or produced by a treatment, but in observational studies the assignment of experimental units to treatment or control is not based on randomization (Cochran, 1965). As a consequence of the nonrandomized assignment of units, the treated and control groups may differ systematically prior to the start of the treatment, so that differences in outcome in treated and control groups can reflect either effects caused by the treatment, or inherent pretreatment differences, or both. When relevant pretreatment differences are accurately measured, statistical adjustments can in many cases reduce the bias in the estimated treatment effect, for example, Cochran & Rubin (1973). Still, there is often reason to be concerned that treated and control groups differ in ways that have not been measured, and therefore that biases may remain after adjustment for the measured or observed covariates. A sensitivity analysis in an observational study is an attempt to display and clarify the extent to which inferences about a treatment effect vary over a range of plausible assumptions about unmeasured pretreatment differences; for example, Rosenbaum & Rubin (1983b).

The current section reviews briefly notation and terminology for observational studies that has been developed in detail by Rubin (1974, 1977, 1978), Hamilton (1979), Holland & Rubin (1983), Rosenbaum & Rubin (1983a, b; 1985a), Rosenbaum (1984a, b, c) and Holland (1986). This adapts and extends that used in the traditional literature on

randomization inference in randomized experiments (Fisher, 1935, § 2; Kempthorne, 1952, § 8; Cox, 1958a, § 2). In § 1·5, related work on sensitivity analysis in observational studies is reviewed and contrasted with the current approach, and two assumptions that structure the sensitivity analysis are stated.

1·2. *Treatment effects*

A treatment is a potential intervention that can be applied to or withheld from any experimental unit under study. A treatment effect is a comparison of the two potential outcomes a unit can exhibit: the outcome, say R_1, that would be observed from the unit if the treatment were applied, and the outcome, say R_0, that would be observed if the treatment were withheld. The null hypothesis of no treatment effect states that $R_1 = R_0$ for each unit, and the model of an additive effect states that $R_1 = R_0 + \tau$ for some fixed τ.

Each sampled unit receives either the treatment or the control. The treatment actually received is indicated by the binary variable Z which equals one for units receiving the treatment and equals zero for units receiving the control. The response to the treatment, R_1, is observed only from the treated units, i.e. those with $Z = 1$, and the response to the control is observed only from the control units, i.e. those with $Z = 0$.

1·3. *Estimating an additive treatment effect by matched sampling*

When treatments are not randomly assigned to units, adjustments for pretreatment variables are commonly used, the simplest type of adjustment being exact matching on a pretreatment, possibly vector, variable, X (Cochran, 1953). The following is a slightly idealized description of the construction of a sample of S treatment-control pairs matched on X. The idealization lies in assuming that we sample directly from an infinite population in which exact matching on X is always feasible; see § 6 below for discussion of matching from finite reservoirs. For a practical example of matching that is a fair approximation to this description see Cohn et al. (1981).

For $s = 1, \ldots, S$, sample without replacement a treated unit at random from among treated units in the population, i.e. sample a unit conditionally given $Z = 1$, and note its value x of X. Then sample at random without replacement a control subject having the same value of X, that is sample a unit conditionally given $(Z = 0, X = x)$. Finally, assign noninformative labels or subscripts, $(s, 1)$ and $(s, 2)$, at random to the treated and control units in pair s.

For notational convenience, the units within a pair are assigned noninformative random labels or subscripts: the treated and control unit within pair s are identified by the sign of $V_s = Z_{s1} - Z_{s2}$, with $V_s = 1$ if unit $(s, 1)$ is the treated unit, and $V_s = -1$ if unit $(s, 2)$ is the treated unit. This notation is convenient because throughout the paper reference is made to the joint distribution of treatment assignment and response, and it is, therefore, convenient to have the treatment assignment explicitly indicated by a random variable, namely by $V_s = Z_{s1} - Z_{s2}$, rather than by a subscript. Let X_s denote the common value of X for the two units in pair s. The random variables defined by matched sampling from the population, namely $\{(R_{1s1}, R_{0s1}, R_{1s2}, R_{0s2}, Z_{s1}, Z_{s2}, X_s); s = 1, \ldots, S\}$, have distributions that are determined in an obvious way by the matching algorithm and by the distribution of (R_1, R_0, Z, X) in the population, but the two distributions are not identical. Perhaps the most conspicuous example concerns the distribution of the observed covariate in the population and in the matched sample: in general $\text{pr}(X \leqslant c) \neq \text{pr}(X \leqslant c | Z = 1) = \text{pr}(X_s \leqslant c)$ $(s = 1, \ldots, S)$ and any constant vector c; that is the distribution of the observed

covariates in the matched sample is the distribution of the observed covariates among treated ($Z=1$) units in the population.

Assuming the treatment has an additive effect, see § 1·2, the difference in responses of units $(s, 1)$ and $(s, 2)$ is

$$D_s = (R_{0s1} - R_{0s2}) + (Z_{s1} - Z_{s2})\tau, \tag{1·1}$$

and the treated-minus-control difference in responses in pair s is

$$D_s V_s = (Z_{s1} - Z_{s2})(R_{0s1} - R_{0s2}) + \tau. \tag{1·2}$$

The obvious estimates of τ are summary measures of the typical treated-minus-control difference within the S pairs, such as $\text{mean}_s\,(D_s V_s)$, $\text{median}_s\,(D_s V_s)$ or a trimmed mean of the $D_s V_s$'s. In observational studies, these estimates do not generally estimate τ because $(Z_{s1} - Z_{s2})(R_{0s1} - R_{0s2})$ need not have a distribution centred at zero: the reason is that nonrandom assignment of treatments, Z, can lead to dependence between $(Z_{s1} - Z_{s2})$ and $(R_{0s1} - R_{0s2})$. However, $(Z_{s1} - Z_{s2})(R_{0s1} - R_{0s2})$ does have a distribution centred at zero whenever treatment assignment is strongly ignorable given X.

