HETEROGENEITY AND CAUSALITY: UNIT HETEROGENEITY AND DESIGN SENSITIVITY IN OBSERVATIONAL STUDIES

PAUL R. ROSENBAUM

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

1. NOTATION AND REVIEW

1.1. **Review: Randomization Inference in a Paired Experiment.** Observed covariate $\mathbf{x}$ and an unobserved covariate u. I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, one control, matched for $\mathbf{x}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but not matched for u, so typically $u_{i1} \neq u_{i2}$. $Z_{ij} = 1$ if j received the treatment in pair i, and $Z_{ij} = 0$ if j received the control, so $Z_{i1} + Z_{i2} = 1$. Subject (i, j) has two potential responses, (r_{Tij}, r_{Cij}), r_{Tij} observed under treatment, $Z_{ij} = 1$, r_{Cij} observed under control, $Z_{ij} = 0$, so the effect of the treatment is $r_{Tij} - r_{Cij}$; Neyman [16] and Rubin [25]. 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 [5] theory of randomization inference. Randomization within pairs ensures $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \frac{1}{2}$, $i = 1, \ldots, I$, with independent assignments in distinct pairs. Observed response is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$. If treatment effect is constant, $\tau = r_{Tij} - r_{Cij}$, then $R_{ij} = r_{Cij} + Z_{ij}\tau$, and the treated-minus-control difference is $D_i = (2Z_{i1} - 1)(R_{i1} - R_{i2}) = \tau + \epsilon_i$ 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 [8] estimate $\widehat{\tau}$ of τ is (essentially) the solution to $W_{\widehat{\tau}} = I(I+1)/4 = \frac{1}{2}(1 + 2 + \ldots + I)$. 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 has $(r_{Ci1} - r_{Ci2})/\sigma \sim_{iid} F(\cdot)$ were $\sigma > 0$ and $F(\cdot)$ is continuous and symmetric about zero, so randomization ensures $\epsilon_i/\sigma \sim_{iid} F(\cdot)$.

June 2005. Supported by Grant SES-0345113 from NSF.

1.2. **Review: Sensitivity to Departures from Random Assignment in Observational Studies.** (i) In population, before matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$, (ii) subjects with same *observed* $\mathbf{x}_{ij}$ may differ in *unobserved* u_{ij} and hence in odds of treatment by factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ik})}{\pi_{ik}(1-\pi_{ij})} \leq \Gamma, \ \ \forall\, i, j, k \tag{1.1}$$

and (iii) the distribution of treatments within treated/control matched pairs $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F})$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$. Here, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$. If $\Gamma = 1$, then $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ 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: bias solely due to observed $\mathbf{x}$ can be eliminated by matching on $\mathbf{x}$. If $\Gamma > 1$ in (1.1), then matching on $\mathbf{x}$ may fail to equalize the π_{ij} in pair i. Γ is unknown. A sensitivity analysis calculates, for several values of Γ, the range of possible inferences. How large must Γ be before qualitatively different causal interpretations are possible?

1.3. **Review: Sensitivity Analysis with the Signed Rank Statistic.** If (1.1) 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 $\theta = \Gamma/(1+\Gamma)$ so $\theta \geq \frac{1}{2}$ because $\Gamma \geq 1$. Write $\overline{\overline{W}}$ for the sum of I independent random variables taking value i with probability θ 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 (1.1) and $H_0 : \tau = \tau_0$ imply the sharp bounds

$$\Pr\left(\overline{W} \geq w\right) \leq \Pr\left(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}\right) \leq \Pr\left(\overline{\overline{W}} \geq w\right), \ \ \forall w; \tag{1.2}$$

e.g., [18]. If $\Gamma = 1$, then equality in (1.2); otherwise bounds (1.2) widen as Γ increases. For $H_0 : \tau = \tau_0$ vs $H_A : \tau > \tau_0$, the upper bound on the one-sided significance level is at most 0.05 for all π_{ij} satisfying (1.1) if $W_{\tau_0} \geq \widetilde{w}$ where $0.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$.

