Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies

Paul R. ROSENBAUM

Before R. A. Fisher introduced randomized experimentation, the literature on empirical methods emphasized reducing heterogeneity of experimental units as the key to inference about the effects caused by treatments. To what extent is heterogeneity relevant to causal inference when ethical or practical constraints make random assignment infeasible?

KEY WORDS: Causal effect; Observational study; Randomization inference; Randomized experiment; Sensitivity analysis; Wilcoxon's signed rank test.

1. J. S. MILL AND R. A. FISHER: DIFFERENCE AND RANDOMIZATION

The statistical theory of causal inference in experiments traces its long history back through the British empirical tradition. Two voices in that tradition were those of the philosopher and economist, John Stuart Mill, and the statistician and geneticist, Sir Ronald Fisher. Mill knew nothing of randomized experimentation; so he proposed a highly influential theory of causal inference that reflected the laboratory tactics of his time, many of which continue in laboratories today. In advocating randomized experiments, in his discussion of his famous example, "the lady tasting tea," Fisher rejected—"rejected" is too gentle, "dismissed"—a certain element in Mill's theory as "totally impossible." I would like to revisit this exchange, with the hope of learning something about observational studies in which ethical or practical considerations prevent randomization. Mill's method included a fanatical effort to drive out heterogeneity of experimental material—the use of genetically engineered, nearly identical mice in the modern biology laboratory would have met with his approval—whereas Fisher proposed a method that dealt with heterogeneity without eliminating it, a method applicable, for instance, to clinical trials with human subjects.

In 1864, in his *System of Logic: Principles of Evidence and Methods of Scientific Investigation*, Mill proposed "four methods of experimental inquiry," including the "method of difference:"

> If an instance in which the phenomenon ... occurs and an instance in which it does not ... have every circumstance save one in common ... [then] the circumstance [in] which alone the two instances differ is the ... cause or a necessary part of the cause (III, sec. 8)

Notice the emphasis on the complete absence of heterogeneity, "have every circumstance save one in common," that is, on treated and control units that are identical but for the treatment. For Mill, homogeneity and sound causal inference were closely linked: he wanted "two instances ... exactly similar in all circumstances except the one" under study. See Holland (1986) for a discussion of Mill. Fisher had his doubts about the "method of difference." In 1935, in Chapter 2 of his *Design of Experiments*, discussing the lady and her tea, Fisher (1935, 1949, p. 18) wrote:

> It is not sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation ... These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences ... because [they] ... are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment ...

The first omission, "...," in this quote contains a thoroughly interesting, albeit somewhat lengthy, discussion of the many ways two cups of tea may differ.

Fisher was, of course, correct. Sick patients in a hospital cannot be made homogeneous, but they can be randomized. Mill was proposing an ideal, something toward which you could strive, as you strive towards all of your ideals, with success, to some degree, eventually, perhaps; whereas, Fisher was proposing a method, something you could do, and if you did it, it would work.

What if the treatment under study is harmful or not subject to the investigator's control? Is Mill's "method of difference" relevant to causal inference when randomization is infeasible? Does reducing the heterogeneity of experimental units strengthen *causal* claims? Or does reducing the heterogeneity without randomizing simply reduce the *standard error* of a biased estimator?

Since Fisher, understanding of heterogeneity is shaped by randomized experiments, in which there is a consistent estimate, say $\widehat{\delta}$, of the treatment effect, say δ. If there were I matched pair differences, D_i, $i = 1, \ldots, I$, independent and identically distributed (iid) with a Normal distribution, $D_i \sim N\left(\delta, \sigma^2\right)$, then the natural estimator is the mean difference, say $\widehat{\delta}$. In this case, decreasing unit heterogeneity by reducing σ to $\sigma/2$ has exactly the same effect on $\widehat{\delta}$ as increasing the sample size I to $4I$, for in both cases $\widehat{\delta} \sim N\left\{\delta, \sigma^2/(4I)\right\}$. This might lead to the intuition that using nearly identical laboratory mice is not much different from using highly variable field mice but using more of them, providing the treatment affects all mice in the same way.

In fact, that intuition may be correct for randomized experiments but is wrong for observational studies. It began by assuming there is a consistent estimate, but without randomization,

Paul R. Rosenbaum is Robert G. Putzel Professor, Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 (E-mail: rosenbaum@stat.wharton.upenn.edu). This work was supported by grant SES-0345113 from the Methodology, Measurement, and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation.