1·4. *Strongly ignorable treatment assignment*

Treatment assignment is said to be strongly ignorable given pretreatment variables X if in the population, for all x,

$$(R_1, R_0) \perp\!\!\!\perp Z | X, \quad 0 < \text{pr}\,(Z = 1 | X = x) < 1, \tag{1·3}$$

where $A \perp\!\!\!\perp B | C$ is Dawid's (1979) notation for conditional independence of A and B given C. Treatment assignment is strongly ignorable (a) when treatments are randomly assigned so that Z is determined by the toss of a fair coin, (b) when Z is determined by the toss of a biased coin, where the bias is a possibly unknown function of X alone (Rubin, 1977), and (c) when treatment assignment is determined by X and certain other irrelevant covariates (Rosenbaum, 1984a, § 2.3). When treatment assignment is strongly ignorable in an observational study, a wide variety of methods of adjustment for X, including matched sampling, subclassification, model based adjustments, and conditional permutation tests, yield consistent inferences about treatment effects; see Theorem 4 of Rosenbaum & Rubin (1983a) with their $b(X) = X$, and Theorem 1 of Rosenbaum (1984b). In particular, when the treatment effect is constant, strongly ignorable treatment assignment implies that the S treated-minus-control differences $D_s V_s$ in § 1·3 are symmetrically distributed about τ, so any consistent estimate of the centre of symmetry of the distribution of treated-minus-control differences is also a consistent estimate of τ.

In practice, however, treatment assignment is rarely if ever known to be ignorable in observational studies, so that possible departures from ignorable assignment need to be considered. One approach involves testing ignorable assignment by contrasting observed data with the predictions of a causal theory (Rosenbaum, 1984a; 1986, § 6). A second approach involves studying the sensitivity of conclusions to violations of ignorable treatment assignment.

1·5. *Sensitivity analyses in observational studies*

A sensitivity analysis in an observational study begins with the assumption that treatment assignment is not strongly ignorable given the observed covariate X, but would

have been strongly ignorable if an additional unobservable pretreatment variable U could have been included with X; that is the assumption that, for all x, u,

$$(R_1, R_0) \perp\!\!\!\perp Z|(X, U), \quad 0<\text{pr}(Z=1|X=x, U=u)<1. \tag{1·4}$$

By the argument of § 1·4, when (1·4) holds, appropriate adjustment for (X, U) would yield a consistent estimate of the treatment effect; however, such an adjustment is not feasible since U is unobserved.

Cornfield et al. (1959) and Bross (1966, 1967) considered the case of binary responses (R_1, R_0) and a binary unobserved covariate U, in the absence of observed covariates X. They developed inequalities for population relative risks showing what one would need to believe about U in order to believe the treatment had no effect, so that the entire apparent difference in response in treated and control groups resulted from imbalances in the distribution of the unobserved U. Continuing the case of binary responses and a binary unobserved covariate U, but incorporating a categorical observed covariate X, Rosenbaum & Rubin (1983b) estimated the average treatment effect $E(R_1-R_0)$ after adjusting for (X, U) under a range of assumptions about U.

The current paper develops methods for studying the sensitivity of certain permutation inferences to assumptions about U. Unlike previous approaches, the methods are applicable to both continuous and discrete responses (R_1, R_0), and, moreover, the methods examine the sensitivity of permutational significance levels and confidence intervals rather than simply the sensitivity of point estimates. Additionally, the unobserved variable is no longer assumed at the outset to be binary. Rather, the case of binary U is derived as the least favourable or most conservative case over one class of possible values for the unobserved covariate, see § 2, whereas certain nonbinary U's are found to be least favourable for another class, § 4. Finally, where the method for binary responses (R_1, R_0) described by Rosenbaum & Rubin (1983b) involved varying four sensitivity parameters, the current approach is simpler, involving only a single sensitivity parameter. The price of this gain in simplicity is that the joint distribution of (R_1, R_0, Z, X, U) is only partially specified: significance levels and confidence intervals are bounded using the least favourable covariate values subject to the partial specification of the joint distribution. The net result is a simpler, more general, though also more conservative statement about the sensitivity of an inference to departures from ignorable treatment assignment.

In addition to (1·4), it is also assumed that the distribution of Z given (X, U) follows a logit model (Cox, 1970, § 2.3),

$$\log\frac{\text{pr}(Z=1|X=x, U=u)}{\text{pr}(Z=0|X=x, U=u)}=\kappa_x+\gamma u, \tag{1·5}$$

where κ_x is an unknown parameter for each x, and γ is an unknown sensitivity parameter, that is, a parameter which will be varied systematically to study the sensitivity of the resulting inferences. Expression (1·5) states that two units with the same value of X but different values of U, say u and u', will have odds of assignment to treatment 1 that differ by the factor $\exp\{\gamma(u-u')\}$. Note that if (1·4) and (1·5) hold and $\gamma=0$, then treatment assignment is strongly ignorable given X in the sense that (1·3) holds; in this case, the procedures described in the present paper yield conventional permutation inferences, such as Wilcoxon (1945) signed rank tests and confidence intervals. As γ is varied, we see the sensitivity of conventional inferences to violations of ignorable treatment assignment given X.

Since U is unobserved, its scale is not fixed. If the notion of varying the sensitivity parameter γ is to be meaningful, some restriction on the spread of U will eventually be required. Otherwise, if the random variable cU is permitted for all fixed real $c>0$ whenever U is permitted, then the value of γ in (1·5) conveys no information about the impact of the unobserved covariate, since $\gamma(cU)=(c\gamma)U$. A natural and convenient restriction will follow from the development in § 2·2.

Write $D=(D_1,\ldots,D_S)^{\mathrm{T}}$, $V=(V_1,\ldots,V_S)^{\mathrm{T}}$ and $W=(W_1,\ldots,W_S)^{\mathrm{T}}$, where $W_s=U_{s1}-U_{s2}$ is the difference in values of the unobserved covariate U for the two units in matched pair s. Moreover, write X^* and U^* for the matrices with S rows containing the X_s's and (U_{s1}, U_{s2})'s.

