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

II. Causal inference based on the assignment mechanism

4. "Fisherian" significance levels in CR experiment
5. "Neymanian" repeated sampling evaluations in CR experiment
6. Extension to studies with variable but known propensities
7. Extension to studies with unknown propensities

III. Causal inference based on predictive distributions of potential outcomes

8. Predictive inference – intuition under ignorability
9. Formal predictive inference – Bayesian [Rubin,1978]
10. Examples when reliance on predictive approach is needed
11. Assessment of critical assumptions

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.

Part I: Framework

Example 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 2: Gain Scores

Potential Outcomes and Causal Effect with One Unit: In hypothetical example, the unit is you at a particular point in tie 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 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 - $\frac{76}{81}$ = 11%	68% - 11% = 57%

Example 4: Potential Outcomes with Two Units Allowing Interference Between Units

Potential Outcomes and Values in Example

You take:	Asp	Not	Asp	Not
I take:	Asp	Not	Asp	Not
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 5: 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 $\{Y_i$(Asp)$\}$ - Median $\{Y_i$(Not)$\}$

Perfect Doctor Example

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

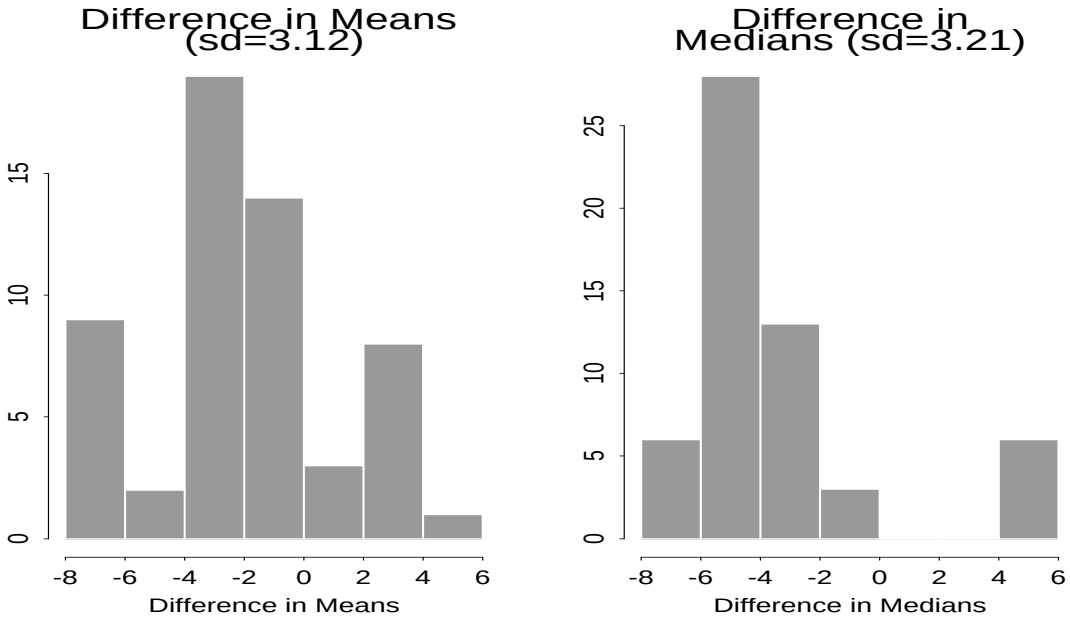
	Y(0)	Y(1)
	13	14
	6	0
	4	1
	5	2
	6	3
	6	1
	8	10
	8	9
True averages	7	5

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

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

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

w	$\overline{y_1} - \overline{y_0}$	median(y_1) - median(y_0)
11100000	-1.6	-6
11010000	-1.07	-6
11001000	-0.53	-6
11000100	-1.2	-6
11000010	2.2	-6
11000001	1.87	-6
10110000	-1.13	-5
10101000	-0.6	-5
10100100	-1.27	-5
10100010	2.13	-5
10100001	1.8	-5
10011000	-0.07	-4
10010100	-0.73	-4
10010010	2.67	-4
10010001	2.33	-4
10001100	-0.2	-2
10001010	3.2	-2
10001001	2.87	-2
10000110	2.53	-4
10000101	2.2	-4
10000011	5.6	5
01110000	-7.2	-5
01101000	-6.67	-5
01100100	-7.33	-5
01100010	-3.93	-5
01100001	-4.27	-5
01011000	-6.13	-4
01010100	-6.8	-4
01010010	-3.4	-4
01010001	-3.73	-4
01001100	-6.27	-2
01001010	-2.87	-2
01001001	-3.2	-2
01000110	-3.53	-4
01000101	-3.87	-4
01000011	-0.47	5
00111000	-6.2	-4
00110100	-6.87	-4
00110010	-3.47	-4
00110001	-3.8	-4
00101100	-6.33	-2
00101010	-2.93	-2
00101001	-3.27	-2
00100110	-3.6	-4
00100101	-3.93	-4
00100011	-0.53	5
00011100	-5.8	-1
00011010	-2.4	-1
00011001	-2.73	-1
00010110	-3.07	-3
00010101	-3.4	-3
00010011	0	6
00001110	-2.53	-3
00001101	-2.87	-3
00001011	0.53	6
00000111	-0.13	6
Average	-2	-2.8



