An overview of statistical methods for removing selection bias in observational studies

Nathan Taback, PhD
observational-science.com
Insecurityinsight.org
University of Toronto

Outline

- Causal inference in randomized experiments
- Causal inference in nonrandomized studies
- Methods: (1) Propensity score, (2) Multivariable modeling, (3) Instrumental variable
- Comparing Methods
- Sensitivity analysis: People Who Look Comparable May Differ
- Planning the Analysis

Causal Inference in Randomized Experiments

- The following data were obtained in a study comparing the standard treatment A to a new treatment B.
- The investigators had 11 patients: 5 were given A, 6 were given B.

Patient	1 2 3 4 5 6 7 8 9 10 11
Treatment	A A B B A B B B A A B

Causal Inference in Randomized Experiments

- The arrangement was obtained by taking 11 playing cards, 5 red corresponding to A and 6 black corresponding to B.
- The cards were thoroughly shuffled and dealt to give the sequence in the table.

Patient	1 2 3 4 5 6 7 8 9 10 11
Treatment	A A B B A B B B A A B

Causal Inference in Randomized Experiments

- The experimental arrangement is one of the 462 possible ways of allocating 5 A's and 6 B's to the 11 patients.
- Null hypothesis: A and B have the same effect on the outcome.
- If the null hypothesis is true:

- A and B are labels and don't affect the outcome.
- e.g. patient 1 would have the same outcome whether it had been labeled A or B.

Causal Inference in Randomized Experiments

- Suppose that the observed difference in the proportion of responders in group B minus group A is 0.097.
- If the null hypothesis is assumed to be true then all the potential differences that could have been generated from the 462 rearrangements of the labels can be calculated.
- This is called the randomization distribution.

Causal Inference in Randomized Experiments

- Fisher spoke of randomization as the “reasoned basis” for causal inference.
- Provides a valid test of the hypothesis of no effect caused by treatment.
- Randomization based on two considerations: (1) the experimenter used cards/random numbers to assign treatments, picking one of the 462 possible treatment assignments; (2) the null hypothesis of no treatment effect.

Causal Inference in Randomized Experiments

- Fisher's randomization inference concerns finite population of n subjects
- E.g. 76,693 people randomized in prostate cancer screening trial: 38,350 men randomized to usual care (Control); 38,343 men randomized to annual screening (Treatment) (Andriole, et al., NEJM, 2009).
- The inference is how these 76,693 people would have responded under treatments they did not receive.

Causal Inference in Randomized Experiments

- In principle each person has two potential responses, a response under 'treatment', T, and a response under 'control', C.
- Refs: Neyman (1923), Rubin (1974).

$$r_{T_i} = \begin{cases} 1, \text{ if man i has prostate cancer with screening} \\ 0, \text{ if man i does not have prostate cancer with screening} \end{cases}$$

$$r_{C_i} = \begin{cases} 1, \text{ if man i has prostate cancer with usual care} \\ 0, \text{ if man i does not have prostate cancer with usual care} \end{cases}$$

- We can only observe one response!

Outline

- Causal inference in randomized experiments
- **Causal inference in nonrandomized studies**
- Methods: (1) Propensity score, (2) Multivariable modeling, (3) Instrumental variable
- Comparing Methods
- Sensitivity analysis: People Who Look Comparable May Differ
- Planning the Analysis

Causal inference in nonrandomized studies

- Nonrandomized/observational study is an empiric investigation of the effects of a treatment in which individuals are not assigned at random to treatment or control, as they would be in a randomized clinical trial (Cochran, 1965).
- Most of the theory for the analysis of nonrandomized studies attempts to "fix" nonrandomized treatment assignment.

Causal inference in nonrandomized studies

- The rationale behind "fixing" this problem is to eliminate bias due to treatment and control groups differing in prognostic variables other than treatment.
- The danger that such a bias will be mistaken for a treatment effect is the main reason that randomized experiments are preferred to observational studies (Rosenbaum, 1991).