2. Sensitivity of matched pair permutation inferences

2·1. *Null distributions of test statistics*

Let $T=t(D, V)$ be a statistic such as (i) the sign-test statistic $t(D, V)=\Sigma\, I(V_sD_s>0)$, where I denotes an indicator function, or (ii) the Wilcoxon (1945) signed-rank statistic with $t(D, V)=\Sigma\, I(D_sV_s>0)q_s(D)$, where $q_s(D)$ is the rank of $|D_s|$, with average ranks in case of ties, or (iii) the sample mean treated versus control difference $(\Sigma\, D_sV_s)/S$, or (iv) the sample midmean of the S differences D_sV_s. Let B be the set containing the 2^S S-dimensional vectors whose coordinates are either 1 or -1; that is B is the set of all possible treatment assignments in the S matched pairs. The following theorem shows that if the vector of unobserved covariate differences, W, were in fact observed, then the null permutation distribution of the test statistic T would have a comparatively simple form. Note in particular that when $\gamma=0$, the conventional randomization distribution is obtained in (2·2).

Theorem 1. *Assume* (a) *that treatment assignment is strongly ignorable given* (X, U) *so* (1·4) *holds,* (b) *that treatment assignment* Z *and* (X, U) *are related by the logit model* (1·5), *and* (c) *that* S *treated/control pairs matched on* X *but not on* U *have been constructed from an infinite population of units as described in* § 1·3. *Then under the null hypothesis of zero treatment effect, we have, for each* k,

$$\mathrm{pr}\{t(D, V)\geqslant k\,|\,D=d, X^*, U^*\}=\mathrm{pr}\{t(d, V)\geqslant k\,|\,W=w\} \tag{2·1}$$

$$=\sum_{v\in B} I\{t(d, v)\geqslant k\}\prod_{s=1}^{S}\frac{\exp(\frac{1}{2}\gamma v_sw_s)}{\exp(\frac{1}{2}\gamma w_s)+\exp(-\frac{1}{2}\gamma w_s)} \tag{2·2}$$

$$=h_k(d, w),$$

say.

Proof. Here (2·1) states that we may ignore X^* and treat D as a constant fixed at its observed value $D=d$ in calculating the conditional distribution of the test statistic (Rosenbaum, 1984b, Th. 1), whereas (2·2) gives the form of the distribution. Under the null hypothesis of no treatment effect, we have $D_s=R_{0s1}-R_{0s2}$. By (1·4) and the matching procedure in § 1·3, we have

$$(R_{1s1}, R_{0s1})\perp\!\!\!\perp(R_{1s2}, R_{0s2})\perp\!\!\!\perp V_s\,|\,(X_s, U_{s1}, U_{s2}),$$

so that, under the null hypothesis, $D_s\perp\!\!\!\perp V_s\,|\,(X_s, U_{s1}, U_{s2})$. Now by (1·5)

$$\mathrm{pr}(V_s=1\,|\,X_s=x_s, U_{s1}=u_{s1}, U_{s2}=u_{s2})=\frac{\exp\{\frac{1}{2}\gamma(u_{s1}-u_{s2})\}}{\exp\{\frac{1}{2}\gamma(u_{s1}-u_{s2})\}+\exp\{-\frac{1}{2}\gamma(u_{s1}-u_{s2})\}}. \tag{2·3}$$

Since (2·3) depends on (x_s, u_{s1}, u_{s2}) only through $u_{s1}-u_{s2}$, and since $W_s = U_{s1}-U_{s2}$, we have (2·1). Now (2·2) follows immediately from (2·3). □

The following corollary is useful in setting confidence limits for τ when the treatment effect is additive. The proof uses (1·1) and is otherwise parallel to Theorem 1.

COROLLARY 1. *Assuming* (a), (b) *and* (c) *of Theorem* 1, *and also assuming the treatment has an additive effect for some* τ, *it follows that* $T' = t(D - V\tau, V)$ *has the conditional distribution given by*

$$\text{pr}\,(T' \geqslant k | D - V\tau = d',\ W = w) = h_k(d', w).$$

2·2. *Least favourable values for unobserved covariates*

As it stands, (2·2) cannot be used for direct calculations since W is unobserved. Instead, a bound on (2·2) is found by identifying the 'least favourable' or most conservative value of w. Intuition suggests that in (2·2) a least favourable w would be highly correlated with d, and that the more extreme w is in the same direction as d, the larger (2·2) would become. This intuition is correct, as the following formal discussion demonstrates.

For any S-dimensional vector d, define the preorder $\leqslant_d$ on R^S by

$$a \leqslant_d b \quad \text{if}\ \{(b_s - a_s)d_s \geqslant 0;\ s = 1, \ldots, S\}. \tag{2·4}$$

If each coordinate of d were strictly positive, then $\leqslant_d$ would define the usual coordinate-wise partial order on R^S; in general, however, $\leqslant_d$ defines 'up' or 'down' for each coordinate based on the sign of the corresponding coordinate of d. If $d_s = 0$ then the sth coordinate is, in effect, ignored in (2·4). For each fixed d, the several statistics $t(d, v)$ mentioned at the beginning of § 2·1 are isotonic or order preserving on B with respect to $\leqslant_d$; that is

$$(v, v' \in B;\ v \leqslant_d v') \rightarrow \{t(d, v) \leqslant t(d, v')\}. \tag{2·5}$$

In words, (2·5) says that for each fixed d, the test statistic $t(d, v)$ increases as the signs of coordinates of v are changed to agree with those of d.

LEMMA 1. *For each fixed* d, *let* $t(d, v)$ *be isotonic with respect to* $\leqslant_d$ *on* B. *Then, for each* $\gamma \geqslant 0$ *and for each fixed* d, *the function* $h_k(d, w)$ *in* (2·2) *is isotonic with respect to* $\leqslant_d$; *that is* $(w \leqslant_d w')$ *implies*

$$\{h_k(d, w) \leqslant h_k(d, w')\}. \tag{2·6}$$

Proof. Fix d. By assumption, (2·5) holds. To prove (2·6), assume $w \leqslant_d w'$ so that $(w'_s - w_s)d_s \geqslant 0$ for $s = 1, \ldots, S$, and hence

$$\left\{\frac{\exp(\frac{1}{2}\gamma w'_s)}{\exp(\frac{1}{2}\gamma w'_s) + \exp(-\frac{1}{2}\gamma w'_s)} - \frac{\exp(\frac{1}{2}\gamma w_s)}{\exp(\frac{1}{2}\gamma w_s) + \exp(-\frac{1}{2}\gamma w_s)}\right\} d_s \geqslant 0. \tag{2·7}$$