 DOI: 10.1198/000313005X42831

there may be no such estimate. This article makes two claims. (i) Without randomization, reducing heterogeneity, even purely random heterogeneity, confers benefits that cannot be achieved by increasing the sample size. Reducing heterogeneity reduces sensitivity to unobserved biases, whereas increasing the sample size does not. This is *demonstrated* in Section 3. after a brief review in Section 2.. (ii) Careful choice of experimental units can sometimes remove key sources of heterogeneity that would otherwise be present. This is *illustrated* in Section 4. with several clever observational studies. Experiments are made, but observational studies are assembled from available materials (Rosenbaum 1999), so the tactics in Section 4. are choices that reduce heterogeneity.

2. BACKGROUND: RANDOMIZATION INFERENCE; SENSITIVITY ANALYSIS

2.1 Review: A Randomized Paired Experiment; Randomization Inference

Before treatment, each subject has an observed covariate $\mathbf{x}$ and an unobserved covariate u. There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, matched for observed, pretreatment covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, but typically matched subjects differ in terms of unobserved covariates, so $u_{i1} \neq u_{i2}$. In a randomized, paired experiment, one subject in each pair is picked at random for treatment, the other receiving control. Write $Z_{ij} = 1$ if j received the treatment in pair i, and $Z_{ij} = 0$ if j received the control in pair i, so $Z_{i1} + Z_{i2} = 1$ for $i = 1, \ldots, I$.

Subject (i, j) has two potential responses, (r_{Tij}, r_{Cij}), with r_{Tij} observed under treatment, $Z_{ij} = 1$, and r_{Cij} observed under control, $Z_{ij} = 0$, so the effect of the treatment on (i, j) is $r_{Tij} - r_{Cij}$; see Neyman (1923) and Rubin (1974). Write $\mathcal{F}$ for $\{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, 2\}$ and $\mathcal{Z}$ for the event $\{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\}$; then $\mathcal{F}$ and $\mathcal{Z}$ are fixed by conditioning in Fisher's theory of randomization inference. Randomization ensures that $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \frac{1}{2}$, $i = 1, \ldots, I$, with independent assignments in distinct pairs. The response R_{ij} observed from (i, j) is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$. If the treatment effect is constant, $\tau = r_{Tij} - r_{Cij}$, then $R_{ij} = r_{Cij} + Z_{ij}\, \tau$. The treated-minus-control difference in responses in pair i is $D_i = (2Z_{i1} - 1)\,(R_{i1} - R_{i2})$, and with a constant effect, τ, this is $D_i = \tau + \epsilon_i$ where $\epsilon_i = (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2})$.

Test $H_0 : \tau = \tau_0$ by ranking $|D_i - \tau_0|$ from 1 to I; then Wilcoxon's signed rank statistic, W_{τ_0}, is the sum of the ranks for which $D_i - \tau_0 > 0$, where ties are assumed absent. If $H_0 : \tau = \tau_0$ is true, randomization ensures that $D_i - \tau_0 = \epsilon_i$ is $r_{Ci1} - r_{Ci2}$ or $r_{Ci2} - r_{Ci1}$, each with probability $\frac{1}{2}$, independently in different pairs. Given $\mathcal{Z}$, $\mathcal{F}$, if $H_0 : \tau = \tau_0$ is true, then the $|D_i - \tau_0|$ are fixed, the $D_i - \tau_0$ are independent, $\Pr(D_i - \tau_0 > 0) = \frac{1}{2}$, and each D_i is symmetric about τ_0, so W_{τ_0} is the sum of I independent random variables taking values i or 0 each with probability $\frac{1}{2}$, $i = 1, \ldots, I$. A confidence interval for τ is obtained by inverting the test, and the Hodges–Lehmann (1963) or HL estimate of τ is (essentially) the value $\widehat{\tau}$ such that $W_{\widehat{\tau}}$ is as close as possible to its null expectation, $I\,(I + 1)/4$. Constant effects τ are convenient but unneeded; see Rosenbaum (2003).

