DESIGN SENSITIVITY IN OBSERVATIONAL STUDIES

PAUL R. ROSENBAUM

Abstract. If an observational study were free of unobserved bias, we would not know this from the observed data. The best we could hope to say is that only very large unobserved biases could alter the conclusions of the study. What features of the design of an observational study reduce its sensitivity to unobserved biases? How can designs for observational studies be compared quantitatively in terms of their ability to resist unobserved biases? The talk is from [30], [31] and [39], the last joint with Dylan Small.

1. Notation and Review

1.1. Review: Notation for a Paired Randomized Experiment[11] [21] [36] [2]. Observed covariate $x$ and an unobserved covariate $u$. $I$ pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, one control, matched for $x$, so $x_{i1} = x_{i2}$, 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 [21] and Rubin [36]. Treatments given at doses $(v_{Tij}, v_{Cij})$, possibly $(v_{Tij}, v_{Cij}) = (1, 0)$. Write $F$ for $\{(r_{Tij}, r_{Cij}, v_{Tij}, v_{Cij}, x_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$ and $Z$ for the event $\{Z_{i1} + Z_{i2} = 1, i = 1, \ldots, I\}$; then $F$ and $Z$ are fixed by conditioning in Fisher’s [11] theory of randomization inference. Randomization in pairs ensures $Pr(Z_{i1} = 1 | Z, F) = \frac{1}{2}$, $\forall i$, with independently in distinct pairs. Observed response is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$, observed dose is $V_{ij} = Z_{ij} v_{Tij} + (1 - Z_{ij}) v_{Cij}$.

1.2. Review: Randomization inference for constant treatment effect[18] [27]. If treatment effect 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, $W_{\tau_{0}}$, 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 $Z, F$, 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 $\tau_{0}$, so $W_{\tau_{0}}$ 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 $\tau$ is obtained by inverting the test, and the Hodges-Lehmann (HL) estimate $\hat{\tau}$ of $\tau$ is (essentially) the
solution to $W_r = I (I + 1) / 4 = \frac{1}{2} (1 + 2 + \ldots + I)$. Null distribution of $W_{\tau_0}$ is the same for all (untied) $F$, but the nonnull distribution depends on $F$ or a model that generates $F$. A common model for a randomized experiment has $(r_{C1} - r_{C2}) / \sigma \sim \text{iid } F (\cdot)$ were $\sigma > 0$ and $F (\cdot)$ is continuous and symmetric about zero, so randomization ensures $\epsilon_i / \sigma \sim \text{iid } F (\cdot)$.

1.3. Review: Randomization inference if randomization is an instrument for the dose received$^{[13]} [17] [25] [28] [27]$. Treatment effect is proportional to dose if

$$r_{T_ij} - r_{C_ij} = \beta (v_{T_ij} - v_{C_ij}),$$

which is a constant effect for doses $(v_{T_ij}, v_{C_ij}) = (1, 0)$, $\forall i, j$. If effect is proportional to dose, then $R_{ij} - \beta V_{ij} = r_{T_ij} - \beta v_{T_ij} = r_{C_ij} - \beta v_{C_ij} = a_{ij}$, say, takes the same value $a_{ij}$ whether $Z_{ij} = 1$ or $Z_{ij} = 0$, and $a_{ij}$ is fixed by conditioning on $F$. Test $H_0 : \beta = \beta_0$ using $R_{ij} - \beta_0 V_{ij} = a_{ij} + (\beta - \beta_0) \{Z_{ij} v_{T_ij} + (1 - Z_{ij}) v_{C_ij}\}$ which is fixed at $a_{ij}$ if $H_0 : \beta = \beta_0$ is true, but otherwise varies with $Z_{ij}$. Write $D_i^{\beta_0}$ for the matched pair difference in $R_{ij} - \beta_0 V_{ij}$, treated $(Z_{ij} = 1)$ minus control $(Z_{ij} = 0)$, so that $D_i^{\beta_0} = (\beta - \beta_0) S_i + \epsilon_i$ where $S_i = Z_{i1} (v_{T1} - v_{C1}) + (1 - Z_{i1}) (v_{T2} - v_{C1})$ and $\epsilon_i = (2Z_{i1} - 1) (a_{i1} - a_{i2})$. In a randomized experiment, if $H_0 : \beta = \beta_0$ were true, then $S_i = \pm (a_{i1} - a_{i2})$ each with probability $\Pr (Z_{i1} = 1) = \frac{1}{2}$.