But (2·7) says that

$$\{\text{pr}\,(V_s = 1 | W = w', D = d) - \text{pr}\,(V_s = 1 | W = w, D = d)\}d_s \geqslant 0$$

for $s = 1, \ldots, S$, so $V_s = 1$ is more probable with $W = w'$ than with $W = w$ if $d_s > 0$, but $V_s = -1$ is more probable with $W = w'$ than with $W = w$ if $d_s < 0$. In other words, if $d_s \neq 0$, V_s is more likely to have the same sign as d_s if $W = w'$ than if $W = w$. If, on the other hand, $d_s = 0$ then, by (2·5), $t(d, V)$ takes the same value regardless of the value of V_s. The result now follows from (2·5), or formally from Lemma 3.3 of Ahmed, Leon & Proschan (1981) applied to coordinates of V with $d_s \neq 0$. □

Recall that the probability in (2·1) and (2·2) cannot be calculated from observable data since W is unobserved. Recall also the observation in § 1·5 that, if the sensitivity parameter γ is to be meaningful, some restriction on the spread of the unobserved covariate U is required. The following theorem bounds (2·1) by two quantities that can be calculated when the unobserved covariate U has a finite range, $\text{pr}(U \in [0, 1]) = 1$. As noted in § 1·5, previous work by Cornfield et al. (1959), Bross (1966, 1967) and Rosenbaum & Rubin (1983b) has assumed that U is binary, so that in particular $\text{pr}(U \in [0, 1] = 1)$. In the current context of matched pair permutation inferences, Theorem 2 implies that, when U is confined to $[0, 1]$, the least favourable values of U are binary: no configuration of U values confined to $[0, 1]$ will perturb the inference more than a certain binary vector.

THEOREM 2. *Assume that conditions* (a–c) *of Theorem* 1 *hold, that the null hypothesis of zero treatment effect holds, and that* $\gamma \geqslant 0$. *Suppose, in addition, that the unobserved covariate* U *is confined to the interval* $[0, 1]$. *Define* $\tilde{w}(d)$ *to be the vector with coordinate* s *equal to* $\text{sgn}(d_s)$, *for* $s = 1, \ldots, S$, *where* $\text{sgn}(a) = 1$ *if* a *is positive,* -1 *if* a *is negative, and* 0 *if* a *is zero. Then, for each possible* w,

$$h_k\{d, -\tilde{w}(d)\} \leqslant \text{pr}(T \geqslant k | D = d, W = w) \leqslant h_k\{d, \tilde{w}(d)\} \tag{2·8}$$

for any statistic $T = t(d, V)$ *isotonic on* B *with respect to* $\leqslant_d$.

Proof. Since $U \in [0, 1]$, it follows that $W_s \in [-1, 1]$ for each s. Then

$$\{\text{sgn}(d_s) - W_s\} d_s \geqslant 0, \quad \{-\text{sgn}(d_s) - W_s\} d_s \leqslant 0$$

for each s; hence, for each d,

$$-\tilde{w}(d) \leqslant_d W \leqslant_d \tilde{w}(d). \tag{2·9}$$

The conclusion (2·8) follows from (2·9) and (2·6). □

Restricting the spread of U by restricting its range has two advantages and one potential disadvantage. The first advantage is the conclusion of Theorem 1: when the spread of U is restricted by restricting its range, the least favourable covariate value $\tilde{w}(d)$ has a simple, easily used form that does not depend on the specific statistic $T = t(d, V)$ providing T is isotonic on B with respect to $\leqslant_d$. The second advantage is that the assumptions of Theorem 2 can be stated with reasonable precision in nontechnical language that can be understood by nonstatisticians: specifically, it is assumed that two units matched exactly on X have odds of assignment to the treatment that differ by at most a factor of e^{γ}. The potential disadvantage of restricting the spread of U by restricting its range arises when the response (R_1, R_0) has a long-tailed distribution and $t(D, V)$ is a nonrobust statistic such as a sample mean difference; in this case, the extreme tail behaviour of (R_1, R_0, U) may be critically important, and may not be adequately described by a distribution of U confined to a finite range. A method that allows for a long-tailed distribution of U is discussed in detail in § 4.

As noted formally in § 4·2, for many test statistics including the signed rank statistic and the sign statistic, $h_k(d, w)$ is unchanged if, for some s, the signs of both d_s and w_s are changed to those of $-d_s$ and $-w_s$, respectively. This is convenient in the examples in the next section, since it allows a natural tabulation of data ignoring the random subscripts, without substantive alteration of the problem.

3. SOME EXAMPLES

3·1. *Signed rank test*

Table 1(a) contains data from a paired randomized experiment involving two treatments for wheat seeds. The data were used by Wilcoxon (1945) in introducing the signed rank test, which is, of course, fully appropriate here since its permutation distribution, namely (2·2) with $\gamma = 0$, is the known distribution created by the random assignment of treatments within each pair. The one-sided significance level is 0·0195, suggesting that treatment A is superior.

In contrast, had identical data been obtained by matching in an observational study, use of the conventional reference distribution for the signed rank statistic would depend on the often tenuous assumption of strongly ignorable treatment assignment given the variables used in forming the pairs. Table 1(b) displays the sensitivity of the one-sided significance level as the sensitivity parameter e^γ is varied. Informally, in testing in an observational study against a one-sided alternative that treatment A is superior, an unobserved covariate U would need to increase the odds of assignment to treatment A by more than 50%, that is $e^\gamma \geqslant 1{\cdot}5$, before altering the qualitative impression that treatment A is superior; however, that impression would be open to question if it were plausible that an unobserved U exists which doubled ($e^\gamma = 2$) or tripled ($e^\gamma = 3$) the odds of assignment to treatment A.

