

# Stability in the Absence of Treatment

Paul R. ROSENBAUM

---

When subjects are measured twice, once at each of two symmetrical locations or times, stability of responses in the absence of treatment within subjects, together with comparability of untreated responses between subjects, is often viewed as supporting a conclusion that differences between treated and control responses reflect effects actually caused by the treatment. The degree to which this intuitive argument is formally correct is explored in several related models: a multivariate Normal model, a nonparametric model defined by symmetries, an analogous randomized experiment, and a sensitivity analysis model for observational studies in which treatments are not randomly assigned to subjects, nor to locations within subjects. Card and Krueger's study of the employment effects of the minimum wage is used to illustrate the methods.

KEY WORDS: Aligned rank test; Hodges–Lehmann estimate; Observational studies; Paired data; Quasi experiment; Rank test; Sensitivity analysis; Separability.

---

## 1. STABILITY IN THE STUDY OF TREATMENT EFFECTS

### 1.1 Stability, Comparability, and Treatment Effects

Experimental subjects present two symmetrical locations, either of which may be subjected to treatment, such as two ears or two asthma attacks separated by several weeks. In addition, subjects are paired using pretreatment covariates so that, before treatment, paired subjects appeared comparable in terms of observed covariates. In each pair, one subject is given the treatment at one of the two locations; the other subject is not treated at both locations; and four outcome measures are obtained, one for each subject at each location. This is the simplest design for a study of treatment effects that uses both stability of responses within control subjects and comparability of control responses between subjects.

Suppose elevated responses are typically observed at the treated location for the treated subject, and much lower, very similar typical responses are observed at the other three locations, that is, at the control location for the treated subject and the two control locations for the control subject. Informally and intuitively, this pattern seems to support a claim that the treatment caused the elevated response; however, it is easy to think of other ways the same pattern could be produced. The intuition seems to depend on four issues:

- (1) anticipation that the two locations would be symmetrical in the absence of treatment,
- (2) confirming observation that, in fact, the two responses for the control subject are similar,
- (3) anticipation that the matched subjects are comparable, and
- (4) confirming observation that, in fact, the response for the untreated location for the treated subject is similar to the two outcomes for the control subject.

In what sense, if any, and to what quantitative extent is this intuition formally correct?

This intuition may be convincing or not, depending on whether an alternative explanation is plausible. For instance,

the results would be convincing if there were many pairs, and if randomization had been used twice in each pair, once to pick the treated subject, and again to pick the treated location for that subject. In the most common application of this design, the so-called “control group design with pretest and posttest” or CP design, randomization is not used to assign subjects to treatment or control, and the untreated response for the treated subject is always the first measurement, that is, a baseline measurement. Cook and Campbell (1979) describe the CP design as “the most frequently used design in social science research.” They say that the CP design is “often interpretable,” but they note it is sometimes open to several interpretations besides a treatment effect, including “selection–maturation” in which the treated and control subjects are maturing or changing in different ways. Despite this weakness of the CP design, in the social sciences the CP design is typically viewed as much better than a design lacking either a control group or a baseline measure of the outcome. Aspects of the CP design and related designs are discussed by Koch (1972), Reichardt (1979), Kershner and Federer (1981), Holland and Rubin (1983), Ashenfelter and Card (1985), Allison (1990), Cook, Campbell, and Peracchio (1990), Meyer (1995), Angrist and Krueger (1999), and Salzberg (1999).

### 1.2 Temporal Symmetry?

Time has an inherent direction and an inherent asymmetry. Growth, learning, aging, recuperation, disease progression, deterioration, and many other natural processes work against stability over time. Where these processes are dramatically at work, stability over time cannot be exploited to strengthen causal inferences in observational studies.

And yet, many topics are of interest precisely because of their resistance to change—for instance, drug or alcohol addiction, chronic unemployment, personality disorders. In such cases, stability in the absence of treatment combined with change following treatment does seem, intuitively, to strengthen claims that the treatment is the cause of the change.

Moreover, many natural processes at work over time have gradual consequences, so dramatic change over short time intervals is not anticipated. If a treatment is imposed with sudden intensity, immediate change following treatment,

---

Paul R. Rosenbaum is Professor in the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. (E-mail: [rosenbaum@stat.wharton.upenn.edu](mailto:rosenbaum@stat.wharton.upenn.edu)). This work was supported by a grant from the Methodology, Measurement, and Statistics Program and the Statistics and Probability Program of the National Science Foundation. The hospitality and support of the Center for Advanced Study in the Behavioral Sciences are gratefully acknowledged.

together with stability in the absence of treatment, is not easily explained by gradual evolution.

In short, temporal stability in the absence of treatment is anticipated in some contexts and not in others, and hence is relevant to some observational studies and not to others. Even in studies that anticipate stability, and even when simple departures from stability are tested and not rejected, it is typically necessary to examine the degree to which instability might explain ostensible treatment effects; see Section 4 where sensitivity analyses address this possibility.

### 1.3 Example: Minimum Wage and Employment

Card and Krueger (1994, 1995) studied the effects on employment of increasing the minimum wage. In April 1992, New Jersey raised its state minimum wage from \$4.25 to \$5.05 per hour, approximately 19%. As their outcome, they looked at full-time equivalent employment in fast-food restaurants such as Burger King and Wendy's in February and March 1992 before the increase, and again in November and December 1992 after the increase. They compared restaurants in New Jersey to similar restaurants in adjacent eastern Pennsylvania, where the minimum wage had not increased. See Rosenbaum (1999b, Table 1) for before/after data on 66 such pairs of restaurants from the Card and Krueger study, matched for two baseline covariates, restaurant chain and starting wage, before the increase.

Suppose that restaurants such as Roy Rogers employed similar numbers of workers in New Jersey and eastern Pennsylvania when the two states had the same minimum wage, and that employment stayed stable in Pennsylvania and declined in New Jersey. If this multivariate pattern of responses was found, then because it is precisely the pattern predicted by economic theory in the absence of major disturbances, one senses that the pattern should strengthen evidence that the wage increase was the cause of its anticipated effect. This sense is not consistent with the most common method of analysis but is true of the new method proposed in this article. Actually, however, Card and Krueger found evidence of stability and comparability in the absence of treatment, but "no evidence that the rise in New Jersey's minimum wage reduced employment at fast-food restaurants in the state." The Card and Krueger data is examined later.

## 2. INFERENCE IN EXPERIMENTS

### 2.1 Randomization Inference

*2.1.1 Paired Experiment with Twin Locations.* Dorn (1953), Cochran (1965), and Rubin (1974, 1977) have argued that consideration of an observational study should begin by considering its similarities and differences from an analogous randomized experiment. Here, the goal is a randomized experiment that accords special roles to both baseline measures of response and also to stability in the absence of treatment. Then, Section 3 includes discussion of observational studies with the same structure, but where random assignment was not used.

Each experimental subject presents two more or less symmetrical locations or opportunities to experiment,  $l = 0, 1$ .

Subjects are paired based on observed pretreatment variables or covariates, and one subject in each pair is randomly selected to receive the treatment at one randomly selected location, the other subject remaining untreated. There are  $I$  pairs,  $i = 1, \dots, I$ , of two units,  $j = 1, 2$ , where  $Z_{ij} = 1$  for the treated unit and  $Z_{ij} = 0$  for the control, so  $1 = Z_{i1} + Z_{i2}$  for  $i = 1, \dots, I$ . Write  $\mathbf{Z} = (Z_{11}, Z_{12}, Z_{21}, \dots, Z_{I1}, Z_{I2})^T$  for the  $2I$  dimensional vector of treatment assignments. If the  $j$ th unit in pair  $i$  received the treatment, write  $V_{il} = 1$  if location  $l$  for this unit was treated, so  $1 = V_{i0} + V_{i1}$ . Let  $\mathbf{V} = (V_{10}, V_{11}, V_{20}, \dots, V_{I0}, V_{I1})^T$  for the  $2I$  dimensional vector of treatment locations. The pair  $(\mathbf{Z}, \mathbf{V})$  defines treatment assignment and location for all  $I$  pairs. Within each pair, there are 4 possible treatment/location assignments, so there are  $4^I$  possible assignments in total.

Write  $\Omega$  for the set of  $4^I$  possible values of  $(\mathbf{Z}, \mathbf{V})$ . If treatments are assigned by independent flips of fair coins, so that  $\text{prob}(Z_{ij} = 1) = \frac{1}{2}$  and  $\text{prob}(V_{il} = 1 | Z_{ij} = 1) = \frac{1}{2}$ , then each element of  $\Omega$  has the same chance, namely  $1/4^I$ , of being the actual assignment, and this is defined to be a paired randomized experiment with twin locations.

