Experiments & Observational Studies:

Causal Inference in Statistics

Paul R. Rosenbaum

Department of Statistics

University of Pennsylvania

Philadelphia, PA 19104-6340

1 A Causal Question

- At age 45, Ms. Smith is diagnosed with stage II breast cancer.

- Her oncologist discusses with her two possible treatments: (i) lumpectomy alone, or (ii) lumpectomy plus irradiation. They decide on (ii).

- Ten years later, Ms. Smith is alive and the tumor has not recurred.

- Her surgeon, Steve, and her radiologist, Rachael debate.

- Rachael says: “The irradiation prevented the recurrence — without it, the tumor would have recurred.”

- Steve says: “You can’t know that. It’s a fantasy — you’re making it up. We’ll never know.”

2 Many Causal Questions

- Steve and Rachael have this debate all the time. About Ms. Jones, who had lumpectomy alone. About Ms. Davis, whose tumor recurred after a year.

- Whenever a patient treated with irradiation remains disease free, Rachael says: “It was the irradiation.” Steve says: “You can’t know that. It’s a fantasy. We’ll never know.”

- Rachael says: “Let’s keep score, add ’em up.” Steve says: “You don’t know what would have happened to Ms. Smith, or Ms. Jones, or Ms Davis — you just made it all up, it’s all fantasy. Common sense says: ‘A sum of fantasies is total fantasy.’ Common sense says: ‘You can’t add fantasies and get facts.’ Common sense says: ‘You can’t prove causality with statistics.’”

3 Fred Mosteller's Comment

- Mosteller like to say: "You can *only* prove causality with statistics."

- He was thinking about a particular statistical method and a particular statistician.

- Not Gauss and least squares, or Yule and Yule's Q (a function of the odds ratio), or Wright and path analysis, or Student and the t-test.

- Rather, Sir Ronald Fisher and randomized experiments.

4 15 Pages

- Fisher's clearest and most forceful discussion of randomization as 'the reasoned basis for inference' in experiments came in his book of 1935, *Design of Experiments*.

- In particular, the 15 pages of Chapter 2 discuss what came to be known as Fisher's exact test for a 2×2 table. The hypergeometric distribution is dispatched in half a paragraph, and Fisher hammers away in English for $14\frac{1}{2}$ pages about something else.

- Of Fisher's method of randomization and randomization, Yule would write: "I simply cannot make head or tail of what the man is doing." (Box 1978, p. 150). But Neyman (1942, p. 311) would describe it as "a very brilliant method."

5 Lumpectomy and Irradiation

- Actually, Rachael was right, Steve was wrong. Perhaps not in every case, but in many cases. The addition of irradiation to lumpectomy causes there to be fewer recurrences of breast cancer.

- On 17 October 2002, the *New England Journal of Medicine* published a paper by Bernard Fisher, et al. describing 20 year follow-up of a randomized trial comparing lumpectomy alone and lumpectomy plus irradiation.

- There were 634 women randomly assigned to lumpectomy, 628 to lumpectomy plus irradiation.

- Over 20 years of follow-up, 39% of those who had lumpectomy alone had a recurrence of cancer, as opposed to 14% of those who had lumpectomy plus irradiation ($P<0.001$).

6 Outline: Causal Inference

. . . in randomized experiments.

■ Causal effects. ■ Randomization tests of no effect. ■ Inference about magnitudes of effect.

. . . in observational studies.

■ What happens when randomized experiments are not possible? ■ Adjustments for overt biases: How to do it. When does it work or fail. ■ Sensitivity to hidden bias.

7 Finite Population

- In Fisher's formulation, randomization inference concerns a finite population of n subjects, the n subjects actually included in the experiment, $i = 1, \ldots, n$.

- Say $n = 1,262$, in the randomized experiment comparing lumpectomy (634) vs lumpectomy plus irradiation (628).

- The inference is *not* to some other population. The inference is to how these n people would have responded under treatments they did not receive.

- We are not sampling people. We are sampling possible futures for n fixed people.

- Donald Campbell would emphasize the distinction between *internal* and *external* validity.

8 Causal Effects: Potential Outcomes

- Key references: Neyman (1923), Rubin (1974).

- Each person i has two potential responses, a response that would be observed under the 'treatment' condition T and a response that would be observed under the 'control' condition C.

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

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

- We see r_{Ti} or r_{Ci}, but never both. For Ms. Smith, we saw r_{Ci}.

9 Comparing Potential Outcomes

- r_{Ti} is the response observed from i under lumpectomy alone, and r_{Ci} is observed from i under lumpectomy plus irradiation.