Causal inference in nonrandomized studies

- Causal inference in health research has largely focused on counterfactual or "potential outcome" analysis (e.g. Rosenbaum and Rubin, 1983; Robins, 1986; Angrist et al. 1996).
- Other frameworks include graphical models (Pearl, 1995 and Cox and Wermuth, 1996), and structural equation models (Sanchez et al., 2005).
- Pearl (1995) has presented a formulation which unifies these approaches.

Outline

- Causal inference in randomized experiments
- Causal inference in nonrandomized studies
- **Methods: (1) Propensity score, (2) Multivariable modeling (3) Instrumental variable**
- Comparing Methods
- Sensitivity analysis: People Who Look Comparable May Differ
- Planning the Analysis

Methods

- The potential outcomes approach is a framework to describe the assumptions that form the basis of randomization.
- Randomization is the "gold standard" of assigning treatments to a heterogeneous group of patients.
- Under the null hypothesis treatment assignment is independent of potential response, observed and unobserved covariates. This means that the P-value is a valid measure of departure from null hypothesis.

Methods: Propensity score

- The propensity score is the conditional probability of treatment Z=1 given the observed covariates x (Rosenbaum and Rubin, 1983).
- The balancing property says that treated (Z=1) and control (Z=0) subjects with the same propensity score have the same distribution of observed covariates.

Methods: Propensity score

- Randomization is a more powerful tool for balancing covariates than matching on an estimate of the PS.
- PS balances observed covariates, whereas randomization balances observed covariates, unobserved covariates, and potential responses.
- PS methods lead to unbiased estimates of causal effects under the assumption that receipt of treatment is independent of outcome conditional on observed covariates.

Methods: Propensity score

- Three main ways to implement PS: (1) matching; (2) stratification; (3) regression adjustment.
- Austin & Mamdani (2006) compared the three methods using a sample of patients (11 542) discharged from Ontario hospitals between April 1, 1999 and March 31, 2001.
- Study part of an initiative to improve quality of care for patients with cardiovascular disease in Ontario.
- Focus on estimating treatment effect of statins on reducing all-cause mortality post-AMI.

Methods: Propensity score

- Developed PS model to predict probability that a patient would be given a prescription for a statin at hospital discharge.
- The goal was to estimate the reduction in all-cause mortality attributable to the post-discharge use of statins.
- Patients receiving statins tended to be younger and healthier than patients not receiving statins.
- Strong evidence that treatment status confounded with factors prognostic of AMI mortality.

Methods: Propensity score

Method	OR	P
Crude	0.49	< 0.0001
Stratification	0.77	0.0003
Covariate adjusted	0.84	0.0011
Matching	0.85	0.0372
Multivariable regression	0.75	< 0.0001

Matching vs. stratifying: similar to classic tradeoff between variance and bias. Stratification may result in greater bias due to residual confounding within stratum. Matching may result in treated and untreated patients discarded from the analysis thus diminishing precision of treatment effect.

Methods: Multivariable modeling

- A linear multivariate model relates the outcomeY to treatment and all known confounders, including interactions, into the model.
- Controlling for these covariates produces a risk-adjusted treatment effect and removes bias due to these factors.

Methods: Instrumental Variable Analysis

- An econometric method used to remove effects of hidden bias in observational studies.
- Two key characteristics of an IV are: (1) highly correlates with treatment; and (2) does not independently affect the outcome.
- A prototype of an instrument is a random nudge to accept a treatment.
- The nudge may or may not induce acceptance and can affect outcome only if it succeeds in inducing acceptance of treatment.

Methods: Instrumental Variable Analysis

- In a paired randomized design individuals paired based on measured pretreatment covariates.
- For each pair a coin is flipped to decide which member of pair will be encouraged to accept treatment (Holland (1988), Rosenbaum (2010)).
- Some individuals may decline treatment despite encouragement and others may take treatment in absence of encouragement.

Methods: Instrumental Variable Analysis

- Unbiased estimates of causal effects using IV relies on several key assumptions (Angrist, Imbens, and Rubin, 1996).
- One of the assumptions is the exclusion criteria: encouragement must have no effect on the outcome besides its effect on acceptance of treatment.
- Z ➔ D ➔Y

(Z is encouragement, D is acceptance,Y is outcome)

- E.g. Encouragement to diet or exercise; military draft
- Violations of exclusion principle may result in a mistaken causal effect.