At each location  $l$ , the  $j$ th subject in pair  $i$  exhibits a response,  $Y_{ijl}$ . The model of an additive treatment effect asserts that this location would exhibit a response  $y_{ijl}$  if assigned to control, which would be increased by  $\tau$  if assigned to treatment, so that the observed response is  $Y_{ijl} = y_{ijl} + \tau Z_{ij} V_{il}$ . In randomization inference, probability enters only through the random assignment of treatments  $(\mathbf{Z}, \mathbf{V})$  so that quantities like the observed responses,  $Y_{ijl}$ , which depend on  $Z_{ij} V_{il}$ , are random variables, whereas quantities like the potential responses under control,  $y_{ijl}$ , which do not change with the treatment assignment, are fixed features of the finite population of  $2I$  subjects (Fisher 1935).

*2.1.2 A Statistic and Its Randomization Distribution.* For the  $i$ th pair, for the  $j$ th subject in the pair, location  $l$  for this subject, there is a fixed score  $s_{ijl}$ . Consider the following statistic,  $T = \sum_{i=1}^I \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} V_{il} s_{ijl}$ , which is the sum of the scores for the treated locations. The randomization creates a known permutation distribution for  $T$  with expectation  $E(T) = \sum_{i=1}^I \bar{s}_i$  and variance  $\text{var}(T) = \frac{1}{4} \sum_{i=1}^I \sum_{j=1}^2 \sum_{l=0}^1 (s_{ijl} - \bar{s}_i)^2$ , where  $\bar{s}_i = \frac{1}{4} \sum_{j=1}^2 \sum_{l=0}^1 s_{ijl}$ .

Consider testing the null hypothesis,  $H_0 : \tau = \tau_0$ . Under the null hypothesis, the adjusted responses  $Y_{ijl} - \tau_0 Z_{ij} V_{il} = y_{ijl}$  are fixed, not varying with the treatment assignment, so quantities computed from the adjusted responses are also fixed. For instance, if the scores  $s_{ijl}$  are functions of the adjusted responses, then under the null hypothesis the scores  $s_{ijl}$  are fixed, and  $T$  may be compared to its permutation distribution to test the null hypothesis. From the hypothesis test, a confidence interval is derived by inverting the test (Bauer 1972; Hajek, Sidak, and Sen 1999, Section 9.1; Lehmann 1963; Moses 1965), and a point estimate is obtained by equating  $T$  to its null expectation and solving for the point estimate (Hajek, Sidak, and Sen 1999, Section 9.1; Hodges and Lehmann 1963).

Tukey (1986) argued that randomization inference should be used to ensure "validity" of an inference, for instance, that a test has its nominal level under the null hypothesis, but that

several distributional models should be used in an effort to obtain “stringency” of an inference; for instance, good power when any of the several distributional models are approximately correct. Speaking informally, the decision to compare  $T$  to its randomization or permutation distribution is necessitated by the requirement of validity, but the choice of score function,  $s_{ijl}$ , can be guided by distributional models with a view to stringency.

The particular scores  $s_{ijl}$  that I propose for use in  $T$  are described next. The remainder of Section 2 will motivate this choice of scores with reference to several distributional models, both parametric and nonparametric. In particular, the proposed scores are aligned rank analogs of a maximum likelihood estimate under a multivariate normal model. The aligned rank analog is seen to perform well when the data are multivariate normal, but also when the data are multivariate Cauchy, whereas the normal maximum likelihood estimate performs poorly for the Cauchy. Moreover, if there is stability in the absence of treatment, the aligned rank estimate is much better than a more common difference-in-differences estimator that makes no use of stability. These distributional models serve precisely one purpose in this paper: They guide the choice of scores,  $s_{ijl}$ . The distributional models play no role in formal inferences, along the lines advocated by Tukey (1986).

The proposed scores  $s_{ijl}$  are now be described in a computational fashion, with motivation postponed to the later subsections of Section 2. The scores involve a weight,  $w$ , whose choice is discussed later, although  $w = \frac{2}{5}$  will later turn out to be a sturdy compromise between conflicting objectives. The goal is to test the null hypothesis  $H_0: \tau = \tau_0$ . Under the null hypothesis, compute first the fixed, adjusted responses  $Y_{ijl} - \tau_0 Z_{ij} V_{il} = y_{ijl}$ . Then, align the four adjusted responses in each pair by subtracting their mean within that pair,

$$y_{ijl} - \frac{1}{4} \sum_{a=1}^2 \sum_{b=0}^1 y_{iab} \\ = (Y_{ijl} - \tau_0 Z_{ij} V_{il}) - \frac{1}{4} \sum_{a=1}^2 \sum_{b=0}^1 (Y_{iab} - \tau_0 Z_{ia} V_{ib}),$$

and rank these aligned, adjusted responses from 1 to  $4I$  using average ranks for ties, and write  $q_{ijl}$  for the rank. For the  $i$ th pair, the score  $s_{ijl}$  that enters  $T$  when the  $j$ th subject is treated at location  $l$  compares the rank of the response of this subject at this location to a weighted average of the ranks for the same subject at the untreated location and the average of the two ranks for the paired, untreated subject. Specifically, let  $s_{i11} = q_{i11} - wq_{i10} - \frac{1}{2}(1-w)(q_{i21} + q_{i20})$ ,  $s_{i10} = q_{i10} - wq_{i11} - \frac{1}{2}(1-w)(q_{i21} + q_{i20})$ ,  $s_{i21} = q_{i21} - wq_{i20} - \frac{1}{2}(1-w)(q_{i11} + q_{i10})$ , and  $s_{i20} = q_{i20} - wq_{i21} - \frac{1}{2}(1-w)(q_{i11} + q_{i10})$ . The proposal is to use  $T$  with this choice of scores to test  $H_0: \tau = \tau_0$ , with confidence intervals and point estimates derived from the test.

## 2.2 A Nonparametric Model Giving Rise to the Same Permutation Distribution

As noted in Section 2.1.2, randomization inference is commonly described both in terms of randomization inference for a finite population, with reference to validity, and in terms of permutation inference for a distributional model,

with reference to stringency; see the alternating chapters of Lehmann (1998) for several standard cases. The experiment in Section 2.1.1 is not a standard case, but nonetheless, a certain population model gives rise to the same permutation distribution as the one obtained from randomization inference in Section 2.1. The model accounts for differences between pairs through additive pair effects,  $\mu_i$ , and exhibits symmetry of the two locations for each person in a pair. The model describes the response  $y_{ijl}$  that would be observed under control, which is increased by  $\tau$  if treatment is given, so the observed response is  $Y_{ijl} = y_{ijl} + \tau_0 Z_{ij} V_{il}$ .

Let  $H(\cdot, \cdot, \cdot, \cdot)$  be a continuous, four dimensional distribution which has a special type of invariance or exchangeability, namely:  $H(a, b, c, d) = H(b, a, c, d) = H(a, b, d, c) = H(c, d, a, b)$  for all  $a, b, c, d$ . In words,  $H(a, b, c, d)$  is unchanged by interchanging the first two coordinates, or the second two coordinates, or by swapping the first two coordinates for the second two coordinates. Note carefully that in general  $H(a, b, c, d) \neq H(a, c, b, d)$ , i.e., that the distribution is not invariant with respect to all permutations. Informally, the two measures on one subject may be more strongly related than the measures on different subjects in the same pair. The distribution  $H(\cdot, \cdot, \cdot, \cdot)$  is invariant with respect to a group  $\mathcal{G}$  of eight permutations of four objects, which is a subgroup of all  $4! = 24$  permutations of four objects (specifically  $\mathcal{G}$  is a wreath product of symmetric groups acting on just two objects). The model asserts that the four-dimensional vector  $(y_{i11} - \mu_i, y_{i10} - \mu_i, y_{i21} - \mu_i, y_{i20} - \mu_i)$  describing centered responses under control for the  $I$  matched pairs,  $i = 1, \dots, I$ , are independent and identically distributed (iid) with distribution  $H(\cdot, \cdot, \cdot, \cdot)$ . Because of the iid sampling of matched sets, the joint distribution of the  $4I$  centered responses is invariant with respect to a group of  $I! \cdot 8^I$  permutations. Nonparametric models with related symmetries have been used for other purposes by Bell and Haller (1969) and Wei (1987).

Lehmann and Stein (1949, Lemma 2) show that an optimal test of a hypothesis of distributional invariance is necessarily a permutation test, i.e., a test that permutes the observed responses in accord with the hypothesized invariance, rejecting the null hypothesis at the 5% level for 5% of these permutations. The hypothesis of invariance of the adjusted responses,  $y_{ijl} = Y_{ijl} - \tau_0 Z_{ij} V_{il}$ , leads to the same permutation distribution as the one obtained by random assignment in Section 2.1. In this sense, population models of the form in this section are the analogs of the randomized experiment in Section 2.1.

Notice that the invariance of  $H(\cdot, \cdot, \cdot, \cdot)$  under  $\mathcal{G}$  is not altered by monotone transformations of  $y_{ijl}$ . In this article, the treatment effect is modelled as an additive constant  $\tau$ ; however, similar considerations would apply to other models for treatment effect, such as the model of a dilated effect (Rosenbaum 1999c).