- The effect of the treatment is a comparisons of r_{Ti} and r_{Ci}, such as $\delta_i = r_{Ti} - r_{Ci}$. Possibilities:

r_{Ti}	r_{Ci}	δ_i	
1	1	0	cancer recurrence either way
1	0	1	irradiation prevents recurrence
0	1	-1	irradiation causes recurrence
0	0	0	no recurrence either way

- If someone gave us (r_{Ti}, r_{Ci}), $i = 1, \ldots, n$, causal inference would be arithmetic, not inference. But we never see δ_i for any i. We don't know δ_i for $i = Ms.\ Smith$.

10 Recap

- A finite population of $n = 1,262$ women.

- Each woman has two potential responses, $(r_{Ti},\, r_{Ci})$, but we see only one of them. Never see $\delta_i = r_{Ti} - r_{Ci}$, $i = 1, \ldots, n$.

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

- Estimate the *average treatment effect*: $\frac{1}{n}\sum_{i=1}^{n} \delta_i$.

- How many more women had a recurrence of cancer because they did not receive irradiation? (*Attributable effect*)

- The $(r_{Ti},\, r_{Ci})$ are $2n$ fixed numbers describing the finite population. Nothing is random.

11 Fisher's Idea: Randomization

- Randomization converts impossible arithmetic into feasible statistical inference.

- Pick m of the n people at random and give them treatment condition T. In the experiment, $m = 634$, $n = 1,262$. That is, assign treatments

> "in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collections of random sampling numbers. . ." (Fisher, 1935, Chapter 2)

- 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.

12 Observable Quantities

- Write $Z_i = 1$ if i is assigned to T and $Z_i = 0$ if i is assigned to C. Then $m = \sum_{i=1}^{n} Z_i$.

- Write R_i for the observed response from i. Then:

  $$R_i = \left[\begin{array}{ll} r_{Ti} \text{ if } Z_i = 1 & \text{(randomly assigned to} \\ & \quad \text{lumpectomy)} \\ r_{Ci} \text{ if } Z_i = 0 & \text{(randomly assigned to} \\ & \quad \text{lumpectomy+irradiation)} \end{array} \right.$$

  or formally

  $$R_i = Z_i \, r_{Ti} + (1 - Z_i) \, r_{Ci} = r_{Ci} + Z_i \, \delta_i .$$

- Unlike the causal effect, δ_i, which are fixed but unobservable features of the finite population, the Z_i and R_i are observable random variables.

13 The Observable 2×2 Table

	$Recurrence$ $R_i = 1$	$No\ recurrence$ $R_i = 0$	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

	$Recurrence$ $R_i = 1$	$No\ recurrence$ $R_i = 0$	Total
No rads $Z_i = 1$	220	414	634
Rads $Z_i = 0$	78	550	628
Total	298	964	$1,262$

14 Testing No Effect

- If the treatment has no effect, $H_0 : \delta_i = 0$ for $i = 1, \ldots, n$, then

$$\begin{aligned} 0 &= \delta_i = r_{Ti} - r_{Ci} \\ \text{or } r_{Ti} &= r_{Ci},\ i = 1, \ldots, n. \end{aligned}$$

- The observed response is then

$$R_i = r_{Ci} + Z_i \, \delta_i = r_{Ci}$$

  is just r_{Ci}, which is fixed, not varying with the treatment assignment Z_i.

- If the null hypothesis were true, then irradiation doesn't affect whether cancer recurs — we observe $R_i = r_{Ci}$ with or without irradiation.

If the null hypothesis were true, the responses in the lumpectomy-alone group are just a simple random sample (without replacement) of size m from a finite populations of size n consisting of the n binary r_{Ci}'s.

15 2×2 Table Under No effect: Fisher's Exact Test

- If the treatment has no effect, $H_0 : \delta_i = 0$ for $i = 1, \ldots, n$, then $R_i = r_{Ci} + Z_i \, \delta_i = r_{Ci}$, and the observable table of Z_i by R_i equals the table of Z_i by r_{Ci}:

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

  which has the hypergeometric distribution from the randomization.

- That is, under the null hypothesis, $\sum_{i=1}^{n} Z_i \, r_{Ci}$ is the total in a simple random sample without replacement of size m from a population of size n containing $\sum_{i=1}^{n} r_{Ci}$ 1's and $\sum_{i=1}^{n} (1 - r_{Ci})$ 0's.

16 Fisher's Exact Test

	Recurrence $R_i = 1$	*No recurrence* $R_i = 0$	Total
No rads $Z_i = 1$	220	414	634
Rads $Z_i = 0$	78	550	628
Total	298	964	$1,262$

- If the null hypothesis were true, so the corner cell had the hypergeometric distribution, then $\Pr(T \geq 220) = 2.7 \times 10^{-21}$.

- That is, if irradiation changed nothing, then the experiment randomly split 1,262 people into 634 and 628.

- A random split would produce the 220/78 split (or larger) of recurrences by chance with probability 2.7×10^{-21}.

17 How far have we come?

- We never see any causal effects, δ_i.

- Yet we are $100\left(1 - 2.7 \times 10^{-21}\right)\%$ confident that some $\delta_i > 0$.

