Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies

Donald B. Rubin
Department of Statistics
Harvard University

The following material is a summary of the course materials used in Quantitative Reasoning (QR) 33, taught by Donald B. Rubin at Harvard University, Spring 2002, Spring 2003, and Spring 2004. Prepared with assistance from Samantha Cook and Elizabeth Stuart.

Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies

I . Framework

1. Primitives: Units, treatments, potential outcomes
2. Learning about causal effects: Replication, stability, the assignment mechanism
3. The transition to statistical inference: Introduction to randomized experiments and the Rubin Causal Model
4. Examples of assignment mechanisms

II. Causal inference based on the assignment mechanism

5. "Fisherian" significance levels in CR experiment
6. "Neymanian" repeated sampling evaluations in CR experiment
7. Extension to studies with variable but known propensities
8. Extension to studies with unknown propensities
9. Examples of methods for estimating propensities

III. Causal inference based on predictive distributions of potential outcomes

10. Predictive inference – intuition under ignorability
11. Matching to impute missing potential outcomes
12. Fitting predictive models within each treatment group
13. Formal predictive inference – Bayesian [Rubin,1978]
14. Principal stratification: Dealing with post-treatment variables (noncompliance, surrogate outcomes, censoring due to death, etc.)

Part I: Framework

Example I-1: Potential Outcomes and Causal Effect with One Unit

In a hypothetical example, the unit is you at a particular point in time with a headache; Y is your assessment of your headache pain two hours after taking an aspirin (action Asp) or not taking aspirin (action Not).

Unit	Potential Outcomes		Causal Effect
	Y(Asp)	Y(Not)	Y(Asp) - Y(Not)
you	25	75	-50

Example I-2: Gain Scores

Potential Outcomes and Causal Effect with One Unit: In hypothetical example, the unit is you at a particular point in time with a headache; Y is your assessment of your headache pain two hours after taking an aspirin (action Asp) or not taking aspirin (action Not), and the outcome is headache reduction, Y - X, where X is your assessment of the pain of your initial headache.

Unit	Initial Headache	Potential Outcomes		Causal Effect
	X	Y(Asp) - X	Y(Not) - X	Y(Asp) - X - [Y(Not) - X]
you	80	-55	-5	-50

Example I-3: Percent Change

Potential Outcomes and Causal Effect with One Unit: In hypothetical example, the unit is you at a particular point in time with a headache; Y is your assessment of your headache pain two hours after taking an aspirin (action Asp) or not taking aspirin (action Not), and the outcome is fractional reduction in headache $Y^* = 1 - \frac{Y+1}{X+1}$, where X = intensity of initial headache.

Unit	Initial Headache	Potential	Outcomes, Y	Causal Effect
	X	Y*(Asp)	Y*(Not)	Y*(Asp) - Y*(Not)
you	80	1 - $\frac{26}{81}$=68%	1 - $\frac{76}{81}$ = 6%	68% - 6% = 62%

Example I-4: Legal Examples of Potential Outcomes and Counterfactual World

In the September 22, 1999 news conference held to announce the United States filing of its lawsuit against the tobacco industry, Assistant Attorney General Ogden (1999) stated:

> The number that's in the complaint is not a number that reflects a particular demand for payment. What we've alleged is that each year the federal government expends in excess of $20 billion on *tobacco* related medical costs. What we would actually recover would be our portion of that annual toll that is the result of the illegal conduct that we allege occurred, and it simply will be a matter of proof for the court, which will be developed through the course of discovery, what that amount will be. So, we have not put out a specific figure and we'll simply have to develop that as the case goes forward.

Also, the Federal Judicial Center's "Reference Manual on Scientific Evidence" (1994, Chapter 3, p. 481) states:

> The first step in a damages study is the translation of the legal theory of the harmful event into an analysis of the economic impact of that event. In most cases, the analysis considers the difference between the plaintiff's economic position if the harmful event had not occurred and the plaintiff's actual economic position. The damages study restates the plaintiff's position "but for" the harmful event; this part is often called the *but-for analysis*. Damages are the difference between the but-for value and the actual value.

Example I-5: Potential Outcomes with Two Units Allowing Interference Between Units

Potential Outcomes and Values in Example

You take: I take:	Asp Asp	Not Not	Asp Not	Not Asp
Unit				
1 = you	Y_1([Asp, Asp]) = 0	Y_1([Not, Not]) = 100	Y_1([Asp, Not]) = 50	Y_1([Not, Asp]) = 75
2 = me	Y_2([Asp, Asp]) = 0	Y_2([Not, Not]) = 100	Y_2([Asp, Not]) = 100	Y_2([Not, Asp]) = 0

Note: The causal effect of Asp versus Not for me is well-defined as 100. The reason is that Y_2([Asp, Asp]) - Y_2([Asp, Not]), which is the effect of Asp versus Not for me when you get Asp, is 0 - 100 = -100; and Y_2([Not, Asp]) - Y_2([Not, Not]), which is the causal effect of Asp versus Not for me when you get Not is also 0 - 100 = -100. In contrast, for you the causal effect of Asp versus not depends on what I receive. If I receive Asp, the causal effect for you is Y_1([Asp, Asp]) - Y_1([Not, Asp]) = 0 - 75 = -75, whereas if I receive Not the causal effect is Y_1([Asp, Not]) - Y_1([Not, Not]) = 50 - 100 = -50, a smaller effect.

Example I-6: Potential Outcomes in Aspirin Example for N Units Under the Stability Assumption

Unit	X	Y(Asp)	Y(Not)	Causal effect
1	X_1	Y_1(Asp)	Y_1(Not)	Y_1(Asp) - Y_1(Not)
2	X_2	Y_2(Asp)	Y_2(Not)	Y_2(Asp) - Y_2(Not)
⋮	⋮	⋮	⋮	⋮
i	X_i	Y_i(Asp)	Y_i(Not)	Y_i(Asp) - Y_i(Not)
⋮	⋮	⋮	⋮	⋮
N	X_N	Y_N(Asp)	Y_N(Not)	Y_N(Asp) - Y_N(Not)

Average causal effect of “Asp” vs. “Not” =
Ave[Y_i(Asp) - Y_i(Not)]
= $\frac{1}{N}\sum_{i=1}^{N}[Y_i(\text{Asp}) - Y_i(\text{Not})]$

Median causal effect of “Asp” vs. “Not” =
Median $\{Y_i$(Asp) - Y_i(Not)$\}$

Difference of median potential outcomes =
Median $\{Y_i(\text{Asp})\}$ - Median $\{Y_i(\text{Not})\}$

Example I-7: Perfect Doctor

The data given below shows all potential outcomes under two different treatments: $Y(0)$ represents years lived after standard surgery and $Y(1)$ represents years lived after new surgery.

Potential Outcomes

	Y(0)	Y(1)
	13	14
	6	0
	4	1
	5	2
	6	3
	6	1
	8	10
	8	9
True averages	7	5

The true average causal effect $\overline{Y(1)} - \overline{Y(0)} = -2$.

**Note: $\overline{Y}$ denotes Average of Y.

The perfect doctor chooses the best treatment for each patient, i.e., the treatment under which the patient will live longer. If there is no difference, he chooses by flipping a coin.

Observed Outcomes under Perfect Doctor's Assignment

W	Y(0)	Y(1)
1	?	14
0	6	?
0	4	?
0	5	?
0	6	?
0	6	?
1	?	10
1	?	9
Observed Averages	5.4	11

**Observed $\overline{y_1} - \overline{y_0} = 5.6 \neq -2$.

The average observed causal effect equals -2 on average over all possible assignments, but for some assignments can be very far from -2.

w	$\overline{y_1} - \overline{y_0}$	median(y_1) - median(y_0)
11100000	-1.6	-5
11010000	-1.1	-4
11001000	-0.5	-3
11000100	-1.2	-5
11000010	2.2	4
11000001	1.9	3
10110000	-1.1	-4
10101000	-0.6	-3
10100100	-1.3	-5
10100010	2.1	4
10100001	1.8	3
10011000	-0.1	-3
10010100	-0.7	-4
10010010	2.7	4
10010001	2.3	3
10001100	-0.2	-3
10001010	3.2	4
10001001	2.9	3
10000110	2.5	4
10000101	2.2	3
10000011	5.6	4
01110000	-7.2	-7
01101000	-6.7	-7
01100100	-7.3	-7
01100010	-3.9	-5
01100001	-4.3	-5
01011000	-6.1	-6
01010100	-6.8	-7
01010010	-3.4	-4
01010001	-3.7	-4
01001100	-6.3	-7
01001010	-2.9	-3
01001001	-3.2	-3
01000110	-3.5	-5
01000101	-3.9	-5
01000011	-0.5	3
00111000	-6.2	-6
00110100	-6.9	-7
00110010	-3.5	-4
00110001	-3.8	-4
00101100	-6.3	-7
00101010	-2.9	-3
00101001	-3.3	-3
00100110	-3.6	-5
00100101	-3.9	-5
00100011	-0.5	3
00011100	-5.8	-6
00011010	-2.4	-3
00011001	-2.7	-3
00010110	-3.1	-4
00010101	-3.4	-4
00010011	0.0	3
00001110	-2.5	-3
00001101	-2.9	-3
00001011	0.5	3
00000111	-0.1	3
Average	-2	-2.3

Observed Outcomes under Random Draw 1

W	Y(0)	Y(1)
1	?	14
1	?	0
1	?	1
0	5	?
0	6	?
0	6	?
0	8	?
0	8	?
Observed Averages	6.6	5

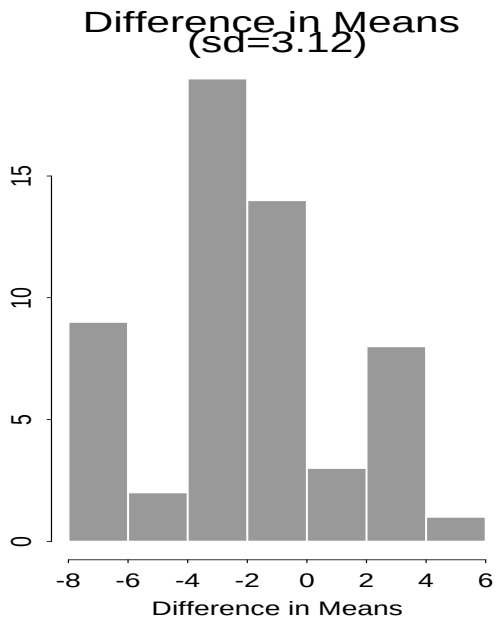
**Observed $\overline{y_1} - \overline{y_0} = -1.6$.

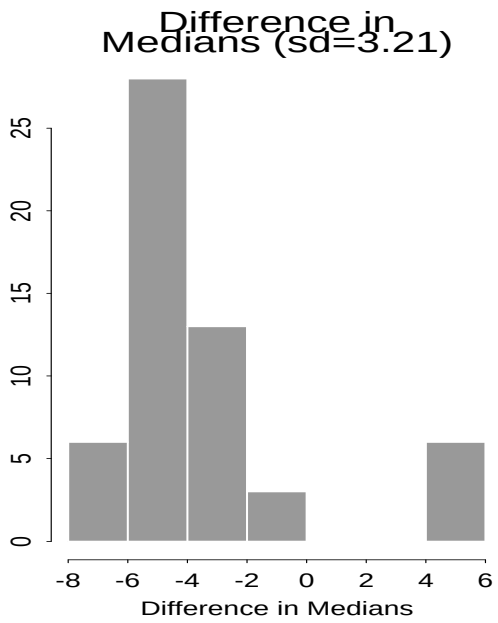
Observed Outcomes under Random Draw 56

W	Y(0)	Y(1)
0	13	?
0	6	?
0	4	?
0	5	?
0	6	?
1	?	1
1	?	10
1	?	9
Observed Averages	6.8	6.7

**Observed $\overline{y_1} - \overline{y_0} = -0.1$.

Summary of All Random Draws



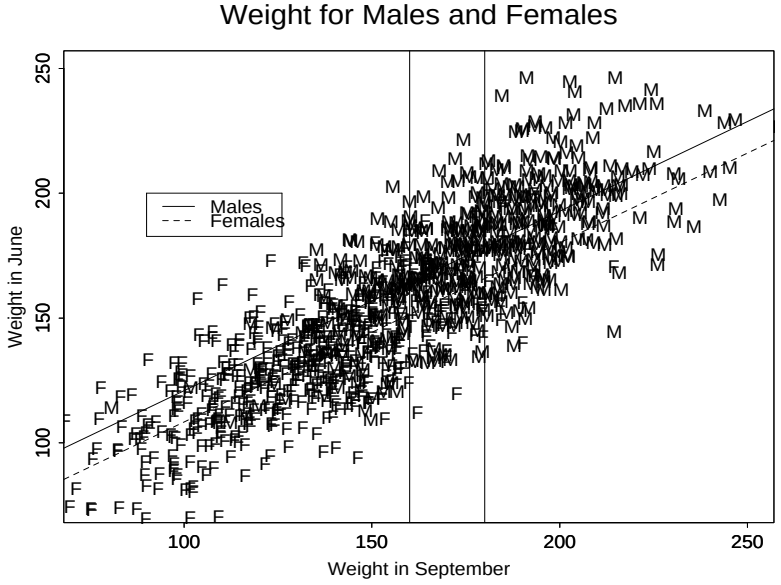


Lord's Paradox

From Holland and Rubin, "On Lord's Paradox," 1983.

"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded."

September weight range (in pounds)	% of Men	% of Women	Male average June weight	Female average June weight	Male Weight Gain - Female Weight Gain
< 100	0.2	12.4	114	102	12
100-109	0.5	10.0	120	108	12
110-119	0.7	10.6	122	110	12
120-129	1.7	14.5	134	122	12
130-139	2.5	13.9	146	134	12
140-149	8.0	15.0	152	140	12
150-159	10.0	10.4	158	146	12
160-169	15.4	5.4	166	154	12
170-179	15.0	4.8	176	164	12
180-189	14.8	1.8	184	172	12
190-199	14.0	1.0	191	179	12
> 200	17.2	0.2	204	192	12



The average weight for Males was 180 in both September and June.
The average weight for Females was 130 in both September and June.

The average weight gain for Males was zero.
The average weight gain for Females was zero.

Statistician 1: Look at gain scores.

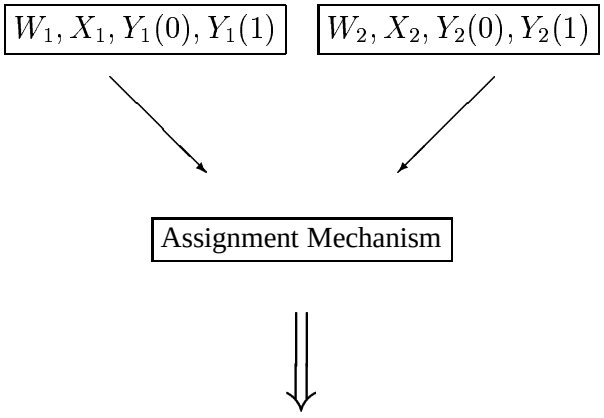
> Thus no effect of diet on weight, and no evidence of differential effect of the two sexes, as no group shows any systematic change.

Statistician 2: Compare June weight (see Figure 1) for males and females with the same weight in September.

> On average, for a given September weight, men weigh more in June than women. Thus, the new diet leads to more weight gain for men.

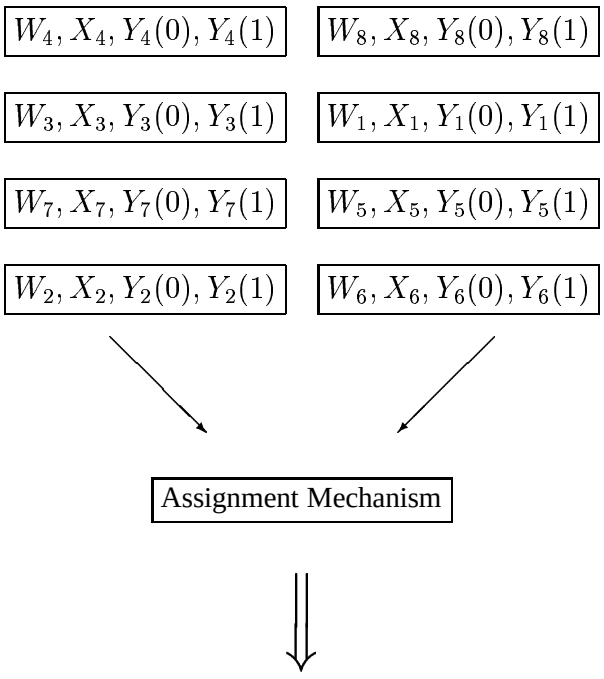
The Assignment Mechanism

Example I-8: Completely Randomized Design with $N = 2$ units, 1 assigned treatment



$\mathbf{W} = (W_1, W_2)$	Prob of $\mathbf{W}$
(0, 0)	0
(0, 1)	0.5
(1, 0)	0.5
(1, 1)	0

Example I-9: Completely Randomized Design with $N = 8$ units, 3 assigned treatment

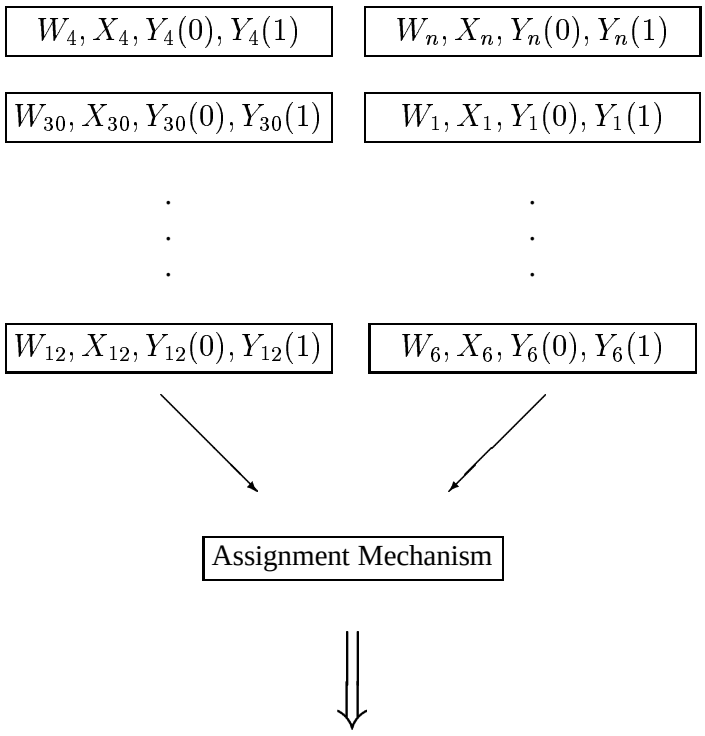


W	Prob of **W**
If $\sum_{i=1}^{8} W_i = 3$ [a]	$\frac{1}{56}$ [b]
If $\sum_{i=1}^{8} W_i \neq 3$	0

[a] i.e., if exactly 3 of the W_i's equal 1

[b] 56 is the number of ways to choose 3 items from 8

Example I-10: Completely Randomized Design with N units, n assigned treatment



W	Prob of $\mathbf{W}^a$
If $\sum_{i=1}^{N} W_i = n$	$\binom{N}{n}^{-1}$
If $\sum_{i=1}^{N} W_i \neq n$	0

[a] $\binom{N}{n}$ is the number of ways to choose n items from N

Examples of the Assignment Mechanism

Example I-11: "Bernoulli" (coin-tossing) assignment, 4 units

Assignment is random, and each individual has the same probability of receiving treatment 1. In this example, this probability is .5, i.e., it is equally likely for each person to receive treatment 0 or treatment 1. Remember that the overall assignment is the vector of all of the individuals' assignments: $\mathbf{W} = (W_1, W_2, W_3, W_4)$.

Since each individual's treatment status is assigned independently of the other individuals, the overall assignment probability is the product of the individual probabilities.

All Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$
$(0,0,0,0)$	$(.5)^4$
$(0,0,0,1)$	$(.5)^4$
$(0,0,1,0)$	$(.5)^4$
$(0,0,1,1)$	$(.5)^4$
$(0,1,0,0)$	$(.5)^4$
$(0,1,0,1)$	$(.5)^4$
$(0,1,1,0)$	$(.5)^4$
$(0,1,1,1)$	$(.5)^4$
$(1,0,0,0)$	$(.5)^4$
$(1,0,0,1)$	$(.5)^4$
$(1,0,1,0)$	$(.5)^4$
$(1,0,1,1)$	$(.5)^4$
$(1,1,0,0)$	$(.5)^4$
$(1,1,0,1)$	$(.5)^4$
$(1,1,1,0)$	$(.5)^4$
$(1,1,1,1)$	$(.5)^4$

Example I-12: "Bernoulli" (coin-tossing) assignment, 4 units

Same as Example I-11, however now the probability of receiving treatment 1 for each individual is $.4$. Again, treatment is assigned independently for each individual.

All Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$
$(0,0,0,0)$	$(.6)^4$
$(0,0,0,1)$	$(.4)^1(.6)^3$
$(0,0,1,0)$	$(.4)^1(.6)^3$
$(0,0,1,1)$	$(.4)^2(.6)^2$
$(0,1,0,0)$	$(.4)^1(.6)^3$
$(0,1,0,1)$	$(.4)^2(.6)^2$
$(0,1,1,0)$	$(.4)^2(.6)^2$
$(0,1,1,1)$	$(.4)^3(.6)^1$
$(1,0,0,0)$	$(.4)^1(.6)^3$
$(1,0,0,1)$	$(.4)^2(.6)^2$
$(1,0,1,0)$	$(.4)^2(.6)^2$
$(1,0,1,1)$	$(.4)^3(.6)^1$
$(1,1,0,0)$	$(.4)^2(.6)^2$
$(1,1,0,1)$	$(.4)^3(.6)^1$
$(1,1,1,0)$	$(.4)^3(.6)^1$
$(1,1,1,1)$	$(.4)^4$

Example I-13: Randomized within blocks

Blocks contain individuals that are grouped together on the basis of some covariate. In this case we consider two blocks: males and females.

In this example, there are 4 males, $i = 1, 2, 3, 4$, and 4 females, $i = 5, 6, 7, 8$: $\mathbf{W} = (\mathbf{W_M}, \mathbf{W_F})$. 2 males and 2 females are chosen randomly to receive treatment 1. The other 2 males and 2 females receive treatment 0.

Males: $X_i = M$

$\mathbf{W_M}$	Prob of $\mathbf{W_M}$
If $\sum_{i=1}^4 W_i = 2$	$\binom{4}{2}^{-1}$
If $\sum_{i=1}^4 W_i \neq 2$	0

Females: $X_i = F$

$\mathbf{W_F}$	Prob of $\mathbf{W_F}$
If $\sum_{i=5}^8 W_i = 2$	$\binom{4}{2}^{-1}$
If $\sum_{i=5}^8 W_i \neq 2$	0

Overall

$\mathbf{W} = (\mathbf{W_M}, \mathbf{W_F})$	Prob of $\mathbf{W}$
If $\sum_{i=1}^4 W_i = 2$ and $\sum_{i=5}^8 W_i = 2$	$\binom{4}{2}^{-1} * \binom{4}{2}^{-1}$
Anything else	0

Example I-14: "Bernoulli" (coin-tossing) assignment within blocks

There are 4 men and 4 women. In the notation below, the covariate $\mathbf{X} = (M, M, M, M, F, F, F, F)$.
For males, the probability of receiving treatment 1 is .2.
For females, the probability of receiving treatment 1 is .7.

Unit	Sex	Prob of $W_i = 1$	Prob of $W_i = 0$
1	M	.2	.8
2	M	.2	.8
3	M	.2	.8
4	M	.2	.8
5	F	.7	.3
6	F	.7	.3
7	F	.7	.3
8	F	.7	.3

Again, since each individual's treatment status is assigned independently of the other individuals, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's.

Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$		
$(0,0,0,0,0,0,0,0)$	$(.8)^4(.3)^4$	=	.003
$(1,0,0,0,0,0,0,0)$	$(.2)^1(.8)^3(.3)^4$	=	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.2)^1(.8)^3(.3)^4$	=	$\cdots$
...	...		
$(0,0,0,0,0,0,1,0)$	$(.8)^4(.3)^3(.7)^1$	=	$\cdots$
$(0,0,0,0,0,0,0,1)$	$(.8)^4(.3)^3(.7)^1$	=	.008
$(1,0,1,0,0,0,0,0)$	$(.2)^2(.8)^2(.3)^4$	=	$\cdots$
...	...		
$(1,1,1,0,0,0,0,0)$	$(.2)^3(.8)^1(.3)^4$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,1,1)$	$(.8)^4(.7)^3(.3)^1$	=	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.8)^2(.2)^2(.7)^2(.3)^2$	=	.001
...	...		
$(0,0,1,1,1,1,1,1)$	$(.8)^2(.2)^2(.7)^4$	=	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.2)^4(.7)^4$	=	.0004

Example I-15: Probability of treatment depends on age, Prob of $W_i = 1$ is $\frac{\text{age}_i}{\text{age}_i+10}$

In the notation below, $\mathbf{Age} = (15, 22, 18, 54, 34, 77, 38, 91)$.

Unit	Age	Prob of $W_i = 1$	Prob of $W_i = 0$
1	15	.60	.40
2	22	.69	.31
3	18	.64	.36
4	54	.84	.16
5	34	.77	.23
6	77	.89	.11
7	38	.79	.21
8	91	.90	.10

Again, since each individual's treatment status is assigned independently of the other individuals, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's.

Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$		
$(0,0,0,0,0,0,0,0)$	$(.40)(.31)(.36)(.16)(.23)(.11)(.21)(.10)$	=	.000004
$(1,0,0,0,0,0,0,0)$	$(.60)(.31)(.36)(.16)(.23)(.11)(.21)(.10)$	=	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.40)(.69)(.36)(.16)(.23)(.11)(.21)(.10)$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,0,1)$	$(.40)(.31)(.36)(.16)(.23)(.89)(.21)(.90)$	=	$\cdots$
$(0,0,0,0,0,0,1,1)$	$(.40)(.31)(.36)(.16)(.23)(.11)(.79)(.90)$	=	.0001
$(1,1,1,0,0,0,0,0)$	$(.60)(.69)(.64)(.16)(.23)(.11)(.21)(.10)$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,1,1)$	$(.40)(.31)(.36)(.16)(.23)(.89)(.79)(.90)$	=	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.60)(.69)(.36)(.16)(.77)(.89)(.21)(.10)$	=	$\cdots$
...	...		
$(0,0,1,1,1,1,1,1)$	$(.40)(.31)(.64)(.84)(.77)(.89)(.79)(.90)$	=	.03
...	...		
$(1,1,1,1,1,1,1,0)$	$(.60)(.69)(.64)(.84)(.77)(.89)(.79)(.10)$	=	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.60)(.69)(.64)(.84)(.77)(.89)(.21)(.90)$	=	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.60)(.69)(.64)(.84)(.77)(.89)(.79)(.90)$	=	.11

Example I-16: "Bernoulli" (coin-tossing) assignment, 4 units

(Mini) School choice example. Program to give vouchers to students to attend private schools. Probability of receiving a voucher depends on quality of current school ($Q = 0$ means good, $Q = 1$ means bad). $W_i = 1$ means that they get a voucher.