1.4. Review: Sensitivity to Departures from Random Assignment in Observational Studies $^{[10]} [33] [23] [24] [12] [16] [19] [1] [9] [27] [32]$. (i) In population, before matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \Pr (Z_{ij} = 1 \mid F)$, (ii) subjects with same observed $x_{ij}$ may differ in unobserved $u_{ij}$ and hence in odds of treatment by factor of $\Gamma \geq 1$,

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

and (iii) the distribution of treatments within treated/control matched pairs $\Pr (Z_{i1} = 1 \mid Z, F)$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$. Here, $\pi_{ij} = \Pr (Z_{ij} = 1 \mid F)$. If $\Gamma = 1$, then $x_{ij} = x_{ik}$ ensures $\pi_{ij} = \pi_{ik}, i = 1, \ldots, I$, whereupon $\Pr (Z_{i1} = 1 \mid Z, F) = \pi_{i1} / (\pi_{i1} + \pi_{i2}) = \frac{1}{2}$, and the distribution of treatment assignments is again the randomization distribution: bias solely due to observed $x$ can be eliminated by matching on $x$. If $\Gamma > 1$ in (1.1), then matching on $x$ may fail to equalize the $\pi_{ij}$ in pair $i$. $\Gamma$ is unknown. A sensitivity analysis calculates, for several values of $\Gamma$, the range of possible inferences. How large must $\Gamma$ be before qualitatively different causal interpretations are possible?

1.5. Review: Sensitivity Analysis with the Signed Rank Statistic $^{[23]} [24]$. If (1.1) and $H_0 : \tau = \tau_0$ are true, then the null distribution of $W_{\tau_0}$ 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 $\overline{W}$ for the sum of $I$ independent random variables taking value $i$ with probability $\theta$ 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 $\theta$. Then (1.1) and $H_0 : \tau = \tau_0$ imply the sharp bounds

$$\Pr (\overline{W} \geq w) \leq \Pr (W_{\tau_0} \geq w \mid Z, F) \leq \Pr (\overline{W} \geq w), \quad \forall w;$$
Table 1. Power of the Sensitivity Analysis: Heterogeneity vs Sample Size.

<table>
<thead>
<tr>
<th>Errors</th>
<th>$I$ Pairs</th>
<th>$\tau$</th>
<th>$\sigma$</th>
<th>$\sigma^2/I$</th>
<th>$\Gamma = 1$</th>
<th>$\Gamma = 1.5$</th>
<th>$\Gamma = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>120</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td>1/120</td>
<td>1.00</td>
<td>0.96</td>
<td>0.60</td>
</tr>
<tr>
<td>Normal</td>
<td>30</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>1/120</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>Cauchy</td>
<td>200</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td>1/200</td>
<td>0.98</td>
<td>0.32</td>
<td>0.02</td>
</tr>
<tr>
<td>Cauchy</td>
<td>50</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>1/200</td>
<td>0.95</td>
<td>0.60</td>
<td>0.28</td>
</tr>
</tbody>
</table>

e.g., [23][27]. If $\Gamma = 1$, then equality in (1.2); otherwise bounds (1.2) widen as $\Gamma$ 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 $\pi_{ij}$ satisfying (1.1) if $W_{\tau_0} \geq \tilde{w}$ where $0.05 = \Pr \left( \frac{\bar{W}}{W} \geq \tilde{w} \right)$.

For each $\pi = (\pi_{11}, \ldots, \pi_{I2})$, there is an HL estimate $\tilde{\tau}_\pi$ (essentially) solving $W_{\tilde{\tau}} = \mu_{\pi}$ where the expectation $\mu_{\pi} = E_{\pi} (W_{\tilde{\tau}} | Z, F)$ is computed using $\pi$. Then (1.1) implies $(1 - \theta) I (I + 1)/2 \leq \mu_{\pi} \leq \theta I (I + 1)/2$, yielding an interval of HL point estimates, $[\tilde{\tau}_{\min}, \tilde{\tau}_{\max}]$. With $\Gamma = 1$, $\mu_{\pi} = I (I + 1)/4$, and $\tilde{\tau}_{\min} = \tilde{\tau}_{\max}$ is the usual HL estimate.

2. Design Sensitivity

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. How do aspects of study design affect the chance that this hope will be realized? Does a dose-response relationship strengthen causal claims [15] [41] [26] [29]? (§2.5) Does multivariate coherence or pattern specificity strengthen causal claims [15] [6] [40] [8] [26] [27] [37]? (§2.5) Does reducing heterogeneity strengthen causal claims? Is reducing heterogeneity any better than increasing the sample size? Or do both just reduce the standard error of a biased estimate [20] [31]? (§2.2-2.3) Is a weak but nearly valid instrumental variable better or worse than a stronger but possibly somewhat biased instrument [34] [4] [39]? (§2.6)

2.2. Heterogeneity: Power of a sensitivity analysis [30] [31]. 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 $\tilde{w}$ so $0.05 = \Pr \left( \frac{\bar{W}}{W} \geq \tilde{w} \right)$ in (1.2); then calculate the probability that $W_{\tau_0} \geq \tilde{w}$ under some specific alternative hypothesis. For $\Gamma = 1$, this is the usual concept of power. If the treatment has an effect and there is no hidden bias, we would not know this, but hope to report results that are insensitive to unobserved bias; the power is the probability this will happen. Therefore, we compute the power under models in which the treatment has an effect (e.g., additive effect $\tau$) and there is no hidden bias (e.g., random assignment within pairs).