## 2.3 A Normal Model for Stability and Comparability

**2.3.1 Model and MLE.** Insight is provided by the following multivariate normal version of the general nonparametric model  $H(\cdot, \cdot, \cdot, \cdot)$  in Section 2.2. The model for the centered responses under control,  $(y_{i11} - \mu_i, y_{i10} - \mu_i, y_{i21} - \mu_i,$

$y_{i20} - \mu_i$ ), is:

$$\begin{bmatrix} y_{i11} - \mu_i \\ y_{i10} - \mu_i \\ y_{i21} - \mu_i \\ y_{i20} - \mu_i \end{bmatrix} \sim N \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \nu & \rho\nu & 0 & 0 \\ \rho\nu & \nu & 0 & 0 \\ 0 & 0 & \nu & \rho\nu \\ 0 & 0 & \rho\nu & \nu \end{bmatrix} \right\}. \quad (1)$$

Control responses  $y_{ijl}$  in pair  $i$  are similar because of  $\mu_i$ , but beyond this, the two responses ( $y_{ij1}, y_{ij0}$ ) of the same person are correlated. Treatment increases  $y_{ijl}$  by  $\tau$ . Write  $\tilde{Y}_{iT1}$  for the treated response of the treated subject, i.e.,  $\tilde{Y}_{iT1} = \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} V_{il} Y_{ijl} = \tau + \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} V_{il} y_{ijl}$ . Similarly, write  $\tilde{Y}_{iT0}$  for the control response of the treated subject,  $\tilde{Y}_{iT0} = \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} (1 - V_{il}) y_{ijl}$ , and write  $\tilde{Y}_{iC1}$  and  $\tilde{Y}_{iC0}$  for the parallel control responses of the control subject,  $\tilde{Y}_{iC1} = \sum_{j=1}^2 \sum_{l=0}^1 (1 - Z_{ij}) V_{il} y_{ijl}$  and  $\tilde{Y}_{iC0} = \sum_{j=1}^2 \sum_{l=0}^1 (1 - Z_{ij}) (1 - V_{il}) y_{ijl}$ .

Estimate the treatment effect  $\tau$  by  $\hat{\tau}_w = (1/I) \sum_{i=1}^I \{ \tilde{Y}_{iT1} - w \tilde{Y}_{iT0} - (1-w)(\tilde{Y}_{iC1} + \tilde{Y}_{iC0})/2 \}$ , which, for all  $w$ , is free of the nuisance parameters  $\mu_i$  and is unbiased,  $E(\hat{\tau}_w) = \tau$ . The estimator  $\hat{\tau}_w$  has variance  $\nu\{(3 + \rho)(1 + w^2) - 2(1 + 3\rho)w\}/(2I)$ , so  $\hat{\tau}_w$  is consistent as  $I \rightarrow \infty$ . If  $w = (3\rho + 1)/(3 + \rho)$ , then straightforward manipulations show  $\hat{\tau}_w$  is the maximum likelihood estimate of  $\tau$ .

Table 1 gives the form of the maximum likelihood estimate for several values of  $\rho$ . As intuition suggests, if  $\rho = 0$ , then the one treated measurement is compared with the average of the three untreated measurements in the pair. If  $\rho > 0$ , then measurements on the same person are positively correlated, and greater weight is given to the treated subject's own control measurement than to a control measurement from a control subject. For instance, with  $\rho = \frac{1}{5}$ ,  $\tilde{Y}_{iT0}$  receives twice the weight that a measurement from the control subject receives. Conversely, when  $\rho = -\frac{1}{3}$ , the treated subject's own control measurement is ignored. The limiting cases,  $\rho = -1$  and  $\rho = 1$ , are intuitive in form but not relevant in actual practice.

In practice, the choice of  $w$  is not the same as estimating  $\rho$ , for several reasons. First, estimators using means, like  $\hat{\tau}_w$ , may exhibit Gaussian behavior with non-Gaussian distributions because of the central limit theorem, whereas estimators that use second moments, like sample correlations, may be much more dependent on Gaussian assumptions; see Scheffe (1959, Section 10.2). Second, even if data were precisely multivariate normal, the choice  $w = (3\rho + 1)/(3 + \rho)$  for the MLE

will largely ignore the responses of control subjects when  $\rho$  is near 1, and this is likely to seem inappropriate even if it is efficient in a purely technical sense. For example, with  $\rho = .8$ , the weight  $w$  is almost .9, so that  $\tilde{Y}_{iT0}$  receives about 18 times as much weight as the control's response  $\tilde{Y}_{iC1}$  at the treated location. In practice, one will not want to ignore any of the three untreated responses. The weight  $w = \frac{3}{5}$  strikes a balance, because one is unlikely to use a symmetric location if the correlation  $\rho$  is much less than  $\frac{1}{3}$ , but  $w = \frac{3}{5}$  gives only a little more weight to  $\tilde{Y}_{iT0}$  than to  $\frac{1}{2}(\tilde{Y}_{iC1} + \tilde{Y}_{iC0})$ . All of the calculations in this article suggest that medium-sized discrepancies between  $w$  and  $\frac{3\rho+1}{3+\rho}$  have only slight impact on the behavior of procedures. In those rare instances where  $\rho$  is near 1, the differences  $\tilde{Y}_{iT1} - \tilde{Y}_{iT0}$  and  $\tilde{Y}_{iC1} - \tilde{Y}_{iC0}$  will have small variances, so one may wish to refrain from using  $\hat{\tau}_w$  altogether, replacing it with the somewhat less efficient estimate  $D$  described next.

### 2.3.2 A Familiar, Unbiased, but Inefficient Estimate.

More familiar than  $\hat{\tau}_w$  is the estimator that compares the mean change among the treated subjects to the mean change among the controls, namely  $D = (1/I) \sum_{i=1}^I (\tilde{Y}_{iT1} - \tilde{Y}_{iT0}) - (\tilde{Y}_{iC1} - \tilde{Y}_{iC0})$ , which is also unbiased,  $E(D) = \tau$  with variance  $\text{var}(D) = 4\nu(1 - \rho)/I$ . Table 1 also gives  $\text{var}(D)/\text{var}(\hat{\tau}_w) = (3 + \rho)/(1 + \rho)$  for the maximum likelihood estimate with  $w = (3\rho + 1)/(3 + \rho)$ , showing that  $D$  has much larger variance. So the familiar estimator,  $D$ , looks quite poor in comparison with the maximum likelihood estimate. This comparison is not entirely fair, in two senses. First, the true correlation,  $\rho$ , is unknown. To address this, Table 2 compares  $\hat{\tau}_{3/5}$  and  $D$  for several values of the true  $\rho$ ; that is, the maximum likelihood estimate when  $\rho = \frac{1}{3}$  is used incorrectly when in fact  $\rho$  is not  $\frac{1}{3}$ , and  $\text{var}(D)/\text{var}(\hat{\tau}_{3/5}) = 25(1 - \rho)/(9 - 7\rho)$  is tabled. For  $\frac{2}{3} \geq \rho \geq 0$ , the MLE at  $\rho = \frac{1}{3}$ , namely  $\hat{\tau}_{3/5}$ , is much better than  $D$ , suggesting that precise knowledge of  $\rho$  is not critical for good performance of the maximum likelihood estimate relative to  $D$ . Notice that  $D$  is better than  $\hat{\tau}_{3/5}$  but not better than the MLE  $\hat{\tau}_w$  with  $w = (3\rho + 1)/(3 + \rho)$  when  $\rho$  is near 1.

The second sense in which Table 1 is somewhat unfair to the familiar estimator,  $D$ , is that  $D$  is sometimes unbiased for  $\tau$  when stability and comparability do not hold, so that  $\hat{\tau}_w$  is not unbiased. This is discussed in detail in Section 2.4.