In the case of the signed rank statistic, $t(d, V) = \Sigma\, I(d_s V_s > 0) q_s(d)$, the upper bound $h_k\{d, \tilde{w}(d)\}$ in (2·9) may be approximated for large S by $1 - \Phi\{(k - \mu - \frac{1}{2})/\sigma\}$, providing $k > \mu + \frac{1}{2}$, where $\Phi(.)$ is the standard normal cumulative distribution function, and μ and σ^2 are the null expectation and variance of T when $W = \tilde{w}(d)$, namely

$$\mu = p \sum_{s:d_s \neq 0} q_s(d), \quad \sigma^2 = p(1-p) \sum_{s:d_s \neq 0} \{q_s(d)\}^2, \tag{3·1}$$

Table 1(a). *Wilcoxon's* (1945) *data on treatments of wheat seeds*

Pair	Treatment A	Treatment B	Signed rank	Least favourable w
1	209	151	8	1
2	200	168	7	1
3	177	147	6	1
4	169	164	1	1
5	159	166	−3	−1
6	169	163	2	1
7	187	176	5	1
8	198	188	4	1

Table 1(b). *Sensitivity analysis: individual rank sum probabilities and the upper bound on the one-sided significance level for the observed sample*

	Sensitivity parameter e^γ				
$T(D, V)$	1·0	1·1	1·5	2·0	3
36	0·0039	0·0057	0·0168	0·0390	0·1001
35	0·0039	0·0052	0·0112	0·0195	0·0334
34	0·0039	0·0052	0·0112	0·0195	0·0334
33	0·0078	0·0098	0·0187	0·0293	0·0445
Significance level	0·0195	0·0258	0·0579	0·1073	0·2113

$e^\gamma = 1$ leads to the conventional randomization distribution.

where $p = \exp(\frac{1}{2}\gamma)\{\exp(\frac{1}{2}\gamma) + \exp(-\frac{1}{2}\gamma)\}^{-1}$. In (3·1), μ and σ^2 are the exact permutational expectation and variance, allowing for both tied/averaged ranks in the $|d_s|$'s, and concordant pairs with $d_s = 0$ which do not enter into the permutation distribution. If D_s has a continuous distribution, then ties and concordant pairs occur with probability zero, and we have

$$\mu = p\frac{S(S+1)}{2}, \quad \sigma^2 = p(1-p)\frac{S(S+1)(2S+1)}{6},$$

which reduce to familiar expressions when $\gamma = 0$ and $p = \frac{1}{2}$ (Lehmann, 1975, p. 128). Since $\{p(1-p)\}^{\frac{1}{2}}$ varies slowly with p in the neighbourhood of $p = \frac{1}{2}$, we may expect modest changes in σ as γ is varied near $\gamma = 0$, compared with more sizeable changes in μ. If this large-sample approximation is applied to the small example in Table 1(a), the significance levels for $e^\gamma = 1, 1{\cdot}1, 1{\cdot}5, 2$ and 3 are, respectively, 0·021, 0·028, 0·059, 0·104 and 0·187.

3·2. *Binary responses*

In the case of binary responses (R_1, R_0), the quantity D_s takes values $-1, 0$ or 1, and the sign statistic $t(D, V) = \Sigma\, I(V_s D_s > 0)$ is the number of matched pairs in which the treated unit responded with a 1 response and the control unit responded with a 0 response. A pair is concordant if $D_s = 0$ and is discordant otherwise. Let $S^* \leqslant S$ be the number of discordant pairs. For concordant pairs, $I(V_s D_s > 0)$ equals zero regardless of the value of V_s, so concordant pairs do not contribute a stochastic summand to the conditional distribution of $t(D, V)$ in (2·2). When $\gamma = 0$, the distribution in (2·2) of the sign statistic is binomial with sample size S^* and probability of success $\frac{1}{2}$, and $t(D, V)$ is the conventional test statistic (McNemar, 1947; Cox, 1958b). For general γ, the distribution of $t(D, V)$ in (2·2) is that of the sum of S^* independent but not identically distributed Bernoulli random variables with probabilities of success

$$\exp\{\tfrac{1}{2}\operatorname{sgn}(d_s)w_s\}/\{\exp(\tfrac{1}{2}\gamma w_s) + \exp(-\tfrac{1}{2}\gamma w_s)\}$$

for the $S^* \leqslant S$ discordant pairs. With $\gamma \geqslant 0$, the least favourable probabilities in (2·8) are once again binomial:

$$\sum_{i=k}^{S^*}\binom{S^*}{i}q^i(1-q)^{S^*-i} \leqslant \operatorname{pr}(T \geqslant k | D = d,\ W = w) \leqslant \sum_{i=k}^{S^*}\binom{S^*}{i}p^i(1-p)^{S^*-i},$$

where p is defined in § 3·1, and $q = 1 - p$.

In his prospective study of the effects of cigarette smoking, Hammond (1964) pair-matched 36975 heavy smokers to nonsmokers on the basis of (i) age, (ii) race, (iii) height, (iv) nativity, (v) rural or urban residence, (vi) occupational exposures to dusts, fumes, etc., (vii) religion, (viii) education, (ix) marital status, (x) alcohol consumption, (xi) sleep duration, (xii) exercise, (xiii) severe nervous tension, (xiv) use of tranquilizers, (xv) current health, (xvi) history of cancer other than skin cancer, and (xvii) history of heart disease, stroke, or high blood pressure. In these 36975 pairs, 122 were apparently discordant for death from lung cancer, with 12 pairs in which the nonsmoker died of lung cancer, and 110 pairs in which the smoker died of lung cancer. In a one-sided test of the null hypothesis of zero effect of smoking, the upper bounds on the significance levels are less than 10^{-4} for $e^\gamma \leqslant 3$, and 0·0036, 0·03, 0·1 at $e^\gamma = 4, 5, 6$. The qualitative conclusion that smoking causes an increase in the risk of lung cancer would not be

altered even if it were assumed that the matching had failed to control an unobserved covariate U associated with a five-fold increase in the conditional odds of smoking, and with an arbitrarily strong association with the development of lung cancer. Conclusions that are similar in substance, though different in detail, were reached by Cornfield et al. (1959).