Student	School Quality	Prob of $W_i = 1$	Prob of $W_i = 0$
1	1	.7	.3
2	0	.4	.6
3	0	.4	.6
4	1	.7	.3

We consider three assignment mechanisms:

a. Independent assignment (Bernoulli)

b. Only have money for 2 vouchers so 2 given vouchers, 2 not given vouchers

c. Only have money for 3 vouchers so 3 given vouchers, 1 not given voucher

All Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}^a$	Prob of $\mathbf{W}^b$	Prob of $\mathbf{W}^c$
$(0,0,0,0)$	$(.3)(.6)(.6)(.3) = .03$	0	0
$(0,0,0,1)$	$(.3)(.6)(.6)(.7) = .08$	0	0
$(0,0,1,0)$	$(.3)(.6)(.4)(.3) = .02$	0	0
$(0,0,1,1)$	$(.3)(.6)(.4)(.7) = .05$	$.05/.39 = .13$	0
$(0,1,0,0)$	$(.3)(.4)(.6)(.3) = .02$	0	0
$(0,1,0,1)$	$(.3)(.4)(.6)(.7) = .05$	$.05/.39 = .13$	0
$(0,1,1,0)$	$(.3)(.4)(.4)(.3) = .01$	$.01/.39 = .02$	0
$(0,1,1,1)$	$(.3)(.4)(.4)(.7) = .03$	0	$.03/.30 = .10$
$(1,0,0,0)$	$(.7)(.6)(.6)(.3) = .08$	0	0
$(1,0,0,1)$	$(.7)(.6)(.6)(.7) = .18$	$.18/.39 = .46$	0
$(1,0,1,0)$	$(.7)(.6)(.4)(.3) = .05$	$.05/.39 = .13$	0
$(1,0,1,1)$	$(.7)(.6)(.4)(.7) = .12$	0	$.12/.30 = .40$
$(1,1,0,0)$	$(.7)(.4)(.6)(.3) = .05$	$.05/.39 = .13$	0
$(1,1,0,1)$	$(.7)(.4)(.6)(.7) = .12$	0	$.12/.30 = .40$
$(1,1,1,0)$	$(.7)(.4)(.4)(.3) = .03$	0	$.03/.30 = .10$
$(1,1,1,1)$	$(.7)(.4)(.4)(.7) = .08$	0	0

[a] Independent assignment
[b] Constrained so that 2 given vouchers
[c] Constrained so that 3 given vouchers

Example I-17: "Bernoulli" (coin-tossing) assignment, 4 units (Extension of Example I-12)

We consider three assignment mechanisms:

a. Independent assignment (Bernoulli), prob of $W_i = 1$ is .4 for each unit

All Possible Assignments

W	Prob of **W**
$(0,0,0,0)$	$(.6)^4$
$(0,0,0,1)$	$(.4)^1(.6)^3$
$(0,0,1,0)$	$(.4)^1(.6)^3$
$(0,0,1,1)$	$(.4)^2(.6)^2$
$(0,1,0,0)$	$(.4)^1(.6)^3$
$(0,1,0,1)$	$(.4)^2(.6)^2$
$(0,1,1,0)$	$(.4)^2(.6)^2$
$(0,1,1,1)$	$(.4)^3(.6)^1$
$(1,0,0,0)$	$(.4)^1(.6)^3$
$(1,0,0,1)$	$(.4)^2(.6)^2$
$(1,0,1,0)$	$(.4)^2(.6)^2$
$(1,0,1,1)$	$(.4)^3(.6)^1$
$(1,1,0,0)$	$(.4)^2(.6)^2$
$(1,1,0,1)$	$(.4)^3(.6)^1$
$(1,1,1,0)$	$(.4)^3(.6)^1$
$(1,1,1,1)$	$(.4)^4$

b. Constrained so that 2 assigned to treatment 1 and 2 assigned to treatment 0, prob of $W_i = 1$ is .4

All Possible Assignments

W	Prob of **W**
$(0,0,0,0)$	0
$(0,0,0,1)$	0
$(0,0,1,0)$	0
$(0,0,1,1)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(0,1,0,0)$	0
$(0,1,0,1)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(0,1,1,0)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(0,1,1,1)$	0
$(1,0,0,0)$	0
$(1,0,0,1)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(1,0,1,0)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(1,0,1,1)$	0
$(1,1,0,0)$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$
$(1,1,0,1)$	0
$(1,1,1,0)$	0
$(1,1,1,1)$	0

c. Constrained so that 3 assigned to treatment 1 and 1 assigned to treatment 0, prob of $W_i = 1$ is .4

All Possible Assignments

W	Prob of **W**
$(0,0,0,0)$	0
$(0,0,0,1)$	0
$(0,0,1,0)$	0
$(0,0,1,1)$	0
$(0,1,0,0)$	0
$(0,1,0,1)$	0
$(0,1,1,0)$	0
$(0,1,1,1)$	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,0,0,0)$	0
$(1,0,0,1)$	0
$(1,0,1,0)$	0
$(1,0,1,1)$	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,0,0)$	0
$(1,1,0,1)$	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,1,0)$	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,1,1)$	0

Example I-18: Randomized within matched pairs

In a trial for a new cholesterol reducing drug, subjects were paired on the basis of covariates (pre-treatment cholesterol level, age, income level, race). Within each pair, 1 subject was randomly assigned treatment and the other was assigned control. Thus, within each pair, each subject had a $.5$ chance of receiving the new treatment (1), as well as a $.5$ chance of receiving placebo (treatment 0). We consider 3 pairs. In the notation below, units 1 and 2 form a pair, 3 and 4 form a pair, and 5 and 6 form a pair.

W	Prob of $\mathbf{W}$
(1,0),(1,0),(1,0)	$(.5)(.5)(.5) = .125$
(1,0),(1,0),(0,1)	$(.5)(.5)(.5) = .125$
(1,0),(0,1),(1,0)	$(.5)(.5)(.5) = .125$
(1,0),(0,1),(0,1)	$(.5)(.5)(.5) = .125$
(0,1),(1,0),(1,0)	$(.5)(.5)(.5) = .125$
(0,1),(1,0),(0,1)	$(.5)(.5)(.5) = .125$
(0,1),(0,1),(1,0)	$(.5)(.5)(.5) = .125$
(0,1),(0,1),(0,1)	$(.5)(.5)(.5) = .125$
Anything else	0

Example I-19: Bernoulli assignment, but probability depends on unobserved covariate ($U_i = Y_i(0)$)

A teacher randomly assigns children in her class to a new reading program (treatment 1). Since she wants motivated children in this new program, in her mind she judges each student's motivation on a scale from 1 to 10 and assigns children to the program such that students with higher motivation are more likely to be put into the new program. To ensure confidentiality, she does not write down or disclose to anyone the students' motivation levels (and she promptly forgets them). Like the perfect doctor, the teacher has great insight and the motivation score is essentially equal to what the child would get without the new program.

For each student, the assignment to treatment 1 is done using the following rule: the probability that $W_i = 1$ is $.1 * U_i$, where U_i is the student's (unobserved) motivation level. ($\mathbf{U}$ is the vector of the motivation levels of all of the students). It is not surprising that U_i is highly correlated with both potential outcomes, thereby inducing a dependence of W_i on the potential outcomes. For students with motivation level 10, the probability of assignment to treatment 1 is .95, and for students with motivation level 0, the probability of assignment to treatment 1 is .05:

$$\text{Probability of receiving Treatment} = \begin{cases} .05 & \text{if} \quad U_i = 0 \\ .1 * U_i & \text{if} \quad 0 < U_i < 10 \\ .95 & \text{if} \quad U_i = 10 \end{cases}$$

Student	U	Prob of $W_i = 1$	Prob of $W_1 = 0$
1	4	.4	.6
2	8	.8	.2
3	2	.2	.8
4	7	.7	.3
5	8	.8	.2
6	10	.95	.05
7	5	.5	.5
8	0	.05	.95

If treatment is assigned independently to each unit, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's.

Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$		
$(0,0,0,0,0,0,0,0)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.95)$	=	.0001
$(1,0,0,0,0,0,0,0)$	$(.4)(.2)(.8)(.3)(.2)(.05)(.5)(.95)$	=	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.6)(.8)(.8)(.3)(.2)(.05)(.5)(.95)$	=	$\cdots$
...	...		
$(0,0,0,0,0,0,1,1)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.05)$	=	$\cdots$
$(1,1,1,0,0,0,0,0)$	$(.4)(.8)(.2)(.3)(.2)(.05)(.5)(.95)$	=	.00009
...	...		
$(1,1,1,1,1,1,1,0)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.95)$	=	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.05)$	=	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.05)$	=	.0009

Example I-20: Ignorable but Confounded Treatment Assignment

The following is based on "Investigating Therapies of Potentially Great Benefit: ECMO" by Jim Ware (1989).

- Persistent pulmonary hypertension of the newborn (PPHN) is an acute lung disease in newborns that results in the newborn being unable to oxygenate their blood. PPHN is highly fatal in the first days of life, however infants who survive have a good long-term prognosis.
- Conventional medical therapy (CMT) mortality rate: approximately 80%.
- Extracorporeal membrane oxygenation (ECMO) treatment mortality rate: less than 20%.
  - ECMO is an extreme therapy that routes the blood out of the jugular vein, oxygenates the blood outside the body, heats it, and then replaces the blood in the body through the carotid artery. It is essentially a simplified heart-lung machine.
- Three randomized studies of ECMO have been done in the treatment of PPHN.

1. Randomized "play-the-winner" (confounded but ignorable)
   - Probability of each newborn receiving ECMO depends on the previous outcomes.
   - 12 infants enrolled sequentially (one after another in time).
   - Assignment: Think of an urn that contains 2 balls: one representing ECMO, one representing CMT. The first infant was randomly given ECMO and future assignment was as follows: "When a treatment was selected and the infant survived, a ball representing that treatment was added to the urn. When the infant died, a ball representing the other treatment was added." To determine the assignment of the next infant, a ball was drawn out of the urn.
   - $Y = 1$ if the patient died and $Y = 0$ otherwise.

Newborn (time order)	Prob of $W_i = \text{ECMO}$ [a]	Prob of $W_i = \text{CMT}$ [b]	W	Y(ECMO)	Y(CMT)
1	1/2	1/2	ECMO	0	?
2	2/3	1/3	CMT	?	1
3	3/4	1/4	ECMO	0	?
4	4/5	1/5	ECMO	0	?
5	5/6	1/6	ECMO	0	?
6	6/7	1/7	ECMO	0	?
7	7/8	1/8	ECMO	0	?
8	8/9	1/9	ECMO	0	?
9	9/10	1/10	ECMO	0	?
10	10/11	1/11	ECMO	0	?
11	11/12	1/12	ECMO	0	?
12	12/13	1/13	ECMO	0	?

[a] Prob of $W_i = \text{ECMO}$ given previous assignments and observed outcomes
[b] Prob of $W_i = \text{CMT}$ given previous assignments and observed outcomes

   - 11 infants received ECMO and all survived. 1 infant received CMT and died.

- Note that we can calculate the Prob of $W_i = \text{ECMO}$ for the observed assignment for each individual given the previous assignments and observed outcomes. Let 1 represent the ECMO treatment and 0 represent the CMT treatment. Then

$$\begin{aligned} P(W_{1-3} = 101) &= P(W_1 = 1) * P(W_2 = 0 | W_1 = 1, Y_1(1) = Y_{1,\text{obs}} = 0) \\ &\quad * P(W_3 = 1 | W_1 = 1, Y_1(1) = Y_{1,\text{obs}} = 0, W_2 = 0, Y_2(0) = Y_{2,\text{obs}} = 1) \\ &= \tfrac{1}{2} * \tfrac{1}{3} * \tfrac{3}{4} \end{aligned}$$

However, we can not calculate the Prob of W for other, unobserved, values of W. For example,

$$P(W_{1-3} = 011) = P(W_1 = 0) * P(W_2 = 1 | W_1 = 0, Y_1(0)) * P(W_3 = 1 | W_1 = 0, Y_1(0), W_2 = 1, Y_2(1))$$

cannot be calculated since we do not know the unobserved potential outcomes $Y_1(0)$ and $Y_2(1)$. There are thus four different possibilities for the value of this probability.

(a) If $Y_1(0) = 0$ and $Y_2(1) = 0$ then

$$\begin{aligned} P(W_{1-3} = 011) &= P(W_1 = 0) * P(W_2 = 1 | W_1 = 0, Y_1(1) = 0) \\ &\quad * P(W_3 = 1 | W_1 = 0, Y_1(1) = 0, W_2 = 1, Y_2(1) = 0) \\ &= \tfrac{1}{2} * \tfrac{1}{3} * \tfrac{2}{4} \end{aligned}$$

(b) If $Y_1(0) = 1$ and $Y_2(1) = 0$ then

$$\begin{aligned} P(W_{1-3} = 011) &= P(W_1 = 0) * P(W_2 = 1 | W_1 = 0, Y_1(1) = 1) \\ &\quad * P(W_3 = 1 | W_1 = 0, Y_1(1) = 1, W_2 = 1, Y_2(1) = 0) \\ &= \tfrac{1}{2} * \tfrac{2}{3} * \tfrac{3}{4} \end{aligned}$$

(c) If $Y_1(0) = 0$ and $Y_2(1) = 1$ then

$$\begin{aligned} P(W_{1-3} = 011) &= P(W_1 = 0) * P(W_2 = 1 | W_1 = 0, Y_1(1) = 0) \\ &\quad * P(W_3 = 1 | W_1 = 0, Y_1(1) = 0, W_2 = 1, Y_2(1) = 1) \\ &= \tfrac{1}{2} * \tfrac{1}{3} * \tfrac{1}{4} \end{aligned}$$

(d) If $Y_1(0) = 1$ and $Y_2(1) = 1$ then

$$\begin{aligned} P(W_{1-3} = 011) &= P(W_1 = 0) * P(W_2 = 1 | W_1 = 0, Y_1(1) = 1) \\ &\quad * P(W_3 = 1 | W_1 = 0, Y_1(1) = 1, W_2 = 1, Y_2(1) = 1) \\ &= \tfrac{1}{2} * \tfrac{2}{3} * \tfrac{2}{4} \end{aligned}$$

2. Randomized with cut-off design (confounded but ignorable)
   - Concerns about small size of earlier study (esp. since only 1 infant received CMT)
   - New design: treatment assigned randomly (probability 0.5) until a set number of deaths (4) were recorded under one of the treatments.
   - After that point, only the other (more successful) treatment was given.

	Phase 1: Randomized		Phase 2: Non-randomized	
	ECMO	CMT	ECMO	CMT
Lived	9	6	19	0
Died	0	4	1	0

- Randomized phase, 4 deaths in the CMT group (out of 10). By that point 9 patients had received ECMO and all survived.
- In non-randomized phase, only ECMO was given.
- By the end of the study, 19 of 20 (97%) ECMO patients survived, compared with 6 of 10 (60%) CMT patients surviving.

3. Completely randomized design (unconfounded and ignorable)
   - UK Collaborative ECMO Trial Group, "UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation," *The Lancet*, July 13, 1996, 75-82.
   - The randomized with cut-off design (#2) was also criticized because not all of the subjects had been randomly assigned
   - New study done in the UK starting in 1996: completely randomized design
     * Probability of receiving ECMO depended on observed covariates, to ensure balance on them: primary diagnosis, disease severity, referral center. Similar to biased coin proposed by Efron (B. Efron, "Forcing a sequential experiment to be balanced", *Biometrika*, 1971, 403-417).
     * Five ECMO centers in the UK. For patients randomized to ECMO they would be transferred to the closest ECMO center; patients not randomized to ECMO would receive CMT from the center at which they were already located.
     * Importance of stability assumption (SUTVA): "All treating hospitals were considered able to provide similar state-of-the-art therapy short of ECMO, an essential condition for the results to be valid." (P.J. Wolfson, "The development and use of extracorporeal membrane oxygenation in neonates", *Annals of Thoracic Surgery*, 2003, S2224-S2229)
   - Study planned for 300 infants, but stopped early after clear answer emerged after 185 infants treated
     * ECMO survival rate (measured at discharge): 70% (out of 93 infants)
     * CMT survival rate (measured at discharge): 41% (out of 92 infants)
     * One-year survival rates showed a similar difference

Example I-21: Perfect Doctor.

Outcome of interest is years lived after surgery. Doctor assigns each patient whichever surgery (old or new) will cause the patient to live longer. If the choice of surgery will have no effect on the patient's lifespan, the doctor flips a (fair) coin and assigns new surgery if heads and old surgery if tails.

$Y(1)$	$Y(0)$
1	5
11	8
3	6
5	4
7	7
10	4

Assignment Mechanism

W	Prob of **W**
(0,1,0,1,1,1)	0.5
(0,1,0,1,0,1)	0.5
Anything else	0

$Y(1)$	$Y(0)$
7	7
3	6
10	4
1	5
11	8
5	4

Assignment Mechanism

W	Prob of **W**
(1,0,1,0,1,1)	0.5
(0,0,1,0,1,1)	0.5
Anything else	0

* Note that in this example the only thing that has changed between the left and right sides is the ordering of the units. The probabilities of assignment do not change between the left and right sides, as the assignment mechanism can not depend on the labeling of the units.

Example I-22: Almost perfect doctor, Version 2

Doctor tosses a biased coin for each individual, based on $Y(0)$ and $Y(1)$. $Y(1)$ is number of years lived past surgery if given new surgery (treatment 1). $Y(0)$ is number of years lived past surgery if given traditional surgery (treatment 0).

If $Y(1) > Y(0)$, the probability of receiving the new treatment is .8: $P(W_i = 1|Y_i(0), Y_i(1)) = .8$
If $Y(1) \leq Y(0)$, the probability of receiving the new treatment is .3: $P(W_i = 1|Y_i(0), Y_1(1)) = .3$

Unit	Y(1)	Y(0)	$P(W_i = 1\|Y_i(0), Y_i(1))$	$P(W_i = 0\|Y_i(0), Y_i(1))$
1	15	9	.8	.2
2	22	27	.3	.7
3	18	10	.8	.2
4	5	7	.3	.7
5	3	3	.3	.7
6	17	12	.8	.2
7	8	10	.3	.7
8	9	11	.3	.7

Again, since each individual's treatment status is assigned independently of the other individuals, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's.

Possible Assignments

$\mathbf{W}$	$P(\mathbf{W}\|\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}))$		
$(0,0,0,0,0,0,0,0)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.7)(.7)$	=	.001
$(1,0,0,0,0,0,0,0)$	$(.8)(.7)(.2)(.7)(.7)(.2)(.7)(.7)$	=	$\cdots$
...	...		
$(0,0,0,0,0,0,1,0)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.3)(.7)$	=	$\cdots$
$(0,0,0,0,0,0,0,1)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.7)(.3)$	=	$\cdots$
$(1,1,0,0,0,0,0,0)$	$(.8)(.3)(.2)(.7)(.7)(.2)(.7)(.7)$	=	.002
$(1,0,1,0,0,0,0,0)$	$(.8)(.7)(.8)(.7)(.7)(.2)(.7)(.7)$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,0,1)$	$(.2)(.7)(.2)(.7)(.7)(.8)(.7)(.3)$	=	$\cdots$
$(0,0,0,0,0,0,1,1)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.3)(.3)$	=	$\cdots$
$(1,1,1,0,0,0,0,0)$	$(.8)(.3)(.8)(.7)(.7)(.2)(.7)(.7)$	=	.009
...	...		
$(1,1,1,1,1,1,1,0)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.3)(.7)$	=	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.7)(.3)$	=	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.3)(.3)$	=	.001

Confounded/Unconfounded and Ignorable/Nonignorable Assignments

Unconfounded Treatment Assignment The probability of assignment to a particular treatment does not involve the values of any potential outcomes.

Confounded Treatment Assignment The probability of assignment does involve the potential outcomes.

Ignorable Treatment Assignment The probability of assignment to a particular treatment involves only observed values of the potential outcomes. It does not depend on the unobserved potential outcomes.

Nonignorable Treatment Assignment The probability of assignment involves unobserved potential outcomes.

Summary of earlier examples

Example		Unconfounded?	Ignorable?
Bernoulli design, $P(W_i = 1) = .5$	Ex. I-10	Yes	Yes
Bernoulli design, $P(W_i = 1) = .4$	Ex. I-11	Yes	Yes
Completely randomized w/in blocks (gender)	Ex. I-12	Yes	Yes
Bernoulli w/in blocks (gender)	Ex. I-13	Yes	Yes
Bernoulli, depends on age	Ex. I-14	Yes	Yes
Bernoulli, School choice	Ex. I-15	Yes	Yes
Bernoulli, Set # treated	Ex. I-16	Yes	Yes
Randomized w/in matched pairs	Ex. I-17	Yes	Yes
Bernoulli, Teacher	Ex. I-18	No	No
ECMO	Ex. I-19	No	Yes
Perfect Doctor	Ex. I-20	No	No
Almost Perfect Doctor	Ex. I-21	No	No

What makes this Rubin Causal Model perspective novel?

1. Potential outcomes define causal effects in all cases
    - Randomized experiments and observational studies
    - Break from the tradition before the 1970's
    - Allows stability (SUTVA) to be stated formally
2. Explicit model for the assignment mechanism
    - Special process for creating missing data in the potential outcomes
    - Allows possible dependence on potential outcomes
    - Randomized experiments a special case whose benefits for causal inference can be formally stated
3. Allows specification of joint distribution of the potential outcomes
    - Framework can thus accommodate both assignment-mechanism-based (randomization-based) and model-based (Bayesian) inference
    - One unified perspective for distinct methods of causal inference traditionally used for randomized experiments and those traditionally used for observational studies
    - Creates firm foundation for (Bayesian) methods for dealing with complications such as noncompliance and dropout

Part II: Causal Inference Based on the Assignment Mechanism

Proof by Contradiction

Steps

1. Start out by assuming the opposite of what you want to prove.
2. Working from this assumption, arrive at a contradiction.
3. Conclude that your initial assumption was wrong, and the proof is complete.

Example II-1: Word Problem

Jane is 23 years younger than her mother.
Jane's parents' ages sum to 58.
Jane's mother is two years younger than Jane's father.
How old is Jane?

We can solve this problem by using the above method many times:

1. Start by assuming Jane is 30.
2. This means Jane's mother must be 53 (since Jane is 23 years younger than her mother), which means Jane's father must be 5 (since her parents' ages sum to 58). However, Jane's mother is then 48 years older than her father. We've reached a contradiction, since the problem says Jane's mother is two years younger than her father.
3. Our assumption that Jane is 30 must be wrong.

Try again:

1. Assume Jane is 10.
2. This means that Jane's mother must be 33, which means Jane's father must be 25. Another contradiction, since Jane's mother is not two years younger than Jane's father here.

3. Our assumption that Jane is 10 must be wrong.

Keep repeating the process until you don't arrive at a contradiction. Eventually you'll guess that Jane is five years old:

1. Assume Jane is five years old.
2. If Jane is five, her mother must be 28, so her father must be 30. Now Jane's father is two years older than her mother. No contradiction!
3. We cannot reject the assumption that Jane is five years old, i.e., Jane being five years old is a solution to the problem.

Example II-2: Irrationality of $\sqrt{2}$

A rational number is one that can be expressed as $\frac{p}{q}$, where p and q are both integers with no common divisors. We want to prove that the square root of two is irrational, i.e., it cannot be expressed this way.

Assume $\sqrt{2}$ is rational: $\sqrt{2} = \frac{p}{q}$.

$\Rightarrow \frac{p^2}{q^2} = 2$

$\Rightarrow p^2 = 2q^2$

$\Rightarrow p^2$ is an even integer

$\Rightarrow p$ is an even integer

$\Rightarrow p = 2k$, where k is an integer

$\Rightarrow p^2 = 4k^2$

$\Rightarrow 4k^2 = 2q^2$

$\Rightarrow 2k^2 = q^2$

$\Rightarrow q^2$ is an even integer

$\Rightarrow q$ is an even integer

We assumed p and q had no common divisors, **BUT** since p and q are both even, they have a common divisor of 2. Thus we have arrived at a contradiction, and so our original assumption that $\sqrt{2}$ is rational must be wrong.

Fisher Test in a Completely Randomized Experiment

Steps

1. Specify null hypothesis (hypothesis regarding the size of the treatment effect).

   Usually use hypothesis of no effect of treatment ($Y_i(0) = Y_i(1)$ for all individuals).

2. Fill in missing potential outcomes using the null hypothesis and the observed values of the potential outcomes.

3. Calculate the observed estimate of the treatment effect (the test statistic of interest).

   Often use the difference in observed sample means of the treated and control groups ($\overline{y(1)} - \overline{y(0)}$).

4. For each possible assignment, calculate the value of the test statistic of interest that would have been observed under that assignment (the same calculation as in Step 3, with different "observed" values).

5. Determine how rare the value observed in Step 3 is. This is called the significance level or probability value (p-value).

   Add up the probabilities of all assignments that lead to a test statistic value as or more extreme than the value observed.

Example II-3: Children's Television Workshop

An experiment was done to examine the effect of watching Children's Television Workshop programs (such as the Electric Company) on children's reading ability. We consider just 6 observations, with 3 given treatment and 3 given control (completely randomized). The treatment is watching the programs, control is not watching them. Post-program test scores of the children are given below. The missing potential outcomes (in parentheses) are filled in using the null hypothesis of no treatment effect ($Y_i(0) = Y_i(1)$ for all individuals).

1. Null hypothesis: There is no effect of the treatment ($Y_i(0) = Y_i(1)$ for all individuals).

2. Fill in missing potential outcomes:

Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(55.0)
2	0	72.0	72.0	(72.0)
3	0	72.7	72.7	(72.7)
4	1	70.0	(70.0)	70.0
5	1	66.0	(66.0)	66.0
6	1	78.9	(78.9)	78.9

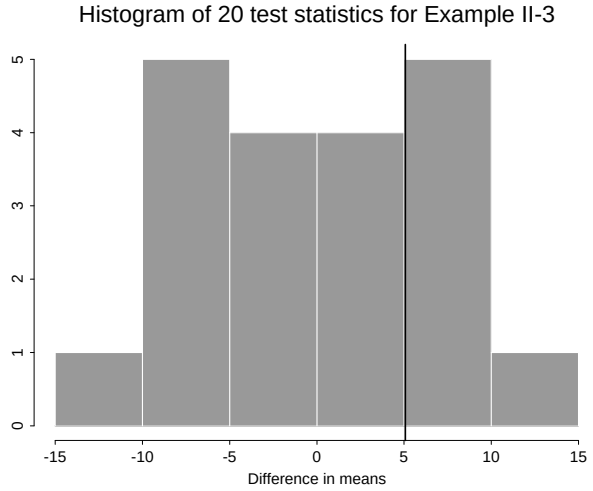
3. Observed estimate of the treatment effect: $\overline{y(1)} - \overline{y(0)} = 5.1$

4. The following table lists all possible randomizations of this data with the corresponding test statistics (difference in means) that would have been observed under each assignment. The observed randomization and outcome are in bold.