### 2.3.3 A Small Simulation Comparing the Methods.

For testing  $H_0 : \tau = 0$ , the aligned rank test statistic  $T$  proposed in §2.1.2 would equal  $I$  times the Normal MLE  $\hat{\tau}_w$  if instead of using integer ranks from 1 to  $4I$  one used the observations  $Y_{ijl}$  as the rank scores. Table 3 reports a small simulation comparing three tests, namely the conventional normal-theory z-test based on  $D$ , the test based on  $\hat{\tau}_{3/5}$  and the aligned rank test again using  $w = \frac{3}{5}$ . Recall that  $w = \frac{3}{5}$  is best when  $\rho = \frac{1}{3}$ ; how-

Table 1. The MLE  $\hat{\tau}_w$  for Several  $\rho$

$\rho$	$\hat{\tau}_w$	$\frac{\text{var}(D)}{\text{var}(\hat{\tau}_w)}$
-1	$\frac{1}{7} \sum_{i=1}^I (\tilde{Y}_{iT1} + \tilde{Y}_{iT0}) - (\tilde{Y}_{iC1} + \tilde{Y}_{iC0})$	$\infty$
$-\frac{1}{3}$	$\frac{1}{7} \sum_{i=1}^I \tilde{Y}_{iT1} - \frac{1}{2}(\tilde{Y}_{iC1} + \tilde{Y}_{iC0})$	4
0	$\frac{1}{7} \sum_{i=1}^I \tilde{Y}_{iT1} - \frac{1}{3}(\tilde{Y}_{iT0} + \tilde{Y}_{iC1} + \tilde{Y}_{iC0})$	3
$\frac{1}{5}$	$\frac{1}{7} \sum_{i=1}^I \tilde{Y}_{iT1} - \frac{1}{4}(2\tilde{Y}_{iT0} + \tilde{Y}_{iC1} + \tilde{Y}_{iC0})$	$\frac{8}{3}$
$\frac{1}{3}$	$\frac{1}{7} \sum_{i=1}^I \tilde{Y}_{iT1} - \frac{1}{5}(3\tilde{Y}_{iT0} + \tilde{Y}_{iC1} + \tilde{Y}_{iC0})$	$\frac{5}{2}$
1	$\frac{1}{7} \sum_{i=1}^I \tilde{Y}_{iT1} - \tilde{Y}_{iT0}$	2

Table 2. Comparison of  $\hat{\tau}_{3/5}$  and  $D$  for Varied  $\rho$

$\rho$	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	.8889	.90	1
$\frac{\text{var}(D)}{\text{var}(\hat{\tau}_{3/5})}$	$\frac{25}{9}$	$\frac{5}{2}$	$\frac{25}{11}$	$\frac{25}{13}$	1	$\frac{25}{27}$	0

Table 3. Simulations Comparing Three Tests With  $w = \frac{3}{5}$ 

Situation	Test	Deviate > 1.65	Deviate > 1.96
Normal $\tau = 0, \rho = \frac{1}{3}$	$\hat{\tau}_{3/5}$	.050	.023
	$D$	.044	.023
	Rank	.055	.028
Normal $\tau = 0, \rho = \frac{2}{3}$	$\hat{\tau}_{3/5}$	.048	.023
	$D$	.052	.025
	Rank	.046	.023
Normal $\tau = \frac{1}{2}, \rho = \frac{1}{3}$	$\hat{\tau}_{3/5}$	.70	.59
	$D$	.40	.28
	Rank	.67	.55
Normal $\tau = \frac{1}{2}, \rho = \frac{2}{3}$	$\hat{\tau}_{3/5}$	.84	.76
	$D$	.61	.48
	Rank	.82	.72
Normal $\tau = \frac{1}{2}, \rho = \frac{4}{5}$	$\hat{\tau}_{3/5}$	.92	.86
	$D$	.80	.70
	Rank	.89	.81
Cauchy $\tau = 0, \rho = \frac{1}{3}$	$D$ -rank	.051	.026
	Rank	.049	.024
Cauchy $\tau = 1, \rho = \frac{1}{3}$	$D$ -rank	.42	.30
	Rank	.70	.59

ever, other values of  $\rho$  are also considered in Table 3. When the distribution is Cauchy,  $\hat{\tau}_{3/5}$  and  $D$  are neither useful nor well-defined, so they are not reported, and the signed rank statistic analogous to  $D$ , labeled  $D$ -rank, is reported instead. Since all of the tests eliminate  $\mu_i$  by differencing, the  $\mu_i$  are not used in the simulation.

Two distributions are considered in Table 3. First is the four-variate normal distribution. The second distribution is a four-variate Cauchy distribution derived from the first distribution. Specifically, a four-variate normal observation with  $\mu_i = 0$  and  $\tau = 0$  is sampled and is divided by a single independent observation from the standard normal distribution. See Johnson and Kotz (1972, Section 37.3, p. 134) for discussion of this distribution, where it is the special case of a multivariate  $t$  distribution having one degree of freedom.

The simulations were based on 3,000 independent samples of size  $I = 20$  pairs with two subjects and four locations in each pair. Standardized deviates were calculated for each test, using the true  $\rho$  for the Gaussian tests and the permutation distribution for the aligned rank test and signed rank test. The proportions of standardized deviates above 1.65 and 1.96 are reported. When  $\tau = 0$ , these proportions estimate the actual level of tests that are attempting to have levels of .05 and .025. When  $\tau > 0$ , these proportions estimate the power of the tests.

In the normal cases considered, the aligned rank test has power estimated to be just slightly below that of the normal theory test based on  $\hat{\tau}_{3/5}$ , but both of these tests have much higher power than the conventional test based on  $D$ . This is true even when  $\rho$  is far from  $\frac{1}{3}$  so  $w = \frac{3}{5}$  is a poor choice of weight. In the Cauchy cases considered, the normal tests are not useful, but the aligned rank test analogous to  $\hat{\tau}_{3/5}$  is estimated to have much higher power than the signed rank test analogous to  $D$ , although both tests are estimated to have the correct level.

In short, in the cases considered in Table 3, the aligned rank test was never substantially inferior to any of the other tests, whereas each of the other tests was substantially inferior to the aligned rank test in at least one case. Table 3 makes two points, one familiar, the other less so. The familiar point is

that rank procedures can often be designed to perform well for normal data and to perform much better than least squares estimates when data are long-tailed. The less familiar point is that methods that exploit stability and comparability can have much higher power than more conventional methods which do not.

#### 2.4 Simple Violations of Stability and Comparability

The familiar estimate  $D$  is less efficient than the maximum likelihood estimate  $\hat{\tau}_w$ ; however,  $D$  remains unbiased for  $\tau$  under a simple violation of stability and comparability, while  $\hat{\tau}_w$  becomes biased. Suppose the centered responses  $\tilde{Y}_{iT1} - (\mu_i + \tau + \zeta + \nu)$ ,  $\tilde{Y}_{iT0} - (\mu_i + \zeta)$ ,  $\tilde{Y}_{iC1} - (\mu_i + \nu)$ ,  $\tilde{Y}_{iC0} - \mu_i$ , have a distribution  $H(\cdot, \cdot, \cdot, \cdot)$  with the symmetries described in Section 2.2, so there is a constant bias  $\zeta$  due to group (e.g., New Jersey vs. Pennsylvania) and a constant bias  $\nu$  due to location (e.g., later time vs. earlier time). Then,  $D$  is unbiased for  $\tau$  but  $\hat{\tau}_w$  is not.

With sufficiently large sample sizes, it is straightforward to detect such additive biases,  $\zeta$  and  $\nu$ . For each pair  $i$ , set aside the treated response  $\tilde{Y}_{iT1}$ , and compute the bivariate difference between the control response of the treated subject,  $\tilde{Y}_{iT0}$ , and the two control responses of the control subject,  $\tilde{Y}_{iC1}$ ,  $\tilde{Y}_{iC0}$ . Under the nonparametric model of stability and comparability in Section 2.2 in which  $H(\cdot, \cdot, \cdot, \cdot)$  invariant under  $\mathcal{G}$ , the bivariate observations  $(\tilde{Y}_{iT0} - \tilde{Y}_{iC1}, \tilde{Y}_{iT0} - \tilde{Y}_{iC0})$  are independent and identically distributed for  $i = 1, \dots, I$  with marginal median vector  $(0, 0)$ . In contrast, with additive biases,  $\tilde{Y}_{iT0} - \tilde{Y}_{iC1}$  is symmetric about  $\zeta - \nu$  and  $\tilde{Y}_{iT0} - \tilde{Y}_{iC0}$  symmetric about  $\zeta$ . If either  $\zeta \neq 0$  or  $\nu \neq 0$ , then  $(\zeta - \nu, \zeta) \neq (0, 0)$ , in which case,  $(\tilde{Y}_{iT0} - \tilde{Y}_{iC1}, \tilde{Y}_{iT0} - \tilde{Y}_{iC0})$  does not have marginal median vector  $(0, 0)$ . Chatterjee (1966) developed a consistent and unbiased test of the hypothesis that iid bivariate observations have marginal median vector  $(0, 0)$  against the alternative that the marginal median vector is not  $(0, 0)$ , so this provides a consistent and unbiased test of stability and comparability against the alternative of additive biases.