- Causal inference is impossible at the level of an individual, i, but it is straightforward for a population of n individuals if treatments are randomly assigned.

- Mosteller's comment: "You can only prove causality with statistics."

18 Testing other hypotheses

- Recall that $\delta_i = r_{Ti} - r_{Ci}$, and Fisher's exact test rejected $H_0 : \delta_i = 0$, $i = 1, \ldots, n = 1262$.

- Consider testing instead $H_0 : \delta_i = \delta_{0i}$, $i = 1, \ldots, n = 1262$ with the δ_{0i} as *possible* specified values of δ_i.

- Since $R_i = r_{Ci} + Z_i \, \delta_i$, if the hypothesis H_0 were true, then $R_i - Z_i \, \delta_{0i}$ would equal r_{Ci}.

- But R_i and Z_i are observed and δ_{0i} is specified by the hypothesis, so if the hypothesis were true, we could calculate the r_{Ci}.

- Under the null hypothesis, the 2 × 2 table recording r_{Ci} by Z_i has the hypergeometric distribution, yielding a test.

19 Procedure

- If $H_0 : \delta_i = \delta_{0i}$, $i = 1, \ldots, n = 1262$ were true, then $r_{Ci} = R_i - Z_i \, \delta_{0i}$, so the the 2×2 table recording r_{Ci} by Z_i would be:

	$Recurrence$ $R_i = 1$	$No\,recurrence$ $R_i = 0$
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 $Z_i = 0$	$\sum (1 - Z_i) \, R_i$	$\sum (1 - Z_i) \, (1 - R_i)$

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

which would have the hypergeometric distribution.

20 Attributable effect

- The procedure shifts a count of $A_0 = \sum Z_i \, \delta_{0i}$, which, if the null hypothesis is true, equals

$$A = \sum Z_i \, \delta_i = \sum Z_i \, (r_{Ti} - r_{Ci}) \, ,$$

  that is the net number of additional women caused to have a recurrence by the use of lumpectomy alone rather than lumpectomy plus irradiation.

- Although I can calculate $A_0 = \sum Z_i \, \delta_{0i}$ from the hypothesis and the data, the true $A = \sum Z_i \, \delta_i$ is an unobservable random variable.

21 Example

- If a possible hypothesis $H_0 : \delta_i = \delta_{0i}$, $i = 1, \ldots, n = 1262$ yields $A_0 = \sum Z_i \, \delta_{0i} = 119$, compute:

	$Recurrence$ $R_i = 1$	$No\ recurrence$ $R_i = 0$	Total
No rads $Z_i = 1$	$220 - 119$	$414 + 119$	634
Rads $Z_i = 0$	78	550	628
Total	179	$1,083$	$1,262$

and the hypergeometric tail probability $\Pr(T \geq 220 - 119) = \Pr(T \geq 101) = 0.0438$, so H_0 is not quite plausible. If we do the same for a possible hypothesis $H_0 : \delta_i = \delta_{0i}$, $i = 1, \ldots, n = 1262$ yielded $A_0 = \sum Z_i \, \delta_{0i} = 120$, then the tail probability is 0.0514, and so barely plausible.

- That is, we are 95% confident that, net, at least 120 more of the 634 women treated with lumpectomy alone had recurrence of cancer caused by the failure to combine lumpectomy with irradiation.

22 Wilcoxon's Signed Rank Statistics

- Partly to illustrate, partly as a transition to observational studies, will illustrate randomization inference with Wilcoxon's signed rank statistic.

- Do with data from an observational study, a nonrandomized study of treatment effects, at first acting as if it were a randomized experiment, then considering the absence of randomization.

- Matched pairs: treated, control. Rank the absolute differences in responses within pairs. Sum ranks of positive differences.

23 Example: A Matched Observational Study

- From Morton, et al. (1982) Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, 115, 549-

- Study of one child of each of 33 workers in a battery factory in Oklahoma in 1978. Concern was that they might bring lead home, exposing their children.

- 33 control children were individually selected and matched to the exposed children. They were matched for neighborhood and age ($\pm 1\ year$). Neighborhood: (i) if an apartment, then another apartment from same complex, (ii) if facing a main road, then a nearby house facing the same road, etc.

- Outcome: child's blood lead level, μg of lead per dl blood.

