SOME COUNTERCLAIMS UNDERMINE THEMSELVES IN OBSERVATIONAL STUDIES

PAUL R. ROSENBAUM

ABSTRACT. Claims based on observational studies that a treatment has certain effects are often met with counterclaims asserting that the treatment is entirely without effect, that all associations with treatment are produced by biased treatment assignment. Some counterclaims undermine themselves in the following specific sense: presuming the counterclaim to be true may strengthen the support that the original data provide for the original claim, so that the counterclaim fails in its role as a critique of the original claim. In mathematics, a proof by contradiction supposes a proposition to be true en route to proving that the proposition is false. Analogously, the supposition that a particular counterclaim is true may justify an otherwise unjustified statistical analysis, and this added analysis may interpret the original data as providing even stronger support for the original claim. More precisely, the original study is sensitive to unmeasured biases of a particular magnitude, Γ, but an analysis that supposes the counterclaim to be true may be insensitive to much larger unmeasured biases, $\Gamma' > \Gamma$. Illustrated using data from the US Fatality Analysis Reporting System

1. NOTATION

There are I matched sets, $i \in \{1, \ldots, I\} = \mathcal{I}$, where set $i \in \mathcal{I}$ contains subjects $\mathcal{J}_i = \{1, \ldots, J_i\}$, one treated with $Z_{ij} = 1$, the rest untreated controls with $Z_{ij} = 0$, so $1 = \sum_{j \in \mathcal{J}_i} Z_{ij}$ for each i. Write $n = \sum_{i \in \mathcal{I}} J_i$ and $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots Z_{IJ_I})^T$, and let $\mathcal{Z}$ be the set containing the $\prod_{i \in \mathcal{I}} J_i$ possible values of $\mathbf{Z}$. Denote by $|\mathcal{A}|$ the number of elements in a finite set $\mathcal{A}$ so $|\mathcal{Z}| = \prod_{i \in \mathcal{I}} J_i$. For a matched pair, $\mathcal{J}_i = \{1, 2\}$ and $J_i = 2$. The example uses this notation 4 times for 4 parallel studies, for instance, with treatments belted drived versus unbelted passenger. Conditioning on $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. Subject ij has measured covariate $\mathbf{x}_{ij}$ and unmeasured covariate u_{ij}. Matching has controlled the measured covariate, so that $\mathbf{x}_{ij} = \mathbf{x}_{ik} = \mathbf{x}_i$, $\forall i, j, k$, but quite possibly $u_{ij} \neq u_{ik}$. Subject ij has two potential responses for the outcome of primary interest, r_{Tij} if assigned to treatment or r_{Cij} if assigned to control, so the observed response of ij is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$ and the effect of the treatment on ij, namely $r_{Tij} - r_{Cij}$ is not observed; see Neyman (1923) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$ for all ij. In the example, (r_{Tij}, r_{Cij}) records the injury scores that subject ij would suffer under treatment and control, R_{ij} is the injury ij actually suffered, and Fisher's H_0 says that swapping the treatments in pair i would not alter the injury suffered by individual ij. Write $\mathbf{R}$, $\mathbf{r}_C$, $\mathbf{r}_T$, and $\mathbf{u}$ for the n dimensional vectors. Each subject may have a K-dimensional row vector of secondary outcomes, $\mathbf{s}_{Tij}$ or $\mathbf{s}_{Cij}$, with observed value $\mathbf{S}_{ij} = Z_{ij} \mathbf{s}_{Tij} + (1 - Z_{ij}) \mathbf{s}_{Cij}$, and associated $n \times K$ matrices $\mathbf{S}$, $\mathbf{s}_C$ and $\mathbf{s}_T$ whose rows are in the lexical order. In one FARS example, the secondary outcome $\mathbf{S}$ indicates whether exactly one of the two people was ejected. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{s}_{Tij}, \mathbf{s}_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, \ldots, J_i\}$.

The subscripts ij are unique but noninformative identifiers, perhaps randomly assigned, and all information about individual ij is in observed or unobserved variables that describe ij. A matched pair, $J_i = 2$, yields a single treated-minus-control pair difference $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ in outcomes, as plotted in various Figures, and if H_0 is true, $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$.