The null distribution of W_{τ_0} is the same for all (untied) $\mathcal{F}$, but the nonnull distribution depends on $\mathcal{F}$ or a model that generates $\mathcal{F}$. A common model for a part of $\mathcal{F}$ has $(r_{Ci1} - r_{Ci2})/\sigma \sim_{\text{iid}} F(\cdot)$ were $\sigma > 0$ and $F(\cdot)$ is a continuous distribution symmetric about zero, so that randomization ensures $\epsilon_i/\sigma \sim_{\text{iid}} F(\cdot)$. Lehmann (1998, sec. 3–4) discussed these standard methods, with randomization distributions given $\mathcal{F}$ in his section 3 and population models for $\mathcal{F}$ in his section 4.

2.2 Review: Sensitivity to Departures from Random Assignment

A simple model for sensitivity analysis in observational studies says that: (i) in the population prior to matching, treatment assignments were independent; (ii) subject i, j had a unknown probability, π_{ij}, of receiving the treatment; (iii) two subjects with the same *observed* covariates $\mathbf{x}_{ij}$—that is, two subjects who might be matched, say i, j and i, k, because $\mathbf{x}_{ij} = \mathbf{x}_{ik}$—may differ in their odds of receiving treatment by a factor of $\Gamma \geq 1$ because they differ in ways that were *unobserved*,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\,(1 - \pi_{ik})}{\pi_{ik}\,(1 - \pi_{ij})} \leq \Gamma, \quad \forall\, i, j, k; \tag{1}$$

and (iv) the distribution of treatments within treated/control matched pairs is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$. Here, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$. Write $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \ldots, \pi_{I2})^T$. In (1), if $\Gamma = 1$, then π_{ij} may vary with $\mathbf{x}_{ij}$ but not with other aspects of $\mathcal{F}$, so $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ ensures $\pi_{ij} = \pi_{ik}$, $i = 1, \ldots, I$, whereupon $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \pi_{i1}/(\pi_{i1} + \pi_{i2}) = \frac{1}{2}$, and the distribution of treatment assignments is again the randomization distribution; that is, bias solely due to observed covariates $\mathbf{x}$ can be eliminated by matching on them, $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. If $\Gamma > 1$ in (1), then matching on $\mathbf{x}$ may fail to equalize the π_{ij} in pair i; there may be bias from unobserved covariates. Of course, Γ is unknown. A sensitivity analysis calculates, for several values of Γ, the range of possible inferences—the range of significance levels or point estimates or confidence intervals—thereby displaying how sensitive conclusions are to various assumptions about biased treatment assignment. How large must Γ be before the range of possible conclusions is so long that qualitatively different interpretations are possible? For discussion of this model with examples, see Rosenbaum (1988; 1993; 2002, sec. 4; 2003). For alternative sensitivity analyses in observational studies, see Cornfield et al. (1959); Gastwirth, Krieger, and Rosenbaum (1998); Lin, Psaty, and Kronmal (1998); Copas and Eguchi (2001); and Imbens (2003).

Observational studies vary markedly in sensitivity to unobserved biases. Hammond's (1964) study of smoking and lung cancer is insensitive to $\Gamma = 5$: even with a such a large bias, the upper bound on the significance level for testing no effect is .03. In contrast, Jick et al.'s (1973) study of coffee as a cause of myocardial infarction is quite sensitive to unobserved biases: the significance level from a randomization inference ($\Gamma = 1$) is .0038, but the upper bound on the significance level is above .05 for $\Gamma \geq 1.3$. See Rosenbaum (1993; 2002, sec. 4; 2003) for examples using W_{τ_0}.

2.3 Review: Sensitivity Analysis with the Signed Rank Statistic

Sensitivity analysis using W_{τ_0} is straightforward. If (1) and $H_0 : \tau = \tau_0$ are true, then the null distribution of W_{τ_0} is unknown but is bounded by two known distributions. Write $\overline{\overline{W}}$ for the sum of I independent random variables taking value i with probability $\theta = \Gamma/(1+\Gamma)$ and value 0 with probability $1-\theta$, $i = 1, \ldots, I$; also, write $\overline{W}$ for the sum of I independent random variables taking value i with probability $1-\theta$ and value 0 with probability θ. Then (1) and $H_0 : \tau = \tau_0$ imply

$$\Pr\left(\overline{W} \geq w\right) \leq \Pr\left(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}\right) \leq \Pr\left(\overline{\overline{W}} \geq w\right) \quad (2)$$

for every w, and the bounds are attained for particular $\boldsymbol{\pi}$ satisfying (1); see Rosenbaum (1988; 2002, sec. 4.3) for the elementary details. If $\Gamma = 1$, then there is equality in (2). The bounds (2) become wider as Γ increases, and one is interested in the magnitude of bias Γ that would materially alter the study's conclusions. For a one-sided test, $H_0 : \tau = \tau_0$ versus $H_A : \tau > \tau_0$, the upper bound on the one-sided significance level is at most .05 for all π_{ij} satisfying (1) if $W_{\tau_0} \geq \widetilde{w}$ where the critical value $\widetilde{w}$ solves $.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$, and a simple approximation to $\widetilde{w}$ is provided by the central limit theorem for large I; see Rosenbaum (1988; 2002, sec. 4.3.3).