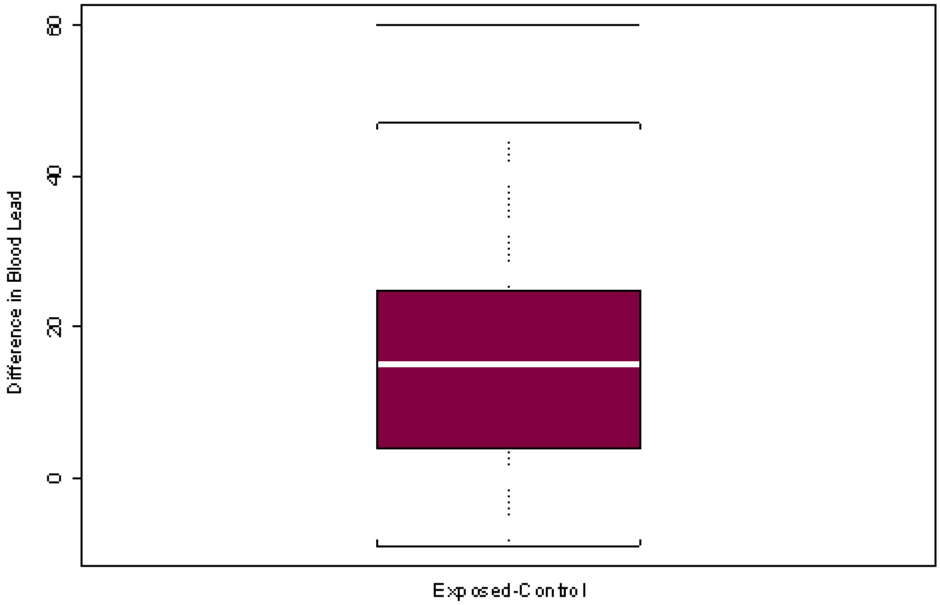


Figure 1: Matched pair differences in lead levels.

24 Notation for a Paired Experiment

Pair s, Subject i: $S = 33$ pairs, $s = 1, \ldots, S = 33$, with 2 subjects in each pair, $i = 1, 2$.

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 $Z_{s2} = 1 - Z_{s1}$. For all $2S$ subjects,

$$\mathbf{Z} = \left(Z_{11}, Z_{12}, \ldots, Z_{S1}, Z_{S2}\right)^T.$$

Random assignment of treatments within pairs: Ω is the set of the $K = 2^S$ possible values $\mathbf{z}$ of $\mathbf{Z}$, and randomization picks one of these at random,

$$\Pr\left(\mathbf{Z} = \mathbf{z}\right) = \frac{1}{K}\, for\ each\, \mathbf{z} \in \Omega.$$

25 Responses, Causal Effects

Potential responses, causal effects, 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.

Finite population, as before. The (r_{Tsi}, r_{Csi}), $s = 1, \ldots, S$, $i = 1, 2$, are again fixed features of the finite population of $2S$ subjects.

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, $R_{si} = r_{Csi} + Z_{si}\, \tau$.

Vectors. $2S-$dimensional vectors $\mathbf{r}_T$, $\mathbf{r}_C$, $\boldsymbol{\delta}$, $\mathbf{R}$; e.g., $\mathbf{R} = (R_{11}, \ldots, R_{S2})^T$.

26 Treated-Minus-Control Differences

Who is treated in pair s? If $Z_{s1} = 1$, then $(s, 1)$ is treated and $(s, 2)$ is control, but if $Z_{s2} = 1$ then $(s, 2)$ is treated and $(s, 1)$ is control.

Treated-minus-control differences with additive effects: If $r_{Tsi} - r_{Csi} = \tau$, then a little algebra shows the *treated-minus-control difference* in observed responses in pair s is:

$$D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2}) + \tau.$$

Signed Rank Test. Wilcoxon's *signed rank statistic* W ranks the $|D_s|$ from 1 to S, and sums the ranks of the positive D_s. (Ties ignored today.)

27 No Effect in an Experiment

Null hypothesis. $H_0 : \delta_{si} = 0$, for $s = 1, \ldots, S$, $i = 1, 2$ where $\delta_{si} = r_{Tsi} - r_{Csi}$.

Differences. If H_0 is true, then the treated-minus-control difference is:

$$D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2})$$

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.

Signed rank statistic. If H_0 is true, D_s is $\pm\,(r_{Cs1} - r_{Cs2})$ with probability $\frac{1}{2}$, 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.

Randomization. Uses just fact of randomization and null hypothesis, so forms the "reasoned basis for inference," in Fisher's phrase.

28 Randomization Test for an Additive Effect

Additive effect. $H_0 : \delta_{si} = \tau_0$, for $s = 1, \ldots, S$, $i = 1, 2$ where $\delta_{si} = r_{Tsi} - r_{Csi}$.

Matched pair differences. If H_0 were true, then

$$D_s = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2}) + \tau_0$$

so the adjusted differences

$$D_s - \tau_0 = (Z_{s1} - Z_{s2})\,(r_{Cs1} - r_{Cs2})$$

satisfy the hypothesis of no effect, and W computed from $D_s - \tau_0$ has the usual null distribution of the signed rank statistic.

Randomization. Again, the inference uses only the fact of randomization and the null hypothesis being tested.