As an illustration, consider applying Chatterjee's test to Card and Krueger's data as recorded in Table 1 of Rosenbaum (1999b). The differences  $(\tilde{Y}_{iT0} - \tilde{Y}_{iC1}, \tilde{Y}_{iT0} - \tilde{Y}_{iC0})$  compare the baseline employment in New Jersey to the two employment measures in Pennsylvania. Counting the patterns of pairs with no zero differences gives 14 pairs with sign pattern  $(-, -)$ , 15 pairs with sign pattern  $(+, +)$ , or  $29 = 14 + 15$  concordant pairs in total, and 13 pairs with sign pattern  $(-, +)$  and 9 with pattern  $(+, -)$ , or  $22 = 13 + 9$  discordant pairs in total. Chatterjee's (1966) exact test compares 14 of 29 and 13 of 22 to two independent binomials both with probability  $\frac{1}{2}$ . Chatterjee's large sample chi-squared statistic on two degrees of freedom,  $.76 = (4/29)(14 - 29/2)^2 + (4/22)(13 - 22/2)^2$  is much less than the .05 critical value of 5.991, so there is not the slightest sign of a departure from stability and comparability. Although it is wise to test for additive biases, one will typically be concerned about hidden biases that may take a different form or otherwise go undetected, and for this sensitivity analyses are needed; see Section 3.

### 3. INFERENCE IN OBSERVATIONAL STUDIES

#### 3.1 Sensitivity Analysis in Observational Studies

In an observational study, treatments are not randomly assigned, so treated subjects may not be comparable to controls. Pretreatment differences visible in the data at hand, or overt biases, are removed by adjustments, such as matching (Cochran 1965; Smith 1997). There is no guarantee, however, that these adjustments will be effective, because there may also be differences in covariates that were not measured, or hidden biases. A sensitivity analysis asks how hidden biases of various magnitudes might alter the conclusions of an observational study. The first sensitivity analysis by Cornfield, Haenszel, Hammond, Lilienfeld, Shimken, and Wynder (1959) showed that for an unobserved binary covariate to explain the strong association between heavy smoking and lung cancer, that covariate would need to be a near-perfect predictor of lung cancer and about nine times more common among smokers than among nonsmokers; see also Greenhouse (1982) and Gastwirth, Krieger, and Rosenbaum (1998a). A sensitivity analysis replaces the correct, completely general, but not very useful logical fact that association does not imply causation. In its place is a quantitative statistical inference specific to the findings of a particular study saying that to explain away the association actually observed, a hidden bias would need to be of such and such a magnitude. Association does not imply causation, but studies vary markedly in their sensitivity to hidden bias; see Rosenbaum (1995a, Section 4) for numerous examples and detailed discussion. Varied methods of sensitivity analysis are discussed by Cornfield et. al. (1959), Rosenbaum and Rubin (1983), Rosenbaum (1987), Gastwirth (1992), Manski (1995), Copas and Li (1997), Gastwirth, Krieger, and Rosenbaum (1998b), and Lin, Psaty, and Kronmal (1998).

In contrast to the randomized experiment in Section 2, an observational study may be biased in the selection of subjects to receive the treatment, and biased in the selection of treated locations for the treated subjects. The sensitivity analysis developed in this section and illustrated in Section 4 explores sensitivity to both forms of hidden bias.

#### 3.2 Departures From Random Assignment

The model of treatment assignment in the observational study is identical to that of the experiment, except that the treatment assignment is not picked from  $\Omega$  at random with equal probabilities, and the treatment assignment probabilities are not known. Treatment assignments in distinct pairs are modelled as mutually independent. Within pair  $i$ , two parameters  $\Gamma \geq 1$  and  $\Lambda \geq 1$  measure the degree of departure from a randomized experiment. Specifically, it is assumed that for  $i = 1, \dots, I, j = 1, 2$ :

$$\Gamma \geq \frac{\text{prob}(Z_{i1} = 1)}{\text{prob}(Z_{i2} = 1)} \geq \frac{1}{\Gamma} \quad (2)$$

and

$$\Lambda \geq \frac{\text{prob}(V_{i1} = 1|Z_{ij} = 1)}{\text{prob}(V_{i0} = 1|Z_{ij} = 1)} \geq \frac{1}{\Lambda}. \quad (3)$$

In words, one subject in a pair is at most  $\Gamma$  times more likely to receive the treatment than the other, and one location for the treated subject is at most  $\Lambda$  times more likely to be treated than the other. This is analogous to the model in Rosenbaum (1987) for treatment assignment in which assignment is affected by an unobserved covariate not controlled by matching, except here the model is applied twice, once to the assignment of treatment to one subject in a pair, and again to the assignment of treatment for that subject. In place of a single known distribution of treatment assignments when  $\Gamma = 1$  and  $\Lambda = 1$ , Model (3) defines a family of departures from equally probable random assignment, specifically, a family that becomes progressively larger and less definite as  $\Gamma$  and  $\Lambda$  increase.

Write  $\pi_i = \text{prob}(Z_{i1} = 1)$  and  $\theta_{ij} = \text{prob}(V_{i1} = 1|Z_{ij} = 1)$ , so (3) implies:

$$\frac{\Gamma}{1 + \Gamma} \geq \pi_i \geq \frac{1}{1 + \Gamma}$$

and

$$\frac{\Lambda}{1 + \Lambda} \geq \theta_{ij} \geq \frac{1}{1 + \Lambda}. \quad (4)$$

Notice that  $\Gamma = 1$  implies  $\pi_i = \frac{1}{2}$  so treatments are randomly assigned to subjects in a pair, whereas  $\Lambda = 1$  implies  $\theta_{ij} = \frac{1}{2}$  so treatments are randomly assigned to locations on the treated subject. When  $(\Gamma, \Lambda) = (1, 1)$ , the sensitivity analysis will reproduce the unique randomization inference from Section 2.1.2, but as  $(\Gamma, \Lambda)$  increases, there will be uncertainty about the treatment assignment probabilities, resulting in a range of inferences, for instance, a range of point estimates or significance levels. For sufficiently large  $\Gamma$  and  $\Lambda$ , any distribution of assignments within a pair satisfies (3), so (3) is not an assumption, but rather a way of measuring the magnitude of the departure from random assignment by indexing it with the two parameters  $\Gamma$  and  $\Lambda$ . If small departures from random assignment produce a broad range of inferences, then the study is highly sensitive to hidden bias, but if only large values of  $(\Gamma, \Lambda)$  can produce a broad range of inferences, then the study is insensitive.

Subject to (3), or equivalently to (4), the four treatment assignment probabilities,  $\pi_i \theta_{i1}$ ,  $\pi_i (1 - \theta_{i1})$ ,  $(1 - \pi_i) \theta_{i2}$ , and  $(1 - \pi_i) (1 - \theta_{i2})$ , may differ by at most a multiplicative factor of  $\Gamma\Lambda$ . For instance, subject to (4), the ratio  $\pi_i \theta_{i1} / \{(1 - \pi_i) (1 - \theta_{i2})\}$  is between  $\Gamma\Lambda$  and  $1/(\Gamma\Lambda)$ . For this reason, one might think of the three conditions,  $(\Gamma, \Lambda) = (2, 1)$ ,  $(\Gamma, \Lambda) = (1, 2)$ , and  $(\Gamma, \Lambda) = (\sqrt{2}, \sqrt{2})$  as representing different patterns but similar magnitudes of hidden bias, as the ratio of treatment assignment probabilities, such as  $\pi_i \theta_{i1} / \{(1 - \pi_i) (1 - \theta_{i2})\}$ , is bounded below by  $\frac{1}{2}$  and above by 2 in all three cases. Notice that  $(\Gamma, \Lambda) = (1, 2)$  implies the assignment of treatments to subjects is randomized but the locations are possibly biased, while  $(\Gamma, \Lambda) = (2, 1)$  implies biased assignments to subjects with symmetrical or randomized locations.

#### 3.3 Approximate Inference Bounds

*3.3.1 Bounds on Expectations of a Test Statistic.* The sensitivity analysis places bounds on inference quantities, such as significance levels or point estimates, subject to (3), and

then varies  $(\Gamma, \Lambda)$  to display the sensitivity of inferences to departures from randomization of various magnitudes. When  $(\Gamma, \Lambda) = (1, 1)$ , and  $H_0$  is true,  $T$  has the randomization distribution in Section 2.1.2. When  $(\Gamma, \Lambda) > (1, 1)$ , many values of  $(\boldsymbol{\pi}, \boldsymbol{\theta})$  satisfy (4), so there are many possible distributions for  $T$ . This section places bounds on  $E(T)$  subject to (4). The relationship between bounds on  $E(T)$  and bounds on  $pr(T \geq k)$  is examined in Section 3.3.3. The material in Section 3.3 is required to implement the procedure; however, the example in Section 4 may be read without Section 3.3.