All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$		
1 1 1 0 0 0	1/20	-5.1		
1 1 0 1 0 0	1/20	-6.9		
1 1 0 0 1 0	1/20	-9.5		
1 1 0 0 0 1	1/20	-0.9		
1 0 1 1 0 0	1/20	-6.4		
1 0 1 0 1 0	1/20	-9.1		
1 0 1 0 0 1	1/20	-0.5	=	$\frac{55+72.7+78.9}{3} - \frac{72+70+66}{3}$
1 0 0 1 1 0	1/20	-10.9		
1 0 0 1 0 1	1/20	-2.3		
1 0 0 0 1 1	1/20	-4.9		
0 1 1 1 0 0	1/20	4.9	=	$\frac{72+72.7+70}{3} - \frac{55+66+78.9}{3}$
0 1 1 0 1 0	1/20	2.3		
0 1 1 0 0 1	1/20	10.9	=	$\frac{72+72.7+78.9}{3} - \frac{55+70+66}{3}$
0 1 0 1 1 0	1/20	0.5		
0 1 0 1 0 1	1/20	9.1		
0 1 0 0 1 1	1/20	6.4		
0 0 1 1 1 0	1/20	0.9		
0 0 1 1 0 1	1/20	9.5		
0 0 1 0 1 1	1/20	6.9		
0 0 0 1 1 1	**1/20**	**5.1**		

5. The probability of observing the value that we did (5.1) or something more extreme is $6/20 = .3$ (i.e., the p-value or significance level is 0.3).



Example II-4: Children's Television Workshop Part II

Same set-up as in Example II-3, however we now use a null hypothesis of a treatment effect of 5 points ($Y_i(1) - Y_i(0) = 5$ for all individuals). This null hypothesis assumes additive treatment effects, i.e., the treatment adds a fixed amount to each control value.

1. Null hypothesis: There is an additive treatment effect of 5 points: $Y_i(1) - Y_i(0) = 5$ for all individuals.
2. Fill in missing potential outcomes:

Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(60.0)
2	0	72.0	72.0	(77.0)
3	0	72.7	72.7	(77.7)
4	1	70.0	(65.0)	70.0
5	1	66.0	(61.0)	66.0
6	1	78.9	(73.9)	78.9

3. The observed estimate of the treatment effect is $\overline{y(1)} - \overline{y(0)} = 5.1$.
4. The following table lists all possible randomizations of this data with the corresponding test statistics (difference in means) that would have been observed under each of the randomizations. The observed randomization and outcome are in bold.

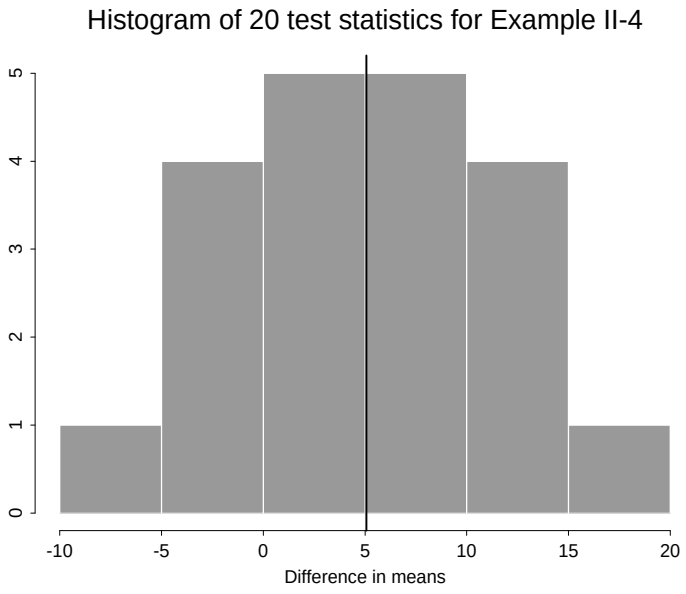
All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$		
1 1 1 0 0 0	1/20	4.9		
1 1 0 1 0 0	1/20	-0.2		
1 1 0 0 1 0	1/20	-2.9		
1 1 0 0 0 1	1/20	5.7		
1 0 1 1 0 0	1/20	0.3		
1 0 1 0 1 0	1/20	-2.4	=	$\frac{60+77.7+66}{3} - \frac{72+65+73.9}{3}$
1 0 1 0 0 1	1/20	6.2		
1 0 0 1 1 0	1/20	-7.5		
1 0 0 1 0 1	1/20	1.1		
1 0 0 0 1 1	1/20	-1.6		
0 1 1 1 0 0	1/20	11.6		
0 1 1 0 1 0	1/20	8.9		
0 1 1 0 0 1	1/20	17.5	=	$\frac{77+77.7+78.9}{3} - \frac{55+65+61}{3}$
0 1 0 1 1 0	1/20	3.8		
0 1 0 1 0 1	1/20	12.4		
0 1 0 0 1 1	1/20	9.7		
0 0 1 1 1 0	1/20	4.3	=	$\frac{77.7+70+66}{3} - \frac{55+72+73.9}{3}$
0 0 1 1 0 1	1/20	12.9		
0 0 1 0 1 1	1/20	10.2		
0 0 0 1 1 1	**1/20**	**5.1**		

5. Ten randomizations give estimated treatment effects as extreme or more extreme than what we observed (under the sharp null hypothesis of a treatment effect of 5 points). These randomizations are listed below.

$\mathbf{W}$	Prob of $\mathbf{W}$	$\overline{y(1)} - \overline{y(0)}$
1 1 0 0 0 1	1/20	5.7
1 0 1 0 0 1	1/20	6.2
0 1 1 1 0 0	1/20	11.6
0 1 1 0 1 0	1/20	8.9
0 1 1 0 0 1	1/20	17.5
0 1 0 1 0 1	1/20	12.4
0 1 0 0 1 1	1/20	9.7
0 0 1 1 0 1	1/20	12.9
0 0 1 0 1 1	1/20	10.2
0 0 0 1 1 1	**1/20**	**5.1**

The probability of observing the value that we did (5.1) or something more extreme is thus $10/20 = 0.5$ (i.e., the p-value or significance level is 0.5).



Constructing Fisher Intervals in Completely Randomized Experiments

Example II-5: Children's Television Workshop (continued)

This continues the example from Handout II-1, of an experiment regarding the effect of Children's Television Workshop programming on children's reading ability. We are now interested in determining a range of plausible values of the treatment effect.

We consider a range of treatment effects, and conduct a Fisher test on each possible value to determine the p-value corresponding to that effect size.

The data is shown below. We now fill in the missing potential outcomes according to a null hypothesis of a treatment effect of size x: $Y_i(1) - Y_i(0) = x$ for all individuals. This method assumes that there is a constant, additive treatment effect (x) for all individuals.

Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(55.0+x)
2	0	72.0	72.0	(72.0+x)
3	0	72.7	72.7	(72.7+x)
4	1	70.0	(70.0-x)	70.0
5	1	66.0	(66.0-x)	66.0
6	1	78.9	(78.9-x)	78.9

A few specific examples of this are shown below.

a. Null Hypothesis: treatment effect size is -6 (i.e. the programming lowers each child's reading score by 6 points.)

Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(49.0)
2	0	72.0	72.0	(66.0)
3	0	72.7	72.7	(66.7)
4	1	70.0	(76.0)	70.0
5	1	66.0	(72.0)	66.0
6	1	78.9	(84.9)	78.9

The following table lists all possible randomizations of this data with the corresponding test statistics (difference in means) that would have been observed under each assignment. The observed randomization and outcome are in bold.

All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$	
1 1 1 0 0 0	1/20	-17.1	
1 1 0 1 0 0	1/20	-14.9	
1 1 0 0 1 0	1/20	-17.5	
1 1 0 0 0 1	1/20	-8.9	
1 0 1 1 0 0	1/20	-14.4	
1 0 1 0 1 0	1/20	-17.1	
1 0 1 0 0 1	1/20	-8.5	$= \frac{49+66.7+78.9}{3} - \frac{72+76+72}{3}$
1 0 0 1 1 0	1/20	-14.9	
1 0 0 1 0 1	1/20	-6.3	
1 0 0 0 1 1	1/20	-8.9	
0 1 1 1 0 0	1/20	-3.1	$= \frac{66+66.7+70}{3} - \frac{55+72+84.9}{3}$
0 1 1 0 1 0	1/20	-5.7	
0 1 1 0 0 1	1/20	2.9	$= \frac{66+66.7+78.9}{3} - \frac{55+76+72}{3}$
0 1 0 1 1 0	1/20	-3.5	
0 1 0 1 0 1	1/20	5.1	
0 1 0 0 1 1	1/20	2.4	
0 0 1 1 1 0	1/20	-3.1	
0 0 1 1 0 1	1/20	5.5	
0 0 1 0 1 1	1/20	2.9	
0 0 0 1 1 1	**1/20**	**5.1**	

The probability of observing the value that we did (5.1) or something more extreme (larger than 5.1) is $3/20 = .15$ (i.e. the p-value or significance level is 0.15).

b. Null Hypothesis: treatment effect size is 12 (i.e., the programming raises children's reading scores by 12 points.)

Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(67.0)
2	0	72.0	72.0	(84.0)
3	0	72.7	72.7	(84.7)
4	1	70.0	(58.0)	70.0
5	1	66.0	(54.0)	66.0
6	1	78.9	(66.9)	78.9

The following table lists all possible randomizations of this data with the corresponding test statistics (difference in means) that would have been observed under each assignment. The observed randomization and outcome are in bold.

All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$	
1 1 1 0 0 0	1/20	18.9	
1 1 0 1 0 0	1/20	9.1	
1 1 0 0 1 0	1/20	6.5	
1 1 0 0 0 1	1/20	15.1	
1 0 1 1 0 0	1/20	9.6	
1 0 1 0 1 0	1/20	6.9	
1 0 1 0 0 1	1/20	15.5	$= \frac{67+84.7+78.9}{3} - \frac{72+58+54}{3}$
1 0 0 1 1 0	1/20	-2.9	
1 0 0 1 0 1	1/20	5.7	
1 0 0 0 1 1	1/20	3.1	
0 1 1 1 0 0	1/20	20.9	$= \frac{84+84.7+70}{3} - \frac{55+54+66.9}{3}$
0 1 1 0 1 0	1/20	18.3	
0 1 1 0 0 1	1/20	26.9	$= \frac{84+84.7+78.9}{3} - \frac{55+58+54}{3}$
0 1 0 1 1 0	1/20	8.5	
0 1 0 1 0 1	1/20	17.1	
0 1 0 0 1 1	1/20	14.4	
0 0 1 1 1 0	1/20	8.9	
0 0 1 1 0 1	1/20	17.5	
0 0 1 0 1 1	1/20	14.9	
0 0 0 1 1 1	**1/20**	**5.1**	

The probability of observing the value that we did (5.1) or something more extreme (here defined as smaller than 5.1) is $3/20 = .15$ (i.e. the p-value or significance level is 0.15).

To determine an interval of likely values for the treatment effect, we now systematically go through hypothesized values until we find effects that are unlikely to lead to the observed data (i.e., have low p-values).

The observed test statistic is 5.1. We first consider values less than 5.1, and determine the corresponding p-value for each hypothesized additive treatment effect. The table below shows the hypothesized treatment effect sizes and the corresponding p-values.

Treatment Effect	p-value	Treatment Effect	p-value
5	0.50	-3	0.20
4	0.40	-4	0.20
3	0.40	-5	0.15
2	0.35	-6	0.15
1	0.35	-7	0.05
0	0.30	-8	0.05
-1	0.30	-9	0.05
-2	0.30	-10	0.05

Note that -6 is the smallest value considered that does not have a p-value less than .05.

We now consider hypothesized treatment effects greater than 5:

Treatment Effect	p-value	Treatment Effect	p-value
6	0.50	16	0.10
7	0.45	17	0.10
8	0.40	18	0.10
9	0.35	19	0.10
10	0.30	20	0.10
11	0.25	21	0.10
12	0.15	22	0.10
13	0.15	23	0.10
14	0.15	24	0.10
15	0.15	25	0.05

Note that +24 is the largest value considered that does not have a p-value less than .05.

A plausible range of values for the treatment effect (a 90% interval) is thus [-6, 24]: the set of values whose p-value is greater than .05 in either direction, i.e. the set of values that are not rejected by a .05 test in either direction.

Neymanian Randomization-Based Inference in Completely Randomized Experiments

Example II-6: Biased/Unbiased Estimates of the Average Treatment Effect

Unbiased estimator If a statistic is an unbiased estimate of the treatment effect, then the average of the value of that statistic over all possible randomizations will equal the true treatment effect.

We have already seen the unbiasedness of $\overline{y(1)} - \overline{y(0)}$ in Chapter 2, such as Example I-6.

Case 1: There is an additive treatment effect of 3 years for each patient.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	7	3
2	3	0	3
3	5	2	3
4	12	9	3
5	8	5	3
6	9	6	3
Average	7.83	4.83	3

A doctor uses a completely randomized design, and assigns 4 patients to treatment and 2 to control. The true data is shown in the table above. The following table lists the 15 ($\binom{6}{4}$) possible assignments and two corresponding estimates of the treatment effect (the observed difference in means and the difference in medians). Note that the doctor will only be able to observe one of these possible randomizations.

All Possible Assignments of 4 to Treatment

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$			median($y(1)$)-median($y(0)$)
1 1 1 1 0 0	1/15	2.00			2.0
1 1 1 0 1 0	1/15	-1.00			-1.0
1 1 1 0 0 1	1/15	-0.25	=	$\frac{10+3+5+9}{4} - \frac{9+5}{2}$	0.0
1 1 0 1 1 0	1/15	4.25			5.0
1 1 0 1 0 1	1/15	5.00			6.0
1 1 0 0 1 1	1/15	2.00			3.0
1 0 1 1 1 0	1/15	5.75			6.0
1 0 1 1 0 1	1/15	6.50	=	$\frac{10+5+12+9}{4} - \frac{0+5}{2}$	7.0
1 0 1 0 1 1	1/15	3.50			4.0
1 0 0 1 1 1	1/15	8.75			8.5
0 1 1 1 1 0	1/15	0.50			0.0
0 1 1 1 0 1	1/15	1.25			1.0
0 1 1 0 1 1	1/15	-1.75	=	$\frac{3+5+8+9}{4} - \frac{7+9}{2}$	-1.5
0 1 0 1 1 1	1/15	3.50			4.0
0 0 1 1 1 1	1/15	5.00			5.0
Average		3.00			3.3
Variance		8.225			8.82

The average of all possible values of $\overline{y(1)} - \overline{y(0)}$ that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect. However, we see that the difference in medians (which may be of clinical interest) is not unbiased for either the true difference in means or the true difference in medians (which is $3 = 8.5 - 5.5$).

Case 2: There is an average treatment effect of 3 years.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	7	3
2	3	2	1
3	5	0	5
4	12	9	3
5	8	6	2
6	9	5	4
Average	7.83	4.83	3

The table below again lists all possible randomizations and the corresponding observed estimates of the treatment effect.

All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$			$\mathrm{median(y(1))} - \mathrm{median(y(0))}$
1 1 1 1 0 0	1/15	2.00			2.0
1 1 1 0 1 0	1/15	-0.50			-0.5
1 1 1 0 0 1	1/15	-0.75	=	$\frac{10+3+5+9}{4} - \frac{9+6}{2}$	-0.5
1 1 0 1 1 0	1/15	5.75			6.5
1 1 0 1 0 1	1/15	5.50			6.5
1 1 0 0 1 1	1/15	3.00			4.0
1 0 1 1 1 0	1/15	5.25	=	$\frac{10+5+12+8}{4} - \frac{2+5}{2}$	5.5
1 0 1 1 0 1	1/15	5.00			5.5
1 0 1 0 1 1	1/15	2.50			3.0
1 0 0 1 1 1	1/15	8.75			8.5
0 1 1 1 1 0	1/15	1.00			0.5
0 1 1 1 0 1	1/15	0.75			0.5
0 1 1 0 1 1	1/15	-1.75	=	$\frac{3+5+8+9}{4} - \frac{7+9}{2}$	-1.5
0 1 0 1 1 1	1/15	4.50			5.0
0 0 1 1 1 1	1/15	4.00			4.0
Average		3.00			3.3
Variance		7.90			8.70

The average of all possible differences in means that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect, even when there is not an additive treatment effect. However, we see that again, the difference in medians is not unbiased for either the true difference in means or the true difference in medians ($3.0 = 8.5 - 5.5$).

Case 3: There is an average treatment effect of 3 years.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	2	8
2	3	9	-6
3	5	7	-2
4	12	0	12
5	8	6	2
6	9	5	4
Average	7.83	4.83	3

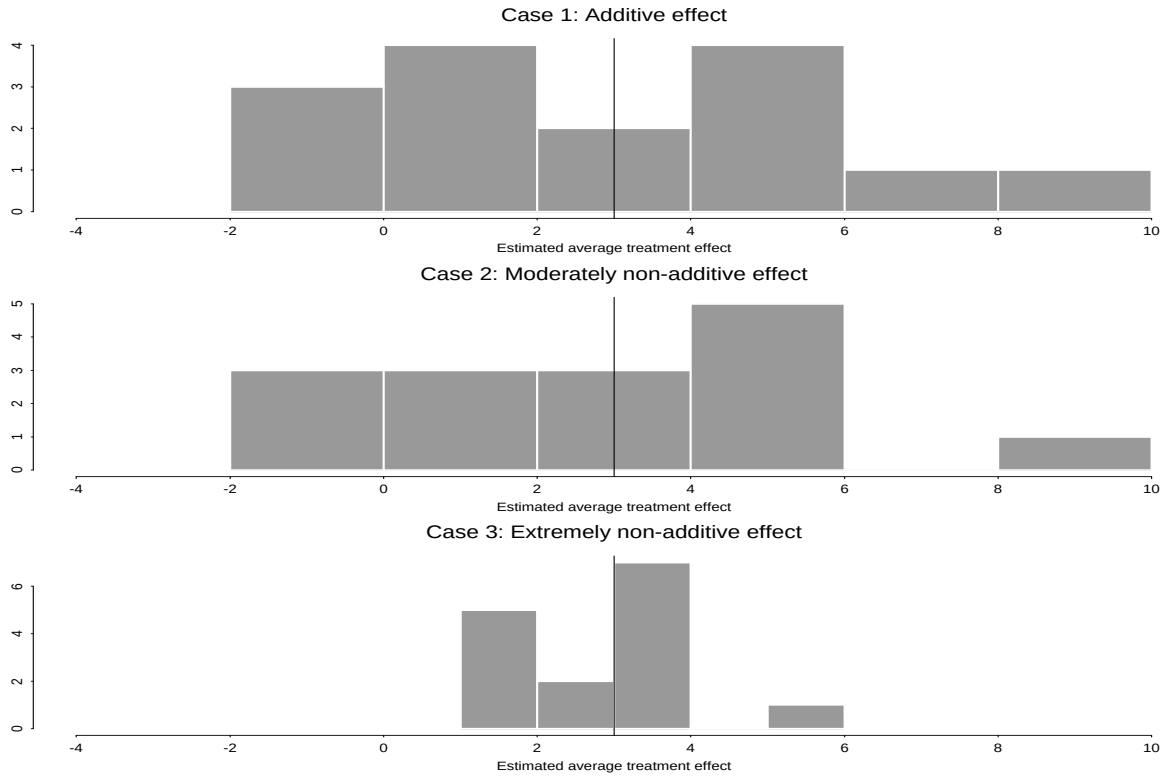
The table below again lists all possible randomizations and the corresponding observed estimates of the treatment effect.

All Possible Assignments

W	Prob of **W**	$\overline{y(1)} - \overline{y(0)}$			$\text{median(y(1))} - \text{median(y(0))}$
1 1 1 1 0 0	1/15	2.00			2.0
1 1 1 0 1 0	1/15	4.00			4.0
1 1 1 0 0 1	1/15	3.75	=	$\frac{10+3+5+9}{4} - \frac{0+6}{2}$	4.0
1 1 0 1 1 0	1/15	2.25			3.0
1 1 0 1 0 1	1/15	2.00			3.0
1 1 0 0 1 1	1/15	4.00			5.0
1 0 1 1 1 0	1/15	1.75	=	$\frac{10+5+12+8}{4} - \frac{9+5}{2}$	2.0
1 0 1 1 0 1	1/15	1.50			2.0
1 0 1 0 1 1	1/15	3.50			4.0
1 0 0 1 1 1	1/15	1.75			1.5
0 1 1 1 1 0	1/15	3.50			3.0
0 1 1 1 0 1	1/15	3.25			3.0
0 1 1 0 1 1	1/15	5.25	=	$\frac{3+5+8+9}{4} - \frac{2+0}{2}$	5.5
0 1 0 1 1 1	1/15	3.50			4.0
0 0 1 1 1 1	1/15	3.00			3.0
Average		3.00			3.3
Variance		1.09			1.23

The average of all possible test statistics that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect, even when there is not an additive treatment effect. However, we see that again, the difference in medians is not unbiased for either the true difference in means or the true difference in medians ($3.0 = 8.5 - 5.5$).

The following histograms show the estimates of the average treatment effect under all possible randomizations, for each of the three cases above.



Note that greater spread occurs as the same average causal effect becomes more additive.

Estimating the Variance of the Average Treatment Effect

Remember that we have defined the average treatment effect as

$$\text{ATE} = \overline{Y(1)} - \overline{Y(0)},$$

the difference in means of the outcomes for the entire population.

Neyman (1923) showed that in a completely randomized experiment, an unbiased estimate of this is

$$\widehat{\text{ATE}} = \overline{y(1)} - \overline{y(0)},$$

the difference in means of the outcomes in the observed samples.

Let n_1 and n_0 be the sizes of the observed treated and control groups, respectively, and $\text{Var}[Y(1)]$ and $\text{Var}[Y(0)]$ be the variance of the true potential outcomes under treatment and control, respectively. Assuming an additive treatment effect, Neyman showed that the true variance of the estimated treatment effect is:

$$\text{VAR} = \frac{\text{Var}[Y(1)]}{n_1} + \frac{\text{Var}[Y(0)]}{n_0}.$$

He also showed that VAR is larger than the true variance of $\widehat{\text{ATE}}$ when additivity does not hold:

$$\text{Var}(\widehat{\text{ATE}}) = \frac{\text{Var}[Y(1)]}{n_1} + \frac{\text{Var}[Y(0)]}{n_0} - \frac{\text{Var}[Y(1) - Y(0)]}{n_1 + n_0}.$$

Now, let s_1^2 and s_0^2 be the sample variances of the observed treated and control groups, respectively. The following is an unbiased estimator of the true variance of the estimated treatment effect, assuming additivity:

$$\widehat{\text{VAR}} = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}.$$

In large samples, a 95% confidence interval can then be formed using the following:

$$(\widehat{ATE} - 1.96 * \sqrt{\widehat{VAR}}, \widehat{ATE} + 1.96 * \sqrt{\widehat{VAR}})$$

with $\widehat{ATE}$ and $\widehat{VAR}$ defined above. This is based on a normal distributional approximation.

The square root of the variance is known as the standard deviation (SD).

Example II-7 (Example II-6, cont.)

Again consider the example from Handout II-3 and the three cases. We now wish to obtain an estimate of the variance of the estimated average treatment effect. For each of the cases, we will calculate the true variance as well as the estimated variances.

Case 1: There is an additive treatment effect of 3 years for each patient.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	7	3
2	3	0	3
3	5	2	3
4	12	9	3
5	8	5	3
6	9	6	3
Average	7.83	4.83	3
Variance	11.0	11.0	0

In the table below, for each randomization we show the sample variance in the treated and control samples, the estimate of the standard deviation (the square root of the variance) that would have been observed under each randomization, and the Neyman large sample 85% confidence interval.

W	Prob of **W**	Var(y(1)) $= s_1^2$	Var(y(0)) $= s_0^2$	$\widehat{\mathrm{SD}(\mathrm{ATE})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$	85% CI	p-value [a] (test of 3)	p-value [b] (test of 0)
1 1 1 1 0 0	1/15	17.7	0.5	2.2	(-1.1, 5.1)	0.64	0.35
1 1 1 0 1 0	1/15	9.7	4.5	2.2	(-4.1, 2.1)	0.06	0.64
1 1 1 0 0 1	1/15	10.9	8.0	2.6	(-4.0, 3.5)	0.21	0.92
1 1 0 1 1 0	1/15	14.9	8.0	2.8	(0.3, 8.3)	0.65	0.13
1 1 0 1 0 1	**1/15**	**15.0**	**4.5**	**2.4**	**(1.5, 8.5)**	**0.41**	**0.04**
1 1 0 0 1 1	1/15	9.7	24.5	3.8	(-3.5, 7.5)	0.79	0.60
1 0 1 1 1 0	1/15	8.9	18.0	3.4	(0.9, 10.6)	0.41	0.09
1 0 1 1 0 1	1/15	8.7	12.5	2.9	(2.3, 10.7)	0.23	0.03
1 0 1 0 1 1	1/15	4.7	40.5	4.6	(-3.2, 10.2)	0.91	0.45
1 0 0 1 1 1	1/15	2.9	2.0	1.3	(6.9, 10.6)	0.00	0.00
0 1 1 1 1 0	1/15	15.3	0.5	2.0	(-2.4, 3.4)	0.22	0.80
0 1 1 1 0 1	1/15	16.3	2.0	2.3	(-2.0, 4.5)	0.44	0.58
0 1 1 0 1 1	1/15	7.6	2.0	1.7	(-4.2, 0.7)	0.01	0.30
0 1 0 1 1 1	1/15	14.0	12.5	3.1	(-1.0, 8.0)	0.87	0.26
0 0 1 1 1 1	1/15	8.3	24.5	3.8	(-0.5, 10.5)	0.60	0.19
Average		11.0	11.0	$\sqrt{8.225}$			

[a] This corresponds to a 2-sided "t-test" of a treatment effect of 3, which is (in large experiments) twice the p-value from a Fisher test of 3

[b] This corresponds to a 2-sided test of a treatment effect of 0, which is (in large experiments) twice the p-value from a Fisher test of 0

Note that 12 out of the 15 intervals contain the true treatment effect (3): $\frac{12}{15} = 0.80$.

Assuming the randomization in bold was observed, we now compare these results with the results from a Fisher test.

86% Fisher Interval:	[-2,11]
p-value for null of no treatment effect:	$\frac{2}{15} = .13$
p-value for null of treatment effect of 3:	$\frac{5}{15} = .33$

Case 2: There is an average treatment effect of 3 years.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	7	3
2	3	2	1
3	5	0	5
4	12	9	3
5	8	6	2
6	9	5	4
Average	7.83	4.83	3
Variance	11.0	11.0	2

In the table below, for each randomization we show the variance in the treated and control samples, the estimate of the standard deviation (the square root of the variance) that would have been observed under each randomization, and the Neyman large sample 85% confidence interval.

