The Annals of Statistics
1989, Vol. 17, No. 2, 643–653

ON PERMUTATION TESTS FOR HIDDEN BIASES IN OBSERVATIONAL STUDIES: AN APPLICATION OF HOLLEY'S INEQUALITY TO THE SAVAGE LATTICE[1]

BY PAUL R. ROSENBAUM

University of Pennsylvania

Randomized experiments and observational studies both attempt to estimate the effects produced by a treatment, but in observational studies, subjects are not randomly assigned to treatment or control. A theory of observational studies would closely resemble the theory for randomized experiments in all but one critical respect: In observational studies, the distribution of treatment assignments is not known. The problems that are special to observational studies revolve around our uncertainty about how treatments were assigned. In this connection, tools are needed for describing distributions of treatment assignments that do not assign equal probabilities to all assignments. Two such tools are a lattice of treatment assignments first studied by Savage and an inequality due to Holley for probability distributions on a lattice. Using these tools, it is shown that certain permutation tests are unbiased as tests of the null hypothesis that the distribution of treatment assignments resembles a randomization distribution against the alternative hypothesis that subjects with higher responses are more likely to receive the treatment. In particular, these tests are unbiased against alternatives formulated in terms of a model previously used in connection with sensitivity analyses.

1. Introduction: Detecting hidden biases in observational studies. To say with Fisher (1935) that randomization forms the "reasoned basis for inference" in randomized experiments is to say that inferences are based entirely on the known distribution of treatment assignments created by the physical act of randomization. In this formulation, the only stochastic element is the known random assignment of subjects to treatments; there is no fictitious sampling from an imagined infinite population of experimental subjects. Inferences are from the observed responses of the subjects in the experiment to the responses these same subjects would have exhibited had they received the alternative treatment, an idea given formal expression by Welch (1937), Section 2.

In contrast, Cochran (1965) defined an observational study as an empirical investigation in which: "... the objective is to elucidate cause and effect relationships ... [in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover or to assign subjects at random to different procedures." A

Received February 1988; revised June 1988.

[1]This work was supported by a grant from the Measurement Methods and Data Improvement Program of the National Science Foundation.

AMS 1980 *subject classifications*. Primary 62G10, 60C05; secondary 06D99.

Key words and phrases. Observational studies, permutation test, lattice theory, Holley's inequality, decreasing in transposition, decreasing reflection function, rank sum test, signed rank test, Mantel–Haenszel test, McNemar–Cox test, unbiased test.

closely argued theory of observational studies would resemble the theory for randomized experiments in all respects but one: In observational studies, the mechanism that assigned subjects to treatments is not known. The problems that arise in observational studies but not in experiments concern this unknown distribution of treatment assignments.

Three broad categories of devices are used in observational studies to address uncertainty about the distribution of treatment assignments. First, we may find that treated and control subjects differed prior to treatment with respect to covariates whose values have been recorded, leading us to adjust for these observed pretreatment differences. Often this entails grouping subjects with the same value of these observed covariates into subclasses or matched pairs. Conventional permutation tests and interval estimates used in observational studies may be derived from the assumption that within these ostensibly homogeneous subclasses or pairs, treatments are assigned essentially at random. For instance, this is true of the Mantel–Haenszel (1959) statistic for subclassified binary data, the Wilcoxon (1945) signed rank statistic for matched pairs and other methods that involve more extensive adjustments [Rosenbaum (1984a, 1988a)]. Of course, this assumption, called adjustable treatment assignment, is quite tenuous: Subjects who appear similar on the basis of observed covariates may differ in ways that have not been observed. The second category of devices used in observational studies entails the collection of information that holds a reasonable prospect of indicating or detecting such unobserved pretreatment differences, thereby providing tests of the assumption of adjustable assignment. Two devices in this category are the use of two control groups [Rosenbaum (1987a)] and the use of outcomes known to be unaffected by the treatment [Rosenbaum (1984a), Section 3.2]. The third category assumes that a relevant unobserved covariate does exist, and investigates the sensitivity of conclusions about treatment effects to a range of assumptions about the unobserved covariate [e.g., Rosenbaum and Rubin (1983) and Rosenbaum (1987b, 1988b)]. The current article contains a fairly general result concerning the performance of devices in the second category, and it relates that result to a model used in the third category.