For each $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{I2})$, there is an HL estimate $\widehat{\tau}_{\boldsymbol{\pi}}$ (essentially) solving $W_{\widehat{\tau}} = \mu_{\boldsymbol{\pi}}$ where the expectation $\mu_{\boldsymbol{\pi}} = E_{\boldsymbol{\pi}}(W_\tau \mid \mathcal{Z}, \mathcal{F})$ is computed using $\boldsymbol{\pi}$. Then (1.1) 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 $\Gamma = 1$, $\mu_{\boldsymbol{\pi}} = I(I+1)/4$, and $\widehat{\tau}_{\min} = \widehat{\tau}_{\max}$ is the usual HL estimate.

2. Heterogeneity and Sensitivity to Unobserved Bias

2.1. **Question.** In the fortunate situation, biases are confined to observed covariates, and adjustments remove these biases, yielding unbiased or consistent estimates of treatment effects. In an observational study, even if the fortunate situation arose, we would not know this from the data. In the fortunate situation, we hope to report insensitivity to small or moderate unobserved biases. In the fortunate situation, how does unit heterogeneity affect the degree of sensitivity to unobserved bias? That is, if the treatment actually worked, and there was no unobserved bias, would we be in a position to assert that there is fairly strong evidence that the treatment worked?

TABLE 1. Power of the Sensitivity Analysis Under Various Assumptions.

Errors	I Matched Pairs	τ	σ	$\frac{\sigma^2}{I}$	Power $\Gamma = 1$	Power $\Gamma = 1.5$	Power $\Gamma = 2$
Normal	120	$\frac{1}{2}$	1	1/120	1.00	0.96	0.60
Normal	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	1.00	1.00	0.96
Logistic	120	$\frac{1}{2}$	1	1/120	0.93	0.31	0.04
Logistic	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	0.93	0.61	0.32
Cauchy	200	$\frac{1}{2}$	1	1/200	0.98	0.32	0.02
Cauchy	50	$\frac{1}{2}$	$\frac{1}{2}$	1/200	0.95	0.60	0.28

2.2. **Power of a sensitivity analysis.** For a fixed $\Gamma \geq 1$, the power of the sensitivity analysis is the probability that the upper bound on the significance level from (1.2) is less than, say, 0.05. Determine $\widetilde{w}$ so $0.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$ in (1.2); then calculate the probability that $W_{\tau_0} \geq \widetilde{w}$ under some specific alternative hypothesis. For $\Gamma = 1$, this is the usual concept of power. The alternatives considered here assume the fortunate situation: the treatment worked, with additive effect τ (but of course we don't know this), with errors $\epsilon_i/\sigma \sim_{iid} F(\cdot)$ that are Normal or logistic or Cauchy (but of course we don't know this), and there actually is no unobserved bias (but of course we don't know this either).

2.3. **Limit as $I \to \infty$.** Whether or not (1.1) is true, for each fixed $\Gamma \geq 1$, as $I \to \infty$, the range of possible HL estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$, 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 (1.1) *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 (1.1) *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 $I \to \infty$.

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 length of the limiting interval $[\tau_{\min}, \tau_{\max}]$ is strongly affected by the heterogeneity of the experimental units σ.

Proposition 1. [23] *If* $(D_i - \tau)/\sigma \sim_{iid} \Phi(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \sigma \Phi^{-1}(\theta)/\sqrt{2}$, *where* $\theta = \Gamma/(1+\Gamma)$. *If* $(D_i - \tau)/\sigma \sim_{iid} \Upsilon(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \sigma \Upsilon^{-1}(\theta)$.