3·3. *Confidence intervals*

For any fixed γ and for each fixed value of W, say $W = w$, conditional confidence sets for τ may be constructed using Corollary 1 by inverting the hypothesis test. One such set is

$$\{\tau: h_{T'}(d', w) \geqslant \alpha \text{ with } T' = t(d', V),\ d' = d - V\tau\}, \tag{3·2}$$

where α is the confidence coefficient. However, since W is unobserved, the confidence set (3·2) cannot be determined from observable data. Still, if τ is contained in the set (3·2), then by Theorem 2 it is also contained in the set

$$\{\tau: h_{T'}(d', \tilde{w}(d')) \geqslant \alpha \text{ with } T' = t(d', V),\ d' = d - V\tau\}.$$

The signed rank statistic varies in discrete jumps as τ is varied, so $h_{T'}\{d', \tilde{w}(d')\}$ with $d' = d - V\tau$ need only be calculated for finitely many values of τ. With Wilcoxon's data in Table 1, the one-sided 95% confidence intervals for τ are respectively $\tau \geqslant 5$, $\tau \geqslant -7$, $\tau \geqslant -7$ and $\tau \geqslant -\infty$ for e^{γ} equal to 1, 1·5, 2 and 3. Alternatively, approximate interval estimates may be obtained by inverting the test based on (3·1), yielding in this particular case the same four confidence intervals.

4. ADDITIONAL INEQUALITY FOR THE SIGNIFICANCE LEVEL

4·1. *Alternative approach to unobserved covariate*

The bounds in § 2·2 on the tail area of the statistic $T = t(D, V)$ assumed that U has a finite range; specifically, that pr $(U \in [0, 1]) = 1$, so pr $(W_s \in [-1, 1]) = 1$ for $s = 1, \ldots, S$. As noted previously, without some restriction on the spread of U, the value of the sensitivity parameter γ would not be meaningful. In many problems, this simple, easily described restriction will provide an adequate indication of the sensitivity of a permutation inference to a range of plausible violations of ignorable treatment assignment. In some instances, however, it may be necessary to allow for the possibility that a small fraction of units have extreme values of U. This can easily be done if the sample contents of certain intervals are assumed; e.g. for some β $(0 \leqslant \beta \leqslant 1)$ and some $0 < a \leqslant 1 < b \leqslant \infty$

$$\sum_{s=1}^{S} I(w_s \in [-a, a]) \geqslant (1-\beta)S, \quad \sum_{s=1}^{S} I(w_s \in [-b, b]) = S. \tag{4·1}$$

In effect, Theorem 1 concerned the special case with $\beta = 0$, $a = 1$. With $a = 1$, $b = \infty$, and $\beta = 0{\cdot}1$, equations (4·1) and (1·5) state that, in 90% of the sample pairs, which have been matched exactly for X, the treated and control units have conditional odds of assignment to the treatment given (X, U) that differ by at most e^{γ}, whereas 10% of pairs have odds that may differ by arbitrarily large factors.

4·2. *An inequality*

For $s = 1, \ldots, S$, let ε_s be the S-dimensional vector whose sth coordinate is 1 and whose remaining coordinates are zero. The reflection group, G, of $S \times S$ orthogonal

matrices generated by the set of matrices M_c such that $M_c = I - 2cc^{\mathrm{T}}$ with either $c = \varepsilon_s$ or $c = (\varepsilon_s - \varepsilon_{s+1})/2^{\frac{1}{2}}$ for some s is the group of coordinate sign changes and coordinate permutations acting on R^S (Eaton & Perlman, 1977; Eaton, 1982, § 4). A function $t(d, v)$ on $R^S \times R^S$ is a decreasing reflection function with respect to G if for all (d, v):

(i) $t(d, v) = t(gd, gv)$ for each $g \in G$;

(ii) if $c = \varepsilon_s$ or $c = (\varepsilon_s - \varepsilon_{s+1})/2^{\frac{1}{2}}$, and if $(c^{\mathrm{T}}d)(c^{\mathrm{T}}v) \geqslant 0$, then $t(d, v) \geqslant t(d, M_c v)$;

for example, Eaton (1982, Definition 4.2). Condition (i) states that the value of $t(d, v)$ is unchanged if the same permutations and sign changes are applied to the coordinates of both d and v. With $c = \varepsilon_s$, (ii) states that if d_s and v_s have the same sign, then $t(d, v)$ would be reduced by changing the sign of v_s alone; this is easily seen to be equivalent to (2·5) if v is restricted to B. With $c = (\varepsilon_s - \varepsilon_{s+1})/2^{\frac{1}{2}}$, (ii) states that, if $(d_s - d_{s+1})(v_s - v_{s+1}) \geqslant 0$, then interchanging v_s and v_{s+1} would reduce $t(d, v)$; that is $t(.,.)$ is decreasing in transposition in the sense of Hollander, Proschan & Sethuraman (1977). As is easily checked, all statistics $t(d, v)$ mentioned in the first paragraph of § 2·1 are decreasing reflection functions with respect to G: informally, the statistics are large when the signs of coordinates of d and v agree and when positive v_s's are paired with large positive differences, i.e. large d_s's. For such statistics, the following lemma shows, in effect, that the tail area under the null hypothesis, $h_k(d, w) = \mathrm{pr}\,(T \geqslant k | D = d, W = w)$, is increased by reordering the coordinates of w and changing their signs to agree with d.

LEMMA 2. *If $t(d, v)$ is a decreasing reflection function with respect to the group, G, of sign changes and permutations, then $h_k(d, w)$ in* (2·2) *is also a decreasing reflection function with respect to G, providing $\gamma \geqslant 0$.*

Proof. It is easily checked that

$$\prod_{s=1}^{S} \frac{\exp\,(\frac{1}{2}\gamma v_s w_s)}{\exp\,(\frac{1}{2}\gamma w_s) + \exp\,(-\frac{1}{2}\gamma w_s)}$$

is a decreasing reflection function in (v, w) with respect to G (Hollander et al., 1977, § 2·7). Also, $I\{t(d, v) \geqslant k\}$ is a decreasing reflection function because $t(d, v)$ is. The result then follows from Theorem 4.3 in Eaton's (1982) review paper; Eaton says that the theorem is due to J. C. Conlon, R. Leon, F. Proschan and J. Sethuraman in unpublished work. □

The least favourable w subject to (4·1) can be determined from Lemmas 1 and 2. Let d^* denote the vector of ordered absolute coordinates of d, so that

$$d_1^* = \min_{1 \leqslant s \leqslant S} |d_s|, \quad d_S^* = \max_{1 \leqslant s \leqslant S} |d_s|,$$

etc. Let L be the least integer greater than or equal to $(1 - \beta)S$. Finally, let w^* be the vector whose first L coordinates equal a, and whose remaining $S - L$ coordinates equal b. In the example in § 3·1 with $\beta = 0{\cdot}125$, $a = 1, b = \infty$, we have $d^* = (5, 6, 7, 10, 11, 30, 32, 58)$, $L = 7$ and $w^* = (1, 1, 1, 1, 1, 1, 1, \infty)$.

