An Observational Study Used to Illustrate Methodology

Paul Rosenbaum, Wharton, U of Pennsylvania

The Fisher Lecture was based on [8,9,12,17] and [11, §6]. These differ in details documented in the articles but not emphasized in the presentation.

What is matching with fine balance? Constrains an optimal (i.e., minimum distance) match to exactly balance the marginal distributions of a nominal covariate, without restricting who is matched to whom. A tool in a toolbox, used with: propensity scores, covariate distances, directional penalties.

Optimal assignment [1] Pairs T rows to T distinct columns in a $T \times C$ distance matrix, $C \geq T$, so the total of the T within-pair distances is minimized. There are $C!/\left(C-T\right)!$ possible pairings, but the best can be found in $O\left(C^3\right)$ arithmetic steps.

Simple implementation of minimum-distance fine-balance. Add $C-T$ rows, making a $C \times C$ matrix, adding 0's and ∞'s to remove required numbers from the control group, leaving behind marginal balance. Still requires $O\left(C^3\right)$ arithmetic steps. Network implementation makes more efficient use of space.

Fine balance: references [8], extensions [6, 19,20,22], R packages Pimentel's `rcbalance`, Yu's `DiPs` and `bigmatch`, Zubizarreta's `designmatch`.

Notation Covariate $(\mathbf{x}, u)$, with $\mathbf{x}$ observed, u unobserved. I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, one control, $Z_{ij} = 0$, matched so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ but perhaps $u_{i1} \neq u_{i2}$. Potential responses (r_{Tij}, r_{Cij}), r_{Tij} observed under treatment, $Z_{ij} = 1$, r_{Cij} observed under control, $Z_{ij} = 0$, so $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$ is observed but the causal effect $r_{Tij} - r_{Cij}$ is not observed [5,16]. 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 $\{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\}$. Randomization [3] would ensure $\Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right) = \frac{1}{2}$, $i = 1, \ldots, I$. Fisher's hypothesis of no effect is $H_0 : r_{Tij} = r_{Cij}$, $\forall i, j$. Treated-minus-control pair i difference is $D_i = \left(2Z_{i1} - 1\right)\left(R_{i1} - R_{i2}\right)$, so that $D_i = \left(2Z_{i1} - 1\right)\left(r_{Ci1} - r_{Ci2}\right)$ if H_0 is true.

Two statistics Let q_i be the rank of $|D_i|$, $q_i = 0$ if $|D_i| = 0$, $s_i = 1$ if $D_i > 0$, $s_i = 0$ otherwise. Wilcoxon's statistic is $W = \sum s_i q_i$, and Stephenson's is $S_m = \sum s_i \cdot \binom{q_i - 1}{m-1}$, where $\binom{a}{b} = 0$ for $a < b$. S_1 is the sign test. S_2 is (almost) W. In an experiment under H_0, randomization creates the null distribution of W and S_m. Invert for CIs and estimates.

Sensitivity to departures from randomization Model: Subjects with the same $\mathbf{x}$ may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$ due to differences in u. Yields $1/\left(1+\Gamma\right) \leq \Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right) \leq \Gamma/\left(1+\Gamma\right)$, and then, for each Γ, sharp bounds on the null distribution of W and S_m. For W, the upper bound is a random variable $\overline{\overline{W}}$ which is the sum of I independent random variables taking the value i with probability $\Gamma/\left(1+\Gamma\right)$ or 0 with probability $1/\left(1+\Gamma\right)$, $i = 1, \ldots, I$. Invert for confidence intervals and point estimates.

Amplification: alternative interpretation of this analysis If unobserved bias led to a Δ-fold increase in the odds of a positive response, $D_i > 0$, and a Λ-fold increase in the odds of treatment, $Z_{i1} - Z_{i2} = 1$, then this is the same as a bias of $\Gamma = \left(\Delta\Lambda + 1\right)/\left(\Delta + \Lambda\right)$; see [10]. For instance, $\Gamma = 1.25$ corresponds with $\Delta = 2$, $\Lambda = 2$, and $\Gamma = 1.5$ corresponds with $\Delta = 4$, $\Lambda = 2$.

Design sensitivity Consider a theoretical situation with a causal effect and no unmeasured biases; however, the investigator cannot know this. In this situation, there a number $\widetilde{\Gamma}$, the design sensitivity, so as $I \to \infty$, the study is sensitive to bias $\Gamma > \widetilde{\Gamma}$ and insensitive to bias $\Gamma < \widetilde{\Gamma}$; see [7,12], [11, Chapter 14], and [15, Chapter 10]. Example, if $D_i \sim N\left(\frac{1}{2}, 1\right)$ and Wilcoxon's W is used, then $\widetilde{\Gamma} = 3.17$; however, switch to a better statistic and $\widetilde{\Gamma} = 4.2$; yet, that statistic has Pitman efficiency 0.98 relative to W in a randomized experiment with Gaussian errors [12, Tables 1, 3]. Increase $\widetilde{\Gamma}$ adaptively [13].