Specifically, it is assumed that there are two groups of subjects and an outcome which is not affected by the treatment that was given to one group and withheld from the other. This situation described here is typically one part of a larger observational study that includes either other outcomes or other groups for which effects are possible. The two groups may be two control groups with any outcome, or they may be the treated and control groups together with an unaffected outcome that was included in the study in an effort to detect hidden biases; for motivation and examples, see Rosenbaum (1984a, 1987a). As developed here, the formal considerations for the two cases—that is, multiple control groups and unaffected responses—are similar, and they are not distinguished in later sections. The null hypothesis states that the distribution of treatment assignment is adjustable given the subclasses, so conventional methods of adjustment yield appropriate corrections. The alternative hypothesis states that treatment assignment is not adjustable given the subclasses because of imbalances in

a relevant unobserved covariate. The unaffected response is the basis for the test, and the question is under what circumstances does such an unaffected response provide useful information about hidden biases.

Previous discussions of the performance of tests for unobserved differences have postulated infinite populations and random samples, neither of which actually exist in a typical observational study. In contrast, in the discussion that follows, the only stochastic element is the distribution of treatment assignments. To describe distributions that do not assign equal probability to all treatment assignments, two tools are used: a lattice of treatment assignments introduced by Savage (1964) and an inequality due to Holley (1974) for probability distributions on a lattice.

2. Notation and definitions. A total of N subjects have been divided into S subclasses on the basis of observed pretreatment characteristics or covariates, with n_s subjects in subclass s, for $s = 1, \ldots, S$. If the ith subject in subclass s receives the treatment then $Z_{si} = 1$, else if he receives the control $Z_{si} = 0$. Also, this subject exhibits a response, r_{si}. Write $\mathbf{Z}$ and $\mathbf{r}$ for the N-dimensional vectors containing, respectively, the Z_{si}'s and r_{si}'s in the lexical order, for example, $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{S, n_S})$. Write $a_s = \sum_{i=1}^{n_s} Z_{si}$ for the number of treated subjects in subclass s. When each $n_s = 2$ and each $a_s = 1$, we have S matched pairs. When $n_s \geq 2$ and $a_s = 1$ for each s, we have matching with multiple controls.

As noted in Section 1, it is assumed that the treatment does not affect this particular response, though it may affect others, so that each subject (s, i) would exhibit response r_{si} whether assigned to treatment or control. Since the response $\mathbf{r}$ is fixed, the only stochastic element being the treatment assignment $\mathbf{Z}$, it is convenient for notation, especially in connection with Section 4, to assume that within each subclass, subjects have been numbered in decreasing order of their responses, so that $r_{si} \geq r_{sj}$ if $i < j$. This notational convenience is without substantive consequences, as all features of the problem will be invariant with respect to permutations within subclasses, that is, with respect to renumbering the n_s subjects in subclass s. It is also convenient at first to assume that there are no ties in the responses within subclasses, so $r_{si} > r_{sj}$ if $i < j$; the case of tied responses will be discussed separately in Section 8.

A test statistic $T = t(\mathbf{Z}, \mathbf{r})$ has been selected to test for higher responses among treated subjects than among controls in the same subclass. Since the treatment does not affect $\mathbf{r}$, systematically higher responses among treated subjects would indicate some way in which treated and control subjects in the same subclass are not comparable. The statistic $t(\mathbf{Z}, \mathbf{r})$ is assumed to be decreasing in transposition within subclasses in a sense closely paralleling the discussion by Hollander, Proschan and Sethuraman (1977). Informally, this means that $t(\mathbf{Z}, \mathbf{r})$ increases in value as the ordering of $\mathbf{Z}$ and $\mathbf{r}$ becomes more similar. Formally, let $\mathbf{z}_{(sij)}$ denote the vector obtained from $\mathbf{z}$ by interchanging coordinates (s, i) and (s, j), that is, interchanging the ith and jth subject in subclass s. Define $\mathbf{r}_{(sij)}$ similarly. A function $t(\cdot, \cdot)$ is *decreasing in transposition within subclasses or* DTS if it is invariant with respect to permutations within subclasses, so that $t(\mathbf{z}, \mathbf{r}) = t\{\mathbf{z}_{(sij)}, \mathbf{r}_{(sij)}\}$ for all (s, i, j), and in addition, $t(\mathbf{z}, \mathbf{r}) \geq$

$t\{\mathbf{z}_{(sij)}, \mathbf{r}\}$ whenever $\{z_{si} - z_{sj}\} \cdot \{r_{si} - r_{sj}\} \geq 0$. When $\mathbf{z}$ is a vector of binary indicators of treatment assignment, $t(\mathbf{z}, \mathbf{r})$ is DTS if exchanging a treated subject with a higher response for a control in the same subclass with a lower response would not increase $t(\mathbf{z}, \mathbf{r})$. The following familiar statistics are all DTS: (i) the Wilcoxon (1945) rank sum statistic; (ii) the treated-minus-control difference in means, midmeans or medians; (iii) in the case of matched pairs—that is, $n_s = 2$, $a_s = 1$ for each s—the Wilcoxon signed rank statistic; (iv) in the case of matching with multiple controls—that is, $a_s = 1$ for each s—most sign/score statistics as defined in Rosenbaum (1988b). Also, many familiar statistics for discrete data are DTS; see Section 8.