THEOREM 3. *Assume that $t(d, v)$ is a decreasing reflection function with respect to G and that $\gamma \geqslant 0$. For each k and for each w satisfying* (4·1)

$$h_k(d, w) \leqslant h_k(d^*, w^*).$$

Proof. By Lemma 2, $h_k(d, w)$ is a decreasing reflection function with respect to G. For any fixed w, we may, therefore, apply a sequence of sign changes to the coordinates of both d and w, yielding $d^{(1)}$ and $w^{(1)}$, with $d_s^{(1)} \geqslant 0$ for all s, and $h_k(d, w) = h_k(d^{(1)}, w^{(1)})$;

this follows from property (i) in § 4·2. Continuing, we may apply a sequence of permutations to the coordinates of both $d^{(1)}$ and $w^{(1)}$, sorting $d^{(1)}$ into order from smallest to largest, yielding d^* and $w^{(2)}$, say, where $h_k(d^{(1)}, w^{(1)}) = h_k(d^*, w^{(2)})$. Now, construct $w^{(3)}$ from $w^{(2)}$ by setting $w_s^{(3)} = |w_s^{(2)}|$ for each s. Since $h_k(.,.)$ is decreasing in sign reflections, we have $h_k(d^*, w^{(2)}) \leqslant h_k(d^*, w^{(3)})$. Since $h_k(.,.)$ is decreasing in transposition, we have $h_k(d^*, w^{(3)}) \leqslant h_k(d^*, w^{(4)})$ where $w^{(4)}$ contains the coordinates of $w^{(3)}$ sorted into order from smallest to largest. If $w^{(4)}$ satisfies (4·1), then $0 \leqslant w_s^{(4)} \leqslant w_s^*$ for each s; it follows from Lemma 1 that $h_k(d^*, w^{(4)}) \leqslant h_k(d^*, w^*)$, proving the theorem. □

4·3. *An example*

To take $\beta = 0{\cdot}125$, $a = 1$, and $b = \infty$ in the example of § 3·1 is to assume, in effect, that within pairs matched exactly on X, the odds of assignment to the treatment varies by at most a factor of e^γ, except perhaps in one pair where the difference may be arbitrarily large. As noted above, this example yields $d^* = (5, 6, 7, 10, 11, 30, 32, 58)$ and $w^* = (1, 1, 1, 1, 1, 1, 1, \infty)$, effectively trimming the $d_8^* = 58$ observation so that it does not affect the permutation distribution. For $e^\gamma = 1, 1{\cdot}5, 2$ and 3, the upper bounds on the one-sided significance levels are 0·0195, 0·0964, 0·1610 and 0·2818, which, as expected from Lemma 1, are noticeably higher than obtained in § 3·1, although not vastly higher. This difference reflects the fact that the bounds obtained in the current section are based on a weaker or more conservative assumption about the impact of the unobserved covariate.

5. MATCHED SAMPLING WITH MORE THAN ONE CONTROL

To increase power and precision, each treated unit is often matched with more than one control; see Miettinen (1969) and Ury (1975) for discussion of power considerations, and Cox (1966) and Cohn et al. (1981) for numerical examples. The results of previous sections do not extend in their full generality to case of multiple controls. A result analogous to Theorem 1 does hold, as does a result analogous to Lemma 2 for the group of permutations within matched sets; however, these two results alone do not generally lead to unique least favourable values of the unobserved covariate for broad classes of test statistics, e.g. for the decreasing reflection functions with respect to permutations within matched sets. Nonetheless, the important special case in which the responses (R_1, R_0) are binary can be dealt with using arguments similar to those in §§ 2 and 4. Assume the null hypothesis, H_0: $R_1 = R_0 = R$, say, and the conditions of Theorem 2, except that in the matching procedure in § 1·3: (i) n_s controls are sampled in step 3 where n_s may be a function of X_s but may not depend on (R_1, R_0), and (ii) random subscripts $(1, \ldots, n_s + 1)$ are assigned. Let f_s be the number of subjects in matched set s for whom $R = 1$. Let $A_s = 1$ if the treated subject in set s has $R = 1$, and $A_s = 0$ otherwise.

Under the null hypothesis, conditionally given the response (R) and covariate (X, U) values, the probability that $Z_{si} = 1$ is $\exp(\gamma u_{si})/\{\Sigma_j \exp(\gamma u_{sj})\}$ from (1·5). It follows that the permutational probability that $A_s = 1$ is $p_s = \{\Sigma' \exp(\gamma u_{si})\}/\{\Sigma_j \exp(\gamma u_{sj})\}$, where Σ' is a sum over units, i, in matched set s with observed responses R equal to 1. Since by assumption pr $(U \in [0, 1]) = 1$, and $\gamma \geqslant 0$, it follows that:

$$p_{*s} = \frac{f_s}{f_s + (n_s + 1 - f_s)e^\gamma} \leqslant p_s \leqslant \frac{f_s e^\gamma}{f_s e^\gamma + (n_s + 1 - f_s)} = p_s^*, \tag{5·1}$$

so the distribution of ΣA_s is bounded by two distributions of sums of independent Bernoulli random variables with probabilities given by the upper and lower bounds in

(5·1). For large S, we may approximate the upper bound on the upper tail area for ΣA_s by referring

$$\left[\left\{\sum_{s=1}^{S}(A_s-p_s^*)\right\}-\tfrac{1}{2}\right]\Big/\left\{\sum_{s=1}^{S}p_s^*(1-p_s^*)^{\frac{1}{2}}\right\} \tag{5.2}$$

to tables of the standard normal distribution. When $\gamma=0$, expression (5·2) is the familiar statistic of Mantel & Haenszel (1959) and Cox (1966).