For each $\boldsymbol{\pi}$, there is an HL estimate $\widehat{\tau}_{\boldsymbol{\pi}}$ formed (essentially) as the solution to the estimating equation $W_{\widehat{\tau}} = \mu_{\boldsymbol{\pi}}$ where the expectation $\mu_{\boldsymbol{\pi}} = E_{\boldsymbol{\pi}}\left(W_\tau \mid \mathcal{Z}, \mathcal{F}\right)$ is computed using these $\boldsymbol{\pi}$. Then (1) implies $(1-\theta)\ I\ (I+1)/2 \leq \mu_{\boldsymbol{\pi}} \leq \theta\, I\ (I+1)/2$, yielding an interval of HL point estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$. With no unobserved bias, $\Gamma = 1$, $\mu_{\boldsymbol{\pi}} = I\,(I+1)/4$, and $\widehat{\tau}_{\min} = \widehat{\tau}_{\max}$ is the usual HL estimate. See Rosenbaum (1993, 2002, sec. 4.3.4) for details and an example.

3. HETEROGENEITY AND CAUSALITY IN OBSERVATIONAL STUDIES

3.1 Power of a Sensitivity Analysis

Suppose (i) in a particular context, randomization is unethical or infeasible; (ii) an observational study is being planned to estimate a treatment effect, τ; (iii) observed covariates are controlled by matching; and (iv) the study will include a sensitivity analysis to address possible bias in treatment assignment due to unobserved covariates. In some settings, it is possible to collect additional data, for instance using multiple control groups, that shed light on plausible unobserved biases (Rosenbaum 2002, secs. 6–9; 2004), but suppose that no such additional data are available for this study, so nothing in the observed data will indicate the magnitude Γ of bias that is present.

If the treatment were effective, say with constant effect $\tau > 0$, and if there were no unobserved bias, $\Gamma = 1$, then we could not be certain of this from the empirical data. We would see that the matched pair differences, D_i, were often positive, but this could be produced by an effect $\tau > 0$ with no bias $\Gamma = 1$, or by no effect $\tau = 0$ with bias, $\Gamma > 1$, or by various combinations. We hope to design the study so that, if the treatment did work, $\tau > 0$, and if there were no unobserved bias, $\Gamma = 1$, then at least the sensitivity analysis would report back that the ostensible effect of the treatment is quite insensitive to bias, that only a large bias $\Gamma > 1$ could explain away the ostensible effect of the treatment.

As a motivating illustration, I created two simulated observational studies, both with constant treatment effect $\tau = \frac{1}{2}$ and, unknown to us, with no unobserved bias. One simulated study, say LM, was larger and more heterogeneous study, with $I = 400$ pairs, and $D_i \sim_{\text{iid}} N\left(\frac{1}{2}, 1\right)$. The second study, say SL, was smaller and less heterogeneous, with $I = 100$ pairs and $D_i \sim_{\text{iid}} N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$. The mean of the I differences is $N\left\{\frac{1}{2}, \frac{1}{400}\right\}$ in both studies. To repeat, there is one sample of $I = 400$ pairs and one sample of $I = 100$ pairs, and these two simulated samples will now be analyzed using the methods in Section 2.3. If we knew there was no hidden bias, if LM and SL were randomized experiments, we would use the conventional randomization distribution of W_{τ_0}, obtaining very similar 95% confidence intervals for τ from the two studies, $[.40,\ .60]$ for LM and $[.43,\ .62]$ for SL, very similar HL point estimates $\widehat{\tau}$ of .50 for LM and .52 for SL, and fairly similar standardized Normal deviates using W_0 to test $H_0 : \tau = 0$, namely 8.96 for LM and 7.59 for SL. However, if LM and SL were observational studies, then SL would be much less sensitive to unobserved bias than LM. Specifically, the upper bound, $\Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$ in (2) is above .05 for $\Gamma \geq 2.5$ for LM, but the upper bound $\Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$ is less than .05 for $\Gamma \leq 6$ for SL. For $\Gamma = 2$, the range of HL point estimates $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ is $[.19,\ .81]$ for LM and $[.37,\ .67]$ for SL. The SL study is less sensitive to unobserved bias than Hammond's (1964) study of smoking and lung cancer, but LM is much more sensitive.

