RANDOMIZED EXPERIMENTS AND OBSERVATIONAL STUDIES: CAUSAL INFERENCE IN STATISTICS

PAUL R. ROSENBAUM

Abstract. This talk describes the theory of causal inference in randomized experiments and nonrandomized observational studies, using two simple theoretical/actual examples for illustration. Key ideas: causal effects, randomized experiments, adjustments for observed covariates, sensitivity analysis for unobserved covariates, reducing sensitivity to hidden bias using design strategies.

1. Seven Key Contributions to Causal Inference

1.0.1. ***Ronald A. Fisher (1935).*** *The Design of Experiments.* Edinburgh: Oliver & Boyd. Although Fisher had discussed his randomized experiments since the early 1920's, his most famous discussion appears in Chapter 2 of this book, in which Fisher's exact test for a 2×2 table is derived from randomization alone in the experiment of the 'lady tasting tea.'

1.0.2. ***Jerzy Neyman (1923).*** On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (In Polish) Roczniki *Nauk Roiniczych,* Tom X, pp1-51. Reprinted in English in *Statistical Science*, 1990, 5, 463-480, with discussion by T. Speed and D. Rubin. In this paper, Neyman writes the effects caused by treatments as comparisons of potential outcomes under alternative treatments.

1.0.3. ***Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959)***. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173-203. This paper contains the first sensitivity analysis in an observational study, replacing the qualitative statement that 'association does not imply causation' by a quantitative statement about the magnitude of hidden bias that would need to be present to explain away the observed association between treatment and response.

1.0.4. ***Donald T. Campbell (1957)***. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312. This is an early paper in Campbell's forty years of highly influential writings about observational studies or quasi-experiments, as he called them. Campbell insisted that the legitimate concern that 'association does not imply causation' must be given tangible form in specific rival explanations or 'threats to validity.' Once specified, a rival explanation led Campbell to study designs with added features to distinguish that rival explanation from an effect of the treatment.

Date: March 2004.

Supported by a grant from the U.S. National Science Foundation.

1.0.5. ***Austin Bradford Hill (1965)***. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine,* **58**, 295-300. Along with Richard Doll, Hill had been an author of some of the most influential observational studies providing evidence of the harmful effects caused by cigarette smoking. This particular paper proposed various considerations intended to aid judgement about whether an observation association between treatment and outcome is causal. The details of some of Hill's suggestions remain controversial, but his general point is not. We approach a study of treatment effects with scientific knowledge that certain patterns of effects are plausible and others are not. That knowledge, combined with expanded study of observed associations, provides evidence that aids in distinguishing actual effects from hidden biases.

1.0.6. ***William G. Cochran (1965)***. The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society,* A,128, 134-155. This paper defined observational studies in parallel with randomized experiments, systematically developing the tasks in research design, adjustments for observed covariates, and addressing hidden bias from unmeasured covariates.

1.0.7. ***Donald B. Rubin (1974)***. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701. Although it had been fairly standard since the 1920's to define treatment effects in randomized experiments as comparisons of potential outcomes under alternative treatments, this important paper began applying the notation systematically in observational studies. Arguably for the first time, the statement 'association does not imply causation' was written down formally, so that the observable association was one population quantity, the effect caused by the treatment was another, and the two were equal in a randomized experiment but not in a nonrandomized study. The paper provides formal insights into adjustments, when they might lead to consistent estimates of treatment effects, when they would fail.

2. A Randomized Experiment

2.1. 2×2 Table in a Randomized Experiment.

2.1.1. *Potential Responses, Causal Effects.* (Reference: Neyman (1923), Rubin (1974) Example: B. Fisher, et al. 2002) n women, $i = 1, \ldots, n$. Each woman i has two potential responses, (r_{Ti}, r_{Ci}), where:

$$r_{Ti} = \left[\begin{array}{l} 1 \text{ if woman } i \text{ would have cancer} \\ \quad \text{recurrence with lumpectomy alone} \\ 0 \text{ if woman } i \text{ would not have cancer} \\ \quad \text{recurrence with lumpectomy alone} \end{array}\right.$$

$$r_{Ci} = \left[\begin{array}{l} 1 \text{ if woman } i \text{ would have cancer} \\ \quad \text{recurrence with lumpectomy+irradiation} \\ 0 \text{ if woman } i \text{ would not have cancer} \\ \quad \text{recurrence with lumpectomy+irradiation} \end{array}\right.$$

but we see only one response or the other; never see the *causal effect*, $\delta_i = r_{Ti} - r_{Ci}$, $i = 1, \ldots, n$.