Example: Observed Outcomes

W	Y(0)	Y(1)
1	?	14
0	6	?
0	4	?
0	5	?
0	6	?
0	6	?
1	?	10
1	?	9
Observed Averages	5.5	11

**Observed $\overline{y_1} - \overline{y_0} = 5.5 \neq -2$.

Example 2: Observed Outcomes

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

Example 3: Observed Outcomes

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

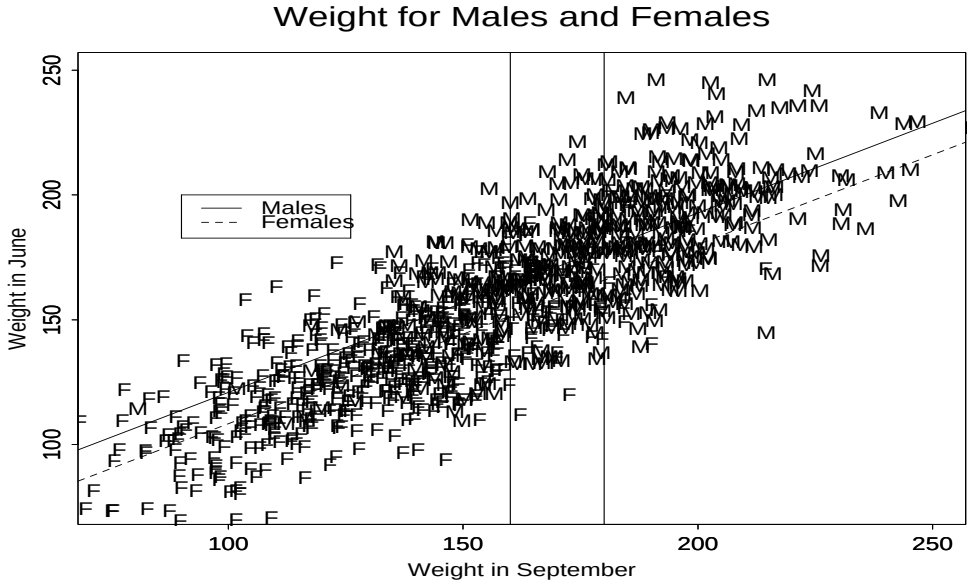
**Observed $\overline{y_1} - \overline{y_0} = -0.1$.

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	Weight Gain = Difference
< 100	1	62	114	101	13
100-119	6	103	122	117	5
120-139	21	142	143	132	12
140-159	90	127	156	142	14
160-179	152	51	172	157	15
180-199	144	14	187	174	13
> 200	86	1	204	171	33



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

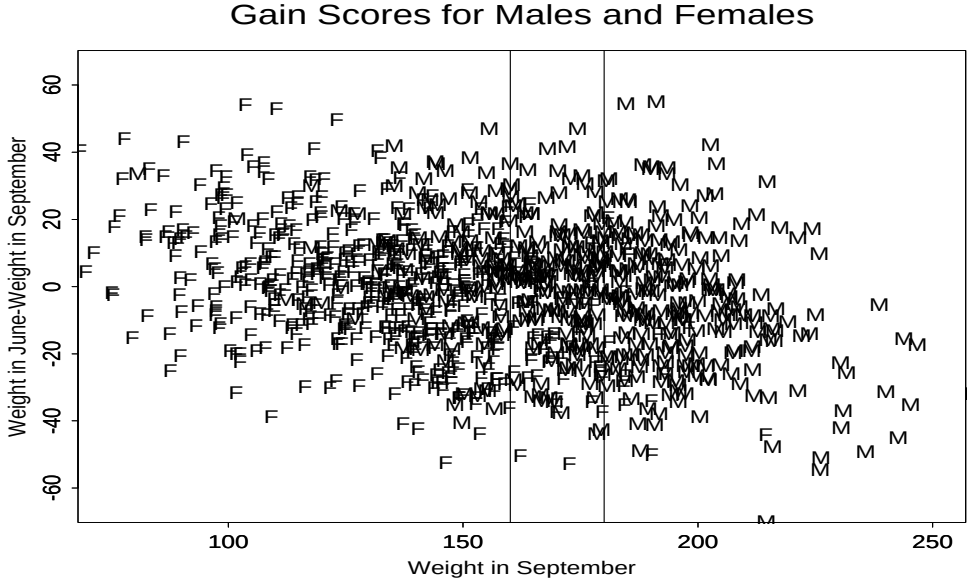
The average weight gain for Males was zero.
The average weight gain for Females was zero.