All Possible Assignments

W	Prob of **W**	Var(y(1)) $= s_1^2$	Var(y(0)) $= s_0^2$	$\widehat{\text{SD}(\text{ATE})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$	85% CI	p-value [a] (test of 3)	p-value [b] (test of 0)
1 1 1 1 0 0	1/15	17.7	0.5	2.2	(-1.1, 5.1)	0.64	0.35
1 1 1 0 1 0	1/15	9.7	8.0	2.5	(-4.2, 3.2)	0.17	0.84
1 1 1 0 0 1	1/15	10.9	4.5	2.2	(-4.0, 2.5)	0.09	0.74
1 1 0 1 1 0	1/15	14.9	12.5	3.2	(1.2, 10.3)	0.38	0.07
1 1 0 1 0 1	1/15	15.0	18.0	3.6	(0.4, 10.6)	0.48	0.12
1 1 0 0 1 1	1/15	9.7	40.5	4.8	(-3.9, 9.9)	1.00	0.53
1 0 1 1 1 0	1/15	8.9	4.5	2.1	(2.2, 8.3)	0.29	0.01
1 0 1 1 0 1	1/15	8.7	8.0	2.5	(1.4, 8.6)	0.42	0.04
1 0 1 0 1 1	1/15	4.7	24.5	3.7	(-2.8, 7.8)	0.89	0.49
1 0 0 1 1 1	1/15	2.9	2.0	1.3	(6.9, 10.6)	0.00	0.00
0 1 1 1 1 0	1/15	15.3	2.0	2.2	(-2.2, 4.2)	0.36	0.65
0 1 1 1 0 1	**1/15**	**16.3**	**0.5**	**2.1**	**(-2.2, 3.7)**	**0.28**	**0.72**
0 1 1 0 1 1	1/15	7.6	2.0	1.7	(-4.2, 0.7)	0.01	0.30
0 1 0 1 1 1	1/15	14.0	24.5	4.0	(-1.2, 10.2)	0.71	0.26
0 0 1 1 1 1	1/15	8.3	12.5	2.9	(-0.2, 8.2)	0.73	0.17
Average		11.0	11.0	$\sqrt{8.225}$			

[a] This corresponds to a 2-sided "t-test" of a treatment effect of 3, which is (in large experiments) twice the p-value from a Fisher test of 3

[b] This corresponds to a 2-sided test of a treatment effect of 0, which is (in large experiments) twice the p-value from a Fisher test of 0

Note that 12 out of the 15 intervals contain the true treatment effect (3): $\frac{12}{15} = 0.80$.

Assuming the randomization in bold was observed, we now compare these results with the results from a Fisher test.

86% Fisher Interval:	[-4,6]
p-value for null of no treatment effect:	$\frac{8}{15} = 0.47$
p-value for null of treatment effect of 3:	$\frac{5}{15} = 0.33$

Case 3: There is an average treatment effect of 3 years.

Patient	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	10	2	8
2	3	9	-6
3	5	7	-2
4	12	0	12
5	8	6	2
6	9	5	4
Average	7.83	4.83	3
Variance	11.0	11.0	42.8

In the table below, for each randomization we show the variance in the treated and control samples, the estimate of the standard deviation (the square root of the variance) that would have been observed under each randomization, and the Neyman large sample 85% confidence interval for the treatment effect.

All Possible Assignments

W	Prob of **W**	Var(y(1)) $= s_1^2$	Var(y(0)) $= s_0^2$	$\widehat{\text{SD}(\text{ATE})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$	85% CI	p-value [a] (test of 3)	p-value [b] (test of 0)
1 1 1 1 0 0	1/15	17.7	0.5	2.2	(-1.1, 5.1)	0.64	0.35
1 1 1 0 1 0	1/15	9.7	12.5	2.9	(-0.2, 8.2)	0.73	0.17
1 1 1 0 0 1	1/15	10.9	18.0	3.4	(-1.2, 8.7)	0.83	0.27
1 1 0 1 1 0	1/15	14.9	2.0	2.2	(-0.9, 5.4)	0.73	0.30
1 1 0 1 0 1	1/15	15.0	0.5	2.0	(-0.9, 4.9)	0.62	0.32
1 1 0 0 1 1	1/15	9.7	24.5	3.8	(-1.5, 9.5)	0.79	0.30
1 0 1 1 1 0	**1/15**	**8.9**	**8.0**	**2.5**	**(-1.8, 5.3)**	**0.62**	**0.48**
1 0 1 1 0 1	1/15	8.7	4.5	2.1	(-1.5, 4.5)	0.48	0.48
1 0 1 0 1 1	1/15	4.7	40.5	4.6	(3.2, 10.2)	0.91	0.45
1 0 0 1 1 1	1/15	2.9	2.0	1.3	(-0.1, 3.6)	0.34	0.18
0 1 1 1 1 0	1/15	15.3	4.5	2.5	(-0.1, 7.1)	0.84	0.16
0 1 1 1 0 1	1/15	16.3	8.0	2.8	(-0.8, 7.3)	0.93	0.25
0 1 1 0 1 1	1/15	7.6	2.0	1.7	(2.8, 7.7)	0.19	0.00
0 1 0 1 1 1	1/15	14.0	12.5	3.1	(-1.0, 8.0)	0.87	0.26
0 0 1 1 1 1	1/15	8.3	24.5	3.8	(-2.5, 8.5)	1.00	0.43
Average		11.0	11.0	$\sqrt{8.225}$			

[a] This corresponds to a 2-sided "t-test" of a treatment effect of 3, which is (in large experiments) twice the p-value from a Fisher test of 3

[b] This corresponds to a 2-sided test of a treatment effect of 0, which is (in large experiments) twice the p-value from a Fisher test of 0

Note that 14 out of the 15 intervals contain the true treatment effect (3): $\frac{14}{15} = 0.93$.

Assuming the randomization in bold was observed, we now compare these results with the results from a Fisher test.

86% Fisher Interval:	[-4,7]
p-value for null of no treatment effect:	$\frac{5}{15} = 0.33$
p-value for null of treatment effect of 3:	$\frac{7}{15} = 0.47$

Effi ciency Benefi ts from Using Covariates

Example II-8: Comparing test statistics in a completely randomized experiment

We are interested in estimating the effect of an SAT prep class on SAT test scores. There are 4 students in the experiment, where 2 will be randomly chosen to receive treatment and the other 2 will receive control (no class). Consider a hypothetical experiment where we know all students' potential outcomes under both treatment and control. For each student, we also observe their SAT score before treatment assignment (X).

Table 1: Data

	Pre-test SAT	Post-test SAT		Causal Effect	Gain Scores		Causal Effect
Unit	(X)	Y(0)	Y(1)	Y(1)-Y(0)	Y(0)-X	Y(1)-X	Y(1)-X-(Y(0)-X)
1	300	350	400	50	50	100	50
2	400	450	550	100	50	150	100
3	500	550	550	0	50	50	0
4	600	650	700	50	50	100	50
Average	450	500	550	50	50	100	50

Table 2: Neyman Estimates for all Possible Assignments

W	$\overline{y_1} - \overline{y_0}$	Std. Dev.	$\overline{y_1 - X} - \overline{y_0 - X}$	Std. Dev. (gain)
1100	$475 - 600 = -125$	90	$125 - 50 = 75$	25
1010	$475 - 550 = -75$	125	$75 - 50 = 25$	25
1001	$550 - 500 = 50$	158	$100 - 50 = 50$	0
0110	$550 - 500 = 50$	150	$100 - 50 = 50$	50
0101	$625 - 450 = 175$	125	$125 - 50 = 75$	25
0011	$625 - 400 = 225$	90	$75 - 50 = 25$	25
Average	50	123	50	25

Table 3: Fisher test results for all Possible Assignments

W	$\overline{y_1} - \overline{y_0}$	p-value	$\overline{y_1 - X} - \overline{y_0 - X}$	p-value (gain)
1100	$475 - 600 = -125$	1	$125 - 50 = 75$	0.17
1010	$475 - 550 = -75$	0.83	$75 - 50 = 25$	0.5
1001	$550 - 500 = 50$	0.5	$100 - 50 = 50$	0.17
0110	$550 - 500 = 50$	0.5	$100 - 50 = 50$	0.5
0101	$625 - 450 = 175$	0.17	$125 - 50 = 75$	0.17
0011	$625 - 400 = 225$	0.17	$75 - 50 = 25$	0.5

Fisher Tests with Unequal Assignment Probabilities

Example II-9: Known assignment probabilities

We are interested in estimating the effect on cholesterol level of going to a step aerobics class twice a week for six months (treatment 1). Control ($W = 0$) is never attending step class. We observe eight individuals. For each individual, we observe their treatment received, their outcome under that treatment, and their gender.

We also know that treatment assignment was done using Bernoulli assignment where the probability of receiving treatment 1 (step class) depended on gender. For females, the probability of receiving treatment 1 was $\frac{3}{4}$, while for males the probability of receiving treatment 1 was $\frac{1}{4}$.

Unit	Gender	Prob of $W_i = 1$	Prob of $W_i = 0$	W	$y(0)$	$y(1)$
1	M	0.25	0.75	0	240	
2	M	0.25	0.75	0	310	
3	M	0.25	0.75	1		250
4	M	0.25	0.75	0	300	
5	F	0.75	0.25	1		180
6	F	0.75	0.25	1		220
7	F	0.75	0.25	0	250	
8	F	0.75	0.25	1		200

We can now carry out a Fisher test based on these assignment probabilities. We will restrict ourselves to only the $\binom{4}{1}\binom{4}{3} = 16$ assignments in which one male and three females receive treatment and three males and one female receive control.

W	$P(\mathbf{W})$	$P^*(\mathbf{W})$
10001110	0.011	0.0625
10001101	0.011	0.0625
10001011	0.011	0.0625
10000111	0.011	0.0625
01001110	0.011	0.0625
01001101	0.011	0.0625
01001011	0.011	0.0625
01000111	0.011	0.0625
00101110	0.011	0.0625
00101101	**0.011**	**0.0625**
00101011	0.011	0.0625
00100111	0.011	0.0625
00011110	0.011	0.0625
00011101	0.011	0.0625
00011011	0.011	0.0625
00010111	0.011	0.0625

As in previous examples, the column $P^*(W)$ is obtained by dividing $P(W)$ by the sum of the values $P(W)$ where W is such that one male and three females receive treatment, three males and one female receive control.

The observed assignment is in bold.

Fisher Test:

For each randomization, we calculate the test statistic for the data that would be observed under the null hypothesis of zero treatment effect. However, since men and women had different assignment probabilities, we shouldn't just look at the difference in means for treatment and control. A better way is to calculate the average treatment effect (difference in means) separately for men and for women and then take the average of these.

Calculating the observed treatment effect:

1. Calculate the difference in means for males: $\overline{y(1)}_M - \overline{y(0)}_M = \frac{250}{1} - \frac{240+310+300}{3} = -33.33.$
2. Calculate the difference in means for females: $\overline{y(1)_F} - \overline{y(0)_F} = \frac{180+220+200}{3} - \frac{250}{1} = -50.$
3. Calculate the mean of these two estimates: Average treatment effect = $\frac{-33.33-50}{2} = -41.67.$

We do the Fisher test the same as usual, but calculate the average treatment effects using the method shown above.

W	$P^*(\mathbf{W})$	Estimated Treatment Effect [a]	$\overline{y(1)} - \overline{y(0)}$
10001110	0.0625	-15.00	-62.50
*10001101	0.0625	-48.33	-80.00
10001011	0.0625	-28.33	-70.00
10000111	0.0625	-1.67	-52.50
01001110	0.0625	31.67	-25.00
01001101	0.0625	-1.67	-42.50
01001011	0.0625	18.33	-32.50
01000111	0.0625	45.00	-15.00
00101110	0.0625	-8.33	-45.00
***00101101**	**0.0625**	**-41.67**	**-62.50**
00101011	0.0625	-21.67	-52.50
00100111	0.0625	5.00	-35.00
00011110	0.0625	25.00	-25.00
00011101	0.0625	-8.33	-42.50
00011011	0.0625	11.67	-32.50
00010111	0.0625	38.33	-15.00
Mean		0.00	-43.13

[a] Calculated by taking the average of the male and female estimated effects, as described above

The observed test statistic is -41.67. There are two assignments (the starred ones) that give test statistics as extreme or more extreme than -41.67. We calculate the p-value by summing the probabilities of these two assignments:

p-value $= 0.0625 + 0.0625 = 0.125.$

The p-value is much larger than .05: the observed data do not provide strong evidence against the null hypothesis of zero treatment effect.

Also note that this average of the male and female treatment effects (the "Estimated Treatment Effect" column above) is unbiased for the average causal effect, under the null hypothesis (the average over all the possible randomizations is 0). However, the straight difference in treated and control means, $\overline{y(1)} - \overline{y(0)}$, is no longer an unbiased estimate of the treatment effect, even under the null hypothesis.

Example II-10: Bernoulli assignment where probability of treatment is known and depends on age

Prob of receiving treatment is $\frac{\text{age}_i}{\text{age}_i+10}$

Unit	Age	Prob of $W_i = 1$	Prob of $W_i = 0$
1	15	0.60	0.40
2	22	0.69	0.31
3	18	0.64	0.36
4	54	0.84	0.16
5	34	0.77	0.23
6	77	0.86	0.14

Treatment ($W = 1$) is a new surgery and control ($W = 0$) is the standard surgery. The outcome measured is years lived after surgery. We observe the following data:

Unit	Age	W	$Y(0)$	$Y(1)$
1	15	1		9
2	22	0	11	
3	18	0	2	
4	54	1		6
5	34	1		10
6	77	1		15

To determine whether the new surgery is significantly better than the standard one, we want to do a Fisher test of the null hypothesis of zero treatment effect. Since this is a Bernoulli experiment, there are $2^6 = 64$ possible assignments. For the Fisher test, we will restrict ourselves to only the $\binom{6}{4}$ = 15 assignments in which four units receive treatment and two units receive control.

W	Prob of $\mathbf{W}$	Prob* of $\mathbf{W}$
111100	0.004	0.01
111010	0.004	0.01
111001	0.01	0.03
110110	0.01	0.03
110101	0.02	0.08
110011	0.02	0.08
101110	0.01	0.02
101101	0.02	0.05
101011	0.02	0.05
100111	**0.04**	**0.13**
011110	0.01	0.03
011101	0.02	0.08
011011	0.02	0.08
010111	0.06	0.20
001111	0.04	0.13

Prob and Prob* of $\mathbf{W}$ are calculated as in previous examples. The bold row is the observed assignment.

Fisher Test:

Calculating the estimated treatment effect:

In Example II-9, we saw that we needed to average the male and female average causal effects to get an unbiased estimate of the overall average causal effect. More generally, when the assignment probabilities are not all equal, we need to calculate a weighted difference in means for the average treatment effect. The general formula for this weighted average is:

$$\text{Average Treatment Effect} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i(1)}{P(W_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i(0)}{1-P(W_i)}.$$

Horwitz and Thompson showed that each term in this quantity is an unbiased estimator of the average over all of the units.

For this problem we have:

$$\hat{ATE} = \frac{1}{6} * (\frac{9}{.60} + \frac{6}{.84} + \frac{10}{.77} + \frac{15}{.86}) - \frac{1}{6} * (\frac{11}{.31} + \frac{2}{.36}) = 1.92.$$

For each randomization, calculate the test statistic (using the above formula) for the data that would be observed under the null hypothesis of zero treatment effect:

W	Prob* of $\mathbf{W}$	Estimated Treatment Effect
111100	0.01	-18.2
111010	0.01	-16.3
* 111001	0.03	4.9
110110	0.03	-10.3
110101	0.08	1.1
* 110011	0.08	3.1
101110	0.02	-17.4
101101	0.05	-6.0
* 101011	0.05	4.1
*** 100111**	**0.13**	**1.9**
011110	0.03	-15.1
011101	0.08	-3.7
011011	0.08	-1.8
* 010111	0.20	4.2
001111	0.13	-2.9

The observed test statistic is 1.92. There are five assignments (the starred ones) that give test statistics as large or larger than 1.92. We calculate the p-value by summing the probabilities of these five assignments:

$$\text{p-value} = 0.03 + 0.08 + 0.05 + 0.13 + 0.20 = 0.49.$$

The p-value is rather large: the observed data do not provide strong evidence against the null hypothesis of zero treatment effect.

A 74% Fisher interval for the estimated treatment effect is (-4, 14). Since the probability of the observed assignment is 0.13, we cannot calculate an interval with coverage greater than 74% ($1 - .13 \times 2 = .74$).

Example II-11: Unknown assignment probabilities but assumed unconfounded

Now consider the same set-up as in Example II-9, but imagine that we do not in fact know the probabilities of assignment for males and females, although we do know that the assignment mechanism was unconfounded for each group. We observe that 1 of the 4 males and 3 of the 4 females went to the step class (received treatment 1). The data is repeated below:

Unit	Gender	W	$y(0)$	$y(1)$
1	M	0	240	
2	M	0	310	
3	M	1		250
4	M	0	300	
4	F	1		180
5	F	1		220
6	F	0	250	
7	F	1		200

By fixing the number of males and females in the experiment (4 males and 4 females), and then fixing the number that receive treatment 1 within the males and within the females, we can estimate the probability that each individual receives treatment 1, given their gender.

Since 1 out of 4 males went to the step class, we estimate that the probability of a male receiving treatment 1 is $\frac{1}{4}$. Since 3 of the 4 females went to the step class, we estimate the probability of a female receiving treatment 1 is $\frac{3}{4}$. We are then back in the situation of Example II-9, with "known" assignment probabilities. We then continue as in that example, and perform a Fisher test using these estimated probabilities.

In this situation, we may also want to run the Fisher test without conditioning on the number of males and females treated. Instead, we may just want to condition on the total number treated and the total number of control. We would thus look at all assignments where 4 people received treatment 1 (the step class) and 4 people received treatment 0 (no step class).

When calculating the average treatment effect, we cannot do the simple averaging of the male and female treatment effects as in Example II-9 since we no longer are conditioning on the number of males and the number of females that receive treatment 1. We thus use the Horwitz-Thompson formula introduced in Example II-10,

$$\text{Average Treatment Effect} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i(1)}{P(W_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i(0)}{1-P(W_i)}.$$

Using this formula to calculate the observed average treatment effect (data from Example II-9), we have the following:

$$\hat{ATE} = \frac{1}{8} * (\frac{250}{0.25} + \frac{180}{0.75} + \frac{220}{0.75} + \frac{200}{0.75}) - \frac{1}{8} * (\frac{240}{0.75} + \frac{310}{0.75} + \frac{300}{0.75} + \frac{250}{0.25} = -41.67$$

Note that this is the same as that calculated in Example II-9, by averaging the male and female estimated treatment effects.

There are now $\binom{8}{4} = 70$ possible assignments. The table below shows some of these, as well as the estimated treatment effects, under the null hypothesis of zero treatment effect.

W	Prob of **W**	Estimated treatment effect
11110000	0.000015	125.0
11001100	0.001236	25.0
11000011	0.001236	58.3
10101010	0.001236	5.0
00111100	0.001236	25.0
00110011	0.001236	58.3
01010101	0.001236	78.3
00001111	0.100113	-41.67

Over all 70 possible assignments, the average of the estimated treatment effects as calculated above (weighted by the Probability of W) would equal 0 since it is unbiased under the null hypothesis. However, if we took the weighted average of the 70 values of $\overline{y(1)} - \overline{y(0)}$, it would equal -15.625, again showing that it is not unbiased when there are unequal probabilities of receiving treatment.

Example II-12: Estimating unknown assignment probabilities

We are now interested in determining which of two types of surgery is better in terms of leading to increased life. The control is a standard surgery and treatment is a new surgery that has just come out. Ethically it is difficult to do a randomized experiment, since the new surgery is thought to be better in general. We thus decide to do an observational study, using data on patient outcomes that hospitals have collected. We have data on 2000 patients, where half of received the standard surgery (treatment 0) and half received the new surgery (treatment 1). The hospital data sets also have information on the patient's ages and cholesterol levels before surgery.

To do a Fisher test on this data, assuming an unconfounded assignment mechanism, we need to estimate the probabilities of treatment assignment for each individual. We could do this in a number of ways.

We first might estimate the probability of receiving treatment 1 as $\frac{1}{2}$ since we observe that 1000 of the 2000 patients received the new treatment.

However, since treatment assignment may vary based on age we would like to include this in the probabilities. We assemble the following data, showing the number of treated and control patients in a variety of age ranges. By dividing the number of treated in each age range by the total number of people in that range, we can estimate the probability of receiving treatment for people in each of the age ranges.

Age range	Total number	Number Treated	Number Control	Estimated Treatment Probability
0-20	137	94	43	0.69
20-40	455	276	179	0.61
40-60	790	393	397	0.50
60-80	479	193	286	0.28
80-100	118	31	87	0.26

We see that the probability of receiving treatment seems to be lower for older patients. Older patients are more likely to receive control, while younger patients are more likely to receive the new treatment.

We could do the same thing for cholesterol levels:

Cholesterol range	Total number	Number treated	Number control	Estimated Treatment Probability
0-200	175	155	20	0.89
200-250	475	354	121	0.75
250-300	704	343	361	0.49
300-350	464	130	334	0.28
350-400	162	16	146	0.10

Here we see that people with low cholesterol are more likely to receive the new treatment, while people with high cholesterol are more likely to receive the old treatment.

We could use any of these estimated probabilities (0.5 for everyone, estimated using age, or estimated using cholesterol level) in a Fisher test, similar to Example II-11. However, since we have two covariates we would like to combine these and use the information in them at the same time. We can do this by looking in a 2 way table with both age and cholesterol level. Each cell shows the number of treated individuals over the total number in that cell, and then the estimated probability.

	0-20	20-40	40-60	60-80	80-100
0-200	11/11 1.00	32/38 0.84	32/49 0.65	17/29 0.59	2/7 0.29
200-250	57/61 0.93	100/119 0.84	75/141 0.53	40/103 0.39	4/25 0.16
250-300	48/57 0.84	145/191 0.76	148/293 0.51	43/177 0.24	7/67 0.10
300-350	28/33 0.85	63/98 0.64	72/172 0.42	28/125 0.22	2/46 0.04
350-400	9/10 0.90	8/22 0.36	11/43 0.26	2/28 0.07	1/13 0.08

Within each “cell” of the table defined by age and cholesterol level, we could treat this as a little randomized experiment, where everyone in that “cell” had the same probability of receiving treatment.

Once we have estimated each individual’s probability of receiving the treatment, we could use the Fisher or Neyman methods already discussed.

Although the estimation of the probabilities is fairly straightforward for this example, it would get very complicated if there were more covariates available. Statistical methods have been developed that model the probability of receiving treatment, given the covariates. One common example is called logistic regression, which essentially “smooths” over the covariate values so people in similar age groups have similar probabilities. More examples of this follow.

Example II-13: Reconstructing hypothetical randomized experiments through subclassification

From Rubin, D.B. (1997). "Estimating Causal Effects from Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*, 127: 757-763. Adapted from Cochran, W.G. (1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics* 24: 295-313.

One of the key benefits of a randomized experiment is the implied balance of all of the background covariates between the treated and control groups. A well designed observational study will also have this feature with respect to the observed covariates, with causal effects being estimated by comparing treated and control units with the same distribution of observed covariates. Here we give an example of how to create a series of hypothetical randomized experiments, using observational data.

The example used is a study of smoking and mortality. The table below shows mortality rates per 1000 person-years for nonsmokers, cigarette smokers, and pipe or cigar smokers, from three large datasets from the United States, United Kingdom, and Canada.

Comparison of Mortality Rates for Three Smoking Groups in Three Databases

	Canada			United Kingdom			United States		
	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers
Mortality rate	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4

These unadjusted mortality rates make it appear that cigarette smoking is good for health, compared to cigar or pipe smoking. In all three data sets, the mortality rates are similar for nonsmokers and cigarette smokers, and higher for cigar or pipe smokers.

An explanation for this surprising result can be found by looking at the average age of the people in each of the smoking categories. The table below shows these ages. Nonsmokers and cigarette smokers tend to be younger than cigar or pipe smokers. Age is highly related to the decision to smoke; in particular, older individuals have a higher probability of making the decision to start smoking cigars or pipes.

Thus, rather than lumping all of the ages together (and thus people with varying probabilities of being in each of the smoking groups), we would like to compare mortality rates among individuals with similar ages and similar probabilities of being nonsmokers versus cigarette smokers versus cigar or pipe smokers. To do this, the population is grouped into age categories of approximately equal size (in this case, equal size based on the number of nonsmokers). Mortality rates are compared within the age categories and an overall result is found by averaging over the age group comparisons.

This can be thought of as trying to recreate mini-randomized experiments within age groups. If we believe that the decision to smoke cigars or pipes is unconfounded given age, people of similar ages will have similar probabilities of smoking cigar or pipe, thus in essence recreating a mini-randomized experiment. For example, people age 20-30 may all have probability 0.2 of smoking a cigar or pipe, and conditional on age, it is random as to who does smoke cigars or pipes and who doesn't. Older individuals, say 50-60, may have a higher probability of smoking cigars or pipes (say 0.5), but again, within this age range, we assume that those who smoke cigars or pipe are only randomly different from those who do not.

The results of this are shown in the table below. We see that the results are much closer to what we would expect, with smokers in general having higher age-adjusted mortality than either nonsmokers or cigar or pipe smokers.

	Canada			United Kingdom			United States		
	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers	Non-Smokers	Cigarette Smokers	Cigar/Pipe Smokers
Mortality rate	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
Average age	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
Adj. mortality rates,									
2 subclasses	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3 subclasses	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9-11 subclasses	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

Cochran (1968) also gives theoretical results for subclassification. He shows that as long as there is reasonable overlap in the distributions of age in the treated and control groups, then subclassification using 5 or 6 subclasses can remove 90% or more of the initial bias due to age.

This example is in terms of a single covariate. Theoretical results have shown that balance of the propensity score (probability of receiving treatment) between the treated and control groups implies balance of all of the covariates that went into the propensity score estimation. Thus, by subclassifying on the propensity score, we can obtain these results (balance between the groups in these mini-randomized experiments) for all of the observed covariates. Within a group of individuals with similar values of the propensity score, we can treat the data as arising from a randomized experiment, assuming we accept that the assignment mechanism is unconfounded given the observed covariates.

A second key feature of a randomized experiment is that the outcome data is not used in the design (randomization phase). We also replicate that here, by forming subclasses using only covariate values. The subclasses are defined and analysis set up without even seeing the outcome variable, although for pedagogical purposes we show the results on the outcome variable here.

Estimating Unknown Propensities: Case Studies on Propensity Scores and Matching

Example II-14: National Supported Work Demonstration

The National Supported Work Demonstration was a program run by the US Government during the 1970's. It was designed to help move disadvantaged workers into the labor market by providing them with work experience and counseling. In order to evaluate the program, applicants were assigned to the program randomly. Baseline measures were obtained on all applicants, and both treatment and control group members were followed for up to four years. However, only the treatment group members received the benefits of the program.