2.1.2. *Finite population.* A finite population of $n = 1,262$ women. The (r_{Ti}, r_{Ci}) are $2n$ fixed numbers describing the finite population. Nothing is random.

TABLE 1. Observable Table: Response by Treatment.

	Recurrence $R_i = 1$	*No recurrence* $R_i = 0$	Total
No rads $Z_i = 1$	$\sum Z_i\, R_i$	$\sum Z_i\ (1 - R_i)$	m
Rads $Z_i = 0$	$\sum (1 - Z_i)\ R_i$	$\sum (1 - Z_i)\ (1 - R_i)$	n-m

TABLE 2. Observable Table in Terms of Potential Reponses: Equals Table 1.

	Recurrence $R_i = 1$	*No recurrence* $R_i = 0$	Total
No rads $Z_i = 1$	$\sum Z_i\, r_{Ti}$	$\sum Z_i\ (1 - r_{Ti})$	m
Rads $Z_i = 0$	$\sum (1 - Z_i)\ r_{Ci}$	$\sum (1 - Z_i)\ (1 - r_{Ci})$	n-m

2.1.3. *No effect.* Is it plausible that irradiation does nothing? *Null hypothesis of no effect.* $H_0 : \delta_i = 0$, $i = 1, \ldots, n$.

2.1.4. *Measures of effect.* Estimate the *average treatment effect*: $\frac{1}{n} \sum_{i=1}^{n} \delta_i$. How many more women A had a recurrence of cancer because they did not receive irradiation? (*Attributable effect*)

2.1.5. *Randomized Experiment.* (Fisher 1935) Pick m of the n people at random and give them treatment condition T. In the experiment, $m = 634$, $n = 1,262$. This means that each of the $\binom{n}{m} = \binom{1,262}{634}$ treatment assignments has the same probability, $\binom{1,262}{634}^{-1}$. The only probabilities that enter Fisher's randomization inference are created by randomization. Write $Z_i = 1$ if i is assigned to T and $Z_i = 0$ if i is assigned to C; then Z_i is a random variable. Also, $m = \sum_{i=1}^{n} Z_i$.

2.1.6. *Observed response.* The observed response, $R_i = Z_i\, r_{Ti} + (1 - Z_i)\ r_{Ci} = r_{Ci} + Z_i\, \delta_i$, is a random variable because it depends on Z_i. That is, Table 1 equals Table 2.

2.1.7. *Attributable effect.* How many more women A had a recurrence of cancer because they did not receive irradiation? $A = \sum Z_i\, \delta_i = \sum Z_i\ (r_{Ti} - r_{Ci})$. Not observed. A random variable. Table 2 and Table 3 differ by the attributable effect, A.

2.1.8. *Testing hypothesized effects.* Consider the hypothesis $H_0 : \delta_i = \delta_{0i}$, $i = 1, \ldots, n = 1262$ with the δ_{0i} as *possible* specified values of δ_i. If H_0 were true, then $R_i - Z_i\, \delta_{0i}$ would equal r_{Ci}, and Table 4 would equal Table 3 and would have the hypergeometric distribution. Basis for test. Table 1 and Table 4 differ by the hypothesized value of the attributable effect, $A_0 = \sum Z_i\, \delta_{0i}$.

3. A MATCHED OBSERVATIONAL STUDY

3.1. **Notation.**

TABLE 3. Table of Responses That Would Have Been Observed Had Treatment Been Withheld. Not observed.

	Recurrence $r_{Ci}=1$	*No recurrence* $r_{Ci}=0$
No rads $Z_i=1$	$\sum Z_i\, r_{Ci}$	$\sum Z_i\, (1-r_{Ci})$
Rads $Z_i=0$	$\sum (1-Z_i)\, r_{Ci}$	$\sum (1-Z_i)\, (1-r_{Ci})$

TABLE 4. Observed Table Adjusted for Hypothesized Treatment Effect. Would Equal Table 3 if the Hypothesis Were True.