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.

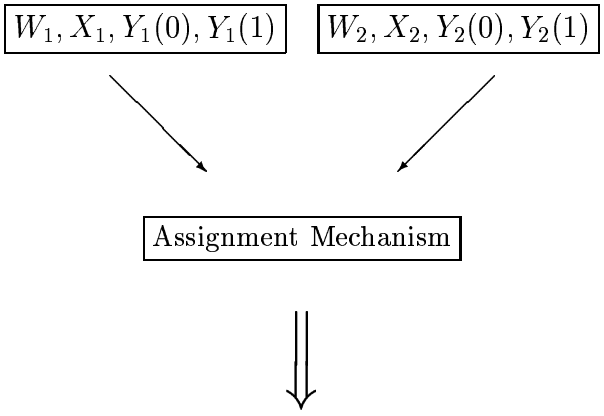
Statistician 2: Compare weight gain for males and females with the same weight in September.

> 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.



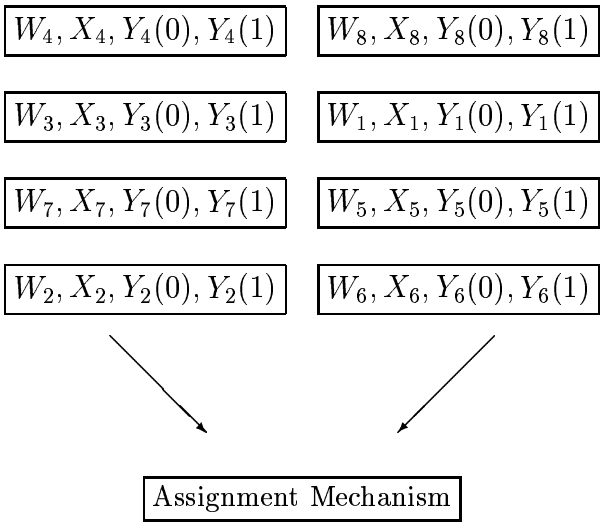
The Assignment Mechanism

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



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

Example 2: Completely Randomized Design with $N = 8$ units, 3 assigned treatment

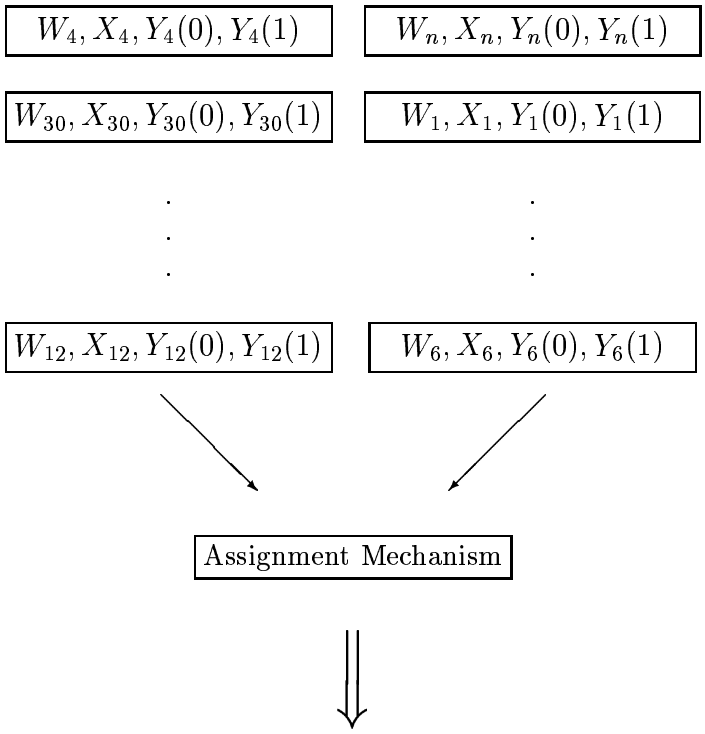


$\mathbf{W}$	$P(\mathbf{W})$
If $\sum_{i=1}^{8} W_i = 3$ [a]	$\frac{1}{56}$ [b]
If $\sum_{i=1}^{8} W_i \neq 3$	0

[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 3: Completely Randomized Design with N units, n assigned treatment



$\mathbf{W}$	$P(\mathbf{W})^a$
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

Example 4: 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

$\mathbf{W}$	$P(\mathbf{W})$
(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

$\mathbf{W}$	$P(\mathbf{W})$
(1,1,0,1,0,1)	0.5
(0,1,0,1,0,1)	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, as the assignment mechanism does not depend on the order of the units.