Is this simulated illustration typical? Is this pattern expected in general? These questions can be answered in various ways, the simplest being in terms of the power of the sensitivity analysis; for example, Rosenbaum (2004). A conventional power calculation determines the critical value, say $\widetilde{w}$, of a statistic, say W_{τ_0}, under certain assumptions, namely under a null hypothesis; then it determines the chance that $W_{\tau_0} \geq \widetilde{w}$ under different assumptions, namely an alternative hypothesis. This is true, also, for the power of a sensitivity analysis. In testing $H_0 : \tau = \tau_0$ versus $H_A : \tau > \tau_0$ at the .05 level, we reject H_0 for all π_{ij} satisfying (1) for specified $\Gamma > 1$ if $W_{\tau_0} \geq \widetilde{w}$, where $.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$ in (2). If the treatment did work in the sense that $\tau > \tau_0$, and if there were no unobserved bias, in the sense that $\Gamma = 1$, then the nonnull distribution of W_{τ_0} is approximately Normal, and the chance that $W_{\tau_0} \geq \widetilde{w}$ is readily approximated using standard results (Lehmann 1998, sec. 4.2) which assume a model for $\mathcal{F}$, commonly that $\epsilon_i/\sigma \sim_{\text{iid}} F(\cdot)$ with F continuous and symmetric about zero.

3.2 Heterogeneity and Causality: Numerical Comparisons of Power

Unknown to us, the treatment worked, with actual effect τ in a study without unobserved bias. Would we then be in a position to report that our results could not be explained away by small unobserved biases? Would we be in a position to say that a bias of magnitude, say, $\Gamma = 2$ could not explain away the observed difference in responses of treated and control subjects?

Table 1. Power of the Sensitivity Analysis Under Various Assumptions

Errors	I Matched Pairs	τ	σ	σ^2/I	Power $\Gamma = 1$	Power $\Gamma = 1.5$	Power $\Gamma = 2$
Normal	120	$\frac{1}{2}$	1	1/120	1.00	.96	.60
Normal	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	1.00	1.00	.96
Logistic	120	$\frac{1}{2}$	1	1/120	.93	.31	.04
Logistic	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	.93	.61	.32
Cauchy	200	$\frac{1}{2}$	1	1/200	.98	.32	.02
Cauchy	50	$\frac{1}{2}$	$\frac{1}{2}$	1/200	.95	.60	.28

Table 1 evaluates the power of the randomization test, $\Gamma = 1$, and the power of the sensitivity analysis with $\Gamma = 1.5$ and $\Gamma = 2$, in three pairs of situations. The pairs have Normal, logistic, and Cauchy error distributions, $F(\cdot)$. In each of the three pairs of situations, one situation has four times as many matched pairs I, whereas the other has the scale of the errors σ reduced in half, so σ^2/I remains the same in each pair. In all situations, the treatment effect is the same, $\tau = \frac{1}{2}$.

In the pair of Normal distributions in Table 1, the first situation has $I = 120$ pairs and $\sigma = 1$, so $D_i \sim_{\text{iid}} N\left(\frac{1}{2}, 1\right)$, $i = 1, \ldots, 120$, but the second has $I = 30$ pairs and $\sigma = \frac{1}{2}$, so $D_i \sim_{\text{iid}} N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$, $i = 1, \ldots, 30$, so the mean difference, $(1/I)\sum_{i=1}^{I} D_i$ is $N\left(\frac{1}{2}, \frac{1}{120}\right)$ in both situations.

In an experiment, randomization would ensure $\Gamma = 1$, and the power would be close to 1 in both of the Normal situations. In an observational study, because treatments were not randomly assigned, one would conduct a sensitivity analysis. With $\Gamma = 2$ the chance that the upper bound on the significance level in (2) is less .05 is 60% for ($I = 120$, $\sigma = 1$), but it is 96% for ($I = 30$, $\sigma = \frac{1}{2}$). In terms of sensitivity to unobserved bias, it is better to reduce the heterogeneity σ than to increase the sample size I. The same pattern is seen with the logistic and Cauchy distributions. Indeed, in both cases, for $\Gamma = 2$, $\sigma = 1$, the power is less than the level of the test, .05.