	Recurrence $R_i=1$	*No recurrence* $R_i=0$
No Rads $Z_i=1$	$\sum Z_i\, (R_i - Z_i\,\delta_{0i})$	$\sum Z_i\, (1-R_i + Z_i\,\delta_{0i})$
Rads $Z_i=0$	$\sum (1-Z_i)\, R_i$	$\sum (1-Z_i)\, (1-R_i)$

TABLE 5. Blood lead levels, in micrograms of lead per decaliter of blood, of exposed children whose fathers worked in a battery factory and age-matched control children from the neigborhood. Exposed father's lead exposure at work (high, medium, low) and hygiene upon leaving the factory (poor, moderate, good) are also given. Adapted for illustration from Tables 1, 2 and 3 of Morton, et al. (1982).

s	Exposure	Hygiene	Exposed Child's Lead Level μg/dl	Control Child's Lead Level μg/dl	Dose Score
1	high	good	14	13	1.0
2	high	moderate	41	18	1.5
3	high	poor	43	11	2.0
⋮	⋮	⋮	⋮	⋮	⋮
33	low	poor	10	13	1.0

3.1.1. *Lead example.* Example is: Morton, et al. (1982). $S = 33$ pairs, $s = 1,\ldots,S = 33$, with 2 subjects in each pair, $i = 1, 2$.

3.1.2. *One treated, one control in each pair.* Write $Z_{si} = 1$ if the i^{th} subject in pair s is treated, $Z_{si} = 0$ if control, so $Z_{s1} + Z_{s2} = 1$ for every s, or $Z_{s2} = 1 - Z_{s1}$. $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{S1}, Z_{S2})^T$. There are 2^S possible $\mathbf{Z}$, and a paired randomized experiment would pick one at random.

3.1.3. *Potential responses, causal effects, finite population, as before.* Each of the $2S$ subjects (s, i) has two potential responses, a response r_{Tsi} that would be seen under treatment and a response r_{Csi} that would be seen under control. (Neyman 1923, Rubin 1974). Treatment effect is $\delta_{si} = r_{Tsi} - r_{Csi}$. *Additive effect,* $r_{Tsi} -$

$r_{Csi} = \tau$ or $\delta_{si} = \tau$ for all s, i. The (r_{Tsi}, r_{Csi}), $s = 1, \ldots, S$, $i = 1, 2$, are again fixed features of the finite population of $2S$ subjects.

3.1.4. *Observed responses, as before.* Observed response is $R_{si} = r_{Tsi}$ if $Z_{si} = 1$ or $R_{si} = r_{Csi}$ if $Z_{si} = 0$, that is, $R_{si} = Z_{si}\, r_{Tsi} + (1 - Z_{si})\, r_{Csi} = r_{Csi} + Z_{si}\, \delta_{si}$. If effect is additive, $R_{si} = r_{Csi} + Z_{si}\, \tau$.

3.1.5. *Treated-minus-control differences.* If $r_{Tsi} - r_{Csi} = \tau$, then the *treated-minus-control difference* in observed responses in pair s is $D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2}) + \tau$. In general, with $\delta_{si} = r_{Tsi} - r_{Csi}$, $D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2}) + Z_{s1}\,\delta_{s1} + Z_{s2}\,\delta_{s2}$.

3.2. **Inference in a Paired Randomized Experiment.**

3.2.1. *Wilcoxon's signed rank statistic.* W ranks the $|D_s|$ from 1 to S, and sums the ranks of the positive D_s. Alternatively, W is the number of positive Walsh averages, $(D_s + D_{s'})/2$, with $1 \le s \le s' \le S$.

3.2.2. *Null distribution of W in a randomized experiment.* If $H_0 : \delta_{si} = 0$ were true for $s = 1, \ldots, S$, $i = 1, 2$ *in a randomized experiment,* then $D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2})$ where $Z_{s1} - Z_{s2}$ is ± 1 where randomization ensures $\Pr(Z_{s1} - Z_{s2} = 1) = \frac{1}{2}$, independently in different pairs, and $r_{Cs1} - r_{Cs2}$ is fixed in Fisher's finite population, so $|D_s| = |r_{Cs1} - r_{Cs2}|$ is fixed, as is its rank, so ranks independently add to W with probability $\frac{1}{2}$, generating W's distribution. Uses just fact of randomization and null hypothesis, so forms the "reasoned basis for inference," in Fisher's phrase.