Examples of the Assignment Mechanism

Example 1: "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: $P(W_i = 1) = .5$. Remember that the overall assignment is the vector of all of the individual's 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:

$$P(\mathbf{W}) = P(W_1) * P(W_2) * P(W_3) * P(W_4) = \prod_{i=1}^{4} P(W_i).$$

All Possible Assignments

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

Example 2: "Bernoulli" (coin-tossing) assignment, 4 units

Same as Example 1, however now the probability of receiving treatment 1 for each individual is .4 ($P(W_i = 1) = .4$). Again, treatment is assigned independently for each individual.

All Possible Assignments

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

Example 3: 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

$\mathbf{W_M}$	$P(\mathbf{W_M})$
If $\sum_{i=1}^{4} W_i = 2$	$\binom{4}{2}^{-1}$
If $\sum_{i=1}^{4} W_i \neq 2$	0

Females

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

Overall $\mathbf{W}$	$P(\mathbf{W}\|\mathbf{Sex})$ [a]
If 2 males and 2 females have $W_i = 1$	$\binom{4}{2}^{-1} * \binom{4}{2}^{-1}$
Anything else	0

[a] Since the males and females are "independent blocks," the probability of an overall treatment assignment is the product of the probability of the males' treatment assignment and the probability of the females' treatment assignment: $P(\mathbf{W}|\mathbf{Sex}) = P(\mathbf{W_M}|\mathbf{Sex}) * P(\mathbf{W_F}|\mathbf{Sex})$.

Example 4: "Bernoulli" (coin-tossing) assignment within blocks

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

Unit	Sex	$P(W_i = 1\|\text{Sex}_i)$	$P(W_i = 0\|\text{Sex}_i)$
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,

$$P(\mathbf{W}|\mathbf{Sex}) = \prod_{i=1}^{8} P(W_i|\text{Sex}_i).$$

Possible Assignments

$\mathbf{W}$	$P(\mathbf{W}\|\mathbf{Sex})$		
$(0,0,0,0,0,0,0,0)$	$(.8)^4(.3)^4$	$=$	.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
$(1,1,0,0,0,0,0,0)$	$(.2)^2(.8)^2(.3)^4$	$=$	$\cdots$
$(1,0,1,0,0,0,0,0)$	$(.2)^2(.8)^2(.3)^4$	$=$	$\cdots$
...	...		
$(0,0,0,0,0,1,0,1)$	$(.8)^4(.7)^2(.3)^2$	$=$	$\cdots$
$(0,0,0,0,0,0,1,1)$	$(.8)^4(.7)^2(.3)^2$	$=$	$\cdots$
$(1,1,1,0,0,0,0,0)$	$(.2)^3(.8)^1(.3)^4$	$=$	$\cdots$
...	...		
$(0,0,0,0,0,1,1,1)$	$(.8)^4(.7)^3(.3)^1$	$=$	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.8)^2(.2)^2(.7)^2(.3)^2$	$=$	.001
...	...		
$(0,0,1,1,1,1,1,1)$	$(.8)^2(.2)^2(.7)^4$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,0)$	$(.2)^4(.7)^3(.3)^1$	$=$	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.2)^4(.7)^3(.3)^1$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.2)^4(.7)^4$	$=$	.0004

Example 5: Probability of treatment depends on age, $P(W_i = 1|\text{Age}_i) = \frac{\text{age}_i}{\text{age}_i+10}$

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

Unit	Age	$P(W_i = 1\|\text{Age}_i)$	$P(W_i = 0\|\text{Age}_i)$
1	15	.6	.4
2	22	.7	.3
3	18	.6	.4
4	54	.8	.2
5	34	.8	.2
6	77	.9	.1
7	38	.8	.2
8	91	.9	.1