Subject to (3), Proposition 1 places bounds on:

$$E(T) = \sum_{i=1}^I \pi_i \{ \theta_{i1} s_{i11} + (1 - \theta_{i1}) s_{i10} \} + (1 - \pi_i) \{ \theta_{i2} s_{i21} + (1 - \theta_{i2}) s_{i20} \}. \quad (5)$$

The bounds have an intuitive structure. For the upper bound, within each subject, one first maximizes the probability of the location having the higher score,  $s_{ijl}$ , raising its  $\theta_{ij}$  to  $\Lambda/(1 + \Lambda)$ , then maximizes the probability of treatment for the subject with the higher expected score,  $\theta_{ij} s_{ij1} + (1 - \theta_{ij}) s_{ij0}$ . Formally, define:

$$\bar{\theta}_{ij} = \begin{cases} \frac{\Lambda}{1+\Lambda} & \text{if } s_{ij1} - s_{ij0} > 0 \\ \frac{1}{1+\Lambda} & \text{otherwise} \end{cases}, \quad \bar{\theta}_{ij} = 1 - \bar{\theta}_{ij},$$

$$\bar{\pi}_i = \begin{cases} \frac{\Gamma}{1+\Gamma} & \text{if } \bar{\theta}_{i1} s_{i11} + (1 - \bar{\theta}_{i1}) s_{i10} > \bar{\theta}_{i2} s_{i21} + (1 - \bar{\theta}_{i2}) s_{i20} \\ \frac{1}{1+\Gamma} & \text{otherwise} \end{cases}$$

$$\bar{\pi}_i = \begin{cases} \frac{\Gamma}{1+\Gamma} & \text{if } \bar{\theta}_{ij} s_{i11} + (1 - \bar{\theta}_{ij}) s_{i10} < \bar{\theta}_{ij} s_{i21} + (1 - \bar{\theta}_{ij}) s_{i20} \\ \frac{1}{1+\Gamma} & \text{otherwise.} \end{cases}$$

*Proposition 1.* If the treatment assignment probabilities satisfy (4), then the expectation  $E(T)$  is bounded by two known quantities:

$$\begin{aligned} & \sum_{i=1}^I \bar{\pi}_i \{ \bar{\theta}_{i1} s_{i11} + (1 - \bar{\theta}_{i1}) s_{i10} \} \\ & + (1 - \bar{\pi}_i) \{ \bar{\theta}_{i2} s_{i21} + (1 - \bar{\theta}_{i2}) s_{i20} \} \\ & \geq E(T) \geq \sum_{i=1}^I \bar{\pi}_i \{ \bar{\theta}_{i1} s_{i11} + (1 - \bar{\theta}_{i1}) s_{i10} \} \\ & + (1 - \bar{\pi}_i) \{ \bar{\theta}_{i2} s_{i21} + (1 - \bar{\theta}_{i2}) s_{i20} \} \end{aligned}$$

Moreover, these bounds are sharp, in the sense that they are attained for treatment assignment distributions that satisfy (4).

*Proof.* Notice that (4) implies

$$\begin{aligned} \bar{\theta}_{ij} s_{ij1} + (1 - \bar{\theta}_{ij}) s_{ij0} & \geq \theta_{ij} s_{ij1} + (1 - \theta_{ij}) s_{ij0} \\ & \geq \bar{\theta}_{ij} s_{ij1} + (1 - \bar{\theta}_{ij}) s_{ij0} \end{aligned}$$

for all  $i, j$ , so that for all  $1 \geq \pi_i \geq 0$ ,

$$\begin{aligned} & \sum_{i=1}^I \pi_i \{ \bar{\theta}_{i1} s_{i11} + (1 - \bar{\theta}_{i1}) s_{i10} \} \\ & + (1 - \pi_i) \{ \bar{\theta}_{i2} s_{i21} + (1 - \bar{\theta}_{i2}) s_{i20} \} \\ & \geq E(T) \geq \sum_{i=1}^I \pi_i \{ \bar{\theta}_{i1} s_{i11} + (1 - \bar{\theta}_{i1}) s_{i10} \} \\ & + (1 - \pi_i) \{ \bar{\theta}_{i2} s_{i21} + (1 - \bar{\theta}_{i2}) s_{i20} \} \quad (6) \end{aligned}$$

Subject to (4), the upper bound in (6) is maximized by taking  $\pi_i = \bar{\pi}_i$  and the lower bound is minimized by taking  $\pi_i = \bar{\pi}_i$ , yielding the stated bounds on  $E(T)$ . The bounds are sharp since  $\bar{\pi}_i, \bar{\theta}_{ij}$ , and  $\bar{\theta}_{ij}$  satisfy (4).

*3.3.2 There Is No Bounding Random Variable.* In the simplest cases of sensitivity analysis (Rosenbaum 1988, 1995a, Section 4.4, 1995b), it is possible to find a set of treatment assignment probabilities that produces a distribution of the test statistic  $T$  that is stochastically larger than for any other set of treatment assignment probabilities. If such a distribution existed for  $T$ , then it would have the largest expectation, so Proposition 1 in the previous section identifies a plausible candidate for that set of treatment assignment probabilities. Alas, the current problem is not one of these simple cases with a bounding random variable—the  $T$  identified in Proposition 1 is not stochastically larger than all others. This section shows by counterexample that  $T$  is not stochastically largest.

Consider the contribution from pair  $i$ , namely  $H_i = \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} V_{il} s_{ijl}$ , and let  $\bar{H}_i$  be the analogous random variables when  $(\pi_i, \theta_{i1}, \theta_{i2}) = (\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2})$ . The Proposition in the previous section showed that  $E(H_i) \geq E(\bar{H}_i)$ . However,  $\bar{H}_i$  need not be stochastically larger than  $H_i$ . For example, suppose  $s_{i11} = 4, s_{i10} = 1, s_{i21} = 3, s_{i20} = 2, \Gamma = 9, \Lambda = 2, (\pi_i, \theta_{i1}, \theta_{i2}) = (1/2, 1/2, 1/2)$ , so that  $(\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2}) = (9/10, 2/3, 2/3)$ . Then  $H_i$  and  $\bar{H}_i$  have the following distributions:

	$s_{i11} = 4$	$s_{i10} = 1$	$s_{i21} = 3$	$s_{i20} = 2$
$H_i$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\bar{H}_i$	$\frac{18}{30}$	$\frac{9}{30}$	$\frac{2}{30}$	$\frac{1}{30}$

so that  $E(\bar{H}_i) = (18/30 \times 4) + (9/30 \times 1) + (2/30 \times 3) + (1/30 \times 2) = 26/9 = 2.89 > 2.5 = 1/4(4 + 1 + 3 + 2) = E(H_i)$ , but  $\text{prob}(\bar{H}_i \geq 2) = 21/30 = 0.7 < 3/4 = \text{prob}(H_i \geq 2)$ . There is, in general, no configuration of  $(\pi_i, \theta_{i1}, \theta_{i2})$  that yields a stochastically largest  $\bar{H}_i$ . Similar considerations apply to the random variable  $\bar{H}_i$  obtained by letting  $(\pi_i, \theta_{i1}, \theta_{i2}) = (\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2})$ , which gave the smallest expectation in Proposition 1.

*3.3.3 Using Asymptotic Separability to Bound  $\text{Pr}(T \geq k)$*

**The Concept of Asymptotic Separability**

When there is no unique bounding random variable for use in sensitivity analysis, Gastwirth, Krieger, and Rosenbaum (2000) showed that one may obtain an approximate upper bound on

a significance level, say  $\Pr(T \geq k)$ , for  $k > E(T)$ , by picking the treatment assignment probabilities to maximize the expectation  $E(T)$ , and if there is a tie among treatment assignment probabilities that maximize the expectation, to resolve the tie by picking the assignment probabilities to maximize the variance. This approximation converges to the true, sharp upper bound on  $\Pr(T \geq k)$  as the number of matched sets increases,  $I \rightarrow \infty$ .

#### A Technical Detail: Ties in the Maximum Expectation

To apply that general approach to the current problem, one must first address the issue of ties in the expectations. Tied expectations occur when several choices of  $(\pi_i, \theta_{i1}, \theta_{i2})$  in a pair all give the maximum  $E(H_i)$ . There are three cases to consider, namely: (i)  $(\Gamma > 1, \Lambda > 1)$ , (ii)  $(\Gamma = 1, \Lambda > 1)$ , and (iii)  $(\Gamma > 1, \Lambda = 1)$ , and it will turn out that cases (i) and (ii) are trivial requiring no adjustments at all. Case (iii) requires a small adjustment. Consider these cases in turn. Keep in mind that a sensitivity analysis entails trying out several specific values for  $(\Gamma, \Lambda)$ , so we always know which case applies.

If  $\Gamma > 1$  and  $\Lambda > 1$ , then there is strict inequality in Proposition 1,  $E(\bar{H}_i) > E(H_i) > E(\bar{H}_i)$ , unless  $s_{i11} = s_{i10} = s_{i21} = s_{i20}$ , in which case  $\bar{H}_i = H_i = \bar{H}_i$  and all three are constant. This means that, when  $\Gamma > 1$  and  $\Lambda > 1$ , there is either a unique  $(\pi_i, \theta_{i1}, \theta_{i2})$ , given by Proposition 1, that maximizes  $E(H_i)$ , or otherwise all values of  $(\pi_i, \theta_{i1}, \theta_{i2})$  give the same distribution for  $H_i$ . Hence no adjustment for ties is needed when  $\Gamma > 1$  and  $\Lambda > 1$ .