3.2.3. *Inference about additive effects in randomized experiments.* If $H_0 : \delta_{si} = \tau_0$ were true for $s = 1, \ldots, S$, $i = 1, 2$ in a randomized experiment, then $D_s - \tau_0 = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2})$, and W computed from $D_s - \tau_0$ has the usual null distribution of the signed rank statistic. This test is inverted for confidence intervals and Hodges-Lehmann point estimates. Again, the inference uses only the fact of randomization and the null hypothesis being tested. Additive effects may be dropped; inference then concerns offsets attributable to treatment.

3.3. **Simple Model for Observational Studies.**

3.3.1. *Unknown treatment assignment probabilities.* An observational study is a study of treatment effects in which each person has an unknown probability of treatment, typically different probabilities for different people.

3.3.2. *Simple model.* In some finite population of people, $j = 1, \ldots, J$, person j has probability $\pi_j = \Pr(Z_j = 1)$ of exposure to treatment, independently, where π_j is not known. Probabilities are always *conditional on things we regard as fixed*, usually measured and unmeasured covariates, potential outcomes, (r_{Tj}, r_{Cj}), etc.

3.3.3. *Covariates.* The people, $j = 1, \ldots, J$, in the finite population have observed covariates $\mathbf{x}_j$ and unobserved covariate u_j. In the example, $\mathbf{x}_j$ describes child's age and neighborhood.

3.3.4. *Exact matching for observed covariates.* Select S pairs, $i = 1, 2$, one treated, one control, from the J people in the population. Match exactly for $\mathbf{x}$, so that $\mathbf{x}_{s1} = \mathbf{x}_{s2}$ for each s, $s = 1, \ldots, S$. In this simplest case, the matching algorithm is permitted to use only $\mathbf{x}$ and $1 = Z_{s1} + Z_{s2}$. Within matched pairs, the relevant treatment assignment probabilities are conditional probabilities $\Pr(Z_{s1} = 1 \,|\, Z_{s1} + Z_{s2} = 1)$.

3.4. **Adjustments for Observed Covariates: When Do They Work?**

3.4.1. *Free of hidden bias.* Treatment assignment is free of hidden bias if π_j is a (typically unknown) function of $\mathbf{x}_j$ — two people with the same $\mathbf{x}_j$ have the same π_j.

3.4.2. *Matching works if free of hidden bias.* If free of hidden bias and we match exactly for $\mathbf{x}$, so $\mathbf{x}_{s1} = \mathbf{x}_{s2}$, then

$$\Pr\left(Z_{s1} = 1 \,|\, Z_{s1} + Z_{s2}\right) = \frac{\pi_{s1}\left(1-\pi_{s2}\right)}{\pi_{s1}\left(1-\pi_{s2}\right)+\pi_{s2}\left(1-\pi_{s1}\right)} = \frac{1}{2} \tag{3.1}$$

because $\pi_{s1} = \pi_{s2}$. A little more work shows that we get the randomization distribution by conditioning. Identifies the key assumption, but of course, doesn't make it true. In contrast, in an experiment, randomization makes the assumption true.

3.4.3. *Divides methods.* Methods of adjustment for $\mathbf{x}$ should work when study is free of hidden bias. Need other methods to address concerns about whether the study is free of hidden bias.

3.5. **Propensity Scores.**

3.5.1. *Many covariates.* If $\mathbf{x}$ is of high dimension, it's hard to match. With just 20 binary covariates, there are 2^{20} or about a million covariate patterns.

3.5.2. *Propensity scores. If the study is free of hidden bias*, then two people with the same $\mathbf{x}_j$ have the same π_j, so π_j is a function of $\mathbf{x}_j$, say $\pi_j = e\left(\mathbf{x}_j\right)$, which is then called the propensity score. *If the study is free of hidden bias*, then don't need to match on high dimension $\mathbf{x}$, just need to match on the scalar $e\left(\mathbf{x}\right)$: if $e\left(\mathbf{x}_{s1}\right) = e\left(\mathbf{x}_{s2}\right)$ then $\pi_{s1} = \pi_{s2}$, and (3.1) is true even if $\mathbf{x}_{s1} \neq \mathbf{x}_{s2}$.