Methods: Instrumental Variable Analysis

Other key assumptions are:

- Potential outcomes unrelated to treatment status of other individuals (Stable unit treatment value)
- Instrumental variable must predict treatment status (non-zero causal effect of instrument on treatment)
- Patients differ according to value of instrument are similar in terms of observed and unobserved covariates (ignorable assignment of instrument).
- Patient with low value of instrument would receive same treatment if they had a high value of instrument (monotonicity) (i.e. rules out the existence of defiers)

Methods: Instrumental Variable Analysis

- Silber et al. (2004) RCT of enalapril to protect hearts of children who had received an anthracycline as part of cancer chemo.
- Randomized 135 children to enalapril or placebo and measured cardiac function.
- Children ‘encouraged’ to take specific dose some took less than required dose.
- Any effect of encouragement on cardiac function is a consequence of biological effects of quantity consumed.
- Randomized encouragement is an instrument for the amount of enalapril consumed.

Methods: Instrumental Variable Analysis

- Are compliant and non-compliant children different?
- Randomization ensures that encouraged children are similar to not-encouraged children, but nothing to ensure that compliant children are comparable to less compliant children.
- If the dose for those encouraged to accept treatment is much larger than the dose for those not encouraged to accept treatment then the instrument is said to be strong.
- If the dose for those encouraged to accept treatment is the same as the dose for those not encouraged to accept treatment then the instrument is said to be weak.
- When instrument is weak the most popular method of analysis (two-stage least squares) tends to give incorrect p-values and confidence intervals (Imbens and Rosenbaum, 2005).

Methods: Instrumental Variable Analysis

- An (intention-to-treat) ITT analysis compares encouraged to not-encouraged group ignoring compliance.
- ITT analysis estimates the effect of encouragement to accept treatment, not the effect of treatment itself.
- E.g. encouragement to quit smoking might be highly ineffective, but quitting might be highly effective. Both are important but they are different.
- IV analysis estimates causal treatment effect in the subset of patients for whom the instrument determines treatment (i.e. compliers or marginal patients).

Outline

- Causal inference in randomized experiments
- Causal inference in nonrandomized studies
- Methods: (1) Propensity score, (2) Multivariable modeling (3) Instrumental variable)
- **Comparing Methods**
- Sensitivity analysis: People Who Look Comparable May Differ
- Planning the Analysis

Comparing methods

- Stukel et al. (2007) compared mortality of elderly Medicare patients with AMI who were eligible for cardiac catheterization.
- 7 year fu to assess association between long-term survival and cardiac cath. within 30 days of admission.
- Comparisons may be biased due to differences in patient prognosis between groups, often due to unobserved treatment selection bias.

Comparing methods

Compared four analytic methods for removing selection bias in observational studies: (1) multivariable model; (2) PS adjustment; (3) PS matching; (4) IV analysis (using regional cath. rates as instrument).

Methods	RR
Multivariable model	0.51 [0.50 - 0.52]
PS adjustment	0.54 [0.54 - 0.55]
PS matching	0.54 [0.52 – 0.56]
IV	0.84 [0.79 – 0.90]

Comparing methods

- Estimated treatment association 3 times higher depending on method used.
- Face validity: survival benefits of routine invasive care from RCTs between 8%-21%.
- RCTs are optimized and tend to overestimate relative benefits achievable in routine clinical practice.

Comparing methods

- Stukel et al. argue that overestimation of standard modeling due to selection of lower risk patients for cath.
- Receiving cath. requires survival from admission to treatment.
- Even controlling for complete information on patients' admission severity could not eliminate this bias.
- Similarity between multivariable and PS estimates expected since both control for measured covariates.

Comparing methods

- IV depends on finding a strong, valid instrument.
- IV estimate measures treatment effect on the subset of patients with uncertain indications whose likelihood of being treated depends on local clinical judgment and cath. lab supply (excludes patients that would always or never receive cardiac cath.).
- Treatment effect interpreted as potentially due to the instrument itself, as well as characteristics of care system associated with instrument.

