Design sensitivity and efficiency in observational studies

Paul R. Rosenbaum*

University of Pennsylvania, Philadelphia

*Address for correspondence: Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: rosenbaum@stat.wharton.upenn.edu.
1 Basis for this talk


Other discussions of design sensitivity:


2  Words in the title I

Efficiency:  The usual notion, Pitman efficiency, also known as asymptotic relative efficiency.

Efficiency is:  an aspect of the stochastic process that generated the data and particular methods of analysis, evaluated when the sample size is large.

Observational study:  Study of treatment effects when subjects are not randomized to treatment or control.

Issue:  Without randomization, treated and control groups may not be comparable.  Adjust for observed covariates, perhaps by matching.

Sensitivity analysis:  Asks what an unobserved covariate would have to be like to alter the conclusions of a naïve analysis that presumes adjustments for observed covariates suffice.
3 Words in the title II

**Design sensitivity:** Speaking informally, the design sensitivity is the limiting sensitivity to unobserved bias as the sample size increases.

**Design sensitivity is:** (like efficiency and unlike sensitivity analysis) an aspect of the stochastic process that generated the data and particular methods of analysis, evaluated when the sample size is large.

**Design sensitivity is:** a number, \( \Gamma \), such that, as the sample size increases, the study will eventually be insensitive to biases smaller than \( \Gamma \) and sensitive to biases larger than \( \Gamma \).

**In particular,** in large samples, the limiting power of a sensitivity analysis is determined by the design sensitivity.
4 Main claim of the talk

• Lacking theoretical guidance, we tend to select statistical methods for use in observational studies based on their efficiency in randomized experiments.

• This turns out to be a mistake.

• A highly efficient method for detecting small treatment effects in randomized experiments need not, and often does not, have the highest power in a sensitivity analysis or the largest design sensitivity.

• That is, the best procedure assuming that an observational study is effectively a randomized experiment need not be the best procedure under more realistic assumptions.

• Or so I shall argue in this talk.
5 Outline

- Very quick review of sensitivity analysis. (See handout and references there for details.)

- Example: A simple, matched pair observational study. Several test statistics.

- A heuristic graph: Where is the evidence that distinguishes effects and biases?

- Very quick review of design sensitivity. (See handout and references there for details.)

- Calculating the design sensitivity for several statistics under several models.

- Simulation of power in finite samples

- Recap
6 Treated-Control Matched Pairs

- There are $I$ pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$. $Z$ is the event $Z_{i1} + Z_{i2} = 1, i = 1, \ldots, I$.

- Matched for observed covariates, so $x_{i1} = x_{i2}$.

- Possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$.

- Randomized paired experiment, $Z_{i1}, i = 1, \ldots, I$, determined by $I$ independent flips of a coin.

- Naïve analysis of an observational study assumes adjustments for $x$ suffice to remove bias.

- Sensitivity analysis asks: What $u$ would have to be like to alter the conclusions of the naïve analysis?
7 Causal effects

- Neyman (1923) and Rubin (1973): Each subject $ij$ has two potential responses, $r_{Tij}$ if treated, $Z_{ij} = 1$, or $r_{Cij}$ if control, $Z_{ij} = 0$.

- Observed response from $ij$ is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$.

- Effect of the treatment, $r_{Tij} - r_{Cij}$, on $ij$ is not observed for any subject.

- Fisher’s sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I, j = 1, 2$.

- $H_0$ is false if the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau$ for all $ij$, $\tau \neq 0$.

- Write $\mathcal{F} = \left\{(r_{Tij}, r_{Cij}, x_{ij}, u_{ij}) \right\}$, $i = 1, \ldots, I, j = 1, 2$. 
8 Treated-minus-control pair differences

- In pair $i$, the observed, treated-minus-control difference in responses is $Y_i = (Z_{i1} - Z_{i2}) (R_{i1} - R_{i2})$.