3.5.3. *Whether or not the study is free of hidden bias,* matching on propensity scores $e = e\left(\mathbf{x}\right)$ tends to balance the observed covariates $\mathbf{x}$ used in the score. Define $e = e\left(\mathbf{x}\right) = \Pr\left(Z = 1 \,|\, \mathbf{x}\right)$, so the study is free of hidden bias if $\pi_j = e\left(\mathbf{x}_j\right)$ for all j, but $e\left(\mathbf{x}\right)$ is defined even if π_j depends on things besides $\mathbf{x}$. Then

$$\Pr\left(\mathbf{x} \,|\, Z = 1, e\right) = \Pr\left(\mathbf{x} \,|\, Z = 0, e\right) \quad \text{or} \quad \mathbf{x} \perp\!\!\!\perp Z \,|\, e\left(\mathbf{x}\right);$$

see Rosenbaum and Rubin (1983).

3.6. **Addressing Bias from Unobserved Covariates: Sensitivity Analysis.**

3.6.1. *Common objection.* Critic says: "Adjusting for $\mathbf{x}_j$ is not sufficient, because there is an unobserved u_j, and adjustments for $(\mathbf{x}_j, u_j)$ were needed."

3.6.2. *Question answered by a sensitivity analysis.* If the critic's objection were true, if the association between treatment Z_j and response R_j were due to hidden bias from u_j, then what would u_j have to be like? What is the critic's counter claim is actually claiming? The answer varies markedly: studies vary markedly in how sensitive they are to hidden bias. First sensitivity analysis by Cornfield, et al. (1959) concerned smoking and lung cancer.

3.6.3. *Sensitivity Model.* Before matching, two subjects, j and k, with the same observed covariates, $\mathbf{x}_j = \mathbf{x}_k$, may differ in terms of u_j and u_k so that their odds of exposure to treatment differ by a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_j\,(1-\pi_k)}{\pi_k\,(1-\pi_j)} \leq \Gamma. \tag{3.2}$$

Free of hidden bias if $\Gamma = 1$. If $\Gamma > 1$, the unknown π_j cannot be eliminated, as before, by matching on $\mathbf{x}_j$, so the randomization distribution is no longer justified. If $\Gamma = 1.001$, the π_j are unknown, but almost the same, but if $\Gamma = 5$, π_j are unknown and could be very different. Plan: For each $\Gamma \geq 1$, find upper and lower bounds on inference quantities, like P-values (or endpoints of confidence intervals), for π_j's satisfying (3.2). Report these for several Γ. When do conclusions begin to change? Replaces qualitative "association does not imply causation," by a quantitative statement based on observed data, "to explain away observed associations as noncausal, hidden biases would have to be of such and such a magnitude."

As before, match on observed covariates $\mathbf{x}$, to form S pairs, $s = 1, \ldots, S$, $i = 1, 2$, with $\mathbf{x}_{s1} = \mathbf{x}_{s2}$, one treated, one control, $Z_{s1} + Z_{s2} = 1$. Then (3.2) implies:

$$\frac{1}{1+\Gamma} \leq \Pr\left(\, Z_{s1} = 1 \,|\, Z_{s1} + Z_{s2}\right) \leq \frac{\Gamma}{1+\Gamma}$$

which places sharp upper and lower bounds on the distribution of W and resulting inferences. Whole argument applies much more generally.

3.7. **Addressing Bias from Unobserved Covariates: Pattern Specificity.**

3.7.1. *Fisher's View.* Cochran (1965, §5) "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.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. ... this multi-phasic attack is one of the most potent weapons in observational studies."

3.7.2. *Pattern Matching and Sensitivity to Hidden Bias.* Can determine whether pattern specificity reduces sensitivity to hidden bias, and if so, by how much. Can appraise strategies for the design of observational studies in terms of the degree to which they reduce sensitivity to hidden bias.

4. Summary

4.0.3. *Causal effects.* Comparison of potential outcomes under competing treatments — not jointly observable (Neyman 1923, Rubin 1974). .

4.0.4. *Randomized experiments.* Permit inference about the effects caused by treatments (Fisher 1935).

4.0.5. *Observational studies: Adjustments.* Without randomization, adjustments are required. Straightforward for observed covariates, but there might be important covariates that you did not observe. (Cochran 1965)

4.0.6. *Observational studies: Sensitivity analysis.* What would unobserved covariates have to be like to alter conclusions? (Cornfield, et al.)