If  $\Gamma = 1$  and  $\Lambda > 1$ , then there is strict inequality,  $E(\bar{H}_i) > E(H_i) > E(\bar{H}_i)$ , unless  $s_{i11} = s_{i10}$  and  $s_{i21} = s_{i20}$ , in which case  $\bar{H}_i = H_i = \bar{H}_i$  and all three random variables equal  $s_{i11}$  or  $s_{i21}$  each with probability  $\frac{1}{2}$ . Again, either there is a unique  $(\pi_i, \theta_{i1}, \theta_{i2})$ , given by Proposition 1, that maximizes  $E(H_i)$ , or else the choice of  $(\pi_i, \theta_{i1}, \theta_{i2})$  does not affect the distribution of  $H_i$ , so again, no adjustment is needed if  $\Gamma = 1$  and  $\Lambda > 1$ .

If  $\Gamma > 1$  and  $\Lambda = 1$ , then  $E(\bar{H}_i) > E(H_i) > E(\bar{H}_i)$  unless  $s_{i11} + s_{i10} = s_{i21} + s_{i20}$ , and in which case  $E(H_i) = \bar{s}_i = E(H_i | Z_i = z)$  for  $z = 0$  and  $z = 1$ . In this case, with  $s_{i11} + s_{i10} = s_{i21} + s_{i20}$ , the expectations for the two subjects in the pair are equal, and one maximizes the variance by taking  $\bar{\pi}_i = \Gamma/(1 + \Gamma)$  if  $|s_{i11} - s_{i10}| > |s_{i21} - s_{i20}|$  and  $\bar{\pi}_i = 1/(1 + \Gamma)$  otherwise.

In summary, to apply asymptotic separability to approximate the upper bound on  $\Pr(T \geq k)$  for  $k > E(T)$  one small change in the definition of  $\bar{\pi}_i$  is required in one special case. Specifically, if  $\Gamma > 1$  and  $\Lambda = 1$ , and in pair  $i$  if  $s_{i11} + s_{i10} = s_{i21} + s_{i20}$ , then redefine  $\bar{\pi}_i$  for this one pair  $i$  as follows:  $\bar{\pi}_i = \Gamma/(1 + \Gamma)$  if  $|s_{i11} - s_{i10}| > |s_{i21} - s_{i20}|$  and  $\bar{\pi}_i = 1/(1 + \Gamma)$  otherwise. Under the same very special circumstances, when approximating the upper bound on  $\Pr(k \geq T)$  for  $k < E(T)$ , redefine  $\bar{\pi}_i = \Gamma/(1 + \Gamma)$  if  $|s_{i11} - s_{i10}| > |s_{i21} - s_{i20}|$ , so that in both cases, tied expectations lead one to increase the probability of the more variable scores. In the example in the Section 5 with  $I = 66$  pairs, this tied situation rarely came up, and when it did come up, it affected only one of the 66 pairs.

#### Procedures for Sensitivity Analysis

As is traditional in nonparametrics, confidence intervals and point estimates are derived from tests, so tests come

first logically, even if in practice it is more useful to report point and interval estimates. Consider testing the hypothesis,  $H_0 : \tau = \tau_0$ , and seeking an upper bound on the one-sided significance level,  $\Pr(T \geq k)$ . Under this hypothesis, one computes the scores  $s_{ijl}$  from the aligned, adjusted responses, and then the test statistic  $T = \sum H_i$  where  $H_i = \sum_{j=1}^2 \sum_{l=0}^1 Z_{ij} V_{il} s_{ijl}$ . One then computes the probabilities,  $(\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2})$ , in Proposition 1, that determine the contributions,  $H_i$ , to  $T$  that maximize  $E(T)$ , making in a few rare instances the slight adjustment to the definition of  $\bar{\pi}_i$  discussed above. Then the observed  $T$  is compared to the maximum expectation  $E(\sum \bar{H}_i) = \sum E(\bar{H}_i) = \sum \bar{\mu}_i$ , say, and the variance at the maximum expectation,  $\text{var}(\sum \bar{H}_i) = \sum \text{var}(\bar{H}_i) = \sum \bar{\sigma}_i^2$  where:

$$\begin{aligned} \bar{\mu}_i &= \bar{\pi}_i \bar{\theta}_{i1} s_{i11} + \bar{\pi}_i (1 - \bar{\theta}_{i1}) s_{i10} \\ &\quad + (1 - \bar{\pi}_i) \bar{\theta}_{i2} s_{i21} + (1 - \bar{\pi}_i) (1 - \bar{\theta}_{i2}) s_{i20} \\ \bar{\sigma}_i^2 &= \bar{\pi}_i \bar{\theta}_{i1} s_{i11}^2 + \bar{\pi}_i (1 - \bar{\theta}_{i1}) s_{i10}^2 \\ &\quad + (1 - \bar{\pi}_i) \bar{\theta}_{i2} s_{i21}^2 + (1 - \bar{\pi}_i) (1 - \bar{\theta}_{i2}) s_{i20}^2 - \bar{\mu}_i^2. \end{aligned}$$

Using Proposition 1 of Gastwirth, Krieger, and Rosenbaum (2000), the sharp upper bound on  $\Pr(T \geq k)$  for  $k > E(T)$  may be approximated with negligible error as  $I \rightarrow \infty$  by  $1 - \Phi\{(T - \sum \bar{\mu}_i) / \sqrt{\sum \bar{\sigma}_i^2}\}$  where  $\Phi(\cdot)$  is the standard normal cumulative distribution. This yields an approximation to the sharp upper bound on the one-sided significance level. In parallel, for the significance level in the opposite tail, one approximates the upper bound on  $\Pr(k \geq T)$  for  $k < E(T)$  using the same procedures but with  $(\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2})$  in place of  $(\bar{\pi}_i, \bar{\theta}_{i1}, \bar{\theta}_{i2})$ , yielding  $\bar{\mu}_i$  and  $\bar{\sigma}_i^2$ .

The two-sided  $100(1 - \alpha)\%$  confidence interval for  $\tau$  is the set of  $\tau_0$ 's not rejected at level  $\alpha/2$  in the above manner by either of the upper bounds on the two, one-sided significance levels. The bounds on the Hodges–Lehmann (1963) point estimate of  $\tau$  are found by solving for  $\tau_0$  twice, once in the equation  $T = \sum \bar{\mu}_i$  and once in  $T = \sum \bar{\mu}_i$ .

#### 4. EXAMPLE: MINIMUM WAGE AND EMPLOYMENT

To illustrate, the sensitivity analysis will be applied to Card and Krueger's (1994, 1995) study, described in Section 1.2.1, of the effects of increasing New Jersey's minimum wage. The goal here is solely to illustrate methodology, not to reach conclusions about minimum wages and their employment effects. As recorded in Table 1 of Rosenbaum (1999b), there are  $I = 66$  pairs of fast food restaurants, one in New Jersey (NJ) and one in eastern Pennsylvania (PA), matched for chain (Burger King, Wendy's, etc.) and for the starting wage before the wage increase. In contrast to economic theory, Card and Krueger found no evidence of a decline in employment following the increase in the minimum wage. What magnitude of hidden bias would need to be present to reconcile economic theory and data?

Table 4 is the sensitivity analysis giving the lower bounds for the Hodges–Lehmann point estimates  $\hat{\tau}$  derived from  $T$

Table 4. Sensitivity of Estimated Number of Employees Lost

$\Gamma$	$\Delta$	Minimum $\hat{\tau}$	Minimum $\hat{\tau}_{low}$
1	1	.37	-1.75
$\sqrt{2}$	$\sqrt{2}$	-1.13	-3.12
2	1	-.38	-2.37
1	2	-1.88	-3.75
$\sqrt{3}$	$\sqrt{3}$	-2.00	-4.00
3	1	-.75	-2.69
1	3	-2.88	-5.12

and for the lower endpoints  $\hat{\tau}_{low}$  of 95% confidence intervals. In all cases, upper bounds, which are not presented, indicate that no effect of the minimum wage is a plausible hypothesis. As discussed in the previous section, biases of  $(\Gamma, \Delta) = (\sqrt{2}, \sqrt{2})$ ,  $(\Gamma, \Delta) = (2, 1)$ , and  $(\Gamma, \Delta) = (1, 2)$  are comparable in magnitude but different in pattern. In particular,  $(\Gamma, \Delta) = (2, 1)$  is a hidden bias with symmetry over time, whereas  $(\Gamma, \Delta) = (1, 2)$  is a hidden bias with comparability of restaurants in New Jersey and eastern Pennsylvania.

As noted in Section 3.2, when  $\Gamma\Delta = 2$ , treatment assignment probabilities differ by at most a factor of 2. For comparison, treatment assignment probabilities would need to differ by a factor of about 6 to explain away the association between heavy smoking and lung cancer found by Hammond (1964), by a factor of about 7 to explain away the association between DES and vaginal cancer found by Herbst, Ulfelder, and Poskanzer (1971), and by about 1.3 to explain away the association between coffee and myocardial infarction found by Jick, Miettinen, Neff, et al. (1973). See Rosenbaum (1995a, Section 4) for detailed discussion.