The definition of DTS functions, given above, is actually more general than the examples just cited, and other DTS functions will be relevant shortly. In particular, there is no need for one of the arguments of a DTS function to be a vector $\mathbf{z}$ with binary coordinates. In general, the DTS condition given above requires that the function increase monotonically as the coordinates of its two vector arguments are permuted into the same order within each subclass. For instance, if for $s = 1, \ldots, S$, the nonnegative functions f_s: $R^2 \to R$ are *totally positive of order* 2 or TP_2 in the sense that $f_s(x, y) \cdot f_s(x', y') - f_s(x', y) \cdot f_s(x, y') \geq 0$ whenever $x \geq x'$ and $y \geq y'$, then $h(\mathbf{x}, \mathbf{y}) = \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(x_{si}, y_{si})$ is DTS; cf. Hollander, Proschan and Sethuraman (1977, Section 2.7). In particular, $f_s(x, y)$ is TP_2 if f_s is the conditional density of y given x under a normal linear regression model, or if f_s is the probability distribution of a binary y given a covariate x under a linear logit model with a nonnegative coefficient for x. When there is only one subclass, so $S = 1$, the DTS functions are the functions "decreasing in transposition" in the sense of Hollander, Proschan and Sethuraman (1977) or the "arrangement increasing" functions of Marshall and Olkin (1979), Section 6.F. For $S \geq 1$, the class of DTS functions is the same as the class of decreasing reflection functions with respect to one particular reflection group, namely the group of permutations within subclasses; see Eaton (1982) for discussion of decreasing reflection functions in general and for related references.

Let $\mathfrak{F}$ be the collection of all

$$L = \prod_{s=1}^{S} \binom{n_s}{a_s}$$

vectors $\mathbf{z}$ with binary coordinates such that $\sum_{i=1}^{n_s} z_{si} = a_s$ for each s. Generally, write $|A|$ for the cardinality of the set A, so that $|\mathfrak{F}| = L$. In a randomized experiment in which a_s of the n_s subjects in subclass s are randomly assigned to the treatment, with independent assignments in different subclasses, the set $\mathfrak{F}$ is the collection of possible treatment assignments, and each $\mathbf{z} \in \mathfrak{F}$ has the same probability, namely $\mathrm{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r}) = 1/L$. These treatment assignment probabilities condition not only on the unaffected response $\mathbf{r}$ but also on the subclass information; however, it is convenient to suppress this in the notation. The conventional randomization significance level is the proportion of treatment assignments $\mathbf{z} \in \mathfrak{F}$ that yield larger values of the test statistic than that observed, namely $|\{\mathbf{z} \in \mathfrak{F}: t(\mathbf{z}, \mathbf{r}) \geq t(\mathbf{Z}, \mathbf{r})\}|/L$.

The distribution $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r}) = 1/L$ will be called the *uniform distribution on* $\mathfrak{F}$. Later sections consider the behavior of $t(\mathbf{Z}, \mathbf{r})$ under alternative distributions for $\mathbf{Z}$ over $\mathfrak{F}$. If $\mathbf{Z}$ does not have the uniform distribution, then treatment assignment is not adjustable given the subclasses or pairs, and conventional adjustments do not suffice. The goal is to use the unaffected outcome $\mathbf{r}$ to detect such a departure from adjustable assignment.

3. DTS statistics yield unbiased tests against DTS alternatives. Recall that a statistical test is unbiased against a class of alternative hypotheses if the power of the test against each alternative in the class exceeds or equals the level of the test. The following theorem says that DTS statistics yield unbiased tests against alternatives in which $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ is DTS.

THEOREM 1. *Let $t(\cdot, \cdot)$ be a DTS statistic, and consider a test of the uniform distribution on $\mathfrak{F}$ which rejects when $t(\mathbf{Z}, \mathbf{r}) \geq c$. This test is unbiased against all alternatives in which* $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ *is DTS.*

In practical terms, this says the following. In observational studies, we may test for hidden biases by applying conventional statistics, such as the subclassified rank sum statistic, to an unaffected response $\mathbf{r}$. Theorem 1 says these tests have a prospect of detecting departures from the randomization distribution in which subjects with higher unaffected responses are more likely to receive the treatment. Typically, we select an observed unaffected response $\mathbf{r}$ to be a proxy or surrogate for an unobserved covariate $\mathbf{u}$ which would have been controlled by subclassification or matching had $\mathbf{u}$ been observed. Sections 6 and 7 discuss the relationship between such a $\mathbf{u}$ and the DTS alternatives in Theorem 1.