2. INFERENCE: TESTING THE HYPOTHESIS OF NO EFFECT

2.1. **Randomization inference in randomized experiments.** In a randomized experiment, one individual in each set is picked at random for treatment independently in distinct matched sets, so

$$(2.1) \qquad \Pr\left(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}\right) = \prod_{i \in \mathcal{I}} J_i^{-1} = |\mathcal{Z}|^{-1} \text{ for each } \mathbf{z} \in \mathcal{Z}.$$

If $t(\mathbf{Z}, \mathbf{R})$ is a test statistic, then in a randomized experiment (2.1), the distribution of $t(\mathbf{Z}, \mathbf{R})$ under Fisher's null hypothesis H_0 of no effect equals its permutation distribution $\Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F},\, \mathcal{Z}\} = |\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}| / |\mathcal{Z}|$, because $\mathbf{R} = \mathbf{r}_C$ when H_0 is true, $\mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$, and $\mathbf{Z}$ is uniform on $\mathcal{Z}$ in a randomized experiment. In an observational study, the counterclaim of selection bias says that the treatment is entirely without effect and $t(\mathbf{Z}, \mathbf{R})$ is large because (2.1) is false.

2.2. **Sensitivity analysis in observational studies.** One model for studying the sensitivity of conclusions to violations of (2.1) says that, in the population prior to matching, treatment assignments are independent and two subjects with the same observed covariates may differ in their odds of treatment, $Z_{ij} = 1$, by at most a factor of $\Gamma \geq 1$; then, the distribution of $\mathbf{Z}$ is returned to $\mathcal{Z}$ by conditioning on $\mathbf{Z} \in \mathcal{Z}$. This is equivalent to assuming that there is an unobserved covariate u_{ij} with $0 \leq u_{ij} \leq 1$ such that

$$(2.2) \qquad \Pr\left(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}\right) = \prod_{i \in \mathcal{I}} \frac{\exp\left(\gamma \sum_{j \in \mathcal{J}_i} z_{ij}\, u_{ij}\right)}{\sum_{j \in \mathcal{J}_i} \exp\left(\gamma\, u_{ij}\right)} = \frac{\exp\left(\gamma \mathbf{z}^T \mathbf{u}\right)}{\sum_{\mathbf{b} \in \mathcal{Z}} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)},\ \mathbf{u} \in [0,1]^n,$$

for each $\mathbf{z} \in \mathcal{Z}$, where $\gamma = \log(\Gamma) \geq 0$; see Rosenbaum (2002, §4.2). For each $\mathbf{u} \in [0,1]^n$, the null distribution of $t(\mathbf{Z}, \mathbf{R})$ under Fisher's H_0 is obtained by summing terms (2.2) over $\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}$. As $\mathbf{u}$ is allowed to range over $[0,1]^n$, the sensitivity analysis determines bounds on this null distribution, yielding for instance the upper and lower bounds on P-values testing Fisher's hypothesis H_0 of no effect. This method with point estimates and confidence intervals is implemented for M-statistics, including the permutational t-test, in the `senm` and `senmCI` functions of the `sensitivitymult` package in R; see Rosenbaum (2007, 2013). The first method of sensitivity analysis is due to Cornfield, et al. (1959).

3. SEGMENTS OF DATA

3.1. **Segments of the data determined by a matrix W.** A segment of the data $\{\mathcal{J}_i,\, i \in \mathcal{I}\}$ is defined to be $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ where $\mathcal{J}_i^{'} \subseteq \mathcal{J}_i$ for each $i \in \mathcal{I}$. Let $\mathfrak{S}$ be the set whose 2^n elements are the 2^n possible segments. For a segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$, write m_i for the random variable that counts the number of treated subjects in $\mathcal{J}_i^{'}$, so $m_i = 0$ if $\mathcal{J}_i^{'} = \emptyset$ and otherwise $m_i = \sum_{j \in \mathcal{J}_i^{'}} Z_{ij}$, so $m_i = 0$ or $m_i = 1$. Write $\mathbf{m} = (m_1, \ldots, m_I)$. The contribution from $\mathcal{J}_i^{'}$ in segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ will be degenerate and uninteresting unless $m_i = 1 < \left|\mathcal{J}_i^{'}\right|$. For matched pairs, $|\mathcal{J}_i| = J_i = 2$ for