29 Confidence Interval for Additive Effect

Additive effects. $\delta_{si} = \tau$, for all s, i where $\delta_{si} = r_{Tsi} - r_{Csi}$

Inverting tests. The 95% interval for τ is the set of all τ_0 not rejected in a 0.05 level test.

Confidence intervals. Test every τ_0 by computing W from the adjusted differences, $D_s - \tau_0$, retaining values τ_0 not rejected at the 0.05 level.

Hodges-Lehmann estimates. Find $\widehat{\tau}$ so that W computed from $D_s - \widehat{\tau}$ equals its null expectation.

30 Example: Lead Exposure

Morton, et al. 33 matched pairs of children, exposed-control, D_s is the difference in blood lead levels.

Not randomized. First, will perform analysis appropriate for a randomized experiment, then return to the example several times to think about consequences of nonrandom assignment to treatment.

Test of no effect. Signed rank statistic is $W = 527$, with randomization based $P - value = 10^{-5}$.

Confidence interval. 95% for an additive effect is [9.5, 20.5] $\mu g/dl$. The two-sided $P - value$ is ≥ 0.05 if W is computed from $D_s - \tau_0$ for $\tau_0 \in (9.5, 20.5)$ and is less than 0.05 for $\tau_0 \notin [9.5, 20.5]$.

HL estimate. $\widehat{\tau} = 15$ $\mu g/dl$ as $D_s - 15$ (effectively) equates W to its null expectation.

31 But the study was not randomized . . .

Not randomized. The analysis would have been justified by randomization in a randomized experiment.

Unknown 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.

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

32 Simple model continued . . .

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.

Absolutely simplest case: 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, $s = 1, \ldots, S$.

Matching algorithm: In this simplest case, the matching algorithm is permitted to use only $\mathbf{x}$ and $1 = Z_{s1} + Z_{s2}$.

33 Free of hidden bias

Definition. 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.

Intuition. A kid j who lives 30 miles from the battery factory is less likely to have a dad working in factory than a kid k who lives two miles from the factory, $\pi_j < \pi_k$, but two kids of the same age who next door are equally likely to have a dad in the factory.

But they didn't match on kid's gender. If gender were not recorded, it would violate 'free of hidden bias' if (roughly) boys were more likely (or less likely) than girls to have a dad working in the battery factor.

34 If free of hidden bias . . .

Problem: Unlike an experiment, π_j are unknown.

If free of hidden bias: Two people with the same $\mathbf{x}_j$ have the same π_j, which is typically unknown.

Eliminate unknowns by conditioning: If we match exactly for $\mathbf{x}$, so $\mathbf{x}_{s1} = \mathbf{x}_{s2}$, then

$$\Pr\left(Z_{s1} = 1 \mid Z_{s1} + Z_{s2}\right) = \frac{\pi_{s1}\left(1 - \pi_{s2}\right)}{\pi_{s1}\left(1 - \pi_{s2}\right) + \pi_{s2}\left(1 - \pi_{s1}\right)} = \frac{1}{2}$$

because $\pi_{s1} = \pi_{s2}$. A little more work shows that we get the randomization distribution by conditioning.

More generally, This argument is quite general, working for matched sets, strata, and more complex problems.

35 Interpretation

If free of hidden bias: Two people with the same $\mathbf{x}_j$ have the same π_j, which is typically unknown.

When do adjustments work? If a study is free of hidden bias, if the only bias is due to observed covariates $\mathbf{x}_j$, even if the bias is unknown, the bias can be removed in various ways, such as matching on $\mathbf{x}_j$, and conventional randomization inferences yield appropriate inferences about treatment effect.

Key, if problematic, assumption. Identifies the key assumption, but of course, doesn't make it true. Focuses attention, frames discussion. In contrast, in an experiment, randomization makes it true.

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.

36 Propensity Scores

Many observed 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.

If free of hidden bias: 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. .

Old argument again: Match exactly for $\mathbf{x}$, so $\mathbf{x}_{s1} = \mathbf{x}_{s2}$, then

$$\Pr\left(Z_{s1} = 1 \mid Z_{s1} + Z_{s2}\right) = \frac{\pi_{s1}\left(1 - \pi_{s2}\right)}{\pi_{s1}\left(1 - \pi_{s2}\right) + \pi_{s2}\left(1 - \pi_{s1}\right)} = \frac{1}{2}$$

because $\pi_{s1} = \pi_{s2}$ or $e\left(\mathbf{x}_{s1}\right) = e\left(\mathbf{x}_{s2}\right)$

Key point: Don't need to match on high dimension $\mathbf{x}$, just need to match on the scalar $e\left(\mathbf{x}\right)$.