The proof of Theorem 1 is given in Section 5 and is based on the ideas reviewed in Section 4.

4. The Savage lattice, Holley's inequality and the composition theorem. Savage (1964) showed that $\mathfrak{F}$ is a finite distributive lattice. It is this fact that permits the application of Holley's inequality to probability distributions on $\mathfrak{F}$. Recall that a set $\mathfrak{F}$ becomes a *lattice* when it is endowed with a partial order $\lesssim$, such that any two elements, $\mathbf{z}$ and $\mathbf{z}^*$ of $\mathfrak{F}$ have a unique least upper bound, denoted by $\mathbf{z} \vee \mathbf{z}^*$, and a unique greatest lower bound, denoted by $\mathbf{z} \wedge \mathbf{z}^*$. In other words, for $\mathbf{z}, \mathbf{z}^* \in \mathfrak{F}$, the least upper bound vector $\mathbf{z} \vee \mathbf{z}' \in \mathfrak{F}$ is larger than both $\mathbf{z}$ and $\mathbf{z}^*$—that is, $\mathbf{z} \lesssim \mathbf{z} \vee \mathbf{z}^*$ and $\mathbf{z}^* \lesssim \mathbf{z} \vee \mathbf{z}^*$—and if $\mathbf{z}^{**}$ is any other vector in $\mathfrak{F}$ larger than both $\mathbf{z}$ and $\mathbf{z}^*$, then $\mathbf{z}^{**}$ is larger than $\mathbf{z} \vee \mathbf{z}^*$. The greatest lower bound is defined analogously. A lattice is *distributive* if the cap $\wedge$ and cup $\vee$ operations are distributive, in the sense that for all $\mathbf{z}, \mathbf{z}^*, \mathbf{z}^{**} \in \mathfrak{F}$, the distributive law holds, namely $\mathbf{z} \wedge (\mathbf{z}^* \vee \mathbf{z}^{**}) = (\mathbf{z} \wedge \mathbf{z}^*) \vee (\mathbf{z} \wedge \mathbf{z}^{**})$. [Actually, Savage showed that $\mathfrak{F}$ is a distributive lattice when there is just one subclass, $S = 1$. However, $\mathfrak{F}$ is still a distributive lattice when $S > 1$ since in this case it is the direct product of S distributive lattices, and is therefore a distributive lattice; see, e.g., Aigner (1979), page 32.]

The partial order in the *Savage lattice* is defined as follows. In $\mathfrak{F}$, the vector $\mathbf{z}^*$ *covers* $\mathbf{z}$, written $\mathbf{z} \lesssdot \mathbf{z}^*$ if, for some (s, i), $z_{si} = 0$ and $z_{s,i+1} = 1$, and $\mathbf{z}^* = \mathbf{z}_{(si,i+1)}$, that is, if $\mathbf{z}^*$ is obtained by interchanging adjacent coordinates of $\mathbf{z}$ in the same subclass in such a way as to move a 1 to the left and a 0 to the right. In other words, the treatment assignment $\mathbf{z}^*$ covers or is immediately above $\mathbf{z}$ if $\mathbf{z}^*$ is obtained from $\mathbf{z}$ by interchanging one treated subject for one control subject from the same subclass such that the control subject had a response whose rank within the subclass was 1 higher than the rank of response of the treated subject. Then $\mathbf{z}$ is less than or equal to $\mathbf{z}^*$ in the partial order, written $\mathbf{z} \lesssim \mathbf{z}^*$, if either $\mathbf{z} = \mathbf{z}^*$ or for some $J \geq 1$, there exists a sequence of $\mathbf{z}_j$'s $\in \mathfrak{F}$ such that $\mathbf{z} = \mathbf{z}_1 \lesssdot \mathbf{z}_2 \lesssdot \cdots \lesssdot \mathbf{z}_J = \mathbf{z}^*$.

The cap $\wedge$ and cup $\vee$ operations are most easily defined and studied by converting the $\mathbf{z}$'s into rank vectors. Specifically, one considers the function $\rho(\mathbf{z})$ which carries the N-dimensional vector $\mathbf{z}$ into a vector of dimension Σa_s containing the ranks of the treated subjects in each subclass, arranged from largest to smallest within subclasses. The reader is referred to Savage for specifics as the form of $\wedge$ and $\vee$ are not needed here.