- If the treatment has an additive effect, $r_{Ti} - r_{Ci} = \tau$ for all $ij$, then $Y_i$ is

  $$Y_i = (Z_{i1} - Z_{i2}) (r_{Ci1} + Z_{i1} \tau - r_{Ci2} - Z_{i2} \tau)$$

  $$= \tau + \epsilon_i$$

  where $\epsilon_i = (Z_{i1} - Z_{i2}) (r_{Ci1} - r_{Ci2})$

- Looking ahead: A sensitivity analysis is an analysis of $Y_1, \ldots, Y_I$. Efficiency, the power of a sensitivity analysis, the design sensitivity refer to a stochastic model that generated the $Y_i$, such as $Y_i \sim iid N(\tau, 1)$. 

9 Test statistics

- Let \((I + 1) q_i\) be the rank of \(|Y_i|\) with average ranks for ties.

- Let \(\text{sgn} (y) = 1\) or 0 for, respectively \(y > 0\) or \(y \leq 0\).

- A general signed rank statistic is of the form \(T = \sum_{i=1}^{I} \text{sgn} (Y_i) \varphi (q_i)\) where \(\varphi : [0, 1] \to [0, \infty)\) is a score function.

- Wilcoxon’s signed rank statistic is essentially the same as taking \(\varphi (q_i) = q_i\), whereas the sign test has \(\varphi (q_i) = 1\).

Randomization creates the null distribution \(\Pr (T \mid \mathcal{F}, \mathcal{Z})\) of \(T\) under Fisher’s \(H_0\).
10 Three other test statistics

- General signed rank statistic is of the form $T = \sum_{i=1}^{I} \text{sgn}(Y_i) \varphi(q_i)$ where $\varphi : [0, 1] \rightarrow [0, \infty)$ is a score function and $(I + 1) q_i$ be the rank of $|Y_i|$ and $\text{sgn}(y) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$.

- Brown (1981). $\varphi(q_i) = 0$ for $0 \leq q_i < \frac{1}{3}$, $\varphi(q_i) = 1$ for $\frac{1}{3} \leq q_i < \frac{2}{3}$, $\varphi(q_i) = 2$ for $\frac{2}{3} \leq q_i \leq 1$. See also Markowski & Hettmansperger (1982).

- Outer-$\lambda$-Wilcoxon. $\varphi(q_i) = 0$ for $0 \leq q_i < 1 - \lambda$, $\varphi(q_i) = (q_i - \lambda) / (1 - \lambda)$ for $1 - \lambda \leq q_i < 1$.

- Central 20%-60% Wilcoxon. $\varphi(q_i) = 0$ for $0 \leq q_i < 0.2$, $\varphi(q_i) = (q_i - .2) / (.6 - .2)$ for $0.2 \leq q_i < 0.6$, $\varphi(q_i) = 1$ for $0.6 \leq q_i \leq 1$. 
Three ways to score the ranks of $|Y_i|$. Wilcoxon’s signed rank statistic uses the ranks themselves as scores (dotted line). Brown’s (1981) statistic takes steps at $1/3$ and $2/3$ (solid line). The outer-2/3-Wilcoxon scores the outer $2/3$ of the ranks (dashed line).
11 Sensitivity analysis

- A sensitivity analysis asks about the magnitude of departure from \( \Pr \left( Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z} \right) = \frac{1}{2} \) that would need to be present to alter the qualitative conclusions of a randomization inference.

- A simple model: In the population prior to matching, subjects have independent treatment assignments with unknown probabilities, \( \pi_{ij} = \Pr \left( Z_{ij} = 1 \mid \mathcal{F} \right) \), such that two subjects, say \( ij \) and \( ij' \), with the same observed covariates, \( x_{ij} = x_{ij'} \), may differ in their odds of treatment by at most a factor of \( \Gamma \geq 1 \),

\[
\frac{1}{\Gamma} \leq \frac{\pi_{ij} \left( 1 - \pi_{ij'} \right)}{\pi_{ij'} \left( 1 - \pi_{ij} \right)} \leq \Gamma \quad \text{whenever} \quad x_{ij} = x_{ij'};
\]

(1) then condition on \( Z_{i1} + Z_{i2} = 1 \).

- For each \( \Gamma \geq 1 \), obtain a range of possible inference quantities, point estimates, p-values, etc.
12 Example: Effects of Benzene

- Tunca and Egeli (1996) examined chromosome damage among 78 individuals in Bursa, Turkey with varied exposures to benzene.