37 Balancing with Propensity Scores

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

That is:

$$\Pr(\mathbf{x} \mid Z = 1, e) = \Pr(\mathbf{x} \mid Z = 0, e)$$

$$\text{or} \quad \mathbf{x} \perp\!\!\!\perp Z \mid e(\mathbf{x})$$

Proof: Suffices to show $\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\}$ equals $\Pr\{Z = 1 \mid e(\mathbf{x})\}$. But $\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\} = \Pr(Z = 1 \mid \mathbf{x})$ which is just $e(\mathbf{x})$. Also, $\Pr\{Z = 1 \mid e(\mathbf{x})\}$ equals $E[\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\} \mid e(\mathbf{x})] = E[\Pr\{Z = 1 \mid \mathbf{x}\} \mid e(\mathbf{x})]$ $= E[e(\mathbf{x}) \mid e(\mathbf{x})] = e(\mathbf{x})$.

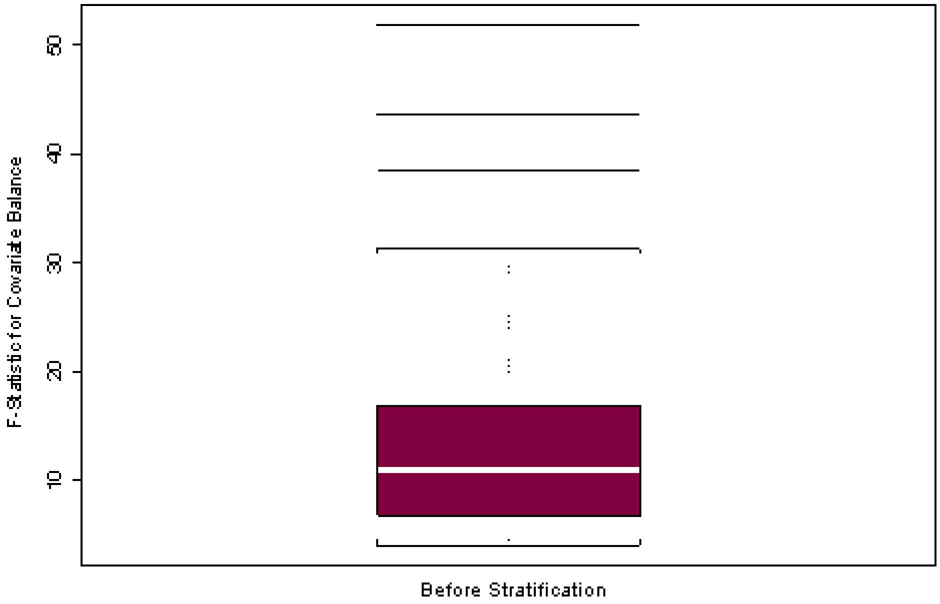
38 Propensity Scores: Example

Source: From Rosenbaum and Rubin (1984) *JASA*.

Data: Database describing 1,515 patients with coronary artery disease, treated either with CABG or drugs. Interest in effects of CABG vs drugs on survival, pain, etc.

Many covariates: CABG and drug patients differed significantly on 74 covariates. Drug patients were either too sick or too healthy for surgery.

Covariate	t-statistic	F-statistic
Ejection fraction	4.4	19.4
Poor left ventricle function	7.2	51.8
Left main artery occluded	4.7	22.1
Progressing Chest Pain	6.6	43.6



39 Boxplot Before Stratification

Covariate Imbalance. Covariate imbalance for 74 covariates before stratification on the propensity score. Display is $F = t^2$ for 74 covariates.

40 Procedure

Propensity score: Estimated using logit regression of treatment (CABG or drugs) on covariates, some quadratics, some interactions.

Five strata: Five groups formed at quintiles of the estimated propensity score.

Counts of Patients in Strata

Propensity Score Stratum	Medical	Surgical
1 $= lowest = most\, medical$	277	26
2	235	68
3	205	98
4	139	164
5 $= highest = most\, surgical$	69	234