The Savage lattice is intimately connected with the class of DTS functions. Keeping in mind the restriction that $r_{si} > r_{sj}$ for $i < j$, a permutation invariant function $t(\mathbf{z}, \mathbf{r})$ is DTS if and only if it is isotonic or order preserving in the Savage lattice; that is, if $\mathbf{z} \lesssim \mathbf{z}^*$, then $t(\mathbf{z}, \mathbf{r}) \leq t(\mathbf{z}^*, \mathbf{r})$.

Holley's (1974) inequality compares two probability distributions, say $\mu_1(\cdot)$ and $\mu_2(\cdot)$, on a finite distributive lattice, say $\mathfrak{F}$. The inequality is a sufficient condition for $\mu_2(\cdot)$ to be stochastically larger than $\mu_1(\cdot)$ in the sense that all functions isotonic in the lattice order have higher expectations under $\mu_2(\cdot)$; that is, all real-valued functions $g(\cdot)$ on $\mathfrak{F}$ such that $\mathbf{z} \lesssim \mathbf{z}^*$ implies $g(\mathbf{z}) \leq g(\mathbf{z}^*)$. Specifically, *Holley's inequality* states that if

$$\mu_2(\mathbf{z} \vee \mathbf{z}^*) \cdot \mu_1(\mathbf{z} \wedge \mathbf{z}^*) \geq \mu_2(\mathbf{z}) \cdot \mu_1(\mathbf{z}^*) \quad \text{for all } \mathbf{z}, \mathbf{z}^* \in \mathfrak{F},$$

then

$$\sum_{\mathbf{z} \in \mathfrak{F}} g(\mathbf{z}) \cdot \mu_2(\mathbf{z}) \geq \sum_{\mathbf{z} \in \mathfrak{F}} g(\mathbf{z}) \cdot \mu_1(\mathbf{z}) \quad \text{for all isotonic } g(\cdot).$$

Holley's (1974) proof of the inequality has been shortened and simplified with the aid of the "four-functions theorem" of Ahlswede and Daykin (1978), though the original proof contains some intermediate results of independent interest. An attractive presentation of the four-functions theorem and Holley's inequality is given by Bollobás (1986), Section 19, especially exercise 8, page 153.

The composition theorem for DTS functions concerns two DTS functions with one common vector argument, say $h_1(\mathbf{w}, \mathbf{x})$ and $h_2(\mathbf{x}, \mathbf{y})$. Let $\mathscr{W}$, $\mathscr{X}$ and $\mathscr{Y}$ be Borel subsets of R^N that are invariant with respect to permutations of coordinates within subclasses; for example, all of R^N and $\mathfrak{F}$ are two such subsets. Let λ be a σ-finite measure on the Borel subsets of R^N which is invariant with respect to permutations within subclasses; for example, Lebesgue measure on R^N and counting measure on $\mathfrak{F}$ are two such measures. The

composition of $h_1(\mathbf{w}, \mathbf{x})$ and $h_2(\mathbf{x}, \mathbf{y})$ is

$$h_3(\mathbf{w}, \mathbf{y}) = \int_{\mathscr{X}} h_1(\mathbf{w}, \mathbf{x}) h_2(\mathbf{x}, \mathbf{y}) \lambda(d\mathbf{x}).$$

The *composition theorem* says that if $h_3(\mathbf{w}, \mathbf{y})$ is well defined and finite for all $\mathbf{w} \in \mathscr{W}$ and $\mathbf{y} \in \mathscr{Y}$, then $h_3(\mathbf{w}, \mathbf{y})$ is DTS on $\mathscr{W} \times \mathscr{Y}$. For $S = 1$, the composition theorem is Theorem 3.3 in Hollander, Proschan and Sethuraman (1977), while for $S \geq 1$, it is a special case of Theorem 4.3 in Eaton (1982), which Eaton says is due to Conlon, Leon, Proschan and Sethuraman in an unpublished work.

5. Proof of Theorem 1: An application of Holley's inequality to the Savage lattice. The following proof of Theorem 1 involves little more than applying Holley's inequality to the Savage lattice.