4.0.7. *Observational studies: Pattern matching, elaborate theories.* Reducing sensitivity to hidden bias. (Campbell 1988, Hill 1965, Cochran 1965)

5. Bibliography

Much of the material in this talk is discussed in my book, Rosenbaum, P. R. (2002) *Observational Studies*, 2^{nd} edition, NY: Springer Verlag.

Key: AE = attributable effects; CE = causal effects; EG = example used in talk; OS = observational studies; PM = pattern matching; PS = propensity score; RE = randomized experiments; RI = randomization inference; SA = sensitivity analysis.

References

[1] Box, Joan Fisher (1978) *R. A. Fisher: The Life of a Scientist.* New York: John Wiley. RE, RI

[2] Braitman, L. E. and Rosenbaum, P. R. (2002). Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of Internal Medicine* **137**, 693-695. PS

[3] Campbell, D. T. (1957) Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312. Reprinted in: In *Methodology and Epistemology for Social Science*, Ed. E. S. Overman, Chicago: University of Chicago Press. OS, PM

[4] Campbell, D. T. (1988). Definitional vs multiple operationalism. In *Methodology and Epistemology for Social Science*, Ed. E. S. Overman, pp. 32-6. Chicago: University of Chicago Press. PM

[5] Cochran, W. G. (1965) The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society,* A,128, 134-155. OS, PM

[6] Cook, T. D., Campbell, D. T. & Peracchio, L. (1990). Quasi-experimentation. In *Handbook of Industrial and Organizational Psychology*, Ed. M. Dunnette and L. Hough, pp. 491—576. Palo Alto, CA: Consulting Psychologists Press. OS, PM

[7] Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology* **45**, 545-80. OS, PM, RE

[8] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173-203. SA

[9] Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments.* New York: CRC Press. RE, RI

[10] Copas, J. B. and Li, H.G. (1997) Inference for non-random samples (with discussion). *JRSS* **B59** 55-96. SA

[11] Dawson, J. D. & Lagakos, S. W. (1993). Size and power of two-sample tests of repeated measures data. *Biometrics* **49**, 1022-35.

[12] Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053-1062. PS

[13] Fisher, Ronald.A. (1935). *The Design of Experiments.* Edinburgh: Oliver & Boyd. RE, RI

[14] Fisher, B., et al. (2002) Twenty year followup of a randomized trial comparing total mastectomy, lumpectomy and lumpectomy plus irradiation for the treatment of invasive breast cancer. *New England Journal of Medicine*, 17 October 2002, **347**, 1233-1241. EG, RE

[15] Friedman, L. M., DeMets, D. L., and Furberg, C. D. (1998) *Fundamentals of Clinical Trials.* New York: Springer-Verlag. RE

[16] Gail, M. H., Tan, W. Y., and Piantadosi, S. (1988) Tests for no treatment effect in randomized clinical trials. *Biometrika* **75** 57-64. RI

[17] Gastwirth, J. L. (1992) Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33** 19-34. SA

[18] Greenhouse, S. (1982) Jerome Cornfield's contributions to epidemiology. *Biometrics* Supplement, 33-45. SA

[19] Greevy, R., Silber, J. H., Cnaan, A. and Rosenbaum, P. R. (2004) Randomization inference with imperfect compliance in the ACE-Inhibitor after anthracycline randomized trial. *JASA*, to appear. CE, RE, RI

[20] Hamilton, M. A. (1979) Choosing the parameter for 2×2 and $2 \times 2 \times 2$ table analysis. *American Journal of Epidemiology* 109, 362-75. CE

[21] Hammond, E. C. (1964) Smoking in relation to mortality and morbidity: Findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, **32**, 1161–1188. EG, OS

[22] Herbst, A., Ulfelder, H., and Poskanzer, D. (1971) Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, **284**, 878–881. EG, OS

[23] Hill, A, B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine,* **58**, 295-300. OS, PM

[24] Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**, 706– 710. PS

[25] Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93**, 126-132. SA

[26] Jick, H., Miettinen, O., Neff, R., et al. (1973) Coffee and myocardial infarction. *New England Journal of Medicine*, **289**, 63–77. EG, OS

[27] Joffe, M. M. and Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology* **150**, 327-333. PS

[28] Kempthorne, O. (1952) *Design and Analysis of Experiments.* New York: John Wiley. RE, RI

[29] Lehmann, E. L. (1998) *Nonparametrics: Statistical Methods Based on Ranks.* Upper Saddle River, NJ: Prentice Hall. RI

[30] Lu, B., Zanutto, E., Hornik, R. & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Am. Statist. Assoc.* **96**, 1245-53. PS

[31] Meyer, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econ. Statist.* **13**, 151—61. OS

[32] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., and Saah, M. (1982) Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, **115**, 549-555. EG, OS

[33] Neyman, J. (1923), On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (In Polish) Roczniki *Nauk Roiniczych,* Tom X, pp1-51. Reprinted in English in *Statistical Science*, 1990, 5, 463-480, with discussion by T. Speed and D. Rubin. CE

[34] Neyman, J. (1942) Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society,* 105, 292-327.