Comparing methods

- IV analyses better suited to inform policy decisions.
- Region or physician is often at level which policy and resource allocation decisions are made.

Outline

- Causal inference in randomized experiments
- Causal inference in nonrandomized studies
- Methods: (1) Propensity score, (2) Multivariable modeling (3) Instrumental variable
- Comparing Methods
- **Sensitivity analysis: People Who Look Comparable May Differ**
- Planning the Analysis

Sensitivity analysis

- Stukel et al. claim that the large RR from the PS analysis were affected by unmeasured covariates such as selecting lower risk patients for catheterization.
- If an unmeasured covariate did exist that is driving the difference in mortality between the groups that received catheterization and the group that did not receive catheterization then how much would the RR change?

Sensitivity analysis

- A sensitivity analysis asks how the conclusion of an argument dependent upon assumptions would change if the assumptions were relaxed (Rosenbaum, 2010).
- Cornfield et al. (1959) asked about the magnitude of the bias from an unobserved covariate *u* that would alter the conclusion from observational studies that heavy smoking causes lung cancer.

Sensitivity analysis

Cornfield showed that if an unobserved non-causal agent A exists that could fully explain the observed risk ratio R0 it would have to satisfy two inequalities:

(1) R0 <= Ru;

(2) R0 <= f1/f0.

R0 = observed risk ratio for the putative agent, B.

Ru = risk ratio for the unobserved variable, A.

f1 = prevalence of A among those with B.

f0 = prevalence of A among those without B.

"if cigarette smokers [B] have 9 times the risk of nonsmokers for developing lung cancer [R0], and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X [A], then the proportion of hormone-X-producers among cigarette smokers [f1] must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect" (Cornfield et al. 1954).

Sensitivity analysis

- This does not mean that such a hormone exists, but limits scope of debate.
- To assert that smoking did not cause lung cancer, but was instead caused by a hidden bias due to hormone X, means you would have argue that hormone X is a near perfect predictor of lung cancer and is at least 9 times more common among smokers than among nonsmokers.

Sensitivity analysis

- PS models "work" if two people with the same observed covariates have the same probability of receiving treatment.
- The sensitivity analysis model says that this might be false.
- Two people with the same observed covariates have odds of treatment that differ by at most a multiplier Γ .

Sensitivity analysis

If two people, k and l, with the same observed covariates $x_k = x_l$, have odds of treatment, $\pi_k/(1-\pi_k)$ and $\pi_l/(1-\pi_l)$, that differ by at most a multiplier of Γ

$$\frac{1}{\Gamma} \le \frac{\pi_k/(1-\pi_k)}{\pi_l/(1-\pi_l)} \le \Gamma, \text{ whenever } x_k = x_l.$$

- If $\Gamma = 1$ when the study is randomized .
- $\Gamma > 1$ indicates bias in failure to control for unobserved covariate.
- E.g. if $\Gamma = 3$ and k and l look the same in terms of observed covariates then k might be three times more likely to receive treatment because of differences in ways not measured.

Sensitivity analysis

- How do inferences (e.g., P-value and CI) change for values of $\Gamma > 1$ when testing for no treatment effect?
- Suppose that we calculated a point estimate with P-value or CI from a paired observational study by applying methods for randomized experiments.
- Inferences would have advertised theoretical properties if $\Gamma = 1$ (i.e. in a randomized study).