- 39 pairs, long vs short/zero exposure, matched for age, alcohol and smoking

- Outcome is relative frequency of chromosome aberrations in blood lymphocytes

- If you do a randomization test in matched pairs, the one-sided $P$-value is $\leq 0.0001$ using the sign test, the Wilcoxon signed rank test, Brown’s test and the outer-$\frac{1}{2}$-Wilcoxon test.
Figure 1: Percent of cells with chromosome damage in 78 individuals in Bursa, Turkey classified by duration of exposure to benzene. The 78 individuals form 39 pairs, with matched pair difference $Y_i, i=1,\ldots,39$. The third plot is an estimate of the $abz(y)$ heuristic function introduced in section 3, the estimate being derived from a nonparametric density estimate with the default settings in R.
Table 1: Sensitivity analysis for the benzene example. The table gives the upper bound on the one-sided $P$-value for several values of $\Gamma$. Although not shown in the table, the upper bound for the sign test is 0.0508 for $\Gamma = 2.7$ and the upper bound for the outer-$(1/2)$ Wilcoxon statistic is 0.0506 for $\Gamma = 4.8$.

<table>
<thead>
<tr>
<th>$\Gamma$</th>
<th>Sign</th>
<th>Wilcoxon</th>
<th>Brown</th>
<th>Outer-W 1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
</tr>
<tr>
<td>3</td>
<td>0.0828</td>
<td>0.0177</td>
<td>0.0132</td>
<td>0.0172</td>
</tr>
<tr>
<td>3.9</td>
<td>0.2182</td>
<td>0.0506</td>
<td>0.0333</td>
<td>0.0332</td>
</tr>
<tr>
<td>4.5</td>
<td>0.3253</td>
<td>0.0809</td>
<td>0.0508</td>
<td>0.0447</td>
</tr>
</tbody>
</table>

$\Gamma = 2.8 \iff$ 5-fold increase in the odds of treatment and a 6-fold increase in the odds of a positive $Y_i$

$\Gamma = 4.5 \iff$ 8-fold increase in the odds of treatment and a 10-fold increase in the odds of a positive $Y_i$

*See Rosenbaum and Silber (2009, *JASA*, 104, 1398-1405) for details about this interpretation of $\Gamma$. 
13 Heuristic Graph I: Where is the evidence that distinguishes effects from unmeasured biases?

- Suppose that the $Y_i$’s are not biased, so each $Y_i$ is telling us about the effects of the treatment. (Of course, we would not know this from the data.)

- In this case, we would like to say that the results are insensitive to small and moderate biases.

- Suppose you could observe an infinite amount of data at any one value of $|Y_i|$, that is, you get to observe $\text{sgn}(Y_i)$.

- What $|Y_i|$ would you pick?
14 Heuristic Graph II: The abz-function

- Suppose that the $Y_i$’s are not biased, so each $Y_i$ is
telling us about the effects of the treatment. (Of
course, we would not know this from the data.)

- Suppose that $Y_i$ are iid from a continuous distribu-
tion $G(\cdot)$ with density $g(\cdot)$.

- Ablers, Bickel and van Zwet (1976) introduced a
  function $abz(y)$ defined for $y > 0$, namely

$$abz(y) = \frac{g(y)}{g(y) + g(-y)} = \Pr \left( Y_i > 0 \mid |Y_i| = y \right)$$

- If $abz(y) > \Gamma / (1 + \Gamma)$, then at $|Y_i| = y$, positive
  $Y_i$ occur with a frequency $abz(y)$ that is too high
to be attributed to a bias of magnitude $\Gamma$. 
Figure 2: Conditionally given various values of $|Y_i|$, the figure shows the probability of a positive treatment-minus-control difference, $Y_i > 0$, for an additive treatment effect $\tau = \frac{3}{4}$ in the standard forms of four distributions.
What do we learn from the heuristic graph?

We actually observe limited data at varied $|Y_i|$. Nonetheless, the heuristic graph suggest little weight should be given to small $|Y_i|$. What you should do with large $|Y_i|$ depends on the distribution $G$ which you typically do not know.
Suppose that the $Y_i$'s are not biased, so each $Y_i$ is telling us about the effects of the treatment. (Of course, we would not know this from the data.)

Suppose that $Y_i$ are iid from a continuous distribution $G(\cdot)$ with density $g(\cdot)$.