6. Matching in practice

As noted previously, the description of matching in § 1·3 is somewhat idealized, for several reasons. First, the reservoir from which controls are sampled, though it may be large, is always finite. Nonetheless, the basic results in previous sections continue to hold with finite reservoirs, providing: (i) the finite treated and control reservoirs used to form matches are random samples from the treated ($Z=1$) and control ($Z=0$) subpopulations; (ii) every treated unit is exactly matched on X with a control; and (iii) if more than one exact pairing of treated and control units is possible, the pairing actually selected is a random draw from the set of all exact pairings (Rosenbaum & Rubin, 1985a, § 1·5).

Unfortunately, with a finite reservoir of controls, an exact matching is likely to exist only if X is discrete and of low dimension, and only if the control reservoir is much larger than the treated sample (Rubin, 1973).

One alternative useful when X is of high dimension is to match not on X but rather on an estimate of the scalar propensity score (Rosenbaum & Rubin, 1983a, 1985b) $e(X)=\text{pr}\,(Z=1|X)$. The results in §§ 2 and 4 do not generally apply to matching on propensity scores because two units with the same $e(X)$ will not necessarily have the same κ_x in (1·5), so (2·3) will not generally hold with X replaced by $e(X)$. However, if the unobserved covariate U is independent of X, then the results in §§ 2 and 4 do apply to matching on $e(X)$, for in this case

$$e(x)=\text{pr}\,(Z=1|X=x)=\int\frac{\exp(\kappa_x+\gamma u)}{1+\exp(\kappa_x+\gamma u)}\,\text{pr}\,(U=u)\,du$$

by (1·5), so $e(x)$ is a $1-1$ function of κ_x. In this case, it follows that units matched exactly on $e(x)$ do have the same κ_x, so (2·3) does hold with X replaced by $e(X)$. The assumption that U and X are independent is fairly innocuous from a practical point of view, since we could always replace U in (1·4) and (1·5) by $U'=U-E(U|X)$ which, though not necessarily independent of X, is at least uncorrelated with X. In short, when exact matching on X is not feasible, the sensitivity analysis described here can reasonably be applied to a close matching on the scalar propensity score, $e(x)=\text{pr}\,(Z=1|X=x)$.

References

Ahmed, A., Leon, R. & Proschan, F. (1981). Generalized association with applications in multivariate statistics. *Ann. Statist.* **9**, 168–76.

Bross, I. D. J. (1966). Spurious effects from an extraneous variable. *J. Chronic Dis.* **19**, 637–47.

Bross, I. D. J. (1967). Pertinency of an extraneous variable. *J. Chronic Dis.* **20**, 487–95.

Cochran, W. G. (1953). Matching in analytical studies. *Am. J. Pub. Health* **43**, 684–91.

Cochran, W. G. (1965). The planning of observational studies of human population (with discussion). *J.R. Statist. Soc.* A **128**, 134–55.

Cochran, W. G. & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā* A **35**, 417–46.

COHN, P. F., HARRIS, P., BARRY, W., ROSATI, R. A., ROSENBAUM, P. R. & WATERNAUX, C. (1981). Prognostic importance of anginal symptoms in angiographically defined coronary artery disease. *Am. J. Cardiol.* **47**, 233–7.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A., SHIMKIN, M. & WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

COX, D. R. (1958a). *The Planning of Experiments.* New York: Wiley.

COX, D. R. (1958b). Two further applications of a model for binary regression. *Biometrika* **45**, 562–5.

COX, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika* **53**, 215–20.

COX, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J.R. Statist. Soc.* B **41**, 1–31.

EATON, M. L. (1982). A review of selected topics in multivariate probability inequalities. *Ann. Statist.* **10**, 11–43.

EATON, M. L. & PERLMAN, M. (1977). Reflection groups, generalized Schur functions, and the geometry of majorization. *Ann. Prob.* **5**, 829–60.

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

HAMILTON, M. A. (1979). Choosing a parameter for 2×2 table or 2×2×2 table analysis. *Am. J. Epidemiol.* **109**, 362–75.

HAMMOND, E. C. (1964). Smoking in relation to mortality and morbidity. Findings in first thirty-four months of follow-up in a prospective study started in 1959. *J. Nat. Cancer Inst.* **32**, 1161–88.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Am. Statist. Assoc.* To appear.

HOLLAND, P. W. & RUBIN, D. B. (1983). On Lord's Paradox. In *Principles of Modern Psychological Measurement*, Eds. H. Wainer and S. Messick, pp. 3–26. Hillsdale, New Jersey: Lawrence-Erlbaum.

HOLLANDER, M., PROSCHAN, F. & SETHURAMAN, J. (1977). Functions decreasing in transposition and their applications in ranking problems. *Ann. Statist.* **5**, 722–33.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden Day.

MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719–48.

MCNEMAR, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* **12**, 153–7.

MIETTINEN, O. S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* **25**, 339–55.

ROSENBAUM, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Am. Statist. Assoc.* **79**, 41–8.

ROSENBAUM, P. R. (1984b). Conditional permutation tests and the propensity score in observational studies. *J. Am. Statist. Assoc.* **79**, 565–74.

ROSENBAUM, P. R. (1984c). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc.* A **147**, 656–66.

ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: An observational study. *J. Educ. Statist.* **11**, 207–24.

ROSENBAUM, P. R. & RUBIN, D. B. (1983a). The central role of the propensity score on observational studies for causal effects. *Biometrika* **70**, 41–55.

ROSENBAUM, P. R. & RUBIN, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B **45**, 212–8.

ROSENBAUM, P. R. & RUBIN, D. B. (1985a). The bias due to incomplete matching. *Biometrics* **41**, 106–16.

ROSENBAUM, P. R. & RUBIN, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statistician* **39**, 33–8.

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159–83. Correction (1974) **30**, 728.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.* **66**, 688–701.

RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.* **2**, 1–26.

RUBIN, D. B. (1978). Bayesian inference for casual effects: The role of randomization. *Ann. Statist.* **6**, 34–58.

URY, H. K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics* **31**, 643–50.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.

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