Mixture of large effects and nonresponders Conover and Salsburg [2] found the locally most powerful rank test for comparing $r_{Cij} \sim_{iid} F$ to $r_{Tij} \sim_{iid} \left(1-p\right)F + pF^m$ as $I \to \infty$ and $p \to 0$, where $F^m = F \times \cdots \times F$ is the maximum of m iid observations from F. This is a Lehmann alternative [4] who discussed $m = 2$. Conover-Salsburg ranks are not easy to interpret, but become indistinguishable from Stephenson's [18] ranks as $I \to \infty$. Stephenson's ranks permit confidence statements for the proportion of extreme responses caused by the treatment [9]. Gaussian version: $r_{Cij} \sim \Phi\left(\cdot\right)$ and $r_{Tij} \sim \left(1-p\right)\Phi\left(\cdot\right) + p\Phi^{\overline{m}}\left(\cdot\right)$ with $p = .25$. For $\overline{m} = 5$, W and S_{10} are close, with $\widetilde{\Gamma} = 1.6$ for W and $\widetilde{\Gamma} = 2.0$ for S_{10}. For $\overline{m} = 500$, $\widetilde{\Gamma} = 2.4$ for W and $\widetilde{\Gamma} = 8.9$ for S_{10}.

Sensitivity references, extensions, R packages References [11, Chapter 16], [15, Chapters 9-10], [9,10]. Extension [12]. Functions `senWilcox` and `senU` in `R` package `DOS`. Function `amplify` in package `sensitivitymult`.

[1] Bertsekas DP. A new algorithm for the assignment problem. Math Prog 1981;21:152-71.

[2] Conover WJ., Salsburg DS. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics 1988;44:189-96.

[3] Fisher RA. Design of Experiments. Edinburgh: Oliver&Boyd 1935.

[4] Lehmann EL. (1953) The power of rank tests. Ann Math Stat 1953;24:23-43.

[5] Neyman J. On the application of probability theory to agricultural experiments. Stat. Sci.1923/1990;5:463-80.

[6] Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. JASA 2015;110:515-27. `R` Package `rcbalance`

[7] Rosenbaum PR. Design sensitivity in observational studies. Biometrika, 2004;91:153-64.

[8] Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer.JASA 2007;102:75-83. Yu's `DiPs` in `R`

[9] Rosenbaum PR. Confidence intervals for uncommon but dramatic responses to treatment. Biometrics 2007;63:1164–71. `senU` in `DOS`

[10] Rosenbaum PR, Silber JH. Amplification of sensitivity analysis in observational studies. JASA 2009;104:1398-1405. `amplify` in `sensitivitymv`

[11] Rosenbaum PR. Design of Observational Studies. NY: Springer, 2010. `R` Package `DOS`

[12] Rosenbaum PR. A new U-statistic with superior design sensitivity in matched observational studies. Biometrics 2011;67:1017-27. `senU` in `DOS`

[13] Rosenbaum PR. Testing one hypothesis twice in observational studies. Biometrika 2012;99:763-74.

[14] Rosenbaum PR. How to see more in observational studies: Some new quasi-experimental devices. Ann Rev Stat Appl 2015;2:21-48.

[15] Rosenbaum PR. Observation and Experiment: An Introduction to Causal Inference. Cambridge, MA: Harvard University Press, 2017. Paperback edition, August 2019.

[16] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psych. 1974;66:688-701.

[17] Silber, JH, Rosenbaum, PR, Polsky, D, Ross, RN, Even-Shoshan, O, Schwartz, S, Armstrong, KA, Randall, TC. Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J Clin Oncol (JCO), 2007;25: 1169-75. Editorial: Cannistra, SA. Gynecologic oncology or medical oncology: What's in a name? JCO 2007;25: 1157-59. 5 letters and 2 rejoinders from S Blank, J Curtin, A Berchuck, M Hoffman, U Iqbal, M Markham, W McGuire, JH Silber, PR Rosenbaum, S Cannistra. JCO 2007;25:1151-58.

[18] Stephenson WR. A general class of one-sample nonparametric test statistics based on subsamples. JASA 1981;76:960-966. `senU` in `DOS`

[19] Yang D, Small DS, Silber JH, Rosenbaum PR. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics 2012;68:628-36. Yu's `DiPs`

[20] Yu R, Silber JH, Rosenbaum PR. Matching methods for observational studies derived from large administrative databases. Stat Sci 2019, to appear. `R` Package `bigmatch`

[21] Yu R, Rosenbaum PR. Directional penalties for optimal matching in observational studies. Biometrics. 2019, to appear, doi:10.1111/biom.13098 Yu's `DiPs`

[22] Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. JASA 2012;107:1360-71. `R` Package `designmatch`