This is in harmony with the claims of both Mill and Fisher. In a randomized experiment, $\Gamma = 1$, bias has been controlled by randomization, and it matters little whether heterogeneity is reduced or the sample size is increased. In contrast, when bias from nonrandom assignment is possible, sharper conclusions are possible from a smaller, less heterogeneous study than from a larger, more heterogeneous study—that is, in the former case, the conclusions would be less sensitive to unobserved bias, so it would take a larger bias Γ to explain away the same effect τ.

3.3 Large Sample Sensitivity of Point Estimates

This section considers the uncertainty about τ that remains after sampling uncertainty has been driven out by letting $I \to \infty$, that is, the uncertainty from unknown, biased treatment assignments satisfying (1). This is a different perspective than Section 3.2, but the conclusions are similar.

As in Section 2.3, for each $\Gamma \geq 1$, there is a range of point estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$, where $\widehat{\tau}_{\min} = \widehat{\tau}_{\max} = \widehat{\tau}$ for $\Gamma = 1$. For each $\Gamma \geq 1$, as $I \to \infty$, this range, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$, converges in probability to a interval, $[\tau_{\min}, \tau_{\max}]$, with $\tau_{\max} = \tau_{\min}$ if $\Gamma = 1$ and $\tau_{\max} > \tau_{\min}$ if $\Gamma > 1$. If (1) *were true* with $\Gamma = 1$, then $\tau = \tau_{\max} = \tau_{\min}$; that is, the HL estimate $\widehat{\tau} = \widehat{\tau}_{\min} = \widehat{\tau}_{\max}$ is consistent for τ in a randomized experiment. If (1) *were true* with a specific $\Gamma > 1$, then $\tau \in [\tau_{\min}, \tau_{\max}]$, but the uncertainty about $\boldsymbol{\pi}$ prevents a more precise statement even as $I \to \infty$.

The length of the limiting interval $[\tau_{\min}, \tau_{\max}]$ is strongly affected by the heterogeneity of the experimental units. Let $\Phi(\cdot)$ and $\Upsilon(\cdot)$ be, respectively, the standard Normal and standard Cauchy cumulative distributions. Proposition 1 indicates what a sensitivity analysis yields, as $I \to \infty$, when, unknown to us, there actually is no unobserved bias. The proof is in the appendix.

Proposition 1. If $(D_i - \tau)/\sigma \sim_{\text{iid}} \Phi(\cdot)$ then $[\tau_{\min}, \tau_{\max}]$ is $\tau \pm \sigma \Phi^{-1}(\theta)/\sqrt{2}$, where $\theta = \Gamma/(1+\Gamma)$. If $(D_i - \tau)/\sigma \sim_{\text{iid}} \Upsilon(\cdot)$ then $[\tau_{\min}, \tau_{\max}]$ is $\tau \pm \sigma \Upsilon^{-1}(\theta)$.

Proposition 1 confirms Mill's method. Proposition 1 describes the situation in which we would like to report as little sensitivity to bias as possible, because in fact the treatment worked with true effect τ and, unknown to us, there was no unobserved bias. In this situation, even after sampling variability has been driven out by letting $I \to \infty$, the uncertainty about unobserved bias in (1), as reflected in $[\tau_{\min}, \tau_{\max}]$, is directly proportion to σ, the heterogeneity in the matched pair differences. In light of this, Mill's fanatical effort to reduce σ is, indeed, directly relevant to the evidence about *causality*, and does not merely reduce the *standard error*.

4. REDUCING HETEROGENEITY IN OBSERVATIONAL STUDIES: EXAMPLES

How can heterogeneity be reduced in observational studies? This section describes relevant features of a few clever observational studies. In experiments, pairing reduces heterogeneity and randomization prevents bias, but in observational studies, pairing reduces both bias and heterogeneity. Empirical examples differ in an important way from the theory in Section 2.. The point made in Section 2. was that even if the heterogeneity was purely random, even if it introduced no bias, we would not know this in an observational study, and reducing heterogeneity reduced sensitivity to unobserved bias, yielding measurably firmer conclusions. In examples, we do not know if pairing reduced bias or just sensitivity to bias; however, both are desirable and produced by the same tactics.