Again, since treatment is assigned independently to each unit,

$$P(\mathbf{W}|\mathbf{Age}) = \prod_{i=1}^{8} P(W_i = 1|\text{Age}_i).$$

Possible Assignments

$\mathbf{W}$	$P(\mathbf{W}\|\mathbf{Age})$		
$(0,0,0,0,0,0,0,0)$	$(.4)(.3)(.4)(.2)(.2)(.1)(.2)(.1)$	=	.000004
$(1,0,0,0,0,0,0,0)$	$(.6)(.3)(.4)(.2)(.2)(.1)(.2)(.1)$	=	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.4)(.7)(.4)(.2)(.2)(.1)(.2)(.1)$	=	$\cdots$
...	...		
$(0,0,0,0,0,0,1,0)$	$(.4)(.3)(.4)(.2)(.2)(.1)(.8)(.1)$	=	$\cdots$
$(0,0,0,0,0,0,0,1)$	$(.4)(.3)(.4)(.2)(.2)(.1)(.2)(.9)$	=	$\cdots$
$(1,1,0,0,0,0,0,0)$	$(.6)(.7)(.4)(.2)(.2)(.1)(.2)(.1)$	=	$\cdots$
$(1,0,1,0,0,0,0,0)$	$(.6)(.3)(.6)(.2)(.2)(.1)(.2)(.1)$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,0,1)$	$(.4)(.3)(.4)(.2)(.2)(.9)(.2)(.9)$	=	$\cdots$
$(0,0,0,0,0,0,1,1)$	$(.4)(.3)(.4)(.2)(.2)(.1)(.8)(.9)$	=	.0001
$(1,1,1,0,0,0,0,0)$	$(.6)(.7)(.6)(.2)(.2)(.1)(.2)(.1)$	=	$\cdots$
...	...		
$(0,0,0,0,0,1,1,1)$	$(.4)(.3)(.4)(.2)(.2)(.9)(.8)(.9)$	=	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.6)(.7)(.4)(.2)(.8)(.9)(.2)(.1)$	=	$\cdots$
...	...		
$(0,0,1,1,1,1,1,1)$	$(.4)(.3)(.6)(.8)(.8)(.9)(.8)(.9)$	=	.03
...	...		
$(1,1,1,1,1,1,1,0)$	$(.6)(.7)(.6)(.8)(.8)(.9)(.8)(.1)$	=	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.6)(.7)(.6)(.8)(.8)(.9)(.2)(.9)$	=	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.6)(.7)(.6)(.8)(.8)(.9)(.8)(.9)$	=	.10

Example 6: 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)	$P(W_i = 1 \mid Y_i(0), Y_i(1))$	$P(W_i = 0 \mid Y_i(0), Y_i(1))$
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 treatment is assigned independently to each unit,

$$P(\mathbf{W}|\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \prod_{i=1}^{8} P(W_i = 1|Y_i(0), Y_i(1)).$$

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
$(1,0,0,0,0,0,0,0)$	$(.8)(.7)(.2)(.7)(.7)(.2)(.7)(.7)$	$=$	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.2)(.3)(.2)(.7)(.7)(.2)(.7)(.7)$	$=$	$\cdots$
...	...		
$(0,0,0,0,0,0,1,0)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.3)(.7)$	$=$	$\cdots$
$(0,0,0,0,0,0,0,1)$	$(.2)(.7)(.2)(.7)(.7)(.2)(.7)(.3)$	$=$	$\cdots$
$(1,1,0,0,0,0,0,0)$	$(.8)(.3)(.2)(.7)(.7)(.2)(.7)(.7)$	$=$	.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
...	...		
$(0,0,0,0,0,1,1,1)$	$(.2)(.7)(.2)(.7)(.7)(.8)(.3)(.3)$	$=$	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.8)(.3)(.2)(.7)(.3)(.8)(.7)(.7)$	$=$	$\cdots$
...	...		
$(0,0,1,1,1,1,1,1)$	$(.2)(.7)(.8)(.3)(.3)(.8)(.3)(.3)$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,0)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.3)(.7)$	$=$	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.7)(.3)$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.8)(.3)(.8)(.3)(.3)(.8)(.3)(.3)$	$=$	.001

Further Examples of the Assignment Mechanism

Example 1: "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	$P(W_i = 1\|Q_i)$	$P(W_i = 0\|Q_i)$
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

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

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

Example 2: "Bernoulli" (coin-tossing) assignment, 4 units (Extension of Example 2 from February 12)

We consider three assignment mechanisms:

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

b. Constrained so that 2 assigned to treatment 1 and 2 assigned to treatment 0, $P(W_i = 1) = .4$

c. Constrained so that 3 assigned to treatment 1 and 1 assigned to treatment 0, $P(W_i = 1) = .4$