Using the upper bound on the $P$-value in a conventional way, we would say that the results of the study are insensitive to a bias of magnitude $\Gamma$ if the upper bound on the $P$-value at this $\Gamma$ was less than 0.05.

The power of the sensitivity analysis at this $\Gamma$ is the probability that this will happen.

For $\Gamma = 1$, this is the power of a randomization test of no treatment effect.
17  To repeat...

- If the $Y_i$'s are not biased, and each $Y_i$ is telling us about the effects of the treatment, then

  (i) we would not know this from the data,
  (ii) we hope to report insensitivity to small and moderate biases

- Call this the ‘favorable situation.’

- If we were in the favorable situation, we would not know it. Hence, we do a sensitivity analysis.

- In the favorable situation, for a specific $\Gamma \geq 1$, the power of an $\alpha$-level test is the probability that the upper bound on the $P$-value is less than $\alpha$. 
18 Design sensitivity

- As $I \to \infty$, there is a value, $\bar{\Gamma}$, such that the power of the sensitivity analysis goes to 1 for $\Gamma < \bar{\Gamma}$ and to 0 for $\Gamma > \bar{\Gamma}$.

- In words: in large samples, you can distinguish a treatment effect from a bias $\Gamma < \bar{\Gamma}$ but not from a bias $\Gamma > \bar{\Gamma}$.

- That number, $\bar{\Gamma}$, is the design sensitivity. It is the limiting sensitivity to unmeasured bias.

- $\bar{\Gamma}$ depends upon the process generating the data and the methods of analysis.
19 Three Pictures of Power

- $Y_i \sim iid \, N(\tau, 1)$ for $\tau = \frac{1}{2}$ or 1 (just $\tau = \frac{1}{2}$ in the second figure)

- $\tilde{\Gamma} = 3.171$ for $\tau = \frac{1}{2}$ and $\tilde{\Gamma} = 11.715$ for $\tau = 1$

- Fig 1: Power in a randomized experiment, plotted as a function of $I$, the number of pairs.

- Fig 2: Power in of a sensitivity analysis, plotted as a function of $I$, for several $\Gamma$.

- Fig 3: Power in of a sensitivity analysis, plotted as a function of $\Gamma$, for several $I$. 
Number of Matched Pairs, I=20,...,200

Power

- \( \tau = 1/2 \)
- \( \tau = 1/4 \)

Number of Matched Pairs, I=20,...,200
Number of Matched Pairs, \( I = 20, \ldots, 2000 \)

Power

\[
\begin{align*}
\Gamma &= 1 \\
\Gamma &= 1.5 \\
\Gamma &= 2 \\
\Gamma &= 2.5 \\
\Gamma &= 3 \\
\Gamma &= 3.5
\end{align*}
\]
19 Calculating the Design Sensitivity

- The paper (and the handout) give a formula for $\tilde{\Gamma}$.

- That formula expresses the design sensitivity $\tilde{\Gamma}$ for $T = \sum_{i=1}^{I} \text{sgn} (Y_i) \varphi(q_i)$ when $Y_i$ is iid from $G(\cdot)$ in terms of $\varphi(\cdot)$ and $G(\cdot)$. 
20 Normal Errors

\[ Y_i = \tau + \epsilon_i \quad \epsilon_i \sim N\left(0, \sigma^2\right) \]

Table 2: Pitman asymptotic relative efficiency versus the Wilcoxon statistic and design sensitivity \(\tilde{\Gamma}\) in the favorable situation with an additive treatment effect and Normal errors.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Efficiency</th>
<th>Design Sensitivity (\tilde{\Gamma}) (\tau/\sigma = 0.5)</th>
<th>Design Sensitivity (\tilde{\Gamma}) (\tau/\sigma = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>1.00</td>
<td>3.2</td>
<td>11.7</td>
</tr>
<tr>
<td>Brown</td>
<td>0.95</td>
<td>3.6</td>
<td>18.0</td>
</tr>
<tr>
<td>Outer 4/5</td>
<td>1.01</td>
<td>3.7</td>
<td>18.7</td>
</tr>
<tr>
<td>Outer 2/3</td>
<td>0.99</td>
<td>4.2</td>
<td>26.4</td>
</tr>
<tr>
<td>Outer 1/2</td>
<td>0.93</td>
<td>5.0</td>
<td>42.6</td>
</tr>
<tr>
<td>Central 20%-60%</td>
<td>0.92</td>
<td>3.3</td>
<td>15.1</td>
</tr>
</tbody>
</table>
21 Logistic Errors