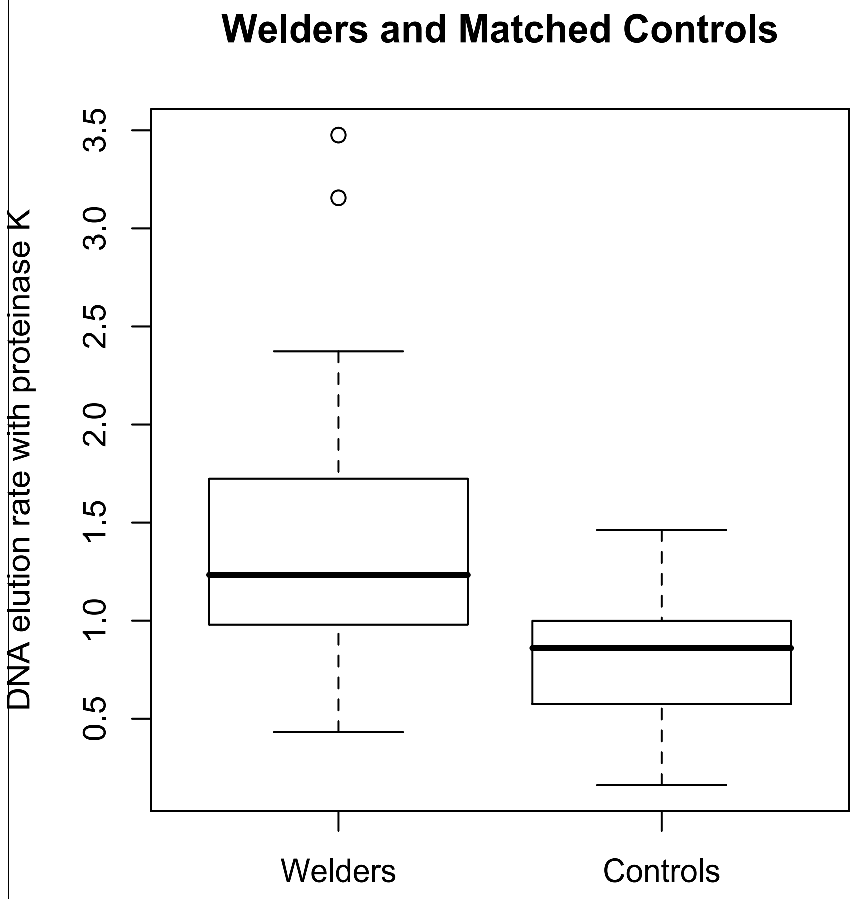
Sensitivity analysis

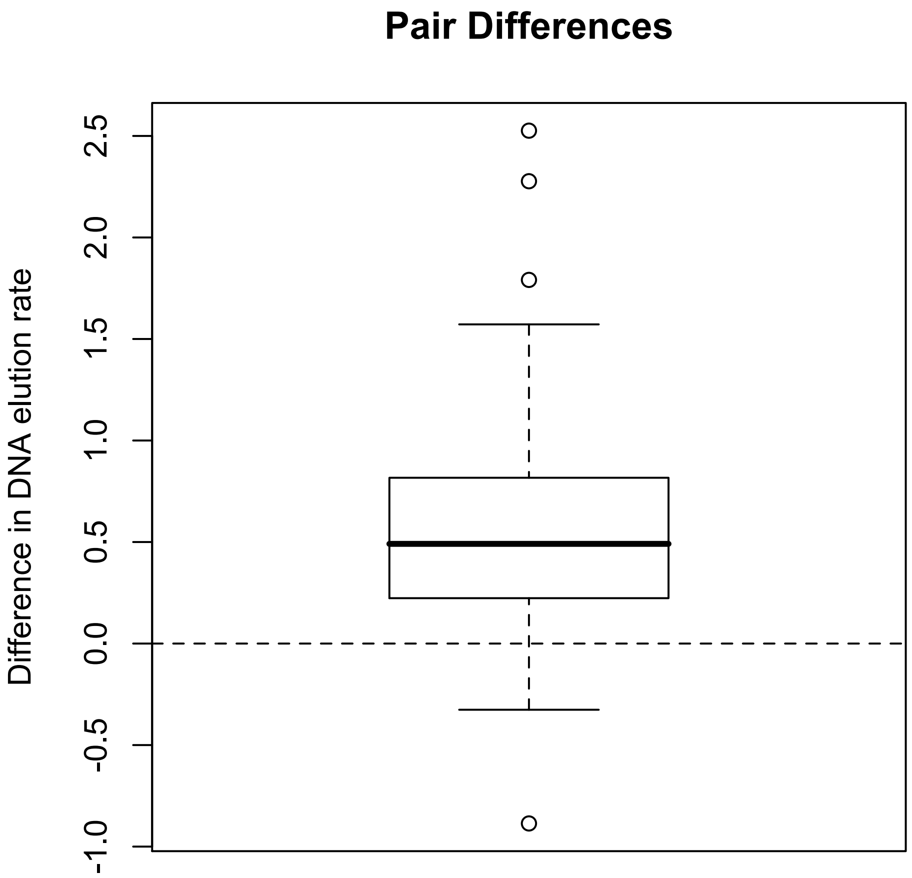
- If Γ =1 led to a P-value of, say, 0.001 and if Γ =2 led to a range of P-values from 0.0001 to 0.02.
- Bias would have to be larger than Γ =2 to make no treatment effect plausible.
- Every study is sensitive to sufficiently large biases.
- There is always a value of Γ such that the interval of possible P-values includes small (e.g. 0.0001) and large values (e.g. 0.1).