The results of this program have been analyzed in several ways. Since it was a randomized experiment, a good estimate of the "true" treatment effect is available. However, as a way to illustrate methods for dealing with observational studies, this data set has also been treated as an observational study, essentially ignoring the control group data. In those cases, a comparison group was found using large national data sets already available. For more information on these analyses, see Lalonde ("Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *The American Economic Review*, September 1986), or Dehejia and Wahba ("Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, December 1999).

Lalonde used standard econometric modeling methods to estimate the treatment effect, and found that the results were very sensitive to the model specification, and that in general they did not replicate the results from the randomized experiment. This is likely due to the fact that most of the individuals in the large national data sets are dramatically different from those in the randomized experiment (Lalonde chose a comparison group from the national databases on the basis of just 1 covariate). Dehejia and Wahba attempted to replicate the randomized experiment results using propensity score and matching methods, which utilize and balance all observed covariates. They had greater success than Lalonde. Here we will summarize their methods and results using one of the large national data sets, the Panel Survey of Income Dynamics (PSID).

The following table summarizes the covariates in the (randomized) treated group and the full (Lalonde) PSID comparison group, composed of all male household heads in the PSID under age 55 who did not classify themselves as retired in 1975.

Covariate	Control Group	Treated Group	PSID Comparison Group
Age	25.05	25.82	34.85*
Education	10.09	10.35	12.12*
Black	0.83	0.84	0.25*
Hispanic	0.1	0.06	0.03
No Degree	0.83	0.71	0.31
Married	0.15	0.19	0.87*
1974 Income	2,107	2,096	19,429*
1975 Income	1,267	1,532	19,063*
Sample Size	260	185	2,490

We see that the treated and control groups are very similar, but that the treated group and the PSID comparison group are actually very different. (Variables that are marked with a * are "significantly" different from each other in the treated group and the PSID comparison group).

To form a better comparison group, propensity scores were estimated and then the treated group members were matched to individuals in the PSID on the basis of their propensity scores (defined as the probability of receiving treatment given the observed covariates). Thus, only comparison group members who looked like the treated group were kept. The following shows the covariate means for the treated group and the new matched comparison group.

Covariate	Treated Group	Matched PSID Group
Age	25.82	26.39
Education	10.35	10.62
Black	0.84	0.86
Hispanic	0.06	0.02
No Degree	0.71	0.55
Married	0.19	0.15
1974 Income	2,096	1,794
1975 Income	1,532	1,126
Sample Size	185	156

The treated and matched comparison groups are now very similar to each other. None of the variables are "significantly" different between these two groups.

By comparing the estimated effects with the effect calculated using the true treated and control groups from the randomized experiment, we also see that this matching improved the estimation of the average treatment effect.

"True" Treatment vs. Control Effect (Standard Error): 1,794 (633)
Estimated Treatment Effect Using Full PSID Sample: -15,205 (1,154)
Estimated Treatment Effect Using Matched PSID Sample: 1,691 (2,209)

Estimating Propensity Scores

Here we give some details on propensity scores, including some of the theory, and an example of how to estimate propensity scores in practice.

Rosenbaum and Rubin (1983) introduced the propensity score as a way to control for all of the observed covariates through one scalar number. The propensity score is defined as the probability of receiving treatment given the observed covariates. It is a balancing score, which means that at each value of the propensity score, the distributions of the covariates (that went into the propensity score specification) in the treated and control groups are the same (the covariates are "balanced"). This implies that within a small range of values of the propensity score, the observations can be thought of as arising from a mini-randomized experiment. In groups with similar propensity scores, each individual will have a similar probability of receiving treatment. This is similar to what we did in the previous example, but instead of subclassifying just on age, we subclassify on a function (the propensity score) of all of the observed covariates. Treatment assignment is assumed to be ignorable given the observed covariates. Later we will discuss methods to assess sensitivity to this assumption.

Formally, let $e(X_i)$ be the probability of individual i being assigned to treatment given covariates X_i: $e(X_i) = P(W_i = 1|X_i)$. Rosenbaum and Rubin (1983) showed that if treatment assignment is independent of the potential outcomes given X, then treatment assignment is independent of the potential outcomes given $e(X_i)$. In other words, we can subclassify or match using just the propensity score rather than all of the covariates. Observations with the same value of the propensity score will have the same distribution of all of the covariates that went into the propensity score specification.

Usually we do not actually know the propensity scores, and so we estimate them. Propensity scores can be estimated in a number of different ways, including discriminant or CART analysis. One of the most popular (and easiest) is logistic regression. Logistic regression is used to model a binary outcome as a function of covariates and can be implemented using software such as SAS, Stata, or S-Plus. The response variable in the logistic regression is treatment received (W) and the observed covariates (possibly including squares and interactions) are used as predictors. The model for logistic regression is:

$$e(X_i) = P(W_i = 1|X_i) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}$$

Matching is often done on the linear propensity score, $\log(\frac{e(X_i)}{1-e(X_i)}) = X'\beta$.

The following procedure summarizes an implementation method. (Note that this is just one possible suggestion; the details are open for interpretation).

1. Start with model (for example, logistic regression with treatment received as the response variable) with main effects for each of the observed covariates.
2. (Optional) Discard control units outside the range of the treated group propensity scores, and/or treated units outside the range of the control group propensity scores.
3. Form 1 block (with propensity scores in the range 0-1), do a t-test of $\hat{e}$ between the treated and control groups. If significant, split into 2 blocks at the median. Continue this process, splitting a block if it has a t-statistic greater than 2 and if there are more than 2 treated and control units in each new block formed.
4. Within each block formed in Step 3, test for equality of means of functions of X (e.g., each covariate, each covariate squared, 2-way interactions of covariates). If any t-statistic is greater than 2.5 in any block, include that term in the new propensity score specification. (Other, similar rules can be used, such as including terms that are significant in more than 1 block).
5. Repeat Steps 1-4 until there are no more (or very few, as few as possible) significant t-statistics. This will imply that within each block the treated and control groups are very well balanced.

Once the propensity score has been estimated, treated and control units can be matched, or subclassified using the propensity score. Analysis can then continue as if the data in each block arose from a randomized experiment.

The balancing property of the propensity score can be used to assess its specification. The main goal is to choose samples of treated and control units with similar distributions of the covariates. Thus the success of the estimation can be easily checked, as we will show below. Note again that at no point is the outcome used! The propensity score is estimated, and assessed, without the outcome variable.

The next few pages show how this method could have been implemented in the previous example, estimating the effects of the National Supported Work Demonstration. The method outlined here is slightly different from that implemented by Dehejia and Wahba, but the spirit is the same.

First, a propensity score was estimated with only main effects (treatment received as the response variable, observed covariates as predictors). Blocks were formed as outlined in Step 3 above. This resulted in 7 blocks, as summarized below:

Lower and Upper Block Boundaries: Specification 1

1 block	2 blocks	3 blocks	5 blocks	6 blocks	7 blocks
0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.77	0.39	0.25	0.09	0.03
	1.00	0.77	0.39	0.25	0.09
		1.00	0.56	0.39	0.25
			0.77	0.56	0.39
			1.00	0.77	0.56
				1.00	0.77
					1.00

The following table shows the corresponding t-statistics for within block average propensity score differences.

T-statistics of propensity score: Specification 1

1 block	2 blocks	3 blocks	5 blocks	6 blocks	7 blocks
25.7	14.5	8.6	4.0	3.9	1.3
	-0.2	2.3	1.3	-0.2	0.4
		-0.2	1.2	1.3	-0.2
			0.9	1.2	1.3
			-0.2	0.9	1.2
				-0.2	0.9
					-0.2

The following table summarizes the results for each block, with specification 1:

Block Results: Specification 1

Block	Lower Bound	Upper Bound	Number Obs	Number Controls	Number Trainees	Mean Controls	Mean Trainees	T-stat Diff
Discard	0.00	0.00	1236	1236	0	0.03	-	-
Block 1	0.00	0.03	929	923	6	0.01	0.01	1.31
Block 2	0.03	0.09	143	137	6	0.05	0.05	0.41
Block 3	0.09	0.25	127	116	11	0.15	0.15	-0.23
Block 4	0.25	0.39	64	41	23	0.31	0.32	1.35
Block 5	0.39	0.56	43	20	23	0.45	0.47	1.19
Block 6	0.57	0.77	30	8	22	0.65	0.67	0.93
Block 7	0.78	0.99	100	9	91	0.91	0.91	-0.17
Discard	0.99	1.00	3	0	3	-	0.65	-

We see that a large number of the PSID individuals were discarded since they had a propensity score lower than that for the lowest treated individual. These PSID individuals are incomparable to anyone in the treated group. Overall, the PSID does not form a valid comparison group for the treated individuals, however a subset of the PSID individuals do look similar to those in the treated group.

The following table shows t-statistics for the covariates, squares, and 2-way interactions of the observed covariates. This is used as a diagnostic, to determine which terms should be added to the propensity score specification. T-statistics less than 2.5 imply that the block is well balanced on that variable.

Block T-statistics: Specifi cation 1

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7
Age	-0.5	1.3	-0.9	0.3	0.9	-0.3	-0.1
Hispanic	-0.5	-0.9	1.2	-0.5	1.4	0.9	-0.9
Black	2.5	-1.7	-0.1	-1.1	-0.5	0.7	0.9
Education	-0.4	0.5	0.0	1.6	-0.8	-0.8	-1.3
Earn ’74	1.3	-1.8	0.6	-0.3	-1.0	-0.4	0.3
Earn ’75	1.1	-2.6	0.7	-2.0	0.5	-0.9	0.6
Unemp ’74	-0.8	1.3	-1.5	1.1	1.1	-1.2	-0.3
Unemp ’75	-1.0	1.3	-1.4	1.0	0.6	0.2	0.6
AgexAge	-0.7	1.2	-0.9	0.1	0.8	-0.2	0.1
AgexHisp	-0.4	-0.8	0.3	-0.8	1.4	0.8	-0.8
AgexBlack	2.1	-1.1	-0.1	-0.5	0.2	0.3	0.5
AgexEduc	-0.4	-0.8	1.1	-0.5	1.4	0.9	-1.5
HispxEduc	-0.4	-0.8	1.1	-0.5	1.4	0.9	-1.5
BlackxEduc	2.6	-1.7	0.4	-0.3	-0.3	0.6	0.2
EducxEduc	-0.6	0.4	-0.4	1.6	-1.0	-1.0	-1.4
AgexEarn’74	1.0	-1.5	0.7	-0.2	-0.9	-0.4	0.3
HispxEarn’74	-0.4	-0.7	0.8	-0.3	1.0	0.6	0.0
BlackxEarn’74	3.2	-1.8	0.1	-0.3	-1.5	-0.4	0.3
EducxEarn’74	1.1	-1.6	0.5	-0.3	-1.0	-0.4	0.3
Earn’74xEarn’74	1.2	-1.4	0.8	-0.2	-0.8	-0.9	0.1
AgexEarn’75	1.3	-2.5	0.9	-2.0	0.8	-0.7	0.6
HispxEarn’75	-0.4	-0.7	2.9	0.3	1.1	0.5	-2.3
BlackxEarn’75	2.5	-1.9	-0.2	-1.7	-0.0	-0.9	0.6
EducxEarn’75	0.7	-2.4	0.8	-2.2	0.1	-1.1	0.7
Earn’74xEarn’75	1.2	-2.2	0.5	-0.6	-1.3	0.3	0.1
Earn’75xEarn’75	1.1	-2.1	0.1	-2.3	1.2	-1.5	0.7
AgexUnemp’74	-0.8	1.3	-1.3	0.7	1.2	-0.8	-0.2
HispxUnemp’74	-0.0	-0.3	-0.7	-1.0	1.0	0.6	-0.9
BlackxUnemp’74	0.0	-0.2	-0.1	0.2	1.2	-0.4	0.7
EducxUnemp’74	-0.8	1.4	-1.3	1.7	0.6	-1.6	-1.3
Earn’75xUnemp’74	-0.3	-0.3	-0.6	-1.7	1.1	-1.5	0.6
AgexUnemp’75	-1.0	1.2	-1.3	0.7	0.7	-0.0	0.3
HispxUnemp’75	-0.1	-0.3	-0.8	-1.0	-0.0	0.6	0.6
BlackxUnemp’75	-0.2	-0.3	-0.2	-0.0	0.4	-0.1	0.3
EducxUnemp’75	-1.0	1.4	-1.2	1.7	0.7	0.1	0.0
Earn’74xUnemp’75	-0.5	-0.2	-0.8	-1.2	0.0	0.0	0.0
Unemp’74xUnemp’75	-0.7	1.4	-1.2	1.6	1.1	0.2	0.6

We see that Black x Education, Black x 1974 Earnings, Age x 1975 Earnings, Hispanic x 1975 Earnings, and Black x 1975 Earnings all have a t-statistic greater than 2.5 in at least 1 block. We thus include these terms in a new propensity score specification. The same blocking procedure is followed, again resulting in 7 blocks. These are summarized below.

Block Results: Specification 2

Block	Lower Bound	Upper Bound	Number Obs	Number Controls	Number Trainees	Mean Controls	Mean Trainees	T-stat Diff
Discard	0.00	0.00	1428	1428	0	0.03	-	-
Block 1	0.00	0.02	702	696	6	0.01	0.01	1.42
Block 2	0.02	0.10	204	198	6	0.05	0.06	1.11
Block 3	0.10	0.23	102	91	11	0.15	0.16	0.28
Block 4	0.23	0.39	67	44	23	0.31	0.31	0.66
Block 5	0.40	0.59	39	16	23	0.47	0.49	1.27
Block 6	0.60	0.80	30	8	22	0.67	0.71	1.89
Block 7	0.81	0.99	100	9	91	0.92	0.92	-0.18
Discard	0.99	1.00	3	0	3	-	0.65	-

We again check the balance of all of the covariates, squares, and interactions within each block. These t-statistics are summarized below.

Block T-statistics: Specifi cation 2

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7
Age	-0.5	-0.4	-1.8	1.7	0.8	0.4	-0.2
Hispanic	-0.6	-0.9	1.2	0.3	0.0	1.3	-1.1
Black	0.9	-0.1	-0.6	-0.6	-1.0	0.5	0.9
Education	0.0	1.0	0.4	-0.3	0.2	-2.0	-1.2
Earn ’74	-0.6	0.5	0.9	-1.2	-0.2	-0.5	0.3
Earn ’75	0.3	-1.0	0.8	-1.4	-0.8	-0.3	0.6
Unemp ’74	0.5	-0.6	-1.8	2.0	1.4	-0.8	-0.3
Unemp ’75	0.1	-0.6	-1.6	1.4	0.6	0.4	0.6
AgexAge	-0.7	-0.4	-1.8	1.8	0.7	0.5	0.0
AgexHisp	-0.5	-0.9	0.2	-0.3	-0.0	1.3	-1.1
AgexBlack	0.6	-0.0	-0.9	0.9	-0.2	0.5	0.5
AgexEduc	-0.2	0.6	-0.9	1.6	0.9	-0.8	-0.9
HispxEduc	-0.5	-0.9	1.8	0.4	-0.0	1.3	-1.9
BlackxEduc	0.7	0.2	-0.6	-0.6	-0.8	0.2	0.3
EducxEduc	-0.2	0.8	0.1	-0.2	0.1	-2.2	-1.2
AgexEarn’74	-0.7	0.1	0.9	-1.3	-0.1	-0.6	0.3
HispxEarn’74	-0.5	-0.8	0.2	1.3	0.0	0.6	-0.0
BlackxEarn’74	0.5	0.8	0.7	-2.1	-0.2	-0.6	0.3
EducxEarn’74	-0.6	1.6	0.6	-1.2	-0.2	-0.6	0.3
Earn’74xEarn’74	-0.7	1.4	0.9	-0.9	0.2	-1.0	0.0
AgexEarn’75	-0.0	-0.8	0.6	-1.8	-1.0	0.1	0.7
HispxEarn’75	-0.5	-0.7	1.3	0.9	-0.0	0.6	-1.7
BlackxEarn’75	0.8	-0.4	0.1	-1.7	-1.4	-0.6	0.6
EducxEarn’75	0.1	-0.5	0.9	-1.8	-1.1	-0.3	0.7
Earn’74xEarn’75	-0.3	0.1	0.9	-0.9	-0.1	-0.1	0.0
Earn’75xEarn’75	0.2	-0.6	0.1	-1.1	-1.1	-0.5	0.7
AgexUnemp’74	-0.3	-0.6	-1.8	2.1	1.2	-0.3	-0.3
HispxUnemp’74	-0.1	-0.2	-0.6	-1.4	0.0	1.1	-1.1
BlackxUnemp’74	0.0	-0.3	-1.3	2.2	0.7	-0.1	0.7
EducxUnemp’74	0.6	-0.5	-1.4	1.9	1.4	-1.5	-1.2
Earn’75xUnemp’74	-0.4	-0.4	-0.6	-1.1	-0.4	-0.4	0.6
AgexUnemp’75	0.1	-0.6	-1.8	1.8	1.0	0.3	0.3
HispxUnemp’75	-0.2	-0.3	-0.7	-1.4	-0.0	0.9	0.6
BlackxUnemp’75	-0.2	-0.4	-1.1	1.3	0.5	0.1	0.3
EducxUnemp’75	0.2	-0.5	-1.3	1.5	0.9	-0.0	0.1
Earn’74xUnemp’75	-0.5	-0.4	-0.8	-1.6	-0.1	-0.0	0.0
Unemp’74xUnemp’75	0.8	-0.4	-1.5	2.4	1.3	0.4	0.6

There are now no terms with t-statistics greater than 2.5. The blocks are well balanced on the observed covariates. An analysis of the outcome can now be done. This can be done in the full matched groups, or within subclasses. Doing the analysis within subclasses is useful if the overall groups are still not well matched. We see that within subclasses the groups are very similar to each other, and methods for randomized experiments can be used. The subclasses could be defined as above, or by quantiles of the propensity score in the treated group, control group, or overall. An overall effect is estimated using a weighted average of the within subclass estimates.

Example II-14: The GAO Breast Conservation Versus Mastectomy Study

The following information is from "Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies," General Accounting Office Document GAO/PEMD-95-9, November 1994.

We are interested in determining the survival rates of breast cancer patients who receive breast conservation (lumpectomy, nodal dissection, and radiation) versus mastectomy. Summarizes results from two types of studies: randomized experiments and an observational study.

1. Randomized Experiments

   - "Gold standard" of medical research.
     - Assignment mechanism is unconfounded. In an observational study, the patients who choose one treatment may be very different from the patients who choose the other treatment. Since treatments are assigned randomly in a randomized experiment, the effects of this are minimized.
     - In some randomized experiments, blinding/double blinding can be used so patients (and possibly doctors) do not know which treatment they are receiving.
   - Day-to-day practice may be different from that in randomized trials.
     - Randomized trials typically are in large, university hospitals (not many "community physicians").
     - Physicians must follow pre-specified procedures in randomized experiments.
     - Patients (and doctors) have to be willing to be randomized.
   - Randomized experiments often use patients and doctors who are not typical of the population of those to be treated, and in obvious ways.
   - Six randomized experiments done around the world.

- Results:

Study	5 year survival rates Breast Conservation	Mastectomy	Difference in rates (Cons-Mast)
US-1	93.9% (n=74)	94.7% (n=67)	−0.8%
Milan	93.5% (n=257)	93.0% (n=263)	0.5%
French	94.9% (n=59)	95.2% (n=62)	−0.3%
Danish	87.4% (n=289)	85.9% (n=288)	1.5%
EORTC	89.0% (n=238)	90.0% (n=237)	−1.0%
US-2	89.0% (n=330)	88.0% (n=309)	1.0%

- Meta-analysis: combine results of the above six studies
  - No statistically significant differences in survival rates

Study	5 year survival rates Breast Conservation	Mastectomy	Difference in rates (Cons-Mast)
Single Center: US-1, Milan, French	93.7%	93.7%	0.0%
Multicenter: Danish, EORTC, US-2	89.0%	88.0%	1.0%
All Six Studies	90.0%	90.0%	0.0%

2. Observational Study: SEER data base

- Goal: To compare outcomes in day-to-day medical practice with results from randomized experiments.
- SEER database
  - National Cancer Institute's Surveillance, Epidemiology, and End Results database.
  - Records for almost all cancer patients in five states (CT, HI, IA, NM, UT) and four metropolitan areas (Atlanta, Detroit, San Francisco-Oakland, Seattle-Puget Sound).
  - Use years 1983-1985 so 5 years follow-up available on all patients.
- Choose patients from SEER who could have been in randomized experiments (similar based on year of treatment, geographic area, tumor size, age, marital status, race or ethnicity).
- Propensity scores estimate the probability of each individual receiving breast conservation based on the covariates.

  In general, young, white, married women, with small tumors, living in San Francisco, Hawaii or Seattle, who were diagnosed late in the time period were more likely to choose breast conservation.

  For example, a woman in her 60's living in Iowa, diagnosed in 1983 was unlikely to receive breast conservation so her propensity score is small.

  A woman under 40, non-Asian, living in San Francisco-Oakland or Seattle-Puget Sound, diagnosed in 1985 with a very small tumor would have a relatively high propensity score.

  However, 2 women with seemingly very different characteristics may have similar probabilities of receiving breast conservation.
  - Woman 1: Asian, divorced woman aged 35 with a large tumor, living in Seattle.
  - Woman 2: White, widowed woman aged 65 with a small tumor, living in Iowa.
  - These two women may have similar probabilities of choosing breast conservation.
- Split all eligible patients in SEER (5,326 women) into five blocks based on their estimated probability of receiving breast conservation (eligible defined according to eligibility for the randomized experiments).
  - Within each block, breast conservation and mastectomy patients had similar values of all of the covariates on average.
  - Consider to be completely randomized within each block. In other words, given the blocking based on these covariates (through the propensity score), treatment assignment is random.

- Results:

Block	Treatment	Number	5 year Survival rate	Difference	Std. Error of Difference
1	Breast Conservation	56	85.6%	−1.1%	4.8%
	Mastectomy	1008	86.7%		
2	Breast Conservation	106	82.8%	−0.6%	3.9%
	Mastectomy	964	83.4%		
3	Breast Conservation	193	85.2%	−3.6%	2.8%
	Mastectomy	866	88.8%		
4	Breast Conservation	289	88.7%	1.4%	2.2%
	Mastectomy	778	87.3%		
5	Breast Conservation	462	89.0%	0.5%	1.9%
	Mastectomy	604	88.5%		
Overall	Breast Conservation	1106	86.3%	−0.6%	1.5%
	Mastectomy	4220	86.9%		

- Overall estimate found by averaging the five blocks.
- Similar results found as in randomized trials. Breast conservation therapy seems, on average, to be similarly effective to mastectomy in day-to-day medical practice.
- Note that on average, survival rates for both therapies in the observational study lower than survival rates in the randomized experiments.
- Note trend in signs.

Example II-15: Matching patients in a randomized trial to historical patients

Fabry disease is a rare disease for which there is currently no commercially available treatment. Double blind clinical trials are currently underway to assess the efficacy of a new drug, Fabrazyme. During the course of the trial it has become apparent that the drug appears to be effective. There is thus desire to approve the drug before the end of the trial, allowing those in the trial (including the placebo control group) to obtain Fabrazyme. However, this would invalidate the traditional use of the placebo control group as a long-term comparison group for those randomized to Fabrazyme.

Longitudinal historical data is also available on approximately 417 patients with Fabry disease. The goal of this part of the analysis was to choose a subset of this historical patient group that looked similar to patients in the randomized trial. This consisted of a series of steps which are briefly outlined here:

- Consider only versions (time points) of historical patients that would have met the eligibility criteria of the randomized trial.
- For each historical patient, choose their version that looks the most like a randomized patient ("most like" defined by their probability of being in the randomized group and by a function of age and their baseline measurement of the outcome variable).
- Re-estimate the probability of being in the randomized group among the selected historical patients and the randomized group, and then discard historical patients that look dissimilar to those in the randomized trial.

The results of this process are summarized in the table below:

	Randomized Group Mean (Standard Deviation)	Standardized Bias[a] – All Versions of All Historical Patients	Standardized Bias[a] – Selected Versions of Historical Patients	Standardized Bias[a] – Plausible Historical Patients	Standardized Bias[a] – Chosen Matched Historical Patients	Standardized Bias [a] Chosen Matched Historical Patients (Subclass Estimates)
N	73	395	100	91	86	86
Sex (Female)	8.2%	4.9%	3.2%	6.0%	5.9%	0.6%
Ethnicity (Caucasian)	89.0%	0.9%	3.0%	3.3%	1.8%	0.2%
Blood Group {B+, B-, AB+, AB-}	10.3%	0.0%	−3.9%	−3.7%	−1.1%	−5.7%
Ace Inhibitor Use	23.3%	4.0%	4.3%	2.4%	3.5%	-6.8%
Age	45.1 (8.68)	0.56	0.86	0.71	0.64	0.10
Weight	70.09 (12.23)	-0.29	-0.17	-0.12	-0.08	-0.05
Height	173.43 (8.13)	-0.21	-0.06	0.01	0.01	NA [b]
Serum Creatinine mg/dL	1.65 (0.52)	0.10	0.41	0.34	0.31	-0.10
Estimated GFR	52.95(17.91)	−0.29	−0.62	−0.52	−0.48	0.09
Estimated creatinine clearance	59.67 (17.54)	-0.65	-0.96	-0.79	-0.68	0.04
Plasma α-GAL	1.02(0.59)	0.36	0.41	0.43	0.37	−0.04
Linear Propensity Score [c]	0 (1)	0.73	0.86	0.99	0.88	0.10

[a] For binary variables (Sex, Ethnicity, Blood Group, Ace Inhibitor), the standardized bias is defined as the difference in proportions. For the remainder of the variables, the standardized bias is defined as the difference in means divided by the standard deviation in the randomized group.

[b] Height missing for all chosen matched historical patients in Block 4.

[c] Propensity score defined with respect to the column

More details on estimating probabilities

- Discrete covariates: Estimate probabilities within cells defined by covariates
- A few continuous covariates: Ad-hoc coarsening to estimate probabilities within cells defined by the covariates (e.g. Example II-12).
- More complicated situations (more covariates):
  - Discriminant analysis: Fits 2 distributions to treated and control groups, compares relative heights of densities at covariate values to estimate probabilities.
  - Logistic or probit regression: Iterative procedure that fits linear regression model to transformation of the probabilities.
- Software available to estimate probabilities using these more complicated methods (e.g. S-Plus, SAS, Stata, SPSS)
- Key point: Real issue is to assess overlap in probabilities and balance of the underlying covariates. This overlap can be easily checked.

Part III: Predictive Inference

The methods below are not formally exactly correct, but convey the essential ideas.

1. Getting started–ignorable treatment assignment with just 1 or 2 covariates

Example III-1: Using donor pools to fill in missing potential outcomes–discrete covariate

Consider again an example of a study attempting to estimate the effect of a new surgery on years lived after the surgery. A completely randomized experiment was done, with 10 patients receiving the new surgery and 10 receiving the old surgery.