Table 3: Pitman asymptotic relative efficiency versus the Wilcoxon statistic in the absence of bias ($\Gamma = 1$) and design sensitivity $\tilde{\Gamma}$ in the favorable situation with an additive treatment effect, $\tau$, and errors $\epsilon_i$ that are logistic with standard deviation $\sigma$.

| Statistic         | Efficiency | Design Sensitivity $\tilde{\Gamma}$  \\ 
|-------------------|------------|--------------------------------------\\ 
|                   |            | $\tau/\sigma = 0.5$ | $\tau/\sigma = 1$  \\ 
| Wilcoxon          | 1.00       | 3.4 | 12.4  \\ 
| Brown             | 0.94       | 3.8 | 17.2  \\ 
| Outer 4/5         | 0.97       | 3.9 | 17.5  \\ 
| Outer 2/3         | 0.91       | 4.2 | 21.0  \\ 
| Outer 1/2         | 0.78       | 4.7 | 25.3  \\ 
| Central 20%-60%   | 0.96       | 3.6 | 15.5  \\ 


## Double Exponential Errors

Table 4: Pitman asymptotic relative efficiency versus the Wilcoxon statistic in the absence of bias ($\Gamma = 1$) and design sensitivity $\tilde{\Gamma}$ in the favorable situation with an additive treatment effect, $\tau$, and errors $\epsilon_i$ that are double exponential with standard deviation $\sigma$.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Efficiency</th>
<th>Design Sensitivity $\tilde{\Gamma}$</th>
<th>$\tau/\sigma = 0.5$</th>
<th>$\tau/\sigma = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>1.00</td>
<td></td>
<td>3.8</td>
<td>13.0</td>
</tr>
<tr>
<td>Brown</td>
<td>0.80</td>
<td></td>
<td>4.1</td>
<td>16.2</td>
</tr>
<tr>
<td>Outer 4/5</td>
<td>0.80</td>
<td></td>
<td>4.1</td>
<td>16.0</td>
</tr>
<tr>
<td>Outer 2/3</td>
<td>0.67</td>
<td></td>
<td>4.1</td>
<td>16.8</td>
</tr>
<tr>
<td>Outer 1/2</td>
<td>0.50</td>
<td></td>
<td>4.1</td>
<td>16.9</td>
</tr>
<tr>
<td>Central 20%-60%</td>
<td>0.90</td>
<td></td>
<td>4.0</td>
<td>15.7</td>
</tr>
</tbody>
</table>
24 Simulation

- Design sensitivity is about what happens as $I \to \infty$.

- What happens for finite $I$?

- Simulate $Y_i \sim iid G(\cdot)$. (The ‘favorable situation.’) For some $\Gamma \geq 1$, calculate the upper bound on the one-sided $P$-value. How often is it less than $0.05$.

- Let’s look at $I = 200$, $\tau/\sigma = 1$, $\Gamma = 8$, $\sigma^2 = \text{var}(Y_i)$.

- Each situation is replicated 10,000 times, so the standard error of a proportion is at most