all i as in the example, the interesting (or nondegenerate) part of a segment $\left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ is simply a subset of the matched pairs. For matched sets with $|\mathcal{J}_i| = J_i > 2$, a segment $\left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ may have nondegenerate parts $\mathcal{J}_i'$ with $m_i = 1 < \left|\mathcal{J}_i'\right| < |\mathcal{J}_i|$ containing the treated subject from $\mathcal{J}_i$ and some but not all of the controls from $\mathcal{J}_i$. For a segment $\left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$, add a prime $'$ to a quantity to denote the value of a quantity confined to the segment, eg $\mathbf{Z}'$ for the vector of Z_{ij} with $i \in \mathcal{I}$ and $j \in \mathcal{J}_i'$, or $n' = \sum_{i \in \mathcal{I}} \left|\mathcal{J}_i'\right|$. Write $\mathcal{Z}_{\mathbf{m}}'$ for the set of possible values of $\mathbf{Z}'$, that is, the set of vectors $\mathbf{z}'$ of dimension n' with 1 or 0 coordinates such that $m_i = \sum_{j \in \mathcal{J}_i'} z_{ij}$. In parallel, write $\mathbf{r}_C'$, $\mathbf{S}'$, etc. Conditioning on the event $\mathbf{Z}' \in \mathcal{Z}_{\mathbf{m}}'$ is abbreviated as conditioning on $\mathcal{Z}_{\mathbf{m}}'$, and generally the conditioning will be on $(\mathcal{Z}, \, \mathcal{Z}_{\mathbf{m}}', \, \mathbf{m})$ jointly.

There is an $n \times M$ matrix $\mathbf{W}$ with rows $\mathbf{w}_{ij}$ in the lexical order. Write $\mathcal{W}$ for the set of possible values for $\mathbf{W}$.

Definition 1. *The phrase "$\mathbf{W}$ determines the segment" means that there is a known function $\mathcal{S}(\mathbf{W})$ that receives $\mathbf{W}$ and returns a segment from $\mathfrak{S}$, that is, $\mathcal{S} : \mathcal{W} \to \mathfrak{S}$.*

This definition 1 needs to be guarded from a natural misinterpretation: Unless $\mathbf{W}$ determines $\mathbf{Z}$, a segment determined by $\mathbf{W}$ cannot make use of the identity of the treated subject.

3.2. **Segments and sensitivity analysis.** When can we select a segment $\left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ using $\mathbf{W}$, yet appropriately analyze this segment as if were an unselected data set? Proposition provides a condition on the segment $\mathcal{S}(\mathbf{W}) = \left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ such that the distribution of treatment assignments in the segment is nothing more than a distribution with the same form as (2.2) confined to the segment by conditioning on $\mathbf{m}$.

Proposition 1. *If a segment $\mathcal{S}(\mathbf{W}) = \left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ is determined by $\mathbf{W}$ in the sense of Definition 1, and if $\mathbf{W}$ is fixed by conditioning on $\mathcal{F}$, then (2.2) implies the distribution of treatment assignments $\mathbf{z}' \in \mathcal{Z}_{\mathbf{m}}'$ within the segment is given by*

$$(3.1) \qquad \Pr\left(\mathbf{Z}' = \mathbf{z}' \mid \mathcal{F}, \, \mathcal{Z}, \, \mathcal{Z}_{\mathbf{m}}', \, \mathbf{m}\right) = \prod_{i \in \mathcal{I} : \left|\mathcal{J}_i'\right| > 0} \frac{\exp\left(\gamma \sum_{j \in \mathcal{J}_i'} z_{ij}' \, u_{ij}\right)}{\sum_{j \in \mathcal{J}_i'} \exp\left(\gamma \, u_{ij}\right)}, \, \mathbf{u}' \in [0,1]^{n'}.$$

Corollary 1. *If a segment $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i', \, i \in \mathcal{I}\right\}$ is determined by the observed value of the supplementary responses $\mathbf{S}$, and if the supplementary responses are unaffected by the treatment, $\mathbf{s}_{Tij} = \mathbf{s}_{Cij}$ for all ij, then (2.2) implies the distribution of treatment assignments within the segment is given by (3.1).*