There is one covariate available for each individual: gender.

Instead of running a Fisher test or using Neyman's method of estimation, we can use the observed values to predict the missing potential outcomes and fill in ("impute") the missing values. Once these missing values are filled in, it is straightforward to calculate the average treatment effect (corresponding to that imputation), average treatment effect within subclasses defined by the covariates, or even values such as the average difference in squared potential outcomes.

The way we will first do this is through the use of "donor pools": individuals in the other treatment group with similar covariate values.

The following are the observed data on the 20 individuals. The outcome is years lived after surgery.

Unit	Gender	W	Y(0)	Y(1)
1	M	1		12
2	M	1		9
3	M	1		9
4	M	1		7
5	M	1		8
6	F	1		12
7	F	1		11
8	F	1		10
9	F	1		14
10	F	1		12
11	M	0	6	
12	M	0	8	
13	M	0	7	
14	M	0	11	
15	M	0	11	
16	F	0	5	
17	F	0	7	
18	F	0	6	
19	F	0	8	
20	F	0	10	

Because this was a completely randomized experiment, we would expect approximately 5 males (out of 10 total) and 5 females (out of 10 total) to be in each treatment group. We see that this is exactly what we have obtained.

For each unit, we create donor pools of people of the same gender. For each male in the treated group, his donor pool will consist of all males in the control group. For each female in the treated group, her donor pool will consist of all females in the control group. Similarly, the donor pool for each male or female in the control group will consist of all males or females (respectively) in the treated group. The result is shown below.

Donor Pools Defined by Gender

Unit	Gender	W	Donor Pool Units
1	M	1	11,12,13,14,15
2	M	1	11,12,13,14,15
3	M	1	11,12,13,14,15
4	M	1	11,12,13,14,15
5	M	1	11,12,13,14,15
6	F	1	16,17,18,19,20
7	F	1	16,17,18,19,20
8	F	1	16,17,18,19,20
9	F	1	16,17,18,19,20
10	F	1	16,17,18,19,20
11	M	0	1,2,3,4,5
12	M	0	1,2,3,4,5
13	M	0	1,2,3,4,5
14	M	0	1,2,3,4,5
15	M	0	1,2,3,4,5
16	F	0	6,7,8,9,10
17	F	0	6,7,8,9,10
18	F	0	6,7,8,9,10
19	F	0	6,7,8,9,10
20	F	0	6,7,8,9,10

To fill in the missing potential outcomes, for each individual, we randomly choose one donor from each's donor pool. The donor's observed potential outcome is then filled in as the value of the missing potential outcome for that individual. This is done for each individual in the study, as illustrated below. This process creates a complete data set, with all of the missing potential outcomes filled in.

The following table shows a set of imputations, with a donor for each individual chosen at random from his or her donor pool. The imputed values are shown in parentheses.

Sample Imputation

Unit	Gender	W	Donor #	Y(0)	Y(1)
1	M	1	12	(8)	12
2	M	1	11	(6)	9
3	M	1	15	(11)	9
4	M	1	12	(8)	7
5	M	1	15	(11)	8
6	F	1	17	(7)	12
7	F	1	19	(8)	11
8	F	1	17	(7)	10
9	F	1	19	(8)	14
10	F	1	20	(10)	12
11	M	0	3	6	(9)
12	M	0	1	8	(12)
13	M	0	2	7	(9)
14	M	0	3	11	(9)
15	M	0	5	11	(8)
16	F	0	8	5	(10)
17	F	0	8	7	(10)
18	F	0	10	6	(12)
19	F	0	6	8	(12)
20	F	0	6	10	(12)

Once this complete data set is created, it is easy to compute an estimate of the difference in means of the potential outcomes under treatment and control, or any other estimate of interest. For example, we could easily estimate the median treatment effect among males or females.

This process should be repeated multiple times, using the same donor pools and randomly drawing a new donor for each individual each time. This gives an estimate of the uncertainty in the estimated treatment effects. Six specific imputations are shown below, and the histograms for the mean causal effect and the median causal effect are also given. The vertical bars in each plot show the bounds for a 95% interval.

Imputation 1:

Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(6)	12	6
2	M	1	(11)	9	-2
3	M	1	(11)	9	-2
4	M	1	(6)	7	1
5	M	1	(11)	8	-3
6	F	1	(10)	12	2
7	F	1	(7)	11	4
8	F	1	(10)	10	0
9	F	1	(10)	14	4
10	F	1	(8)	12	4
11	M	0	6	(9)	3
12	M	0	8	(7)	-1
13	M	0	7	(8)	1
14	M	0	11	(9)	-2
15	M	0	11	(9)	-2
16	F	0	5	(8)	3
17	F	0	7	(10)	3
18	F	0	6	(7)	1
19	F	0	8	(10)	2
20	F	0	10	(5)	-5
Average					**0.85**
Median					**1.0**

Imputation 2:

Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(7)	12	5
2	M	1	(6)	9	3
3	M	1	(6)	9	3
4	M	1	(8)	7	-1
5	M	1	(8)	8	0
6	F	1	(10)	12	2
7	F	1	(5)	11	6
8	F	1	(10)	10	0
9	F	1	(6)	14	8
10	F	1	(6)	12	6
11	M	0	6	(9)	3
12	M	0	8	(9)	1
13	M	0	7	(9)	2
14	M	0	11	(12)	1
15	M	0	11	(8)	-3
16	F	0	5	(5)	0
17	F	0	7	(8)	1
18	F	0	6	(7)	1
19	F	0	8	(7)	-1
20	F	0	10	(6)	-4
Average					**1.65**
Median					**1.0**

Imputation 3:

Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(6)	12	6
2	M	1	(7)	9	2
3	M	1	(11)	9	-2
4	M	1	(11)	7	-4
5	M	1	(8)	8	0
6	F	1	(8)	12	4
7	F	1	(6)	11	5
8	F	1	(10)	10	0
9	F	1	(10)	14	4
10	F	1	(10)	12	2
11	M	0	6	(7)	1
12	M	0	8	(12)	4
13	M	0	7	(12)	5
14	M	0	11	(7)	-4
15	M	0	11	(9)	-2
16	F	0	5	(6)	1
17	F	0	7	(5)	-2
18	F	0	6	(10)	4
19	F	0	8	(7)	-1
20	F	0	10	(5)	-5
Average					**0.9**
Median					**1.0**

Imputation 4:

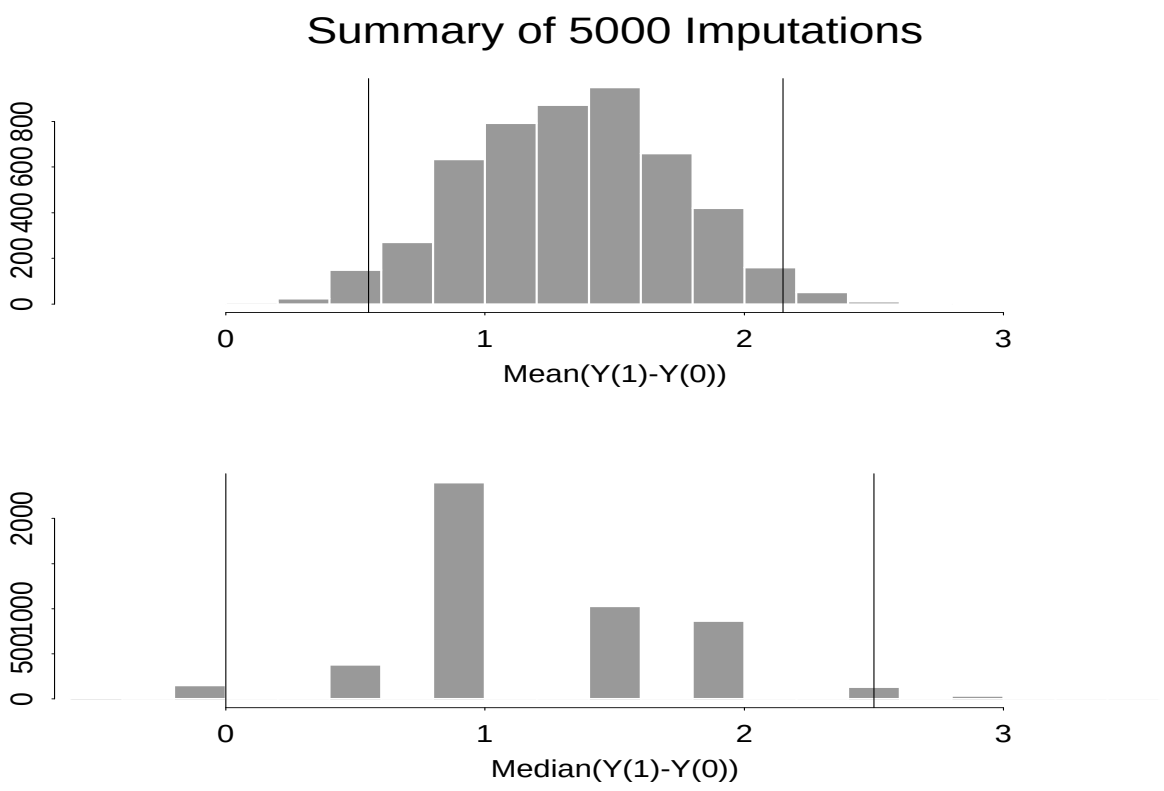
Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(8)	12	4
2	M	1	(6)	9	3
3	M	1	(11)	9	-2
4	M	1	(7)	7	0
5	M	1	(7)	8	1
6	F	1	(6)	12	6
7	F	1	(5)	11	6
8	F	1	(6)	10	4
9	F	1	(8)	14	6
10	F	1	(6)	12	6
11	M	0	6	(12)	6
12	M	0	8	(9)	1
13	M	0	7	(9)	2
14	M	0	11	(12)	1
15	M	0	11	(7)	-4
16	F	0	5	(7)	2
17	F	0	7	(7)	0
18	F	0	6	(5)	-1
19	F	0	8	(7)	-1
20	F	0	10	(10)	0
Average					**2.0**
Median					**1.5**

Imputation 5:

Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(11)	12	1
2	M	1	(11)	9	-2
3	M	1	(11)	9	-2
4	M	1	(8)	7	-1
5	M	1	(11)	8	-3
6	F	1	(8)	12	4
7	F	1	(5)	11	6
8	F	1	(5)	10	5
9	F	1	(8)	14	6
10	F	1	(5)	12	7
11	M	0	6	(9)	3
12	M	0	8	(8)	0
13	M	0	7	(9)	2
14	M	0	11	(12)	1
15	M	0	11	(9)	-2
16	F	0	5	(8)	3
17	F	0	7	(8)	1
18	F	0	6	(8)	2
19	F	0	8	(10)	2
20	F	0	10	(10)	0
Average					**1.65**
Median					**1.5**

Imputation 6:

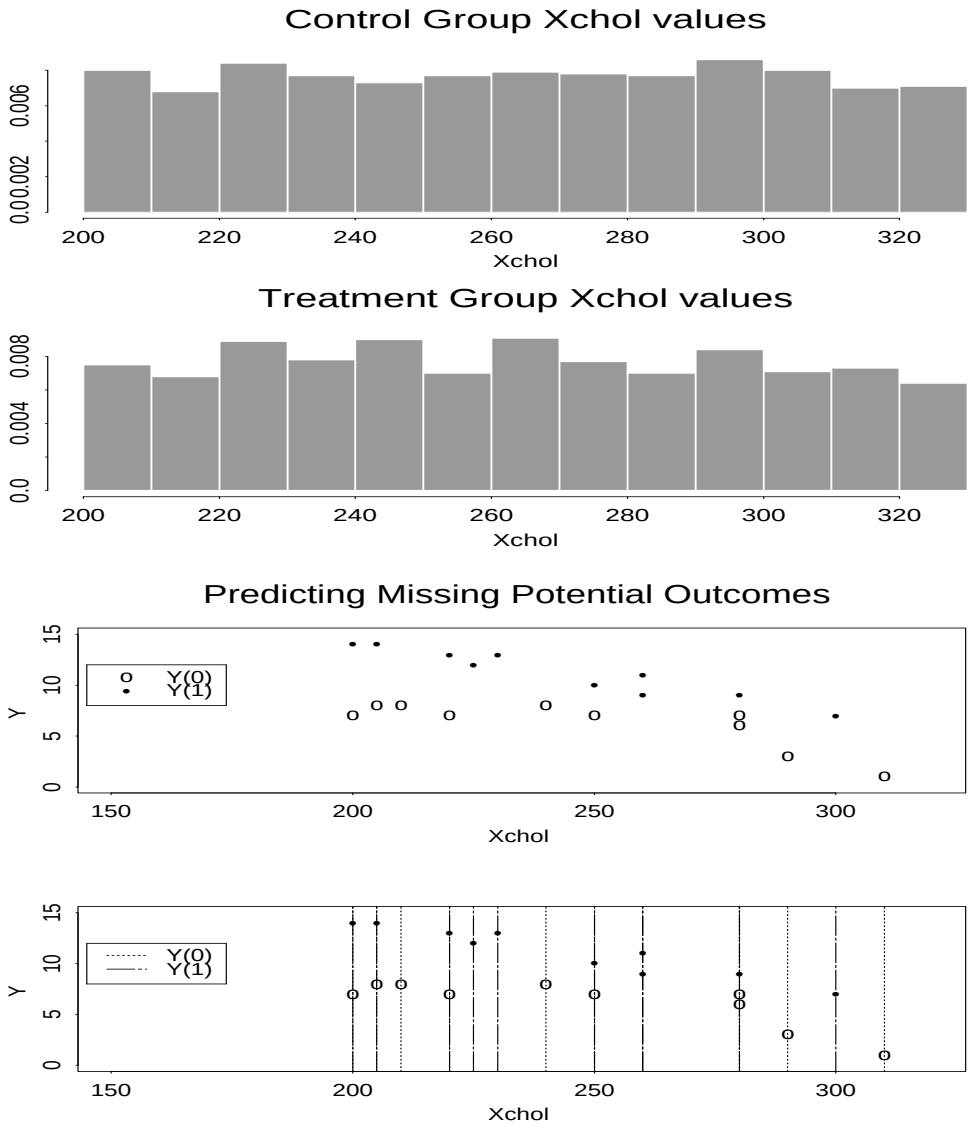
Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(7)	12	5
2	M	1	(11)	9	-2
3	M	1	(7)	9	2
4	M	1	(11)	7	-4
5	M	1	(11)	8	-3
6	F	1	(7)	12	5
7	F	1	(6)	11	5
8	F	1	(6)	10	4
9	F	1	(8)	14	6
10	F	1	(8)	12	4
11	M	0	6	(9)	3
12	M	0	8	(9)	1
13	M	0	7	(8)	1
14	M	0	11	(9)	-2
15	M	0	11	(8)	-3
16	F	0	5	(6)	1
17	F	0	7	(6)	-1
18	F	0	6	(10)	4
19	F	0	8	(8)	0
20	F	0	10	(10)	0
Average					**1.3**
Median					**1.0**


Summary of 5000 Imputations
0
200
400
600
800
0
1
2
3
Mean(Y(1)-Y(0))
0
500
1000
2000
0
1
2
3
Median(Y(1)-Y(0))

Example III-2: Donor pools with a continuous covariate

A doctor is conducting an experiment (again) to determine which of two types of surgery is better in terms of leading to increased life for male patients. The control is the standard surgery and treatment is a new surgery he just developed. He recruits study participants and assigns half to treatment and half to control. The design is completely randomized. There are 1,000 treated and 1,000 control patients, but for simplicity we will focus on the first ten patients in each group. The importance of the large sample size will become clear later. $Y(0)$ and $Y(1)$ represent years of life after the old and new surgery, respectively (age at death minus age at the time of the surgery).

The doctor also has recorded the cholesterol level of each patient prior to the surgery (the covariate X_{chol} represents cholesterol prior to surgery). The histograms below show all values of cholesterol for the 2,000 patients, and the bottom plots show the observed potential outcome data for the first 10 in each group.



The values of the covariate (cholesterol level) overlap between the treatment and control groups, as they should because the study is a completely randomized experiment. We wish to estimate the treatment effect by matching on this covariate. The large sample sizes in both the treatment and control groups ensure that we will have good matches for each patient. For each unit, we define a donor pool of potential units in the other treatment group with "similar" values of the covariate, and will then fill in the missing potential outcomes by drawing randomly from these pools (a unit is chosen randomly out of the donor pool, and that unit's outcome value is used to fill in the missing potential outcome). This method is not quite correct theoretically, but it conveys most of the essential ideas, and is known as "hot-deck" multiple imputation (in survey practice).

Suppose we define the donor pool as the patients in the other treatment group with the four closest values of the covariate. The chosen donor pools for the first 10 treated and control individuals is shown in the table below. Units 1-1000 are control, 1001-2000 are treated. The donor unit numbers are shown, as well as the donors' cholesterol levels. The two columns on the right show the potential outcomes observed for each donor pool.

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	Donor Pool Units	X_{chol} for Donor Units	Donor $Y(0)$	Donor $Y(1)$
1	210	0	8		1440,1616,1703,1902	208,215,210,212		14,13,12,13
2	200	0	7		1234,1476,1692,1952	200,204,205,210		14,14,13,12
3	310	0	1		1348,1678,1872,1925	306,309,300,302		11,9,9,7
4	220	0	7		1112,1382,1883,1956	215,218,222,219		14,13,12,13
5	280	0	7		1088,1112,1199,1560	275,282,280,270		11,9,9,7
6	290	0	3		1063,1282,1345,1882	291,282,288,280		11,9,9,7
7	240	0	8		1192,1253,1488,1828	238,241,240,241		13,12,13,10
8	250	0	7		1097,1138,1452,1782	255,250,251,250		13,10,11,9
9	280	0	6		1234,1274,1451,1919	275,282,280,270		11,9,9,7
10	205	0	8		1156,1291,1333,1814	204,210,208,205		14,14,13,12
1001	250	1		10	214,672,734,982	250,249,252,253	7,8,7,7	
1002	225	1		12	172,367,529,873	226,228,224,225	7,8,7,8	
1003	300	1		7	65,245,673,836	282,295,292,300	7,6,3,4	
1004	260	1		11	293,439,739,992	262,257,260,261	8,7,7,6	
1005	230	1		13	153,373,552,921	228,230,231,233	8,7,8,7	
1006	220	1		13	88,259,462,569	219,222,220,224	7,8,7,8	
1007	200	1		14	388,452,673,881	210,204,205,212	8,7,8,7	
1008	280	1		9	184,222,382,972	279,275,280,282	7,7,6,4	
1009	260	1		9	441,482,731,881	262,257,260,261	8,7,7,6	
1010	205	1		14	257,338,581,871	204,210,212,205	8,7,8,7	

To generate an estimate of the treatment effect, we fill in ("impute") the missing potential outcomes using the potential outcomes of the units in the pool of potential matches. For each unit, a value of its missing potential outcome is drawn from the units in its donor pool. Examples of this are shown below for the first ten patients in each treatment group. The imputed values are in parentheses. After having filled in everyone's missing potential outcome, either $Y(0)$ or $Y(1)$, we can then calculate each individual's treatment effect and estimate the average treatment effect (corresponding to that imputation), or the median treatment effect on log(Y), i.e. $\log(Y(1)) - \log(Y(0))$, etc. This is done repeatedly to display the variability in the calculated treatment effect depending on the imputation.

Six specific imputations are displayed below, and the histograms for the mean causal effect and the median causal effect are also given. The vertical bars in each plot show the bounds for a 95% interval.

What would we do here if the assignment mechanism had been that of the perfect doctor?

Imputation 1:

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	210	0	8	(14)	6
2	200	0	7	(13)	6
3	310	0	1	(9)	8
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(9)	6
7	240	0	8	(11)	3
8	250	0	7	(13)	6
9	280	0	6	(11)	5
10	205	0	8	(14)	6
1001	250	1	(7)	10	3
1002	225	1	(8)	12	4
1003	300	1	(3)	7	4
1004	260	1	(6)	11	5
1005	230	1	(8)	13	5
1006	220	1	(7)	13	6
1007	200	1	(7)	14	7
1008	280	1	(6)	9	3
1009	260	1	(7)	9	2
1010	205	1	(8)	14	6
Average					**4.95**
Median					**5.5**

Imputation 2:

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
1	210	0	8	(13)	5
2	200	0	7	(14)	7
3	310	0	1	(11)	10
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(9)	6
7	240	0	8	(12)	4
8	250	0	7	(11)	5
9	280	0	6	(9)	3
10	205	0	8	(14)	6
1001	250	1	(7)	10	3
1002	225	1	(7)	12	5
1003	300	1	(4)	7	3
1004	260	1	(8)	11	3
1005	230	1	(7)	13	6
1006	220	1	(7)	13	6
1007	200	1	(8)	14	6
1008	280	1	(3)	9	6
1009	260	1	(7)	9	2
1010	205	1	(8)	14	6
Average					**5**
Median					**5.5**

Imputation 3:

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
1	210	0	8	(12)	4
2	200	0	7	(14)	7
3	310	0	1	(7)	6
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(11)	8
7	240	0	8	(9)	1
8	250	0	7	(13)	6
9	280	0	6	(9)	3
10	205	0	8	(13)	5
1001	250	1	(7)	10	3
1002	225	1	(8)	12	4
1003	300	1	(4)	7	3
1004	260	1	(7)	11	4
1005	230	1	(7)	13	6
1006	220	1	(7)	13	6
1007	200	1	(8)	14	6
1008	280	1	(3)	9	6
1009	260	1	(7)	9	2
1010	205	1	(8)	14	6
Average					**4.7**
Median					**5.5**

Imputation 4:

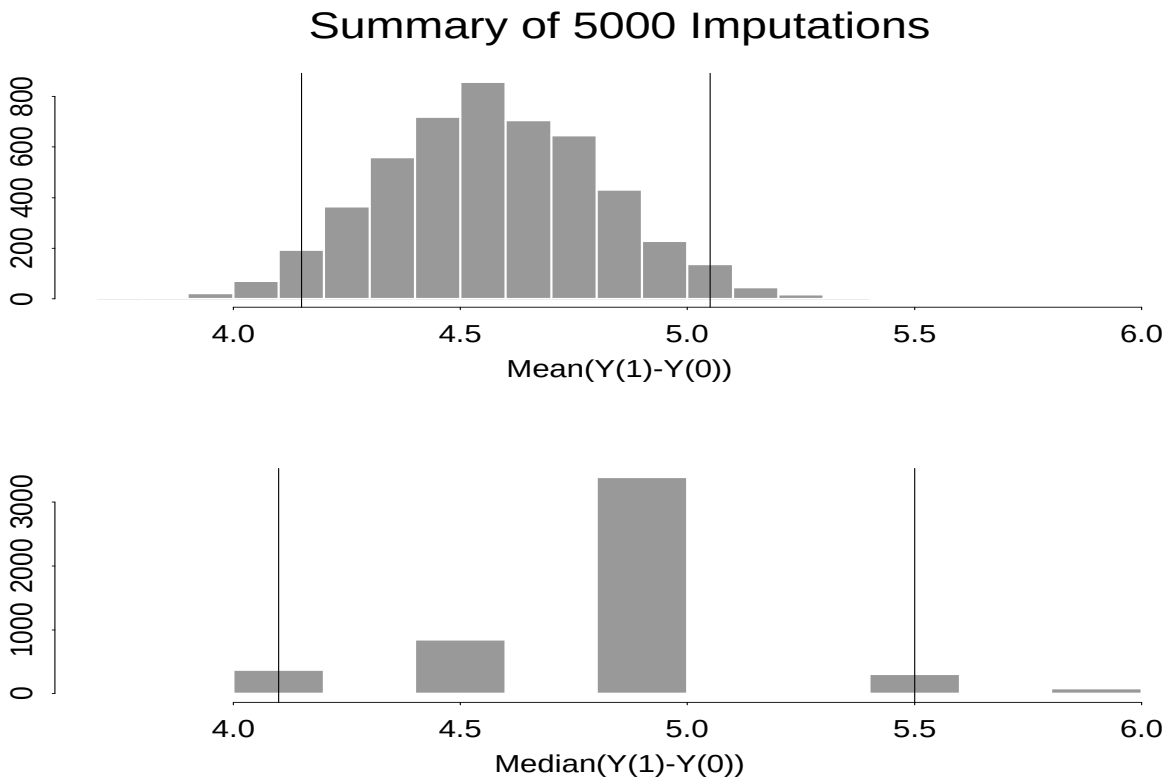
Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
1	210	0	8	(13)	5
2	200	0	7	(14)	7
3	310	0	1	(9)	8
4	220	0	7	(12)	5
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	240	0	8	(9)	1
8	250	0	7	(13)	6
9	280	0	6	(9)	3
10	205	0	8	(14)	6
1001	250	1	(7)	10	3
1002	225	1	(7)	12	5
1003	300	1	(4)	7	3
1004	260	1	(7)	11	4
1005	230	1	(8)	13	5
1006	220	1	(7)	13	6
1007	200	1	(7)	14	7
1008	280	1	(4)	9	5
1009	260	1	(6)	9	3
1010	205	1	(7)	14	7
Average					**4.95**
Median					**5**

Imputation 5:

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
1	210	0	8	(13)	5
2	200	0	7	(13)	6
3	310	0	1	(11)	10
4	220	0	7	(12)	5
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	240	0	8	(9)	1
8	250	0	7	(9)	2
9	280	0	6	(11)	5
10	205	0	8	(13)	5
1001	250	1	(7)	10	3
1002	225	1	(7)	12	5
1003	300	1	(3)	7	4
1004	260	1	(6)	11	5
1005	230	1	(7)	13	6
1006	220	1	(8)	13	5
1007	200	1	(8)	14	6
1008	280	1	(7)	9	2
1009	260	1	(7)	9	2
1010	205	1	(8)	14	6
Average					**4.65**
Median					**5**

Imputation 6:

Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
1	210	0	8	(13)	5
2	200	0	7	(12)	5
3	310	0	1	(7)	6
4	220	0	7	(13)	6
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	240	0	8	(11)	3
8	250	0	7	(9)	2
9	280	0	6	(11)	5
10	205	0	8	(12)	4
1001	250	1	(7)	10	3
1002	225	1	(7)	12	5
1003	300	1	(3)	7	4
1004	260	1	(8)	11	3
1005	230	1	(7)	13	6
1006	220	1	(8)	13	5
1007	200	1	(8)	14	6
1008	280	1	(7)	9	2
1009	260	1	(6)	9	3
1010	205	1	(8)	14	6
Average					**4.45**
Median					**5**



2. More Challenging Examples

Example III-3: Need for Covariate Overlap with One Covariate

In this example, we observe people after receiving either a new surgery ($W = 1$) or the standard surgery ($W = 0$). The outcome, Y, is quality of life six months after surgery, and age (at the time of surgery) is a covariate. Quality of life is measured on a scale of 0 to 100. We observe the following data. Treatments were assigned as some stochastic function of age; i.e., treatment assignment is ignorable given age.