Sensitivity analysis

- E.g. (Rosenbaum, Pg.79-81): Compare DNA damage of 39 male welders to 39 male controls matched for age and smoking habits (Werfel et al.).
- Werfel et al. presented several measures of genetic damage including DNA single strand breakage and DNA-protein cross-links using elution rates through polycarbonate filters with proteinase K. Broken strands are expected to pass through the filters at a higher rate.

Sensitivity analysis





Sensitivity analysis

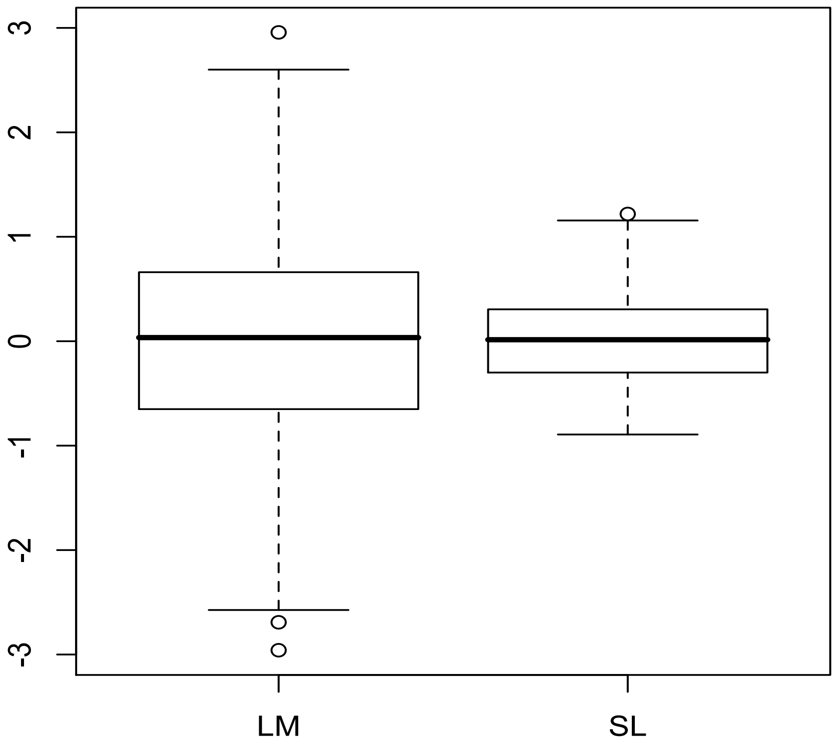
Sensitivity analysis for one-tailed test using Wicoxon signed rank test of no treatment effect of a male welder and control matched for age and smoking

Γ	P_{min}	P_{max}
1	3.1×10^{-7}	3.1×10^{-7}
2	3.4×10^{-12}	0.00064
3	$< 10^{-15}$	0.011
4	$< 10^{-15}$	0.047
5	$< 10^{-15}$	0.108

- $\Gamma = 1$ no uncertainty about treatment assignment (78 men paired for age and smoking and randomized to their careers)
- $\Gamma = 2$ two men that have same smoking status and age may not have same chance of a career as a welder: one man may be twice as likely to choose career as welder, because they differ in terms of an unmeasured covariate.
- Determine every possible P-value that could be produced when $\Gamma = 2$ (NB: In a matched pair one man might have prob. = 2/3 being welder and other might have prob. = 1/3).