41 Checking balance

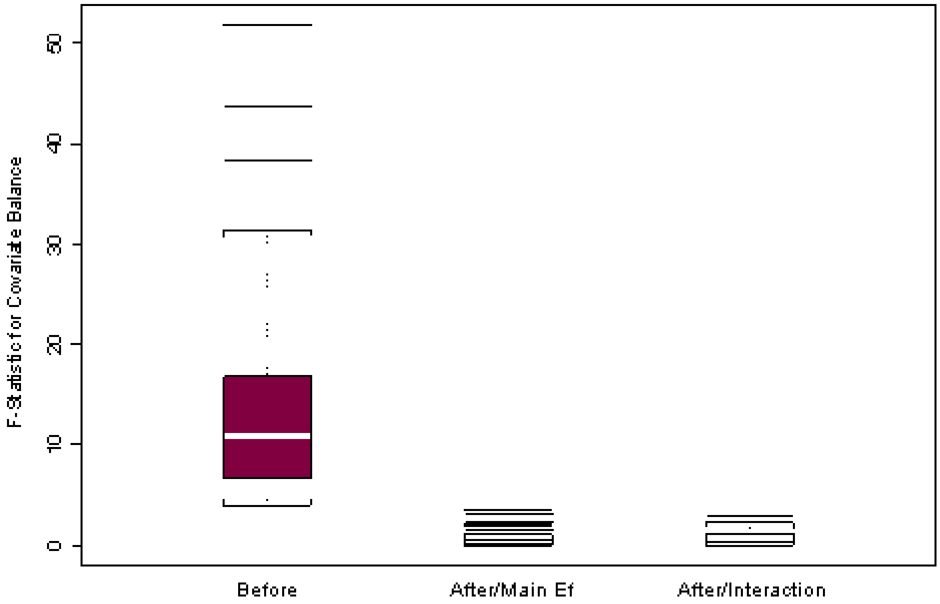
2-Way 5×2 Anova for Each Covariate

Propensity Score Stratum	Medical	Surgical
$1 = lowest = most\, medical$		
2		
3		
4		
$5 = highest = most\, surgical$		

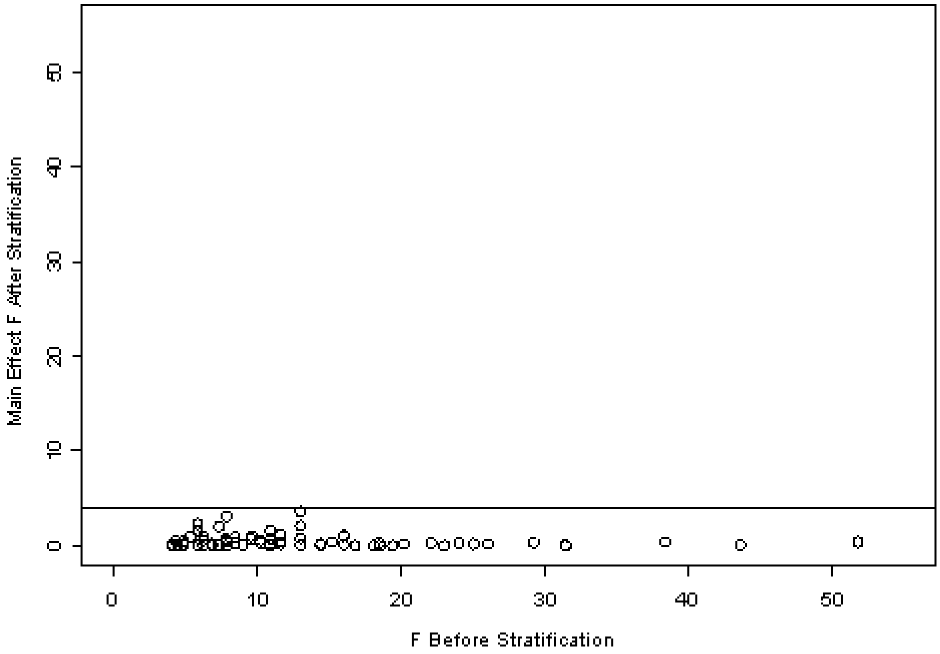
Balance check. Main effect and interaction $F-$statistics.

42 F-statistics Before and After Stratification

Covariate	Before	After Main Effect	After Interaction
Ejection fraction	19.4	0.0	0.3
Poor LV function	51.8	0.4	0.9
Left main occluded	22.1	0.3	0.2
Progressing Pain	43.6	0.1	1.4



43 Is there covariate balance within strata?



44 Covariate balance: Alternative view

45 Last words about propensity scores

Balancing. Stratifying or matching on a scalar propensity score tends to balance many observed covariates.

Effects of estimating the score. Examples, simulations, limited theory suggest estimated scores provide slightly *more* than true propensity scores.

Other methods. Various methods permit explicit acknowledgement of use of estimated scores.

Key limitation. Propensity scores balance only observed covariates, whereas randomization also balances unobserved covariates.

46 Addressing hidden bias

If free of hidden bias: Two people with the same observed $\mathbf{x}_j$ have the same π_j, which is typically unknown. Can remove the overt biases due to $\mathbf{x}_j$.

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

That is, the objection asserts that, or raises the possibility that, the observed association between treatment Z_j and response R_j is not an effect caused by the treatment, but rather due to hidden bias from their shared relationship with u_j.

Formally, treatment assignment Z_j and response $R_j = r_{Cj} + Z_j \left(r_{Tj} - r_{Cj}\right)$ may be associated because $r_{Tj} - r_{Cj} \neq 0$ (a treatment effect) or because $r_{Tj} - r_{Cj} = 0$ but π_j and r_{Cj} both vary with u_j (a hidden bias due to u_j).

47 Sensitivity analysis

Question answered by a sensitivity analysis: If the 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 does the counter-claim actually claim? A sensitivity analysis looks at the observed data and uses it to clarify what the critic's counter claim is actually claiming.

Sensitivity varies. Studies vary markedly in how sensitive they are to hidden bias.