In the absence of hidden bias,  $(\Gamma, \Delta) = (1, 1)$ , the randomization distribution gives a point estimate of  $\hat{\tau} = .37$  or a gain, rather than the anticipate loss, of .37 employees per restaurant; however, from the confidence interval, a loss of 1.75 employees per store is consistent with chance even in the absence of bias. Biases of moderate size are consistent with point estimates of a loss of 1.88 employees per store and lower endpoints of 3.75 employees. As the typical store had about 20 full-time equivalent employees, declines of 2 or 4 employees are not catastrophic, but they are not inconsequential either. Card and Krueger were certainly correct to say the data provide no indication of a decline in employment following the increase in the minimum wage, but the loss of several employees per store cannot be ruled out if moderate biases are plausible.

## 5. CONCLUSION: THE CONTRIBUTION OF STABILITY

Stability and comparability, if present, yield tests for treatment effect with greater power but with no increase in the sample size. This increase in power is due to a larger non-centrality parameter in the normal theory test, so the treatment effect stands out more clearly. Generally, larger treatment effects are less sensitive to hidden biases—only large hidden biases can explain away large ostensible treatment effects. Analyses that exploit stability in the absence of treatment may, in some studies, yield reduced sensitivity to hidden bias.

[Received August 1999. Revised April 2000.]

## REFERENCES

- Allison, P. D. (1990), "Change Scores As Dependent Variables in Regression Analysis," in *Sociological Methodology*, C. C. Clogg, ed., Oxford: Basil Blackwell, pp. 93–114.
- Angrist, J., and Krueger, A. (1999), "Empirical Strategies in Labor Economics," in *The Handbook of Labor Economics, III*, Chapter 23, New York: Elsevier.
- Ashenfelter, O., and Card, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648–660.
- Bauer, D. F. (1972), "Constructing Confidence Sets Using Rank Statistics," *Journal of the American Statistical Association*, 67, 687–690.
- Bell, C. B., and Haller, H. S. (1969), "Bivariate Symmetry Tests," *The Annals of Mathematical Statistics*, 40, 259–269.
- Card, D., and Krueger, A. (1994), "Minimum Wages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–793.
- (1995), *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton, NJ: Princeton University Press.
- Chatterjee, S. K. (1966), "A Bivariate Sign Test for Location," *The Annals of Mathematical Statistics*, 37, 1771–1781.
- Cochran W. G. (1965), "The Planning of Observational Studies of Human Populations (with Discussion)," *Journal of the Royal Statistical Society, Ser. A*, 128, 234–255.
- Cook, T. D., and Campbell, D. T. (1979), *Quasi-Experimentation*. Boston: Houghton Mifflin.
- Cook, T. D., Campbell, D. T. and Peracchio, L. (1990), "Quasi-Experimentation," in *Handbook of Industrial and Organizational Psychology*, eds. M. Dunnette and L. Hough, Palo Alto, CA: Consulting Psychologists Press, pp. 491–576.
- Cook, T. D., and Shadish, W. R. (1994), "Social Experiments: Some Developments Over The Past Fifteen Years," *Annual Review of Psychology*, 45, 545–580.
- Copas, J. B., and Li, H. G. (1997), "Inference for Non-Random Samples (with discussion)," *Journal of the Royal Statistical Society, Ser. B*, 59, 55–96.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.
- Dom, H. F. (1953), "Philosophy of Inferences From Retrospective Studies," *American Journal of Public Health*, 43, 677–683.
- Fisher, R. A. (1935), *The Design of Experiments*. Oliver and Boyd.
- Gastwirth, J. L. (1992), "Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables," *Jurimetrics*, 33, 19–34.
- Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998a), "Cornfield's inequality," in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton eds., New York: Wiley, pp. 952–955.
- (1998b), "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika*, 85, 907–920.
- (2000), "Asymptotic Separability in Sensitivity Analysis," *Journal of the Royal Statistical Society, Ser. B*, No. 3, to appear.
- Greenhouse, S. (1982), "Jerome Cornfield's Contributions to Epidemiology," *Biometrics Supplement*, 33–45.
- Hajek, J., Sidak, Z., and Sen, P. K. (1999), *Theory of Rank Tests* (2nd ed.), New York: Academic Press.
- Hammond, E. C. (1964), "Smoking in Relation to Mortality and Morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.
- Herbst, A., Ulfelder, H., and Poskanzer, D. (1971), "Adenocarcinoma of the Vagina: Association of Maternal Stibestrol Therapy With Tumor Appearance in Young Women," *New England Journal of Medicine*, 284, 878–881.
- Hodges, J. L., and Lehmann, E. L. (1962), "Rank Methods for Combination of Independent Experiments in the Analysis of Variance," *The Annals of Mathematical Statistics*, 33, 482–497.
- (1963), "Estimates of Location Based on Ranks," *The Annals of Mathematical Statistics*, 34, 598–611.
- Holland, P. W., and Rubin, D. B. (1983), "On Lord's Paradox," in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, eds. H. Wainer and S. Messick, Hillsdale, NJ: Lawrence Erlbaum, pp. 3–25.
- Hollander, M., and Wolfe, D. A. (1999), *Nonparametric Statistical Methods* (2nd ed.), New York: Wiley.
- Jick, H., Miettinen, O., Neff, R., Shapiro, S., Heinonen, O. P., and Sloan, D. (1973), "Coffee and myocardial infarction," *New England Journal of Medicine*, 289, 63–77.

- Johnson, N. L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, New York: Wiley.
- Kershner, R. P., and Federer, W. T. (1981), "Two-treatment Crossover Designs for estimate a variety of effects," *Journal of the American Statistical Association*, 76, 612–619.
- Koch, G. G. (1972), "The use of Non-parametric Methods in the Statistical Analysis of the Two-Period Change-over Design," *Biometrics*, 28, 577–584.
- Lehmann, E. L. (1963), "Nonparametric Confidence Intervals for a Shift Parameter," *The Annals of Mathematical Statistics*, 34, 1507–1512.
- (1998), *Nonparametrics: Statistical Methods Based on Ranks* (rev. 1st ed.), Upper Saddle River, NJ: Prentice-Hall.
- Lehmann, E. L., and Stein, C. (1949), "On the Theory of Some Nonparametric Hypotheses," *The Annals of Mathematical Statistics*, 20, 28–45.
- Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963.
- Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.
- Meyer, B. D. (1995), "Natural and Quasi-experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151–161.
- Moses, L. E. (1965), "Confidence Limits From Rank Tests," *Technometrics*, 7, 257–260.
- Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles," (In Polish) *Roczniki Nauk Rolniczych*, Tom X, pp. 1–51. Reprinted in English in *Statistical Science*, 1990, 5, 463–480, with discussion by T. Speed and D. Rubin.
- Reichardt, C. S. (1979), "The Statistical Analysis of Data From Nonequivalent Group Design," in *Quasi-Experimentation*, eds. T. D. Cook and D. T. Campbell, Boston: Houghton Mifflin, pp. 147–206.
- Rosenbaum, P. R. (1987), "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.
- Rosenbaum, P. R. (1988), "Sensitivity Analysis for Matching with Multiple Controls," *Biometrika*, 75, 577–581.
- (1995a), *Observational Studies*. New York: Springer-Verlag.
- (1995b), "Quantiles in Nonrandom Samples and Observational Studies," *Journal of the American Statistical Association*, 90, 1424–1431.
- (1999a), "Choice as an Alternative to Control in Observational Studies (with discussion)," *Statistical Science*, 14, 259–304.
- (1999b), "Using Quantile Averages in Matched Observational Studies," *Applied Statistics*, 48, 63–78.
- (1999c), "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies With Dilated Treatment Effects," *Biometrics*, 55, 560–564.
- Rosenbaum, P., and Rubin, D. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society, Ser. B*, 45, 212–218.
- Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.
- (1977), "Randomization on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.
- Salzberg, A. (1999), "Removable Selection Bias in Quasi-experiments," *The American Statistician*, 53, 103–107.
- Scheffe, H. (1959), *The Analysis of Variance*, New York: Wiley.
- Shadish, W. R., Cook, T. D., and Leviton, L. C. (1995), *Foundations of Program Evaluation*, Thousand Oaks, CA: Sage Publications.
- Smith, H. L. (1997), "Matching With Multiple Controls to Estimate Treatment Effects in Observational Studies," *Sociological Methodology*, 27, 325–353.
- Tukey, J. W. (1986), "Sunset Salvo," *American Statistician*, 40, 72–76.
- Wei, L. J. (1987), "Two Sample Problem With Bivariate Exchangeable Observations," *Journal of the Royal Statistical Society, Ser. B*, 49, 40–45.