2.3. Heterogeneity: Limiting Uncertainty [31]. Whether or not (1.1) is true, for each fixed $\Gamma \geq 1$, as $I \to \infty$, the range of HL estimates, $[\tilde{\tau}_{\min}, \tilde{\tau}_{\max}]$, converges in probability
to \([\tau_{\min}, \tau_{\max}]\), 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 \(\hat{\tau} = \hat{\tau}_{\min} = \hat{\tau}_{\max}\) is consistent for \(\tau\) 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 \(\tau\) 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 \(\sigma\).

**Proposition 1.** [31] 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 \(\tau \pm \sigma \Upsilon^{-1}(\theta)\).

### 2.4. Design sensitivity [30]

There is a value \(\bar{\Gamma}\) of \(\Gamma\) such that the power of the sensitivity analysis tends to 1 for \(\Gamma < \bar{\Gamma}\) and to 0 for \(\Gamma > \bar{\Gamma}\); this value, \(\bar{\Gamma}\), is called the design sensitivity. No matter how large the sample size becomes, the design will always be sensitive to biases larger than \(\bar{\Gamma}\). Akin to Pitman efficiency — competing designs or methods compared for same task in large samples.

### 2.5. Design sensitivity: dose-response and coherence [30]

Table 2. Dose Response and Coherence. Left: Power at \(\Gamma = 2\), \(I = 200\), \(k = 3\). Right: Design Sensitivity \(k = 5\). \(k\) controls per matched set, \(p\) outcomes with correlation \(\rho\). Stratified rank sum statistic.

<table>
<thead>
<tr>
<th>(I = 200)</th>
<th>(\Gamma = 2)</th>
<th>(\Gamma \to \infty)</th>
<th>(\Gamma \to \infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k = 3)</td>
<td>(\rho = 0)</td>
<td>(\rho = 0)</td>
<td>(\rho = \frac{1}{2})</td>
</tr>
<tr>
<td>Doses</td>
<td>(p)</td>
<td>(\rho = 0)</td>
<td>(\rho = \frac{1}{2})</td>
</tr>
<tr>
<td>((\frac{1}{2}, 1, \frac{3}{2}))</td>
<td>1</td>
<td>0.54</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>6.4</td>
</tr>
<tr>
<td>((1, 1, 1))</td>
<td>1</td>
<td>0.28</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>5.1</td>
</tr>
<tr>
<td>((\frac{3}{2}, \frac{3}{2}, \frac{3}{2}))</td>
<td>1</td>
<td>0.98</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>11.7</td>
</tr>
</tbody>
</table>

Consider power for \(\Gamma = 2\), \(k = 3\), \(I = 200\), and design sensitivity for \(k = 5\), \(I \to \infty\), and will use the stratified Wilcoxon rank sum, with dose weights and coherence among outcomes; details in [30]. Power evaluated with \(p = 1\) or \(p = 3\) outcomes, linearly related to dose with slope \(\beta = 1/2\), symmetrically correlated multivariate Normal errors, with variances 1 and intercorrelations \(\rho\). Consider three possible patterns: \((\frac{1}{2}, 1, \frac{3}{2})\), \((1, 1, 1)\), \((\frac{3}{2}, \frac{3}{2}, \frac{3}{2})\), where \((\frac{1}{2}, 1, \frac{3}{2})\) yields dose-response, \((1, 1, 1)\) has the same average dose without dose-response, and \((\frac{3}{2}, \frac{3}{2}, \frac{3}{2})\) has larger doses without dose response.
2.6. Design sensitivity: instrument strength and validity [39]. Instrument strength characterized by compliance probabilities: \( \pi_A, \pi_C, \pi_N \) for, respectively, the proportion of 'always takers' \((v_{Tij}, v_{Cij}) = (1, 1)\), 'compliers' \((v_{Tij}, v_{Cij}) = (1, 0)\), and 'never takers' \((v_{Tij}, v_{Cij}) = (0, 0)\), with no defiers \((v_{Tij}, v_{Cij}) = (0, 1)\).

Design Sensitivity \( \Gamma \) For Instruments with Varying Strength. Effect size \( \lambda = (\beta_0 - \beta) / \sigma \).

<table>
<thead>
<tr>
<th>Compliance</th>
<th>100%</th>
<th>50%</th>
<th>20%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_A, \pi_C, \pi_N )</td>
<td>0, 1, 0</td>
<td>( \frac{1}{4}, \frac{1}{2}, \frac{1}{4} )</td>
<td>( \frac{2}{5}, \frac{1}{2}, \frac{2}{5} )</td>
<td>( \frac{9}{20}, \frac{2}{20}, \frac{9}{20} )</td>
</tr>
<tr>
<td>Normal 1</td>
<td>11.7</td>
<td>2.7</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Normal ( \frac{1}{2} )</td>
<td>3.2</td>
<td>1.7</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Cauchy 1</td>
<td>3.0</td>
<td>1.7</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Cauchy ( \frac{1}{2} )</td>
<td>1.8</td>
<td>1.4</td>
<td>1.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

References