Unit	W	Age	$Y(0)$	$Y(1)$	Unit	W	Age	$Y(0)$	$Y(1)$
1	0	10	99		26	0	30	48	
2	0	12	98		27	0	30	47	
3	0	14	96		28	0	31	46	
4	0	15	97		29	0	31	47	
5	0	16	95		30	0	31	46	
6	0	17	97		31	1	32		48
7	0	18	96		32	0	33	42	
8	0	18	96		33	0	36	40	
9	0	19	93		34	0	40	36	
10	0	20	88		35	1	40		39
11	0	21	90		36	0	42	33	
12	0	22	87		37	0	48	20	
13	0	22	86		38	1	49		32
14	0	23	83		39	0	52	22	
15	0	25	81		40	1	53		27
16	0	25	78		41	1	55		28
17	0	25	75		42	0	55	20	
18	0	26	72		43	0	61	17	
19	0	27	68		44	1	62		18
20	0	28	65		45	1	65		12
21	0	29	58		46	1	68		15
22	0	29	53		47	0	72	3	
23	0	29	51		48	1	73		9
24	0	29	49		49	1	79		2
25	0	30	49		50	0	80	0	

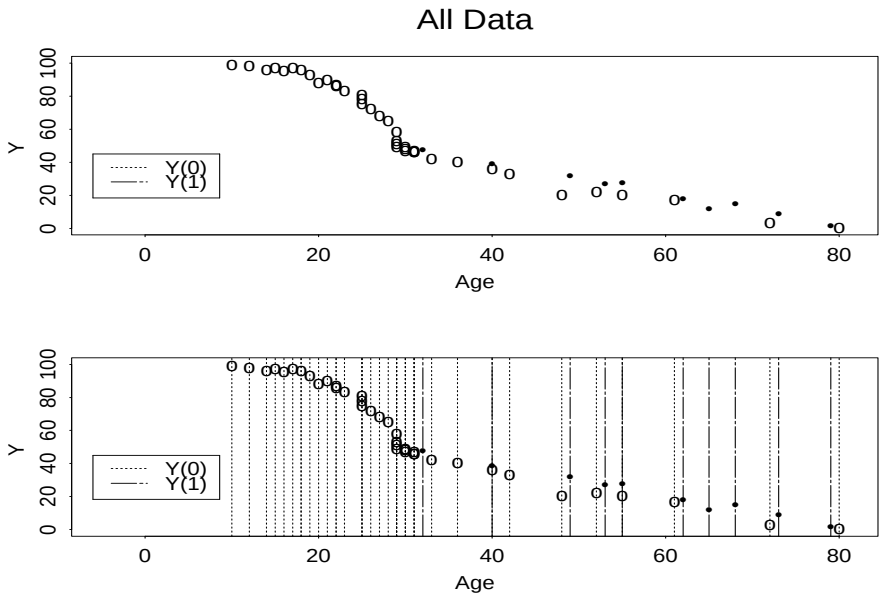
Suppose interest focuses on the effect of the new surgery on those who might receive it. We care only about estimating the effect on the population of people who might receive the new surgery, not on the population in general. A real example of this occurs in estimating the effect of smoking on those who choose to smoke. There is no intention of forcing never smokers to smoke, and thus we estimate the effect of smoking among only those who do smoke.

We thus would like to impute the missing potential outcomes only in the treated group, predicting the missing $Y(0)$ for these units:

Unit	W	Age	$Y(1)$	$Y(0)$
31	1	32	48	
35	1	40	39	
38	1	49	32	
40	1	53	27	
41	1	55	28	
44	1	62	18	
45	1	65	12	
46	1	68	15	
48	1	73	9	
49	1	79	2	

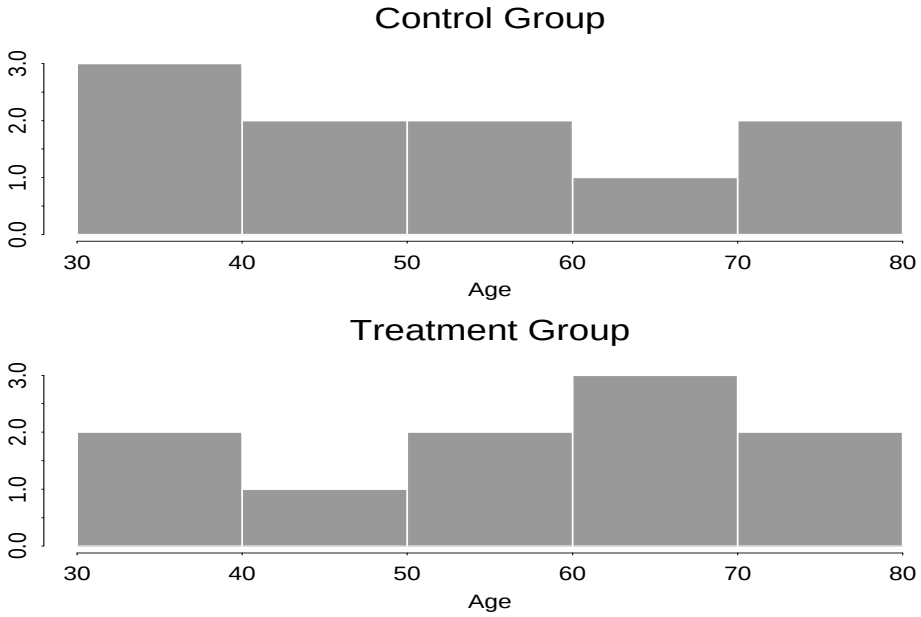
The 32-year-old person is the youngest in the treated group. Notice that most people in the control group were younger than the 32-year-old.

It is not really "fair" to use the 10- to 31-year-olds in the control group to impute $Y(0)$ for the people in the treatment group because their ages are so different from those in the treatment group. In fact, the average age in the control group is 30.5, and the average age in the treated group is 57.6.



The imputation will be more appropriate if we only use the controls who are "like" the treated, in the sense that they have similar ages. In terms of Part II of the course, the simple estimate of the probability of receiving treatment for those under 31 or 32 is zero.

We will use the ideas of matching and donor pools to correct for this bias. For each individual in the treated group, we form a donor pool consisting of the four control individuals closest in age to the treated individual. We thus only use control members with similar ages to individuals in the treated group. As seen in the plot below, there is good overlap in the age distributions in the range of ages of those in the treated group. We should thus be able to form a good donor pool for each treated individual.



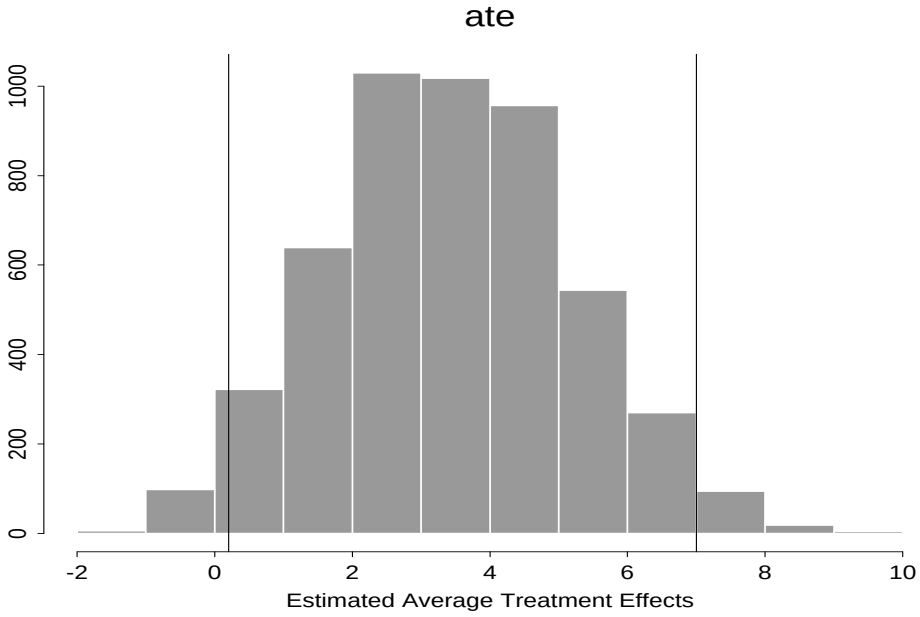
The donor pools (as defined above) are shown below:

Unit	Age	W	$Y(0)$	$Y(1)$	Units for Donor Pool	Donor $Y(0)$
31	32	1		48	28,29,30,32	45,43,43,42
35	40	1		39	32,33,34,36	42,40,36,33
38	49	1		32	36,37,39,42	33,20,22,20
40	53	1		27	37,39,42,43	20,22,20,17
41	55	1		28	37,39,42,43	20,22,20,17
44	62	1		18	39,42,43,47	22,20,17,3
45	65	1		12	39,42,43,47	22,20,17,3
46	68	1		15	42,43,47,50	20,17,3,0
48	73	1		9	42,43,47,50	20,17,3,0
49	79	1		2	42,43,47,50	20,17,3,0

To generate an estimate of the treatment effect, we fill in ("multiply impute") the missing potential outcomes using the potential outcomes of the units in the donor pool. For each unit, a value of its missing potential outcome is drawn from the units in its donor pool. An example of this is shown below. The imputed values are in parentheses. We can then calculate each individual's treatment effect and calculate the average (or median) treatment effect for those in the treated group.

Age	W	$Y(0)$	$Y(1)$	$Y(1)-Y(0)$
32	1	(42)	48	6
40	1	(36)	39	3
49	1	(33)	32	-1
53	1	(17)	27	10
55	1	(22)	28	6
62	1	(22)	18	-4
65	1	(3)	12	9
68	1	(17)	15	-2
73	1	(0)	9	9
79	1	(3)	2	-1
Average				**3.5**

By repeating this process many times, we can estimate the average treatment effect as well as the variance of the estimated treatment effect. We repeated the process 5,000 times, and the histogram below shows the values of the estimated treatment effect. The estimated mean treatment effect is 3.5 and its estimated variance is 3.1. A 95% interval for the treatment effect is (0.2, 7.0).



3. Matching using propensity score (donor pools): multiple covariates

The method described in the previous examples can also be used with multiple covariates. There are many ways to do this; a popular method based on propensity scores (which were previously introduced) is described here. First we estimate the treatment assignment probabilities for each unit. These propensity scores represent a one-dimensional summary of all the covariates. Once we have estimates of the propensity scores, we can match individuals based on their propensity scores and impute missing outcomes exactly the same way we did here. The propensity score method works because within a specific range of propensity score values, the two groups will have similar values of all covariates.

When using covariates for matching, it is generally important that they be TRUE covariates. A true covariate is a characteristic that is not affected by treatment. In this example, age is a true covariate because surgery does not affect age. All pre-treatment assignment variables are also true covariates: for example, treatment could not change someone's pre-treatment cholesterol level. In the surgery example, recovery time would not be a true covariate because it could depend on which treatment was assigned.

Note also that all of this is done without knowing the values of the outcome variables!

4. After balanced as well as possible in design, then include models requiring access to the outcome data
But: Go as far as possible without using Y_{obs}

Example III-4: Predictive Inference to Determine the Effects of in utero Phenobarbital exposure on Intelligence

The following is based on Reinisch, J.M., Sanders, S.A., Mortensen, E.L., and Rubin, D.B. "In Utero Exposure to Phenobarbital and Intelligence Deficits in Adult Men." *Journal of the American Medical Association*, November 15, 1995.

- Medications containing barbiturates are often prescribed to pregnant women for the treatment of a variety of disorders, such as predicted premature delivery or convulsive disorders.
- Some evidence of permanent negative effects of barbiturate exposure in laboratory animals prompted this study, which aimed to examine the effects of in utero exposure in humans.

Two studies done, with very similar designs. We concentrate on the larger one here. Medical records were used to identify the treated and control groups.

- Treated (exposed) group: Men born at the largest hospital in Copenhagen, Denmark between 1959 and 1961 whose mother took phenobarbital while pregnant (determined using medical records).
  - Some screening done based on other medical factors (mother with diabetes, twins, mother less than 16 when child born, etc.).
  - 81 men in final exposed sample (with available outcome data).
- Control group: Potential controls were men born at the hospital between 1959 and 1961 who were not exposed to phenobarbital in utero.
  - Same screening done as in treated group, resulted in over 3000 potential controls.
  - Matching done: "The objective of the matching was to obtain a set of control subjects, approximately the same number as exposed, whose distributions of matching variables were nearly the same as the distributions for exposed subjects."
    - 10 best matches determined for each exposed individual, using Mahalanobis metric matching within calipers defined by the estimated propensity score.
    - This group of matches refined by the senior author (Reinisch).
  - 101 controls selected.

The following table summarizes the effects of the matching:

Variable	Full Set of Controls	Matched Controls	Exposed Subjects
In Prediction Models			
% Firstborn	56.41	50.50	50.62
% Unwanted pregnancy	59.51	48.51	48.00
% Abortion Attempted	7.91	6.93	6.58
% Single Mother	41.09	22.77	22.50
Mean SES	4.07	4.47	4.53
Mean breadwinner's education	3.39	3.44	3.44
Mean predisposing risk score	28.14	26.02	26.52
Mean mother's age	24.76	26.50	27.04
Mean father's age	28.63	29.70	29.62
Potential Confounding Variables			
Mean gestational length (wks)	38.59	38.63	38.73
Mean birth weight (g)	3233	3260	3219
Mean birth length (cm)	51.28	51.64	51.57
Mean # cigarettes in 3rd trimester	6.40	5.26	5.03
Mean maternal weight gain (kg/m^3)	26.88	28.18	27.65
Mean maternal complaint	1.70	3.97	4.95
Sample size	3308	101	81

- We see that the matched sample of controls is much more similar to the exposed group than is the full sample of controls.
- The following variables are "significantly" different between the full set of controls and the exposed individuals: % unwanted pregnancy, % single mother, mean socioeconomic status, mean mother's age, and mean maternal complaint score.
- There are no variables that are "significantly" different between the matched controls and the exposed subjects.

Results:

- Outcome: score on Danish Military Draft Board Intelligence test. Test given to nearly all Danish men. 78 questions covering letter matrices, verbal analogies, number series, and geometric figures. Score is the number of items correct.
- Linear model used for outcome, with model estimated using the matched control subjects.
  - Predictors used: family's socioeconomic status (SES) when child 1 year old, breadwinner's education, sibling position, whether pregnancy was "wanted", whether abortion attempted, maternal marital status, predisposing risk score, mother's age, father's age, subject's age at time of testing, square of the deviation of SES from the mean, square of the deviation of age at testing from the mean.

  - Model then used to predict the potential outcome under control for the treated subjects. The observed treated outcome is then compared with the predicted control outcome.

- Also looked within subgroups.

Group	Sample Size	Mean Observed Score	Mean Predicted Score	Mean Difference	Adjusted p-value
All exposed	81	39.58	44.35	-4.77	0.002
Socioeconomic Status					
Lower	55	36.24	42.25	-6.01	0.002
Higher	21	49.57	47.28	2.29	0.23
Wanted pregnancy?					
Unwanted	36	36.89	42.01	-5.12	0.02
Wanted	39	42.77	45.84	-3.07	0.08
Timing of Exposure					
3rd trimester only	72	40.26	44.64	-4.38	0.006
3rd trimester and earlier	5	23.80	41.22	-17.42	0.001
Prior to 3rd trimester only	4	47.00	43.01	3.99	0.23
Total Dosage					
≤ 5000 mg	71	40.60	44.58	-3.98	0.02
> 5000 mg	10	32.30	42.72	-10.42	0.001

Conclusions:

- Effects of exposure to phenobarbital in utero can be seen well into adulthood even in the absence of physical abnormalities.
- Timing of drug exposure affects the size of the effect.
- Social and psychological factors interact with in utero exposure to affect the size of the effect.
- Physicians should exercise caution in prescribing phenobarbital to pregnant women, particularly those with lower socioeconomic status.

Summary of Prediction Methods

1. Methods that do not use Y_{obs}
   (a) Exact matching on covariates
   (b) Nearest available matching using some distance measure, to form donor pools
   (c) Matching using estimated probabilities of receiving treatment
   (d) Combinations of methods
2. Methods that use Y_{obs}
   (a) Parallel linear regressions in treated and control groups
   (b) Separate linear regressions in treated and control groups
   (c) Non-linear modeling in treated and control groups
   (d) For these methods, for robust results, should include indicator for matched pairs or subclasses and interact with covariates, or do separate analyses within each subclass.

5. Formal methods to deal with nonignorable treatment assignment

Once the observed covariates have been dealt with as well as possible (through matching, subclassification, modeling within subclasses, etc.), then attention can be shifted to consider the impact of possible unobserved covariates.

The first analysis to do this was Cornfield et al. (1959) in a study of the relationship between smoking and lung cancer. This study addressed criticism that the observed relationship could be due to an unobserved covariate (such as a genetic component) that would increase both someone's probability of smoking and their probability of being diagnosed with lung cancer. Cornfield showed that this covariate would have to have a much stronger relationship between both treatment assignment and the outcome than any other covariate already measured in order for the observed relationship to go away. It is unlikely that such a covariate exists, and thus this analysis was seen to provide strong evidence that there is indeed a causal relationship between smoking and lung cancer.

Rosenbaum and Rubin (1983) extend Cornfield's approach, giving methods to assess sensitivity to an unobserved binary covariate, also taking into account the observed covariates. The main approach involves positing an unobserved binary covariate u such that treatment assignment is confounded when u is unobserved, but given u it is unconfounded. The method estimates the treatment effect over ranges of plausible correlations between u and both treatment assignment and the outcome of interest. If the conclusions are insensitive to this range of plausible correlations, then the conclusions assuming unconfounded treatment assignment (not given u) are more believable.

We can also use the ideas of prediction to think about what we would do if assignment were nonignorable. In Example III-2 we assumed that treatment assignment was ignorable given age and pre-treatment cholesterol. What if instead the data had arisen from the perfect doctor? Remember that the perfect doctor could see each individual's potential outcomes under both treatments, and assigned the treatment best for each person (or tossed a fair coin when there was no difference in survival). We would like to match on variables that would make treatment assignment ignorable, but since the doctor did not write down the potential outcomes under both treatments for each individual, we do not have that information.

Assignment is no longer ignorable, but we can still think about imputing the missing potential outcomes. Consider Unit 1, who lived for 8 years after the old surgery ($W_1 = 0, Y_1(0) = 8$). Since the perfect doctor assigned the treatment that would be best for each individual, we know that Unit 1 must have a potential outcome under the new surgery of less than or equal to 8 years! In other words, $Y_1(1) \leq 8$. We could then think about imputing values less than 8 years as $Y_1(1)$. Maybe any value uniformly between 0 and 8? Or perhaps specify a different lower bound?

We may even want to use baseline cholesterol level to help with this imputation. For example, we could find individuals among the treated group who match Unit 1 on baseline cholesterol and lived for less than 8 years. These people would then form a donor pool for Unit 1.

This process could be repeated for each individual, forming donor pools as we did before. The missing potential outcomes would be filled in using these donor pools, and would reflect what we know about the assignment mechanism. This intuitive process of filling in sensible values for the missing potential outcomes is much more reasonable than just looking at the straight difference in means $\overline{y(1)} - \overline{y(0)}$ as the perfect doctor did. We saw that $\overline{y(1)} - \overline{y(0)}$ led to very misleading results. Using this predictive approach, we can use methods that make more sense and use all of the information that we have about the science and the assignment mechanism. Notice that this more sensible method of imputing the missing potential outcomes results in estimated causal effects that have absolutely nothing to do with the observed sample means.

Example III-5: Assessing the Assumption of Unconfounded Treatment Assignment Through Simulation

We have previously discussed general ideas of how to assess sensitivity to an unobserved binary covariate (e.g. Cornfield, Rosenbaum and Rubin). We will now see how to do this through simulation, by predicting the missing unobserved variable and re-estimating the average treatment effect under various scenarios.

Suppose we are interested in estimating the effect of a new surgery on mortality. We have observational data on 100 treated individuals and 100 control individuals. The treated group ($W = 1$) received the new surgery, whereas the control group ($W = 0$) received the old surgery. The outcome of interest is death (1=dead, 0=alive). There are no covariates observed.

The data is summarized in the following table:

Number of Individuals	W	Y(0)	Y(1)
70	0	0	
30	0	1	
40	1		0
60	1		1

We first do the analysis assuming that treatment assignment was unconfounded and use the predictive approach to estimate the average treatment effect and a 95% interval. As before, we use the observed $Y(0)$ to impute the missing potential outcomes under control for the treated group, and use the observed $Y(1)$ to impute the missing potential outcomes under treatment for the control group. We can then calculate $\overline{Y(1)} - \overline{Y(0)}$ for this "complete" data set. We repeat this 1000 times, and get an estimate of 0.300, with a 95% interval of $(0.235, 0.365)$.

A critic then points out that it is possible that the individuals in the treated group were actually more unhealthy to start with, which is why they had a higher mortality rate. He thinks that pre-treatment health status is an unobserved covariate (we will call it U), that is related to both treatment assignment and the outcome. In other words, he asserts that treatment assignment is confounded when we do not know U, but if we did know the value of U for each person, then treatment assignment would be unconfounded. He also believes that there were more sick people among the ones who died than among the ones who lived (in both treatment groups).

To address this criticism, we assess how sensitive the estimated treatment effect is to different scenarios regarding this unobserved covariate. Consider $U = 1$ to mean the person was unhealthy before treatment assignment. $U = 0$ means the person was healthy before treatment assignment.

We can now specify four different probabilities which give the probability of $U = 1$ in each of the four groups defined by the table above. Varying these probabilities will allow us to assess how sensitive the estimated treatment effect is to this possibly important unobserved covariate. More concretely, we have:

Number of Individuals	W	Y(0)	Y(1)	$P(U=1\|W,Y_{obs})$
70	0	0		p_1
30	0	1		p_2
40	1		0	p_3
60	1		1	p_4

We need to relate these 4 probabilities to the critic's assertions. The first part of his assertion was that the probability of being unhealthy (of having $U = 1$) is higher in the treated group than it is in the control group. In terms of the 4 probabilities, this means:

$$\frac{70}{100} * p_1 + \frac{30}{100} * p_2 < \frac{40}{100} * p_3 + \frac{60}{100} * p_4$$

The second assumption is that in both groups, the probability of being unhealthy is higher among the people who died than among the people who lived. This implies:

$$p_1 < p_2$$

$$p_3 < p_4$$

To assess the sensitivity of the average treatment effect to the unobserved covariate U, we try a variety of values of p_1, p_2, p_3 and p_4 that meet these two criteria. For a given set of values of these 4 probabilities, we use the ideas of prediction to calculate an estimated average treatment effect under that scenario. For each type of person defined by the observed data (each row of the table above), we multiply impute values of U. For example, for people with $W_i = 0$ and $Y_i(0) = 0$, we impute $U_i = 1$ with probability p_1 and impute $U_i = 0$ with probability $1 - p_1$. (So of the 70 people in that category, we would expect about $70 * p_1$ would have $U_i = 1$).

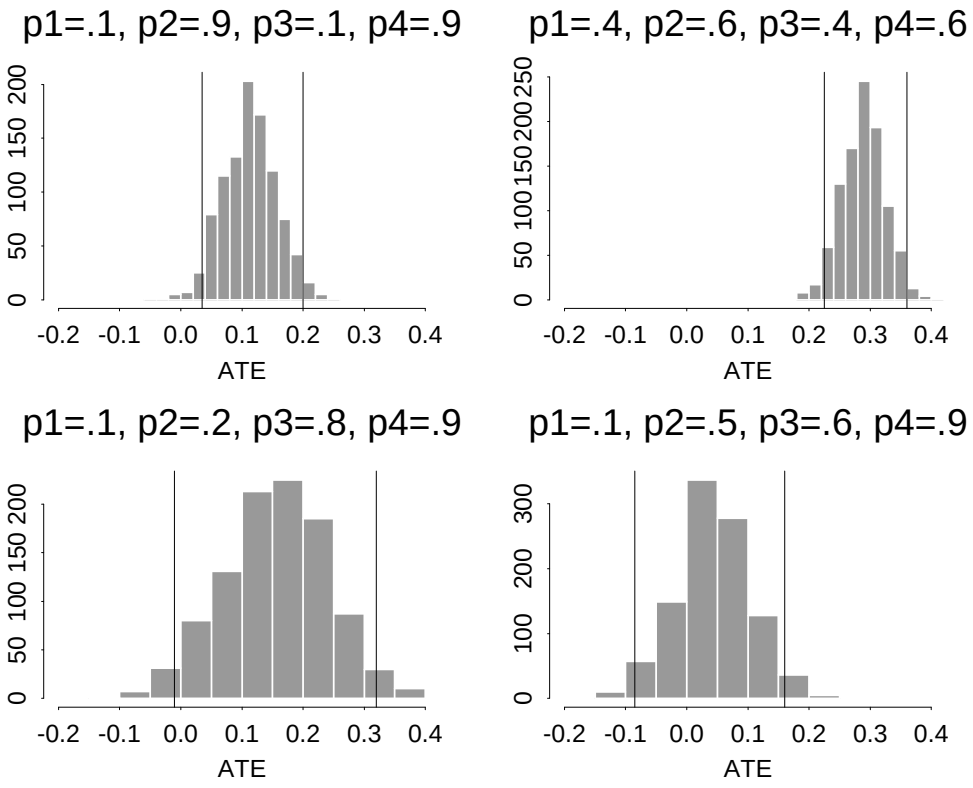
Once these values of U are imputed, we can compute the estimated treatment effect for that data set, treating U as if it was just another observed covariate. We again use the predictive approach to impute the missing potential outcomes, but do this separately for those with $U_i = 1$ and those with $U_i = 0$. This is the same procedure that we would follow if U had been an observed covariate such as gender. We would use the males to impute the males' missing potential outcomes, and the females to impute the females' missing potential outcomes.

Similarly, since we are assuming that U is a covariate related to treatment assignment, we calculate the average treatment effect among the unhealthy people ($U_i = 1$) and the average treatment effect among the healthy people ($U_i = 0$), and then form a weighted average of these to get the overall effect (weighted by the percentage of unhealthy and healthy people, respectively). This is the same procedure we followed when we knew that treatment assignment was random within blocks defined by gender.

Using this same set of values of p_1, p_2, p_3, and p_4, we repeat this process many times (e.g., 1000) to obtain the average treatment effect and its variability.

Below, we estimate the average treatment effect under 4 different scenarios, all of which meet the criteria above. Histograms showing the treatment effect under each scenario are also given.

Scenario	p_1	p_2	p_3	p_4	Estimated ATE	95% Interval
1	0.1	0.9	0.1	0.9	0.11	(0.035, 0.200)
2	0.4	0.6	0.4	0.6	0.29	(0.215, 0.350
3	0.1	0.2	0.8	0.9	0.16	(-0.005, 0.310)
4	0.1	0.5	0.6	0.9	0.04	(-0.075, 0.160)



We see that the estimate of the treatment effect varies quite a bit, depending on the scenario regarding the 4 probabilities. If this unobserved health status does exist, as the critic has suggested, it could dramatically change our view of the new surgery. For example, the probabilities in Scenario 3 imply that people in the treated group are much more likely to be unhealthy than those in the control group, although in both treatment groups the probability of being unhealthy is about the same in the group who lived and the group who died. If this situation were true, the interval for the treatment effect would include 0, and so we would not reject the hypothesis that there is no effect of the new treatment on mortality. We could get outside expert medical opinion to determine if these probabilities are realistic.