Sensitivity analysis – Which is better a large heterogeneous study or a smaller study with less variability?

Suppose that an investigator faces a choice between a larger study (LM) with more heterogeneous responses or a smaller study (SM) with less heterogeneous responses (Rosenbaum, pg. 278-279).



Sensitivity analysis- Which is better a large heterogeneous study or a smaller study with less variability?

- In a simulated example compare P_{max} in a smaller study with 100 matched pairs and half the standard deviation of a larger study with 400 matched pairs.

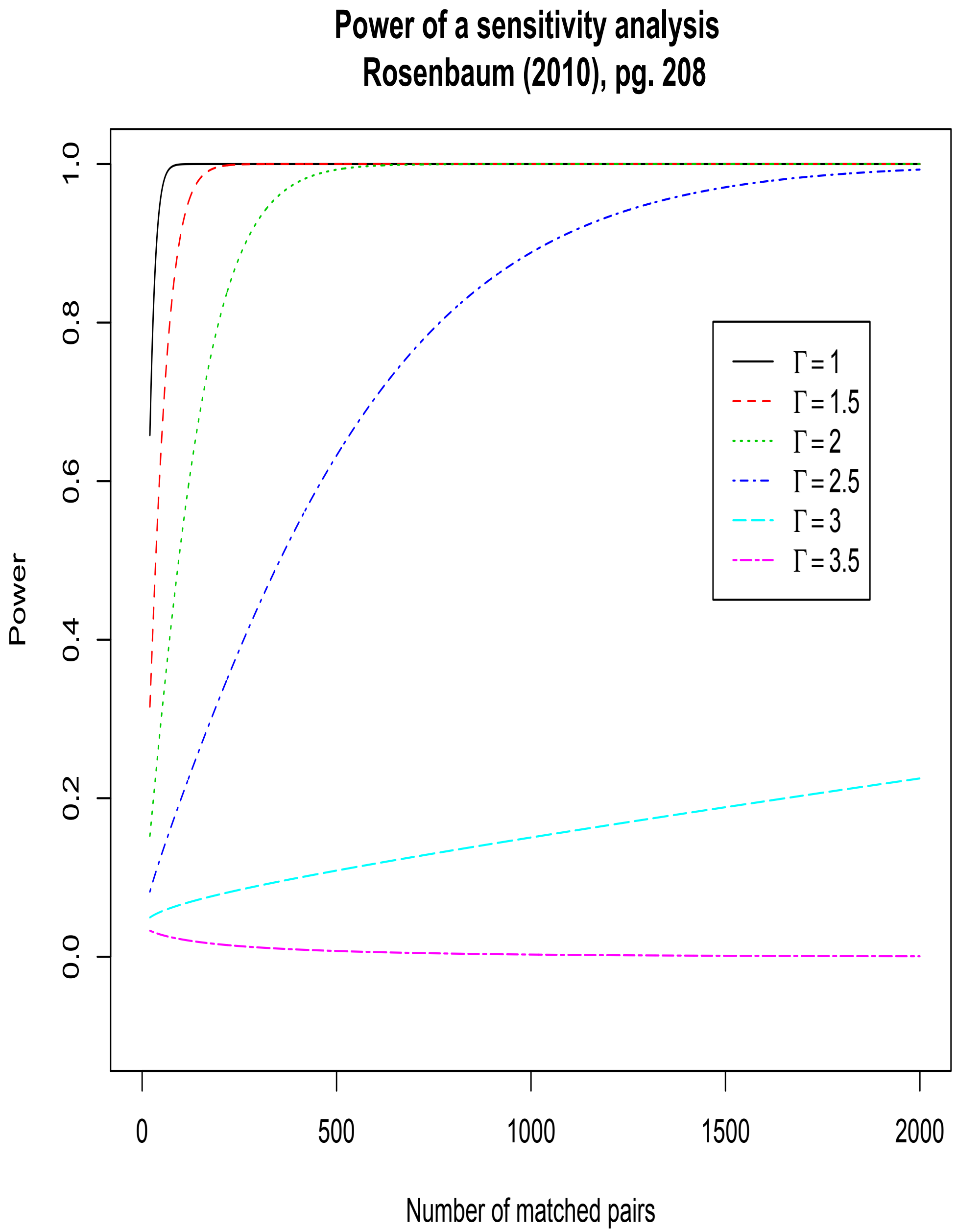
Γ	1	2	3	4	5
Large study	$< 10^{-10}$	0.00046	0.39	0.97	1.00
Small study	$< 10^{-10}$	0.000016	0.0021	0.022	0.083