PROOF OF THEOREM 1. Write $\overline{\text{pr}}(\cdot; \mathbf{r})$ for probabilities computed under the uniform distribution on $\mathfrak{F}$ and write $\text{pr}(\cdot; \mathbf{r})$ for the actual probability distribution. To prove unbiasedness, we need to show that $\text{pr}\{t(\mathbf{Z}, \mathbf{r}) \geq c; \mathbf{r}\} \geq \alpha_c$ whenever $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ is DTS, where $\alpha_c = \overline{\text{pr}}\{t(\mathbf{Z}, \mathbf{r}) \geq c; \mathbf{r}\}$. The probability that $t(\mathbf{Z}, \mathbf{r}) \geq c$ is the expectation of the indicator function $[t(\mathbf{Z}, \mathbf{r}) \geq c]$ which equals 1 if $t(\mathbf{Z}, \mathbf{r}) \geq c$ and equals 0 and otherwise. Now, $[t(\mathbf{Z}, \mathbf{r}) \geq c]$ is DTS because $t(\mathbf{Z}, \mathbf{r})$ is DTS, so to prove the result is to show that a particular DTS function, namely $[t(\mathbf{Z}, \mathbf{r}) \geq c]$, has a higher expectation under $\text{pr}(\cdot; \mathbf{r})$ than under $\overline{\text{pr}}(\cdot; \mathbf{r})$. By Holley's inequality, it suffices to show that for all $\mathbf{z}, \mathbf{z}^* \in \mathfrak{F}$,

$$\text{pr}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*; \mathbf{r}) \cdot \overline{\text{pr}}(\mathbf{Z} = \mathbf{z} \wedge \mathbf{z}^*; \mathbf{r}) \geq \text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r}) \cdot \overline{\text{pr}}(\mathbf{Z} = \mathbf{z}^*; \mathbf{r}).$$

But $\overline{\text{pr}}(\mathbf{Z} = \mathbf{z} \wedge \mathbf{z}^*; \mathbf{r}) = \overline{\text{pr}}(\mathbf{Z} = \mathbf{z}^*; \mathbf{r})$, as this distribution is uniform, so it suffices to show that $\text{pr}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*; \mathbf{r}) \geq \text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$. If $\mathbf{z} \vee \mathbf{z}^* = \mathbf{z}$, then $\text{pr}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*; \mathbf{r}) = \text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$, and we are done. On the other hand, if $\mathbf{z} \vee \mathbf{z}^* \neq \mathbf{z}$, then $\mathbf{z} \lesssim \mathbf{z} \vee \mathbf{z}^*$, and there is a sequence $\mathbf{z} = \mathbf{z}_1 \lesssim \mathbf{z}_2 \lesssim \cdots \lesssim \mathbf{z}_J = \mathbf{z} \vee \mathbf{z}^*$, and since $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ is DTS, its values are monotone increasing over this sequence, proving the result. □

6. How do DTS alternatives arise? Distributions $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ on $\mathfrak{F}$ that are DTS arise naturally when, in addition to the observed covariates used to define the subclasses, there is an unobserved *pretreatment* covariate $\mathbf{U}$ which is positively related to both $\mathbf{Z}$ and $\mathbf{r}$. Informally, if there are systematic differences between treated and control subjects in the same subclass with respect to a relevant unobserved covariate $\mathbf{U}$, then the groups were not comparable prior to treatment, and adjustments for the subclasses are insufficient to make them comparable.

Suppose, for instance, that

$$\text{pr}(\mathbf{Z} = \mathbf{z} | \mathbf{U} = \mathbf{u}, \mathbf{r}) = \frac{\exp(\gamma \mathbf{u}^{\mathrm{T}} \mathbf{z})}{\sum_{\mathbf{z}^* \in \mathfrak{F}} \exp(\gamma \mathbf{u}^{\mathrm{T}} \mathbf{z}^*)}, \tag{6.1}$$

where $\gamma \geq 0$. This is the model for $\text{pr}(\mathbf{Z} = \mathbf{z}|\mathbf{U} = \mathbf{u}, \mathbf{r})$ that arises in sensitivity analyses for permutation inferences, as described in Rosenbaum (1987b, 1988b) and Rosenbaum and Krieger (1988). If $\gamma = 0$, then (6.1) reduces to the uniform distribution on $\mathfrak{F}$. If $\gamma > 0$, a subject (s, i) with a higher value of u_{si} is more likely to receive the treatment than another subject (s, j) in the same subclass with a lower value of u_{sj}. In (6.1), the treatment assignment is unrelated to $\mathbf{r}$ among subjects with the same value of the unobserved covariate, so treatment assignment would have been adjustable given both the observed subclasses and $\mathbf{U}$; this says that adjustments would have sufficed had $\mathbf{U}$ been observed and included in the subclassification or matching, or informally that $\mathbf{U}$ is the relevant unobserved covariate. Also, (6.1) is DTS as a function of $(\mathbf{z}, \mathbf{u})$ for $\gamma \geq 0$.