All Possible Assignments

$\mathbf{W}$	$P(\mathbf{W})^a$	$P(\mathbf{W})^b$	$P(\mathbf{W})^c$
$(0,0,0,0)$	$(.6)^4$	0	0
$(0,0,0,1)$	$(.4)^1(.6)^3$	0	0
$(0,0,1,0)$	$(.4)^1(.6)^3$	0	0
$(0,0,1,1)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(0,1,0,0)$	$(.4)^1(.6)^3$	0	0
$(0,1,0,1)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(0,1,1,0)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(0,1,1,1)$	$(.4)^3(.6)^1$	0	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,0,0,0)$	$(.4)^1(.6)^3$	0	
$(1,0,0,1)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(1,0,1,0)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(1,0,1,1)$	$(.4)^3(.6)^1$	0	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,0,0)$	$(.4)^2(.6)^2$	$\frac{(.4)^2(.6)^2}{6(.4)^2(.6)^2} = \frac{1}{6}$	0
$(1,1,0,1)$	$(.4)^3(.6)^1$	0	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,1,0)$	$(.4)^3(.6)^1$	0	$\frac{(.4)^3(.6)^1}{4(.4)^3(.6)^1} = \frac{1}{4}$
$(1,1,1,1)$	$(.4)^4$	0	0

[a] Independent assignment
[b] Constrained so that 2 given treatment 1, 2 given treatment 0
[c] Constrained so that 3 given treatment 1, 1 given treatment 0

Example 3: 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.

$\mathbf{W}$	$P(\mathbf{W})$
(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 4: Bernoulli assignment, but probability depends on unobserved covariate ($u_i = Y_i(0)$)

A teacher randomly assigns children in her class to a new experimental reading program (treatment 1). Since she wants motivated children in this new program, in her mind she judges each student's motivation level 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 else the motivation levels of each student (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 thus done using the following rule: $P(W_i = 1|U_i) = .1 * U_i$, where U_i is the student's (unobserved) motivation level. ($\mathbf{U}$ is thus 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, their probability of assignment to treatment 1 is .95, and for students with motivation level 0, their probability of assignment to treatment 1 is .05:

$$P(W_i = 1|U_i) = \begin{cases} .05 & \text{if} \quad U_i = 0 \\ .1 * U_i & \text{if} \quad 0 < U_i < 10 \\ .95 & \text{if} \quad U_i = 10 \end{cases}$$

Student	U	$P(W_i = 1\|U_i)$	$P(W_i = 0\|U_i)$
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,

$$P(\mathbf{W}|\mathbf{U}) = \prod_{i=1}^{8} P(W_i|U_i).$$

	Possible Assignments		
$\mathbf{W}$	$P(\mathbf{W}\|\mathbf{U})$		
$(0,0,0,0,0,0,0,0)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.95)$	$=$	.0001
$(1,0,0,0,0,0,0,0)$	$(.4)(.2)(.8)(.3)(.2)(.05)(.5)(.95)$	$=$	$\cdots$
$(0,1,0,0,0,0,0,0)$	$(.6)(.8)(.8)(.3)(.2)(.05)(.5)(.95)$	$=$	$\cdots$
...	...		
$(0,0,0,0,0,0,1,0)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.95)$	$=$	$\cdots$
$(0,0,0,0,0,0,0,1)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.05)$	$=$	$\cdots$
$(1,1,0,0,0,0,0,0)$	$(.4)(.8)(.8)(.3)(.2)(.05)(.5)(.95)$	$=$	.0004
$(1,0,1,0,0,0,0,0)$	$(.4)(.2)(.2)(.3)(.2)(.05)(.5)(.95)$	$=$	$\cdots$
...	...		
$(0,0,0,0,0,1,0,1)$	$(.6)(.2)(.8)(.3)(.2)(.95)(.5)(.05)$	$=$	$\cdots$
$(0,0,0,0,0,0,1,1)$	$(.6)(.2)(.8)(.3)(.2)(.05)(.5)(.05)$	$=$	$\cdots$
$(1,1,1,0,0,0,0,0)$	$(.4)(.8)(.2)(.3)(.2)(.05)(.5)(.95)$	$=$	.00009
...	...		
$(0,0,0,0,0,1,1,1)$	$(.6)(.2)(.8)(.3)(.2)(.95)(.5)(.05)$	$=$	$\cdots$
$(1,1,0,0,1,1,0,0)$	$(.4)(.8)(.8)(.3)(.8)(.95)(.5)(.95)$	$=$	$\cdots$
...	...		
$(0,0,1,1,1,1,1,1)$	$(.6)(.2)(.2)(.7)(.8)(.95)(.5)(.05)$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,0)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.95)$	$=$	$\cdots$
$(1,1,1,1,1,1,0,1)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.05)$	$=$	$\cdots$
...	...		
$(1,1,1,1,1,1,1,1)$	$(.4)(.8)(.2)(.7)(.8)(.95)(.5)(.05)$	$=$	.0009