$$\sqrt{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{10000}} = 0.005.$$
Figure 3: Comparison of 10,000 simulated upper bounds on the one-sided P-value in the favorable situation with $\tau/\sigma = 1$ and $\Gamma = 8$. The statistics are $W_i =$ Wilcoxon’s signed rank statistic, $B_r =$ Brown’s statistic, $O_u =$ outer 2/3 Wilcoxon statistic, and $C_e =$ central 20% - 60% Wilcoxon statistic.
Table 5: Simulated power of a 0.05 level sensitivity analysis in the favorable situation with an additive treatment effect, $\tau$, and errors $\varepsilon_i$ that are Normal with $\tau/\sigma = 1/2$ and either $I = 100$ or $I = 250$ matched pairs. Highest estimated power is in **bold**.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>$\Gamma$</th>
<th>Normal $I = 100$</th>
<th>Normal $I = 250$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td>0.54</td>
<td>0.89</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 2$</td>
<td>0.63</td>
<td>0.94</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td><strong>0.73</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 2$</td>
<td>0.55</td>
<td>0.89</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 2$</td>
<td>0.63</td>
<td>0.95</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 3$</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td><strong>0.21</strong></td>
<td><strong>0.44</strong></td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 3$</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 3$</td>
<td>0.09</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Table 6: Simulated power of a 0.05 level sensitivity analysis in the favorable situation with an additive treatment effect, $\tau$, and errors $\varepsilon_i$ that are logistic with $\tau/\sigma = 1/2$ and either $I = 100$ or $I = 250$ matched pairs. The highest estimated power is in **bold**.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\Gamma$</th>
<th>Logistic $I = 100$</th>
<th>Logistic $I = 250$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td>0.65</td>
<td>0.95</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 2$</td>
<td>0.71</td>
<td>0.97</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td><strong>0.73</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>Central 20%-60%</td>
<td>$\Gamma = 2$</td>
<td>0.67</td>
<td>0.96</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 2$</td>
<td>0.67</td>
<td>0.96</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 3$</td>
<td>0.15</td>
<td>0.30</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td><strong>0.22</strong></td>
<td><strong>0.45</strong></td>
</tr>
<tr>
<td>Central 20%-60%</td>
<td>$\Gamma = 3$</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 3$</td>
<td>0.13</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 7: Simulated power of a 0.05 level sensitivity analysis in the favorable situation with an additive treatment effect, $\tau$, and errors $\epsilon_i$ that are Double Exponential with $\tau/\sigma = 1/2$ and either $I = 100$ or $I = 250$ matched pairs. The highest estimated power is in **bold**.

<table>
<thead>
<tr>
<th></th>
<th>$\Gamma$</th>
<th>Double Exponential $I = 100$</th>
<th>Double Exponential $I = 250$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td>0.78</td>
<td>0.99</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 2$</td>
<td>0.77</td>
<td>0.98</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 2$</td>
<td>0.69</td>
<td>0.97</td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 2$</td>
<td><strong>0.80</strong></td>
<td><strong>0.99</strong></td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 2$</td>
<td>0.19</td>
<td>0.38</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td>0.19</td>
<td>0.38</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 3$</td>
<td>0.20</td>
<td>0.43</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 3$</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 3$</td>
<td><strong>0.24</strong></td>
<td><strong>0.47</strong></td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 3$</td>
<td>0.18</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Table 8: Simulated power of a sensitivity analysis in the favorable situation with an additive effect, $\tau$, and errors $\varepsilon_i$ that are Normal, logistic or double exponential (DE) with $\tau/\sigma = 1$ and $I = 200$ pairs. The highest power is in **bold**.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Logistic</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 6$</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 6$</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 6$</td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 6$</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 6$</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 8$</td>
<td>0.30</td>
<td>0.39</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 8$</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 8$</td>
<td><strong>0.91</strong></td>
<td><strong>0.77</strong></td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 8$</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 8$</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>$\Gamma = 10$</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td>Brown</td>
<td>$\Gamma = 10$</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>Outer 2/3 Wilcoxon</td>
<td>$\Gamma = 10$</td>
<td><strong>0.69</strong></td>
<td><strong>0.49</strong></td>
</tr>
<tr>
<td>Central 20%–60%</td>
<td>$\Gamma = 10$</td>
<td>0.28</td>
<td>0.31</td>
</tr>
<tr>
<td>Permutational t-test</td>
<td>$\Gamma = 10$</td>
<td>0.09</td>
<td>0.12</td>
</tr>
</tbody>
</table>
24 Recap

- If there was a treatment effect and no bias from unmeasured covariates (the favorable situation), we do not know it. In this case, we would hope to report that the conclusions are insensitive to small and moderate biases. The power of a sensitivity analysis and the design sensitivity indicate the chance this hope will be realized.

- The most popular tests for matched pairs — Wilcoxon’s signed rank test, the one-sample t-test, the sign test — have lower power in a sensitivity analysis than other methods which pay little attention to the smallest $|Y_i|$.

- The design sensitivity $\tilde{\Gamma}$ is a better guide to the power of a sensitivity analysis than efficiency of a procedure in a randomized experiment.