The observed unaffected response $\mathbf{r}$ provides information about imbalances in $\mathbf{U}$ when they are positively related; this is true in two senses. Section 7 considers the conditional power for fixed (but unknown) $\mathbf{U}$, while the proposition immediately below concerns the expected power averaging over a distribution for $\mathbf{U}$. Combined with Theorem 1 above, Proposition 1 states that the unaffected outcome $\mathbf{r}$ provides an unbiased test for imbalances in $\mathbf{U}$ providing $\mathbf{U}$ is positively dependent on $\mathbf{r}$, that is, providing the (continuous or discrete) conditional density, $\text{pr}(\mathbf{U} = \mathbf{u}; \mathbf{r})$, of $\mathbf{U}$ given $\mathbf{r}$ is DTS. [Many familiar models for positive dependence in $\text{pr}(\mathbf{U} = \mathbf{u}; \mathbf{r})$ yield a DTS distribution; for instance, this is true if

$$\text{pr}(\mathbf{U} = \mathbf{u}; \mathbf{r}) = \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(u_{si}|r_{si}),$$

where $f_s(\cdot|\cdot)$ is a TP_2 conditional density for each s.]

PROPOSITION 1. *Under model* (6.1) *with* $\gamma \geq 0$, $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ *is DTS whenever* $\text{pr}(\mathbf{U} = \mathbf{u}; \mathbf{r})$ *is DTS.*

PROOF. From (6.1)

$$\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r}) = \int \text{pr}(\mathbf{Z} = \mathbf{z}|\mathbf{U} = \mathbf{u}) \cdot \text{pr}(\mathbf{U} = \mathbf{u}; \mathbf{r})\, d\mathbf{u},$$

so that the result follows from the composition theorem for DTS functions. □

7. The conditional power given U. Instead of introducing a distribution for $\mathbf{U}$ as in Section 6, we may consider the (conditional) power $\beta(\mathbf{r}, \mathbf{u})$ of the test for hidden bias as a function of $\mathbf{U} = \mathbf{u}$, namely,

$$\beta(\mathbf{r}, \mathbf{u}) = \sum_{\mathbf{z} \in \mathfrak{F}} [t(\mathbf{z}, \mathbf{r}) \geq c] \frac{\exp(\gamma \mathbf{u}^{\mathrm{T}} \mathbf{z})}{\sum_{\mathbf{z}^* \in \mathfrak{F}} \exp(\gamma \mathbf{u}^{\mathrm{T}} \mathbf{z}^*)},$$

where $[t(\mathbf{z}, \mathbf{r}) \geq c]$ equals 1 if $t(\mathbf{z}, \mathbf{r}) \geq c$ and equals 0 otherwise. The following proposition says that the power increases steadily as the ordering of $\mathbf{u}$ and $\mathbf{r}$ becomes more similar within each subclass. [D'Abadie and Proschan (1984) gave the name "isotonic power in the arrangement ordering" to this property of a

power function.] This is the second sense, mentioned in Section 6, in which an unaffected response $\mathbf{r}$ provides information about an unobserved covariate with which it is positively related.

PROPOSITION 2. *If $t(\cdot, \cdot)$ is DTS and $\gamma \geq 0$, then the power $\beta(\mathbf{r}, \mathbf{u})$ is also DTS.*

PROOF. Since $[t(\mathbf{z}, \mathbf{r}) \geq c]$ is DTS because $t(\cdot, \cdot)$ is DTS, the result again follows from the composition theorem for DTS functions. □

8. Ties. In Section 2, ties were assumed absent. Of course, ties will be numerous when the response r_{si} is discrete, and especially so when the response is binary. In fact, ties present no fundamental problem, though they do require a few adjustments.

Large parts of the previous argument are unchanged by the presence of ties. DTS functions are still defined as in Section 2. From this definition, permutations that involve tied responses do not change the value of a DTS function; that is, if $r_{si} = r_{sj}$, then $t(\mathbf{z}, \mathbf{r}) = t\{\mathbf{z}_{(sij)}, \mathbf{r}\}$. Many familiar statistics for discrete scores or binary responses are DTS, including: (i) the rank statistics mentioned in Section 2 with average ranks used in case of ties; (ii) the statistic of Mantel and Haenszel (1959) and Birch (1964) for binary responses; (iii) the statistic of Mantel (1963) and Birch (1965) for discrete scores; (iv) in the case of matched pairs—that is, $n_s = 2$, $a_s = 1$ for each s—the statistic of McNemar (1947) and Cox (1958). Propositions 1 and 2 and their proofs are unchanged.