48 First Sensitivity Analysis

Cornfield, et al. (1959): they write:

“If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.

Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers 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.”

49 The Cornfield, et al Inequality

The Cornfield, et al sensitivity analysis is an important conceptual advance:

> "Association does not imply causation
> — hidden bias can produce associations,"

is replaced by

> "To explain away the association actually seen, hidden biases would have to be of such and such a magnitude."

Provides a quantitative measure of uncertainty in light of data.

As a confidence interval measures sampling uncertainty without making it go away, a sensitivity analysis measure uncertainty due to hidden bias without making the uncertainty go away.

50 Alternative sensitivity analysis

Limitations. Cornfield's inequality concerns binary responses only and ignores sampling variability. Not explicit about observed covariates.

Alternative formulation. 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 $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_j\left(1-\pi_k\right)}{\pi_k\left(1-\pi_j\right)} \leq \Gamma.$$

Free of hidden bias is then $\Gamma = 1$.

When bias is present, when $\Gamma > 1$, the unknown π_j cannot be eliminated, as before, by matching on $\mathbf{x}_j$, so the randomization distribution is no longer justified.

51 Alternative sensitivity analysis, continued

Model. Two subjects, j and k, with $\mathbf{x}_j = \mathbf{x}_k$, may differ their odds of exposure to treatment differ by a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_j \left(1 - \pi_k\right)}{\pi_k \left(1 - \pi_j\right)} \leq \Gamma \tag{1}$$

so Γ provides measured departure from "no hidden bias."

Intuition: If $\Gamma = 1.001$, the π_j are unknown, but almost the same. 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 (1). Report these for several Γ. When do conclusions begin to change?

52 Signed Rank Statistic

Model. If $\mathbf{x}_j = \mathbf{x}_k$, then

$$\frac{1}{\Gamma} \leq \frac{\pi_j (1 - \pi_k)}{\pi_k \left(1 - \pi_j\right)} \leq \Gamma. \tag{2}$$

Structure: 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, $Z_{s1} + Z_{s2} = 1$.

Free of hidden bias: If $\Gamma = 1$, obtained the randomization distribution of Wilcoxon's signed rank statistic W, as $\Pr(Z_{s1} = 1 \mid Z_{s1} + Z_{s2}) = \frac{1}{2}$.

Fact: Then (2) implies:

$$\frac{1}{1 + \Gamma} \leq \Pr(Z_{s1} = 1 \mid Z_{s1} + Z_{s2}) \leq \frac{\Gamma}{1 + \Gamma}$$

which places sharp upper and lower bounds on the distribution of W and resulting inferences.

53 Lead Exposure: Significance Levels

Data: $S = 33$ pairs of children matched for age and neighborhood, one having a parent exposed to lead, the other a control. Measured lead levels in the children's blood. Used Wilcoxon's signed rank test, W.

Sensitivity analysis. One sided significance levels for testing no effect.

Γ	min	max
1	<0.0001	<0.0001
2	<0.0001	0.0018
3	<0.0001	0.0136
4	<0.0001	0.0388
4.25	<0.0001	0.0468
5	<0.0001	0.0740

54 One Sided Confidence Intervals

95% CI. For an additive effect, $r_{Tsi} = r_{Csi} + \tau$, the signed rank test may be inverted to yield a one-sided 95% confidence interval.

Range of values: For $\Gamma > 1$, the endpoint $\widehat{\tau}_{low}$ of the one-sided 95% interval $[\widehat{\tau}_{low}, \infty)$ for τ has a range of values. Table gives the smallest value in the range — the smallest plausible effect for the given quantity of hidden bias.

Sensitivity analysis.

Γ	min $\widehat{\tau}_{low}$
1	10.5
2	5.5
3	2.5
4	0.5
4.25	0.0
5	-1.0

55 Comparing Different Studies

Studies vary markedly in their sensitivity to hidden bias.

Treatment	$\Gamma = 1$	$(\Gamma$, max $P - value)$
Smoking/Lung Cancer Hammond 1964	$<$0.0001	(5, 0.03)
DES/vaginal cancer Herbst, et al. 1976	$<$ 0.0001	(7, 0.054)
Lead/Blood lead Morton, et al.1982	$<$ 0.0001	(4.25, 0.047)
Coffee/MI Jick, et al. 1973	0.0038	(1.3, 0.056)

Small biases could explain Coffee/MI association. Very large biases would be needed to explain DES/vaginal cancer association.

56 Sensitivity Analysis: Interpretation

Uses data, says something tangible. 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.”

Measures uncertainty. Measures uncertainty due to hidden bias, but does not dispel it. (As a confidence interval measures sampling uncertainty but does not dispel it.)

Fact of the matter. Your opinion about how much hidden bias is present is your opinion. But the degree of sensitivity to hidden bias is a fact of the matter, something visible in observed data.

57 Summary

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

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

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

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