Power of a sensitivity analysis

- Power of a statistical test is the probability that the test will detect a false null hypothesis and reject it.
- E.g. If test rejects null hypothesis when P-value <= 0.05 then power is the probability that the P-value will be less than or equal to 0.05 when the null is false.
- Rejection of a false null hypothesis highly probable = powerful test!

Power of a sensitivity analysis

- For a given magnitude of departure from random assignment an interval of possible P-values can be determined $[P_{min}, P_{max}]$.
- Power of a sensitivity analysis is the probability of $P_{max} <= \alpha$ (e.g. $\alpha = 0.05$) for a specified Γ.
- If treatment had an effect and no unmeasured bias then the power of sensitivity analysis for specified α and Γ is the probability that we will be able to say that a bias of magnitude Γ would not lead to acceptance of the null hypothesis of no effect when tested at level α.



- Suppose that we study designs where the treatment has an effect compared to control.
- Power ➔ 1 as I ⬆ for Γ < 3
- Power ➔ 0 as I ⬆ for Γ > 3
- Γ =3 is called “design sensitivity”
- Association between treatment and response can be distinguished for biases < 3

Outline

- Causal inference in randomized experiments
- Causal inference in nonrandomized studies
- Methods: (1) Propensity score, (2) Multivariable modeling (3) Instrumental variable)
- Comparing Methods
- Sensitivity analysis: People Who Look Comparable May Differ
- **Planning the Analysis**

Planning the Analysis (Ref. Rosenbaum, 2010, pgs. 353-3540

1. *The planner of an observational study should always ask himself the question,"How would the study be conducted if it were possible to do it by controlled experimentation?"*(Cochran, 1965).

- Typical structure is comparison of treated and control groups that are comparable prior to treatment.

Planning the Analysis

2. *Adjustments for observed covariates should be simple, transparent, and convincing.*

- Major source of uncertainty come from possible failure to control for covariates that were not measured.
- One simple, transparent way to adjust for observed covariates is to compare treated and control with similar distributions of observed covariates.

Planning the Analysis

3. *Plausible alternatives to treatment effect should be anticipated and addressed.*

- Not possible to anticipate every plausible objection to a claim that comparison of matched treated and control groups estimates treatment effect.
- Objections usually claim comparison is ambiguous and is distorted by a specific bias.
- With a specific form of bias in mind design elements are often possible to add, such as two control groups.

Planning the Analysis

4. *Analysis should address possible biases from unmeasured covariates.*

- Sensitivity analysis: how much bias from unmeasured covariates would need to be present to qualitatively alter the conclusions suggested by standard comparison

Planning the Analysis

5. *Design observational studies to be insensitive to biases from unmeasured covariates*

- Measure design sensitivity.

Planning the Analysis

6. *A plan for a primary analysis should exist.*

- "About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied": 'Make your theories elaborate.'
- An elaborate theory predicts a difference here, near equivalence there, the absence of a trend here, etc.

Planning the Analysis

- E.g. an elaborate theory might predict higher responses in the treated groups than in two control groups, with the two control groups differing negligibly from each other.
- Predictions of a theory are only predictions if they precede looking at the data (an elaborate theory is not meant to fit a dataset).

Planning the Analysis

1. How would you conduct the study if it was a controlled experiment?
2. Adjustments for observed covariates should be simple.
3. Plausible alternatives to treatment effect should be considered in addressed.
4. Analysis should address biases from possible from unmeasured covariates.
5. Design observational studies to be insensitive to biases from unmeasured covariates.
6. Plan a primary analysis.