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	Date	Unconfounded?	Ignorable?
Bernoulli design, $P(W_i = 1) = .5$	Feb. 14 (Ex. 1)	Yes	Yes
Bernoulli design, $P(W_i = 1) = .4$	Feb. 14 (Ex. 2)	Yes	Yes
Completely randomized w/in blocks (gender)	Feb. 14(Ex. 3)	Yes	Yes
Bernoulli w/in blocks (gender)	Feb. 14 (Ex. 4)	Yes	Yes
Bernoulli, depends on age	Feb. 14 (Ex. 5)	Yes	Yes
Bernoulli, Perfect Doctor Version 2	Feb. 14 (Ex. 6)	No	No
Bernoulli, School choice	Feb. 19 (Ex. 1)	Yes	Yes
Bernoulli, Set # treated	Feb. 19 (Ex. 2)	Yes	Yes
Randomized w/in matched pairs	Feb. 19 (Ex. 3)	Yes	Yes
Bernoulli, Teacher	Feb. 19 (Ex. 4)	No	No

The following discussion 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.
- Two 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 outcomes of the previous newborns in the study.
   - 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. The following table summarizes this assignment mechanism. $Y(\text{ECMO}) = 1$ means the patient died under ECMO. $Y(\text{CMT}) = 0$ means the patient survived under CMT.

Newborn (time order)	$P(W_i = \text{ECMO})$	$P(W_i = \text{CMT})$	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	?

   - 11 infants received ECMO and all survived. 1 infant received CMT and died.
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 .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. (This description is somewhat simplified from the full design).
   - Data:

	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).
   - In non-randomized phase, only ECMO was given.

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 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 methods 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 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

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 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 1: 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	55	(55)
2	0	72	72	(72)
3	0	72.7	72.7	(72.7)
4	1	70	(70)	70
5	1	66	(66)	66
6	1	78.9	(78.9)	78.9

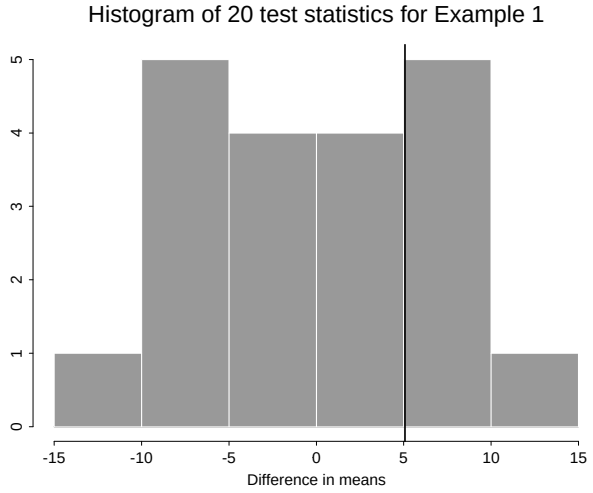
3. Observed estimate of the treatment effect: $\overline{y(1)} - \overline{y(0)} = 5.1$

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	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$		
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 2: Children's Television Workshop Part II

Same set-up as in Example 1, 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 is only appropriate if we assume additive treatment effects, i.e., the treatment adds a fixed amount to each control value.

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

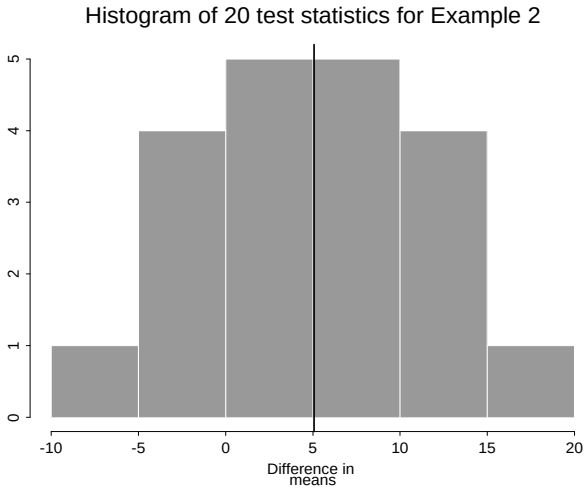
Unit	Actual Treatment (W)	Observed Outcome	Potential Outcomes $Y_i(0)$	$Y_i(1)$
1	0	55	55	(60)
2	0	72	72	(77)
3	0	72.7	72.7	(77.7)
4	1	70	(65)	70
5	1	66	(61)	66
6	1	78.9	(73.9)	78.9

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	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$	
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. The probability of observing the value that we did (5.1) or something more extreme is $10/20 = 0.5$ (i.e. the p-value or significance level is 0.5).



Constructing Fisher Confidence Intervals

Example 1: 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 (a confidence interval).