4.1 Road Hazards

What permanent road hazards increase the risk of fatal

collisions with roadside objects? Besides permanent road hazards, many factors affect accident risk: (i) the driver's skill, aggressiveness, risk tolerance, and sobriety; (ii) the weather, ice, snow, rain, fog; (iii) ambient light; (iv) safety equipment, brakes, tires, traction control, air bags; and (v) use of safety equipment, wearing seat belts. These factors are not unrelated. Sobriety may be more common at noon than at midnight, so sobriety and ambient light may vary together. In the rain or the snow, perhaps one drives on the highway to work, but not on the dirt road to the picnic area or the hiking trail, so weather and road features may vary together. To study permanent road hazards, one wants to compare different road hazards with the same driver, in the same car, in the same weather, with the same ambient light, in the same state of sobriety, with seat belts in the same state of use. Is this possible?

To a close approximation, it is. Wright and Robertson (1976) used the following simple, clever design. They examined 300 fatal accidents in Georgia in 1974 and 1975 in which collision with a roadside object—for example, trees, embankments, ditches, utility poles, guardrails, etc.—was a substantial factor in the fatality. They inspected the crash site and recorded permanent road features. The 300 accidents were compared to 300 nonaccidents involving the same driver, car, weather, light, belting, and sobriety, again recording road features. The nonaccidents occurred one mile back on the same road, a location passed by the driver minutes earlier en route to the crash site. At crash sites, when contrasted with control sites, they found a substantial excess of roads that curved more than six degrees with downhill gradients greater than two percent. [This is, technically, one of Maclure's (1991) case-crossover studies, but an unusual one, defined by geography rather than time, requiring the dual sensitivity analysis in Gastwirth et al. (1998).]

4.2 Reducing Heterogeneity in Economics

What are the economic returns to additional education? The question is not answered by comparing the earnings of high school dropouts and college graduates; these groups differed as children, before high school, in terms of parents' wealth and education, and possibly in terms of genetic endowment. One would like to compare children of the same parents, growing up in the same home at the same time, with the same genes. Comparing identical twins with differing education, Ashenfelter and Rouse (1998) estimated about a 9% increase in earnings from an additional year of schooling.

Legislation that regulates business practices—for instance, minimum wage legislation—typically applies in the same way throughout some region, such as a state, for a period of time. Some economic analyses have compared all businesses in one state to all businesses in another, but one would like to compare nearly identical businesses in states that have different regulations. How can one find nearly identical businesses in different states? In their study of the effects on employment of New Jersey's 19% increase in the minimum wage in 1992, Card and Krueger (1994) compared changes in employment at fast food restaurant chains in New Jersey and eastern Pennsylvania, comparing Burger Kings to Burger Kings, Wendy's to Wendy's, and so on.

4.3 Motorcycle Helmets

To what extent, if at all, do helmets reduce the risk of death in motorcycle crashes? Different crashes occur on different motorcycles, at different speeds, with different forces, on highways or country roads, in dense or light traffic, encountering deer or Hummers. One would like to compare two people, one with a helmet, the other without, on the same type of motorcycle, riding at the same speed, on the same road, in the same traffic, crashing into the same object. Is this possible? It is when two people ride the same motorcycle, a driver and a passenger, one helmeted, the other not. Using data from the Fatality Analysis Reporting System, Norvell and Cummings (2002) performed such a matched pair analysis using a conditional model with numerous pair parameters, estimating approximately a 40% reduction in risk associated with helmet use.

5. SUMMARY

If treatments are randomly assigned, so treatment effects may be estimated without bias, increasing the sample size and decreasing the heterogeneity of experimental units have similar consequences: they reduce the sampling variability of unbiased estimates. The situation is different in observational studies, where estimates of effects may be biased by nonrandom selection into treated and control groups. In observational studies, reducing heterogeneity reduces both sampling variability and sensitivity to unobserved bias—with less heterogeneity, larger biases would need to be present to explain away the same effect. In contrast, increasing the sample size reduces sampling variability, which is, of course useful, but it does little to reduce concerns about unobserved bias. Several observational studies illustrated tactics used to eliminate major sources of heterogeneity.