[35] Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. *Br. J. Cancer* **34**, 585–612. RE

[36] Piantadosi, S. (1997) *Clinical Trials.* New York: Wiley. RE

[37] Pitman, E. J. G. (1937) Significance tests which may be applied to samples from any population. *JRSS*, **4**, 119-130. RI

[38] Raz, J. (1990) Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *JASA* **85** 132-138. RI

[39] Robins, J. M., Rotnitzkey, A. & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, pp. 1-94. New York: Springer. SA

[40] Rosenbaum P. R. (1984) From association to causation in observational studies. *Journal of the American Statistical Association,* **79**, 41-48. OS, PM

[41] Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79**, 565-574. PS

[42] Rosenbaum, P. R. (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74, 13-26. SA

[43] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387-394. PS
[44] Rosenbaum, P. R. (1997). Signed rank statistics for coherent predictions. *Biometrics* **53**, 556-66. PM, SA
[45] Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies (with Discussion). *Statist. Sci.* **14**, 259-304. OS, PM
[46] Rosenbaum, P. R. (2001) Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, **88**, 219-231. AE, RI, SA
[47] Rosenbaum, P. R. (2002) *Observational Studies* (Second Edition). New York: Springer-Verlag. AE, CE, OS, PM, PS, RE, RI, SA
[48] Rosenbaum, P.R. (2002) Covariance adjustment in randomized experiments and observational studies (with Discussion). *Statistical Science* **17,** 286-327. OS, PS, RE, RI, SA
[49] Rosenbaum, P. R. (2003). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1-10. PM, SA
[50] Rosenbaum, P. R. (2003) Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician*, **57**, 132-138. AE, RI, SA
[51] Rosenbaum, P. R. (2004) Design sensitivity in observational studies. *Biometrika*, **91**, 153-164. PM, SA
[52] Rosenbaum, P. R. (2004) The case-only odds ratio as a causal parameter. *Biometrics*, **60**, 233-240. CE, OS
[53] Rosenbaum, P. R., Rubin, D. B. (1983) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *JRSS*, **B45**, 212–218. SA
[54] Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55. PS
[55] Rosenbaum, P. & Rubin, D. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516-524. EG, PS
[56] Rosenzweig, M. R. and Wolpin, K. I. (2000) Natural "natural experiments" in economics. *Journal of Economic Literature*, **38**, 827–874. OS
[57] Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701. CE, RE, OS
[58] Rubin, D. B. (1977) Randomization on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1-26. CE, RE, OS
[59] Rubin, D. B. (1984) William G. Cochran's contributions to the design, analysis and evaluation of observational studies. In: *W. G. Cochran's Impact on Statistics,* eds., P. S. R. S. Rao and J. Sedransk, New York: John Wiley, pp. 37-69. OS
[60] Reynolds, K. D. & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*. **11**, 691-714. PM
[61] Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin. OS, PM
[62] Trochim, W. M. K. (1985). Pattern matching, validity and conceptualization in program evaluation. *Evaluation Review*. **9**, 575-604. PM
[63] Weed, D. L. & Hursting, S. D. (1998). Biologic plausibility in causal inference: current method and practice. *American Journal of Epidemiology* **147**, 415-25. PM
[64] Weiss, N. (1981). Inferring causal relationships: Elaboration of the criterion of 'dose-response.' *American Journal of Epidemiology* **113**, 487–90. PM
[65] Weiss, N. (2002) Can the 'specificity' of an association be reahbilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13, 6-8. PM
[66] Welch, B. L. (1937) On the z-test in randomized blocks and latin squares. *Biometrika*, **29**, 21-52. CE, RI

Department of Statistics, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA.
E-mail address: rosenbaum@stat.wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~rosenbap/index.html