The one substantive change occurs in connection with the Savage lattice and the relationship between the lattice partial order and the class of DTS functions. As in Section 2, subjects are numbered in decreasing order of their responses, so that $r_{si} \geq r_{sj}$ if $i < j$; however it may now happen that $r_{si} = r_{sj}$ for some $i < j$. (The ordering of tied subjects may be done in any arbitrary way that does not involve $\mathbf{Z}$; for instance, it may be done at random.) The Savage lattice is as before; however, the partial order now creates artificial distinctions or orderings associated with the arbitrary arrangement of ties. The DTS functions are now a proper subset of the functions isotonic in the lattice order—actually the interesting subset, namely the isotonic functions that ignore the arbitrary ordering of tied responses. Theorem 1 is true and its proof unchanged for all isotonic $t(\mathbf{z}, \mathbf{r})$ and $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$; hence, it is true in particular if these functions are DTS. Of course, since the ordering of tied subjects is arbitrary, Theorem 1 is of interest only for DTS functions which ignore the arbitrary ordering.

In short, the results of previous sections are true as stated in the presence of tied responses.

9. Other uses of Theorem 1. Theorem 1 may be applied to problems other than detecting biases in observational studies using an unaffected outcome. To sketch one familiar example, consider the subclassified two sample problem: For $s = 1, \ldots, S$, the bivariate response/sample indicator vectors, (R_{si}, Z_{si}), are independent and identically distributed with (discrete or continuous) conditional

densities $f_s(r|z)$. One sense in which $f_s(r|1)$ might be said to be larger than $f_s(r|0)$ is if $f_s(r|1)/f_s(r|0)$ is monotone increasing in r over the possible values of r, that is, if there is a monotone likelihood ratio (MLR), which implies $f_s(r|z)$ is TP_2, and hence that $\text{pr}(\mathbf{Z} = \mathbf{z}; \mathbf{r})$ is DTS. It follows from Theorem 1 that any DTS statistic $t(\mathbf{Z}, \mathbf{r})$ yields an unbiased test of the hypothesis of no difference within each subclass against MLR alternatives within subclasses. This covers some familiar and some less familiar cases. For instance, if R takes on K ordered values, Theorem 1 says that the Mantel (1963)–Birch (1965, Section 5), test for $2 \times K \times S$ contingency tables is unbiased against MLR alternatives, providing the scores attached to the K columns of each of the S $2 \times K$ tables are monotone.

REFERENCES

AHLSWEDE, R. and DAYKIN, D. (1978). An inequality for weights of two families of sets, their unions and intersections. *Z. Wahrsch. verw. Gebiete* **43** 183–185.

AIGNER, M. (1979). *Combinatorial Theory*. Springer, New York.

BIRCH, M. W. (1964). The detection of partial association. I. The 2 × 2 case. *J. Roy. Statist. Soc. Ser. B* **26** 313–324.

BIRCH, M. W. (1965). The detection of partial association. II. The general case. *J. Roy. Statist. Soc. Ser. B* **27** 111–124.

BOLLOBÁS, B. (1986). *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability*. Cambridge Univ. Press, New York.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 134–155.

COX, D. R. (1958). Two further applications of a model for binary regression. *Biometrika* **45** 562–564.

D'ABADIE, C. and PROSCHAN, F. (1984). Stochastic versions of rearrangement inequalities. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) 4–12. IMS, Hayward, Calif.

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

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

HOLLANDER, M., PROSCHAN, F. and SETHURAMAN, J. (1977). Functions decreasing in transposition and their applications in ranking problems. *Ann. Statist.* **5** 722–733.

HOLLEY, R. (1974). Remarks on the FKG inequalities. *Comm. Math. Phys.* **36** 227–231.

MANTEL, N. (1963). Chi-square tests with one degree of freedom. Extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Assoc.* **58** 690–700.

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22** 719–748.

MARSHALL, A. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic, New York.

MCNEMAR, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* **12** 153–157.

ROSENBAUM, P. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Amer. Statist. Assoc.* **79** 41–48.

ROSENBAUM, P. R. (1984b). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574.

ROSENBAUM, P. R. (1987a). The role of a second control group in an observational study (with discussion). *Statist. Sci.* **2** 292–316.

ROSENBAUM, P. R. (1987b). Sensitivity analysis for certain permutation tests in matched observational studies. *Biometrika* **74** 13–26.

ROSENBAUM, P. R. (1988a). Permutation tests for matched pairs with adjustments for covariates. *Appl. Statist.* **37** 401–411.

ROSENBAUM, P. R. (1988b). Sensitivity analysis for matching with multiple controls. *Biometrika* **75** 577–581.

ROSENBAUM, P. R. and KRIEGER, A. (1988). Sensitivity analysis for subclassified permutation inferences. Unpublished manuscript.

ROSENBAUM, P. and RUBIN, D. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

SAVAGE, I. R. (1964). Contributions to the theory of rank order statistics: Applications of lattice theory. *Rev. Internat. Statist. Inst.* **32** 52–63.

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

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

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302