APPENDIX: PROOF OF PROPOSITION 1

Proof: Write $H_{\tau_0} = 2W_{\tau_0}/\{I(I+1)\}$, and note that $\{I(I-1)\}/\{I(I+1)\} \to 1$ as $I \to \infty$. If the D_i are iid from a continuous distribution, then letting $I \to \infty$ in a standard result (e.g., Lehmann 1998, expression 4.25, p. 165) yields $E(H_{\tau_0}) \to \Pr\{(D_i + D_j)/2 > \tau_0\} = \lambda(\tau_0)$, say, as $I \to \infty$. If $(D_i - \tau)/\sigma \sim_{iid} \Phi(\cdot)$, then $\lambda(\tau_0) = \Phi\left\{\sqrt{2}(\tau - \tau_0)/\sigma\right\}$, so $\lambda(\tau_0)$ is continuous and strictly decreasing as a function of τ_0; moreover, for each τ_0, H_{τ_0} converges in probability to $\lambda(\tau_0)$ as $I \to \infty$. As noted in Section 2.3, $(1-\theta) \le E_{\boldsymbol{\pi}}(H_\tau \mid \mathcal{Z}, \mathcal{F}) \le \theta$ for every $\boldsymbol{\pi}$ satisfying (1), and $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ is found as the solutions to $H_{\widehat{\tau}_{\min}} = \theta$ and $H_{\widehat{\tau}_{\max}} = 1-\theta$. It follows that $\widehat{\tau}_{\min} \to \tau_{\min}$ and $\widehat{\tau}_{\max} \to \tau_{\max}$ in probability where $\theta = \Phi\left\{\sqrt{2}(\tau - \tau_{\min})/\sigma\right\}$ and $1-\theta = \Phi\left\{\sqrt{2}(\tau - \tau_{\max})/\sigma\right\}$ and rearranging gives $[\tau_{\min}, \tau_{\max}]$ as $\tau \pm \sigma\,\Phi^{-1}(\theta)/\sqrt{2}$. In parallel, for the Cauchy, if $(D_i - \tau)/\sigma \sim_{\text{iid}} \Upsilon(\cdot)$, then $\lambda(\tau_0) = \Pr\{(D_i + D_j)/2 > \tau_0\} = \Upsilon\{(\tau - \tau_0)/\sigma\}$ and the rest of the argument is the same.

[Received April 2004. Revised January 2005.]

REFERENCES

Ashenfelter, O., and Rouse, C. (1998), "Income, Schooling and Ability: Evidence From a New Sample of Identical Twins," *Quarterly Journal of Economics*, 113, 253–284.

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.

Copas, J., and Eguchi, S. (2001), "Local Sensitivity Approximations for Selectivity Bias," *Journal of the Royal Statistical Society*, Ser. B, 63, 871–896.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959), "Smoking and Lung Cancer," *Journal of the National Cancer Institute*, 22, 173–203.

Fisher, R. A. (1935, 1949), *Design of Experiments*, Edinburgh: Oliver and Boyd.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998), "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika*, 85, 907–920.

Hammond, E. C. (1964), "Smoking in Relation to Mortality and Morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

Hodges, J. L., and Lehmann, E. L. (1963), "Estimates of Location Based on Ranks," *Annals of Mathematical Statistics*, 34, 598–611.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

Jick, H., Miettinen, O., Neff, R., et al. (1973), "Coffee and Myocardial Infarction," *New England Journal of Medicine*, 289, 63–77.

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

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.

Maclure, M. (1991), "The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events," *American Journal of Epidemiology*, 133, 144–152.

Mill, J. S. (1867), *A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation*, New York: Harper & Brothers.

Neyman, J. (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments," *Statistical Science*, 5, 463–480.

Norvell, D. C., and Cummings, P. (2002), "Association of Helmet Use with Death in Motorcycle Crashes: A Matched-Pair Cohort Study," *American Journal of Epidemiology*, 156, 483–487.

Rosenbaum, P. R. (1988), "Sensitivity Analysis for Matching with Multiple Controls," *Biometrika*, 75, 577–581.

——— (1993), "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies," *Journal of the American Statistical Association*, 88, 1250–1253.

——— (1999), "Choice as an Alternative to Control in Observational Studies" (with discussion), *Statistical Science*, 14, 259–304.

——— (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag.

——— (2003), "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test," *The American Statistician*, 57, 132–138.

——— (2004), "Design Sensitivity in Observational Studies," *Biometrika*, 91, 153–164.

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

Wright, P. H., and Robertson, L. S. (1976), "Priorities for Roadside Hazard Modification," *Traffic Engineering*, 46, 24–30.