3. ANNOTATED BIBLIOGRAPHY

KEY TO ANNOTATION: DS = design sensitivity, the two papers most closely related to the talk [23], [22]. SEN = methods of sensitivity analysis used in the talk [18], [19], [20], [21]. AS = alternative methods of sensitivity analysis [4], [24], [6], [13], [3], [10]. EG = illustrative examples with tactics to reduce heterogeneity [26], [2], [1], [17]. CE = causal effects [16], [25]. RI = randomization inference [5], [12], [20]. WSR = Wilcoxon's signed rank statistic [12], [21]. HL = Hodges-Lehmann estimate [8], [12], [19]. HIS = history of role of heterogeneity in causality: John Stuart Mill thought it important [15], [9], Fisher vehemently disagreed [5]. MISC = miscellaneous references [7], [14], [11].

References

[1] Ashenfelter, O. & Rouse, C. (1998), "Income, schooling and ability: Evidence from a new sample of identical twins," *Quart. J. .Econ.*, 113, 253-284. EG

[2] Card, D. & Krueger, A. (1994), "Minimum wages and employment," *Am. Econ. Rev.,* 84 772-793. EG

[3] Copas, J. & Eguchi, S. (2001), "Local sensitivity approximations for selectivity bias," *JRSS,* B 63, 871-96. AS

[4] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959), "Smoking and lung cancer," *J. Nat. Cancer Inst.,* 22, 173-203. AS

[5] Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd. RI, HIS

[6] Gastwirth, J. L., Krieger, A. M. & Rosenbaum, P. R. (1998), "Dual and simultaneous sensitivity analysis for matched pairs," *Biometrika*, 85, 907-920. AS

[7] Hammond, E. C. (1964), "Smoking in relation to mortality and morbidity," *J. Nat. Cancer Inst.*, 32, 1161–1188. MISC

[8] Hodges, J. L. & Lehmann, E. L. (1963), "Estimates of location based on ranks," *Ann. Math. Stat.,* 34, 598-611. HL

[9] Holland, P. W. (1986), "Statistics and causal inference," *JASA*, 81, 945-960. HIS

[10] Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation," *Am. Econ. Rev.*, 93, 126-132. AS

[11] Jick, H., et al. (1973), "Coffee and myocardial infarction," *NEJM*, 289, 63–77. MISC

[12] Lehmann, E. L. (1998), *Nonparametrics,* Upper Saddle River, NJ: Prentice Hall. RI, HL, WSR

[13] Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998), "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, 54, 948-963. AS

[14] Maclure, M. (1991), "The case-crossover design," *Am. J. Epidem.,* 133, 144-152. MISC

[15] Mill, J. S. (1867), *A System of Logic*, NY: Harper & Brothers. HIS

[16] Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Stat. Sci.*, 5, 463-480. CE

[17] Norvell, D. C. & Cummings, P. (2002), "Association of helmet use with death in motorcycle crashes: A matched-pair cohort study," *Am. J. Epidem.*, 156, 483-487. EG.

[18] Rosenbaum, P. R. (1988), "Sensitivity analysis for matching with multiple controls," *Biometrika*, 75, 577-81. SEN, WSR

[19] Rosenbaum, P. R. (1993), "Hodges-Lehmann point estimates of treatment effect in observational studies," *JASA,* 88 1250-1253. SEN, HL, WSR

[20] Rosenbaum, P. R. (2002), *Observational Studies* (2^{nd} ed). NY: Springer. SEN, RI, WSR, HL

[21] Rosenbaum, P. R. (2003), "Exact confidence intervals for nonconstant effects by inverting the signed rank test," *Am. Statist.*, 57, 132-138. SEN, WSR

[22] Rosenbaum, P. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164. DS

[23] Rosenbaum, P. R. (2005), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *Am. Statist.*, 59, 147-152. DS, WSR, HL, HIS

[24] Rosenbaum, P. R. & Rubin, D. B. (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *JRSS,* B **45**, 212–8. AS

[25] Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Ed. Psych.*, 66, 688-701. CE

[26] Wright, P. H., & Robertson, L. S. (1976), "Priorities for roadside hazard modification," *Traffic Engineering*, 46, 24-30. EG

Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104-6430 USA
E-mail address: rosenbaum@stat.wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~rosenbap/index.html