We also note that if p_1 is close to p_2 and p_3 is close to p_4, as in Scenario 2, the estimated treatment effect is similar to that found before. Under this scenario, the probability of being unhealthy is about the same among the people who lived and among the people who died, and so the unobserved covariate is not highly related to the outcome of interest. If $p_1 = p_2$ or $p_3 = p_4$, this would imply that U is unrelated to the outcome of interest, and thus treatment assignment would be unconfounded. Similarly, if the probability of being unhealthy is the same in the two groups ($\frac{70}{100} * p_1 + \frac{30}{100} * p_2 = \frac{40}{100} * p_3 + \frac{60}{100} * p_4$), then U is unrelated to treatment assignment and so again treatment assignment would be unconfounded.

6. Non-ignorable assignment mechanisms and other complications (more advanced topics)

The following list provides a summary of more advanced topics. In the remainder of the course handouts, we will provide examples of only a few of these, due to time and space limitations. More information on these topics can be found in some of the readings provided in the reference list.

- Speculation about a non-ignorable assignment mechanism
- Missing data
- Dropout
- Noncompliance
- Multiple treatments
- Interest in many subgroups
- Longitudinal data
- Principal stratification examples (adjusting for a post-treatment variable)

Example III-6: Noncompliance

Sommer and Zeger (1991) analyzed data from a study of the effects of vitamin A on child mortality. The article, "On Estimating Efficacy from Clinical Trials," is in the course pack. The study took place in Indonesia, where villages were randomized to receive either vitamin A supplements or control (no supplements). Out of 450 villages, 225 were chosen to receive treatment, while the other villages received control. Children who lived in the treatment villages received large oral doses of vitamin A, and the outcome (death) was measured in all villages one year after treatment was received. Because of Indonesian government policy, placebos could not be used.

Some individuals in the treatment group did not actually take the vitamin A supplements; we call these people noncompliers. No one in the control group took vitamin A because the supplements were only available in those villages randomized to treatment. The data recorded for each child were treatment assigned (W=1 for vitamin A and W=0 for control), treatment received, and the outcome.

The people in this study can be classified into one of two types: true compliers (C) or true noncompliers (N). True compliers are those who would take vitamin A if assigned to it, and true noncompliers are those who would not take vitamin A if assigned to it. We only observe compliance status in those people assigned to treatment; we do not know what people assigned to control *would have done* had they been assigned to treatment, so we do not know their compliance status. Like the treatment group, the control group is a mixture of compliers and noncompliers; unlike the treatment group, we do not know which individuals in the control group are compliers and which are noncompliers.

All the data from the study are given in the following table. Treatment assigned (W) equals 1 for vitamin A and 0 for control. Treatment received equals 1 if vitamin A was taken and 0 otherwise. Y_{obs} equals 0 if the child was alive at the end of the study and 1 otherwise.

Compliance Type	Treatment Assigned	Treatment Received	Y_{obs}	Number of Children
?	0	0	0	11514
?	0	0	1	74
N	1	0	0	2385
N	1	0	1	34
C	1	1	0	9663
C	1	1	1	12
				23682

The standard analysis for randomized studies with noncompliance is called Intention to Treat (ITT). This method ignores compliance information and compares those assigned to treatment to those assigned to control. This gives a valid estimate of the effect of treatment assignment on outcome.

As-treated and per protocol are two other ways that data of this type could be analyzed. An as-treated analysis compares those who received treatment with those who received control, ignoring treatment assignment. Per protocol analysis compares people who were assigned to and received treatment with those who were assigned to and received control.

The estimates from these methods are given below. The "treatment effect" is defined as the difference in mortality rates between the two groups being compared.

Method	Estimate	
ITT	-.0026	$= \frac{12+34}{9663+2385+12+34} - \frac{74}{11514+74}$
As-treated	-.0065	$= \frac{12}{9663+12} - \frac{34+74}{11514+2385+34+74}$
Per protocol	-.0052	$= \frac{12}{9663+12} - \frac{74}{11514+74}$

As stated above, the ITT estimate is a true causal effect estimate; it represents the effect of assignment on mortality. It does not, however, estimate the effect of taking vitamin A on mortality. The as-treated and per protocol estimates do not even estimate true causal effects because they are comparing groups of people that are fundamentally different. [This difference is evident from the data: note that the death rate for the noncompliers in the treatment group is 34/(34+2385) = .014, much higher than in the control group (74/74+11514 = .006), even though both received the same treatment.] The as-treated estimate compares those who received treatment with those who received control. Those who received treatment are all compliers, but those who received control are a mixture of compliers and noncompliers.

The per protocol estimate ignores non-compliers in the treatment group, and compares those who complied in the treatment group with those who complied in the control group (in our case the whole control group). This also compares compliers with a mixture of compliers and noncompliers, because the control group contains both compliers and noncompliers.

The ITT estimate compares two groups which are both mixtures of compliers and noncompliers, and because treatment was assigned randomly the proportion of compliers and noncompliers should be the same in the treatment and control groups.

None of these estimates, therefore, is estimating what we are really interested in: the effect of taking vitamin A on child mortality. Using a method similar to instrumental variables from economics, we can estimate the effect of treatment on compliers, i.e., the causal effect of receiving treatment on outcome.

Let ACE (average causal effect) denote the causal effect of treatment assignment on outcome. The ITT estimate is an unbiased estimate of the ACE. Since there are two distinct types of people (compliers and noncompliers) in our example, the ACE is a weighted average of the ACE for each group (weighted by the proportion of the population in each group):

$$\mathrm{ACE} = p_N \cdot \mathrm{NACE} + p_C \cdot \mathrm{CACE}.$$

Here p_N and p_C denote the proportion of noncompliers and compliers, respectively, in the population. NACE and CACE denote the average causal effect of assignment for noncompliers and compliers, respectively.

ACE, p_N, and p_C can all be estimated from the data. The ITT estimate is unbiased for the ACE, and p_N and p_C can be estimated as the proportion of compliers and noncompliers in the treatment group, since treatment was assigned randomly. 9663+12=9675 people in the treatment group complied, and 2385+34=2419 did not comply. Thus we estimate $\hat{p}_C$ = 9675/(2419+9675) =.8 and $\hat{p}_N$ = 2419/(2419+9675) = 0.2. This leaves two unknowns, NACE and CACE, in a single equation:

$$-0.0025 = .2 \cdot \text{NACE} + .8 \cdot \text{CACE}.$$

Suppose we assume that NACE is equal to zero: since noncompliers do not take treatment regardless of treatment assignment, we assume assignment has no effect on outcome. This gives

$$-0.0025 = .8 \cdot \text{CACE} \Rightarrow \text{CACE} = -0.0025/.8 = -0.0031.$$

We call this the instrumental variables (IV) estimate of the complier average causal effect (CACE). Note that this does estimate the effect of *treatment*, since treatment assigned and treatment received are the same for compliers. The IV estimate is a valid estimate of the effect of treatment on outcome if the following four criteria/assumptions are met:

1. SUTVA. SUTVA (or some other assumption) is required for all causal inference.
2. Random assignment. Random assignment to treatment allows us to estimate the proportion of compliers and noncompliers in the population using only the individuals in the treatment group.
3. $p_C > 0$. We divide by p_C to obtain the estimate, so p_C cannot equal zero.
4. NACE = 0. We assume that since behavior cannot be changed by assignment for noncompliers, neither can outcome. This assumption is called the Exclusion Restriction, and must be considered carefully for each experiment, as it does not always hold.

Example III-7: The New York School Choice Scholarships Program

Consider the New York School Choice Scholarships Program, which has been discussed earlier. The evaluation was interested in the effects of school vouchers on test scores. Students entered a lottery to receive a voucher to help them pay for private school. The voucher did not cover the full costs of private school tuition. Thus, in reality there are two types of noncompliers: students who received a voucher and did not attend private school, and students who did not receive a voucher but did attend private school. Here we give a hypothetical and simplified example where we assume that the families are of such low income that they would not be able to attend private school without the voucher, so we only have the first type of noncomplier. For more information on the full study, see Hill, Rubin, and Thomas (2000) or Barnard, Du, Hill, and Rubin (1998).

We are interested in estimating the complier average causal effect: the effect for students who received a voucher and went to private school.

A summary of the simulated data is shown below. The test score is the score on a standardized exam at the end of the first full school year after the lottery.

Treatment Assigned	Treatment Received	N	Mean Test Score
No voucher	Public school	400	88
Voucher	Public school	220	85
Voucher	Private school	180	95

To estimate the complier average causal effect (CACE), we need to calculate the percent compliers as well as the Intention to Treat (ITT) estimate. The ITT estimate ignores the noncompliance, and compares the outcomes for the students assigned to treatment and assigned to control. It is a valid estimate of the causal effect of assignment on the outcome.

To estimate the percent compliers (p_C), we look within the treatment group since we observe compliance status for these individuals. The compliers are those individuals who receive a voucher and go to private school. Noncompliers are those who receive a voucher but go to public school. Since assignment was randomized, the percent compliers in the treated and control groups should be the same.

$$\widehat{p_C} = \frac{180}{180 + 220} = 0.45$$

We now calculate the ITT estimate by comparing the outcomes for the voucher and no voucher groups:

$$\begin{aligned} \widehat{ITT} &= \overline{y_1} - \overline{y_0} \\ &= \frac{220 * 85 + 180 * 95}{220 + 180} - 88 \\ &= 89.5 - 88 \\ &= 1.5 \end{aligned}$$

We can then use the following formula to estimate the CACE. The average causal effect of treatment (ACE) is the same as the overall Intention to Treat estimate (ITT).

$$\widehat{ACE} = \widehat{ITT} = p_N NACE + p_C CACE.$$

We assume the exclusion restriction (that the effect of assignment to treatment for non-compliers is zero; in other words, for students who will attend public school regardless of treatment assignment, giving them a voucher does not affect their test scores). In other words, we assume $NACE = 0$. We then get the following estimate of the CACE:

$$\begin{aligned} \widehat{CACE} &= \frac{\widehat{ITT}}{p_C} \\ &= \frac{1.5}{0.45} \\ &= 3.33 \end{aligned}$$

Example III-8: The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT)

The following study was reported in Goetghebeur and Molenberghs (1996). We have simplified it somewhat, but the main ideas are below.

This study estimated the effect on cholesterol reduction of six daily packets of cholestyramine or placebo over a period of years. For each subject, the percentage of prescribed dose taken was estimated based on packet count and the opinion of the subject's doctors. We have summarized compliance status into a binary variable: compliers and non-compliers. Compliers are defined to have taken over 60% of their prescribed doses, while non-compliers took less than 60% of their prescribed doses. Note that it is assumed that subjects could only obtain the medication prescribed for them (e.g., placebo group members could not obtain cholestyramine). Success is defined as a cholesterol reduction of over 20 points. The data are shown below.

Treatment	Percent Doses Taken	Success	Failure
Placebo	$<= 60\%$	4	42
Placebo	$> 60\%$	28	98
Cholestyramine	$<= 60\%$	27	50
Cholestyramine	$> 60\%$	72	16

Example III-9: Encouragement Designs

From Hirano, Imbens, Rubin, Zhou. "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics*, 2000.

One of the ethical concerns in randomized studies is the issue of denying some individuals the treatment of interest. When it is not known if the new treatment is in fact better than the old (control) treatment, the experimenters are justified in randomly assigning individuals to receive treatment or control. However, when it is known that the new treatment is better for at least some individuals, and the interest is in examining the effect for a different group of people or in estimating the size of the effect, it is unethical to refuse the new, better treatment to some individuals.

To get around this, encouragement designs are used. In an encouragement design, one group is particularly encouraged to take the treatment of interest. It thus increases the use of the treatment in one group, without affecting the use of the treatment in the other group. An example of this might include an after school program for students, where all students have access to the program but only some receive a personalized letter encouraging them to attend. Encouragement designs are then analyzed in ways similar to randomized studies with noncompliance since subjects may or may not take the treatment that is being encouraged.

In this case, we are interested in estimating the effect of the influenza vaccine on flu-related hospitalizations for elderly patients. Since the flu vaccine is known to be effective, the experiment could not randomly assign some individuals to not receive this treatment. An encouragement design was thus implemented. Physicians were randomly selected to receive a computer generated reminder encouraging them to give their at risk patients a flu vaccine. The outcome of interest is flu related hospital visits.

There are also two covariates available: patient's age and whether they have chronic obstructive pulmonary disease (COPD). A summary of the data is shown below.

	No letter	Letter	No flu shot	Flu shot
Letter	0	1	0.475	0.631
Flu shot	0.19	0.307	0	1
Hospitalization	0.092	0.078	0.085	0.084
Age	65.0	65.4	64.7	66.8
COPD	0.29	0.277	0.264	0.343

We see that since receipt of the letter was randomized, the two covariates are well balanced between patients whose doctor received the letter and patients whose doctor did not. However, the covariates are not well balanced between patients who received a flu shot and those who didn't, due to noncompliance. We thus cannot simply compare the outcomes by flu shot status to obtain a reasonable estimate of the effect of the vaccine.

First we estimate the intention to treat (ITT) effect. This is an estimate of the causal effect of encouragement to get a flu shot on hospitalization and is estimated by comparing hospitalization rates among patients whose doctor received a letter and those whose doctor didn't receive a letter:

$$\widehat{ITT} = 0.092 - 0.078 = .014$$

This represents a $15\% = \frac{.078-.092}{.092}$ reduction in hospitalization rates due to encouragement to get flu shots.

Note that patients who have COPD are more likely to receive the vaccine than patients who do not have COPD. This implies that there is a link between treatment (vaccine) status and health, thus invalidating both an as treated analysis and a per protocol analysis.

To determine the causal effect of the vaccine on hospitalizations, we need to make a few assumptions. We define the following types of people:

Type	Assigned to (Z)	Treatment Received (D(Z))
Complier	Letter	Flu Shot
	No Letter	No flu shot
Never-taker	Letter	No flu shot
	No Letter	No flu shot
Always-taker	Letter	Flu shot
	No Letter	Flu shot
Defier	Letter	No flu shot
	No Letter	Flu shot

We do not observe each individual's full compliance status. We only observe their behavior under the observed assignment. To simplify the calculations, we make the assumption that there are no defiers. We are then able to identify some people as specific types. For example, someone whose doctor receives the letter and who does not get a flu shot must be a never-taker. Similarly, someone whose doctor does not receive the letter but who does get a flu shot must be an always-taker. For individuals who are not identified as a specific type, their compliance status is imputed using a model for compliance status, to be described below.

There are two other assumptions that make inference easier, but are not necessary. They are the following:

1. Exclusion restriction for never-takers: for never-takers, assignment to treatment does not affect their probability of flu related hospitalization.
2. Exclusion restriction for always-takers: for always-takers, assignment to treatment does not affect their probability of flu related hospitalization.

In this case, exclusion for never-takers seems more reasonable than exclusion for always-takers. For the always-takers, they get the shot either way, but their doctor receiving the letter might prompt them to receive other health benefits and greater awareness of the risks of the flu. They tend to be sicker than compliers and never-takers, and so receiving these extra benefits may affect their outcome. In addition, they may receive the vaccine earlier than they would have otherwise.

The never-takers are unlikely to receive other benefits from their doctor, since they aren't even receiving the flu vaccine. Assignment to letter is thus unlikely to directly affect their outcomes.

Under the predictive framework, either or both of these assumptions can be relaxed. The predictive framework involves positing a model for the "complete" data and using this model to predict individual's missing potential outcomes, i.e., outcomes under the treatment they didn't receive.

The complete data here are the observed outcomes Y_i (hospitalizations) and covariates X_i (X_{i1} = age, X_{i2} = COPD), as well as each individual's compliance type C_i (c = complier, a = always-taker, n = never-taker). The complete data model is formulated in terms of a model for the compliance types and a model for the outcomes, conditional on compliance types. Since the outcome is binary, it is modeled as a logistic regression with age and COPD as predictors. A separate logistic regression is posited for each combination of compliance type and treatment assigned. In other words, for each of the three possible compliance types, two logistic regressions are estimated: one for outcomes of people whose doctors received a letter, and one for outcomes of people whose doctors did not receive a letter.

$$P(Y_i(Z_i, D_i(Z_i)) = 1 | C_i = t, Z_i = z, X_{i1} = x_1, X_{i2} = x_2, \pi) = \Lambda(x_1, x_2, \beta_{tz})$$

where $\beta_{tz} = (\beta_{tz0}, \beta_{tz1}, \beta_{tz2})'$, $\Lambda(x_1, x_2, \beta_{tz}) = \frac{\exp(\beta_{tz0}+\beta_{tz1}\cdot x_1+\beta_{tz2}\cdot x_2)}{1+\exp(\beta_{tz0}+\beta_{tz1}\cdot x_1+\beta_{tz2}\cdot x_2)}$, for all $t \in \{c, n, a\}$ and $z = 0, 1$, and π represents all model parameters.

Compliance type is modeled as a multinomial logit with age and COPD as predictors.

$$\begin{aligned}
P(C_i = c | X_{i1} = x_1, X_{i2} = x_2, \pi) &= \Psi(c, x_1, x_2, \psi_c, \psi_n, \psi_a) \\
P(C_i = n | X_{i1} = x_1, X_{i2} = x_2, \pi) &= \Psi(n, x_1, x_2, \psi_c, \psi_n, \psi_a) \\
P(C_i = a | X_{i1} = x_1, X_{i2} = x_2, \pi) &= \Psi(a, x_1, x_2, \psi_c, \psi_n, \psi_a),
\end{aligned}$$

where, for $t \in \{c, n, a\}$, we have $\Psi(t, x_1, x_2, \psi_c, \psi_n, \psi_a) = \frac{\exp(\psi_{t0}+\psi_{t1}x_1+\psi_{t2}x_2)}{\sum_{v\in\{c,n,a\}} \exp(\psi_{v0}+\psi_{v1}x_1+\psi_{v2}x_2)}$.

These probabilities are normalized by setting $\psi_n = (\psi_{n0}, \psi_{n1}, \psi_{n2})$ equal to the three-dimensional vector of zeros.

Since compliance type is unknown for some subjects (i.e., those who received both a letter and a flu shot and those who received no letter and no flu shot), we treat the analysis as a missing data problem. Estimating the model parameters involves iterating between drawing compliance type for those patients whose compliance type is unknown, conditional on a current draw of the model parameters, and drawing the model parameters (i.e., logistic regression coefficients and multinomial logit coefficients) conditional on a current draw of missing compliance types. Model parameters are drawn from their posterior distribution conditional on compliance type. For individuals whose compliance type is unknown, there are two possible values of compliance type (individuals who received the letter and the flu shot are either always-takers or compliers and individuals who received no letter and no flu shot are either never-takers or compliers) and therefore compliance type is treated as a Bernoulli random variable. The probability of being either of the possible compliance types is proportional to the overall probability of being that compliance type (conditional on covariates) times the likelihood of the individual's data for that compliance type. For example, the probability of being a complier for someone whose doctor received the letter and who received a flu shot is proportional to $\Psi(c, X_{i1}, X_{i2}, \psi_c, \psi_n, \psi_a)\Lambda(X_{i1}, X_{i2}, \beta_{cZ_i})^{Y_i}(1 - \Lambda(X_{i1}, X_{i2}, \beta_{cZ_i}))^{1-Y_i}$ and the probability of being an always-taker is proportional to $\Psi(a, X_{i1}, X_{i2}, \psi_c, \psi_n, \psi_a)\Lambda(X_{i1}, X_{i2}, \beta_{aZ_i})^{Y_i}(1 - \Lambda(X_{i1}, X_{i2}, \beta_{aZ_i}))^{1-Y_i}$.

After the model is fit in this way, draws of the model parameters can be used to estimate the ITT effects. For each draw of the model parameters, missing potential outcomes are drawn, and the estimated ITT effects are simply the differences in means under the two assignments, classified by compliance type. The different exclusion restrictions are incorporated by constraining certain sets of logistic regression coefficients to be equal. For example, to include the exclusion restriction for never-takers, the logistic regression coefficients for never-takers assigned treatment and never-takers assigned control are set equal to each other, i.e., $\beta_{n0} = \beta_{n1}$. The following table summarizes the results obtained. The standard errors are shown in parentheses.

	Both excl. rest.	Excl. for never-takers	Excl. for always-takers	Neither excl. rest.
ITT_C	-0.082 (0.068)	-0.037 (0.078)	-0.196 (0.147)	-0.168 (0.161)
ITT_N	0	0	0.022 (0.026)	0.025 (0.027)
ITT_A	0	-0.053 (0.032)	0	-0.058 (0.033)
ITT	-0.010 (0.008)	-0.014 (0.008)	-0.009 (0.007)	-0.013 (0.008)

These results lead to a few interesting conclusions. Encouragement seems to have a similar beneficial effect on the always-takers as it does on the compliers. The exclusion restriction does not seem to hold for always-takers. This indicates that it may be encouragement to get the shot rather than the shot itself that is reducing flu related hospitalizations.

Causal Inference

Suggested Readings

1. Angrist, J. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review* 80: 313–335.

2. Angrist, J., Imbens, G.W. and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables (with discussion and rejoinder). *Journal of the American Statistical Association* 91: 444-472.

3. Angrist, J. and Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings. *Quarterly Journal of Economics* 106: 979–1014.

4. Barnard, J., Du, J., Hill, J. and Rubin, D.B. (1998). A Broader Template for Analyzing Broken Randomized Experiments. *Sociological Methods and Research* 27: 285–318.

5. Cornfield, J. et al. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22: 173–200.

6. Cox, D.R. (1958). *Planning of Experiments.* New York: Wiley. Chapters 1–3.

7. D'Agostino, R., Jr. and Rubin, D.B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association* 95: 749–759.

8. 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.

9. Ettner, S.L. (1996). The Timing of Preventive Services for Women and Children: The Effect of Having a Usual Source of Care. *American Journal of Public Health,* 86: 1748–1754.

10. Frangakis, C., and Rubin, D.B. (1999). Addressing Complications of Intention-To-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes. *Biometrika* 86: 365–379.

11. Frangakis, C. and Rubin, D.B. (2002). Principal Stratification in Causal Inference. *Biometrics* 58: 21-29.

12. Frangakis, C., Rubin, D.B. and Zhou, X. (2002), Clustered Encouragement Designs with Individual Noncompliance: Bayesian Inference with Randomization, and Application to Advance Directive Forms. *Biostatistics* 3: 147-164.

13. Goetghebeur, E. and Molenberghs, G. (1996). Causal Inference in a Placebo-Controlled Clinical Trial with Binary Outcome and Ordered Compliance. *Journal of the American Statistical Association* 435: 928–934.

14. Hill, J.L., Rubin, D.B., and Thomas, N. (2000). The Design of the New York School Choice Scholarships Program Evaluation. In *Research Designs: Donald Campbell's Legacy*, L. Bickman (ed.). Thousand Oaks, CA: Sage. Chapter 7, 155–180.

15. Hirano, K., Imbens, G., Rubin, D.B. and Zhou, X. (2000). Assessing the Effect of an Influenza Vaccine in an Encouragement Design. *Biostatistics* 1: 69–88.

16. Holland, P.W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* 81: 945–960.

17. Holland, P.W. and Rubin, D.B. (1983). On Lord's Paradox. Chapter 1 (pages 3–25) in *Principals of Modern Psychological Measurement*, ed. Wainer, H. and Messick, S. Hillsdale, NJ: Lawrence Erlbaum Associates.

18. Imbens, G. and Rubin, D.B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *Review of Economic Studies* 64: 555-574.

19. Imbens, G. and Rubin, D.B. (1997). Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. *The Annals of Statistics* 25: 305–327.

20. Little, R.J. and Rubin, D.B. (2000). Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health* 21:121–145.

21. McKim, V.R. and Turner, S.P. (1997). *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences.* Pages 23–80 (““Net Effects”: A Short History” by Stephen Turner, and “Searching for Causal Relations in Economic Statistics: Reflections from History” by Mary S. Morgan). Notre Dame, IN: University of Notre Dame Press.

22. Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments, Essay on Principles, Section 9. Translated in *Statistical Science* (1990) 5: 465–480.

23. Reinisch, J., Sanders, S., Mortensen, E. and Rubin, D. (1995). In Utero Exposure to Phenobarbital and Intelligence Deficits in Adult Men. *Journal of the American Medical Association* 274: 1518–1525.

24. Reiter, J. (2000). Using Statistics to Determine Causal Relationships. *The American Mathematical Monthly* 107: 24–32.

25. Roberts, S. (2001). Surprises from Self-Experimentation: Sleep, Mood, and Weight (with Discussion). *Chance* 14: 7–18.

26. Rosenbaum, P. and Rubin, D.B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society, Series B* 45: 212–218.

27. Rosenbaum, P. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70: 41–55.

28. Rosenbaum, P. and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 79: 516–524.

29. Rosenbaum, P. and Rubin, D.B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *American Statistician* 39: 33–38.

30. Rosenbaum, P. and Rubin, D.B. (1985). The Bias Due to Incomplete Matching. *Biometrics* 41: 103–116.

31. Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688–701.

32. Rubin, D.B. (1990). Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science* 5: 472–480.

33. Rubin, D.B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics* 46: 1213–1234.

34. Rubin, D.B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine* 127: 757–763.

35. Rubin, D.B. (2000). Statistical Inference for Causal Effects in Epidemiological Studies via Potential Outcomes. In *Atti Della XL Riunione Scientifica della Societa Italiana Di Statistica.* Roma: Societa Italiana di Statistica. Pages 419–430.

36. Rubin, D.B. (2000). Statistical Issues in the Estimation of the Causal Effects of Smoking Due to the Conduct of the Tobacco Industry. *Statistical Science in the Courtroom*, J.L. Gastwirth (ed). New York: Springer.

37. Rubin, D.B. (2001). Estimating the Causal Effects of Smoking. *Statistics in Medicine* 20: 1395–1414.

38. Rubin, D.B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services Outcome Research Methodology* 2: 169–188.

39. Rubin, D.B. and Thomas, N. (1992). Affinely Invariant Matching Methods with Ellipsoidal Distributions. *Annals of Statistics* 20: 1079–1093.

40. Rubin, D.B. and Thomas, N. (1992). Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika* 79: 797–809.

41. Rubin, D.B. and Thomas, N. (1996). Matching Using Estimated Propensity Scores, Relating Theory to Practice. *Biometrics* 52: 249–264.

42. Rubin, D.B. and Thomas, N. (2000). Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association* 95: 573–585.

43. Sommer, A. and Zeger, S.L. (1991). On Estimating Efficacy from Clinical Trials. *Statistics in Medicine* 10:45–52.