4. HUBER-MARITZ M-STATISTICS

In testing H_0, the Huber-Maritz M-statistic uses $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} \psi(Y_i/s)$ where s is the 95% quantile of $|Y_i|$, and $\psi(\cdot)$ is an odd function, $\psi(y) = -\psi(-y)$. Here, $\psi_{\mathrm{t}}(y) = y$ yields the permutational t-test, $\psi_{\mathrm{hu}}(y) = \mathrm{sign}(y) \min(|y|, 1)$ are Huber's scores, and $\psi_{\mathrm{in}}(y) = \mathrm{sign}(y) \max\left\{0, \, \min(|y|, 1) - \frac{1}{4}\right\}$ performs inner trimming. Under H_0, $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \pm(r_{Ci1} - r_{Ci2})$, $|Y_i|$ and hence s are fixed by conditioning on $\mathcal{F}$, so $\psi(Y_i/s) = \pm\psi(|r_{Ci1} - r_{Ci2}|/s)$. The upper bound on the distribution of $\sum_{i=1}^{I} \psi(Y_i/s)$ under (2.2) is the sum of I independent random variables taking

value $\psi\left(\left|r_{Ci1} - r_{Ci2}\right|/s\right)$ with probability $\Gamma/\left(1 + \Gamma\right)$ and value $-\psi\left(\left|r_{Ci1} - r_{Ci2}\right|/s\right)$ with probability $1/\left(1 + \Gamma\right)$, reducing to the randomization distribution for $\Gamma = 1$; see Rosenbaum (2007). For the design sensitivity of M-statistics and $\psi_{\mathrm{in}}\left(y\right)$, see Rosenbaum (2013). If $Y_i \sim_{\mathrm{iid}} N\left(0.5, 1\right)$, one standard implementation yields design sensitivity $\widetilde{\Gamma} = 3.5$ for $\psi_{\mathrm{t}}\left(y\right)$, $\widetilde{\Gamma} = 3.3$ for $\psi_{\mathrm{hu}}\left(y\right)$ and $\widetilde{\Gamma} = 4$ for $\psi_{\mathrm{in}}\left(y\right)$, so $\psi_{\mathrm{in}}\left(y\right)$ reports greater insensitivity in large samples; see Rosenbaum (2013, Table 3).

References

Cornfield, J. et al. (1959), "Smoking and lung cancer," *J. Nat. Cancer Inst.*, 22, 173-2.

Evans, L. (1986), "Effectiveness of safety belts in preventing fatalities," *Accid. Anal. Prevent.*, 18, 229-41.

Fatality Analysis Reporting System, National Highway Traffic Safety Administration, `www.nhtsa.gov`

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

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013), "Effect modification and design sensitivity in observational studies," *J. Am. Statist. Assoc.*, 108, 135-148.

Huber, P. (1980), *Robust Statistics*, New York: John Wiley.

Levitt, S. D. and Porter, J. (2001), "Sample selection in the estimation of air bag and seat belt effectiveness," *Rev. Econ. Stat.*, 83, 603-615.

Maritz, J. S. (1979), "On exact robust confidence intervals for location," *Biometrika*, 66, 163-66. (`R: senmCI` in `sensitivitymult`)

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

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

Rosenbaum, P. R. (2002), *Observational Studies* (2nd edition), New York: Springer.

Rosenbaum, P. R. (2005), "Attributable effects in case^2-studies," *Biometrics*, 61, 246-253.

Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, 63, 456-64. (`R: senm` and `senmCI` in `sensitivitymult`)

Rosenbaum, P. R., Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *J. Am. Statist. Assoc.*, 104, 1398-1405. (Interpretation of Γ, `R: amplify` in `sensitivitymult`)

Rosenbaum, P. R. (2013), "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics*, 69, 118-27.

Rosenbaum, P. R. (2015), "Some counterclaims undermine themselves in observational studies," *J. Am. Statist. Assoc.* 110, 1389-1398.

Rosenbaum, P. R. (2017), *Observation and Experiment*, Cambridge, MA: Harvard University Press.

Rosenbaum, P. R. (2018), "Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels," *Ann. Appl. Statist.*, to appear. `R: senstrat`.

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

Sivak, M., Schoettle, B., Reed, M., Flannagan, M. (2006), "Influence of visibility out of the vehicle cabin on lane-change crashes," *Accident Anal. Prevent.*, 38, 969-72.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002), "Trucated product method of combining P-values," *Genetic Epidemiol.*, 22, 170-85. (`R: truncatedP` in `sensitivitymv`)

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104 US
E-mail address: `rosenbaum@wharton.upenn.edu`
URL: `http://www-stat.wharton.upenn.edu/~rosenbap/`