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	55	(55+x)
2	0	72	72	(72+x)
3	0	72.7	72.7	(72.7+x)
4	1	70	(70-x)	70
5	1	66	(66-x)	66
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	55	(49)
2	0	72	72	(66)
3	0	72.7	72.7	(66.7)
4	1	70	(76)	70
5	1	66	(72)	66
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	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$	
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	55	(67)
2	0	72	72	(84)
3	0	72.7	72.7	(84.7)
4	1	70	(58)	70
5	1	66	(54)	66
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	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$	
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 (smaller than 5.1) is $3/20 = .15$ (i.e. the p-value or significance level is 0.15).

To determine a confidence interval 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	.5	-3	.2
4	.4	-4	.2
3	.4	-5	.15
2	.35	-6	.15
1	.35	-7	.05
0	.3	-8	.05
-1	.3	-9	.05
-2	.3	-10	.05

We now consider hypothesized treatment effects greater than 5:

Treatment Effect	p-value	Treatment Effect	p-value
6	.5	16	.1
7	.45	17	.1
8	.4	18	.1
9	.35	19	.1
10	.3	20	.1
11	.25	21	.1
12	.15	22	.1
13	.15	23	.1
14	.15	24	.1
15	.15	25	.05

A plausible range of values for the treatment effect (a 90% confidence 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.

Biased/Unbiased Estimates of the Average Treatment Effect
Introduction to Neymanian Randomization-Based Inference

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.

Example 1: The all-knowing, but imperfect doctor

Consider a doctor who can look at each patient and know their potential outcome under both the standard surgery and a new experimental surgery (the outcome is the number of years lived past the surgery). The doctor can thus calculate the true treatment effect of the new surgery: $\overline{Y(1)} - \overline{Y(0)}$, the difference in means of the average outcomes under the old and new surgeries for all patients.

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

Although he knows every patient would do better under the new treatment (1) than under control (0), the doctor is interested in obtaining an estimate of the average treatment effect and assigns patients to each treatment randomly. He uses a completely randomized design, and assigns 4 to treatment and 2 to control. The following 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 he will only be able to observe one of these possible randomizations.

All Possible Assignments

W	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$			median$(y(1))$-median$(y(0))$
1 1 1 1 0 0	1/15	2			2
1 1 1 0 1 0	1/15	-1			-1
1 1 1 0 0 1	1/15	-0.2	=	$\frac{10+3+5+9}{4} - \frac{9+5}{2}$	0
1 1 0 1 1 0	1/15	4.2			5
1 1 0 1 0 1	1/15	5			6
1 1 0 0 1 1	1/15	2			3
1 0 1 1 1 0	1/15	5.8			6
1 0 1 1 0 1	1/15	6.5	=	$\frac{10+5+12+9}{4} - \frac{0+5}{2}$	7
1 0 1 0 1 1	1/15	3.5			4
1 0 0 1 1 1	1/15	8.8			8.5
0 1 1 1 1 0	1/15	0.5			0
0 1 1 1 0 1	1/15	1.2			1
0 1 1 0 1 1	1/15	-1.8	=	$\frac{3+5+8+9}{4} - \frac{7+9}{2}$	-1.5
0 1 0 1 1 1	1/15	3.5			4
0 0 1 1 1 1	1/15	5.0			5
Average		3.0			3.3

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. 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	1	2
3	5	1	4
4	12	7	5
5	8	7	1
6	9	6	3
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	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$			median(y(1)) − median(y(0))
1 1 1 1 0 0	1/15	1			1
1 1 1 0 1 0	1/15	0			0
1 1 1 0 0 1	1/15	-0.25	=	$\frac{10+3+5+9}{4} - \frac{7+7}{2}$	0
1 1 0 1 1 0	1/15	4.75			5.5
1 1 0 1 0 1	1/15	4.5			5.5
1 1 0 0 1 1	1/15	3.5			4.5
1 0 1 1 1 0	1/15	5.25	=	$\frac{10+5+12+8}{4} - \frac{1+6}{2}$	5.5
1 0 1 1 0 1	1/15	5			5.5
1 0 1 0 1 1	1/15	4			4.5
1 0 0 1 1 1	1/15	8.75			8.5
0 1 1 1 1 0	1/15	0.5			0
0 1 1 1 0 1	1/15	0.25			0
0 1 1 0 1 1	1/15	-0.75	=	$\frac{3+5+8+9}{4} - \frac{7+7}{2}$	-0.5
0 1 0 1 1 1	1/15	4			4.5
0 0 1 1 1 1	1/15	4.5			4.5
Average		3.0			3.3

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 ($3.5 = 8.5 - 5$).

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	10	0
2	3	3	0
3	5	5	0
4	12	3	9
5	8	8	0
6	9	0	9
Average	7.83	4.83	3

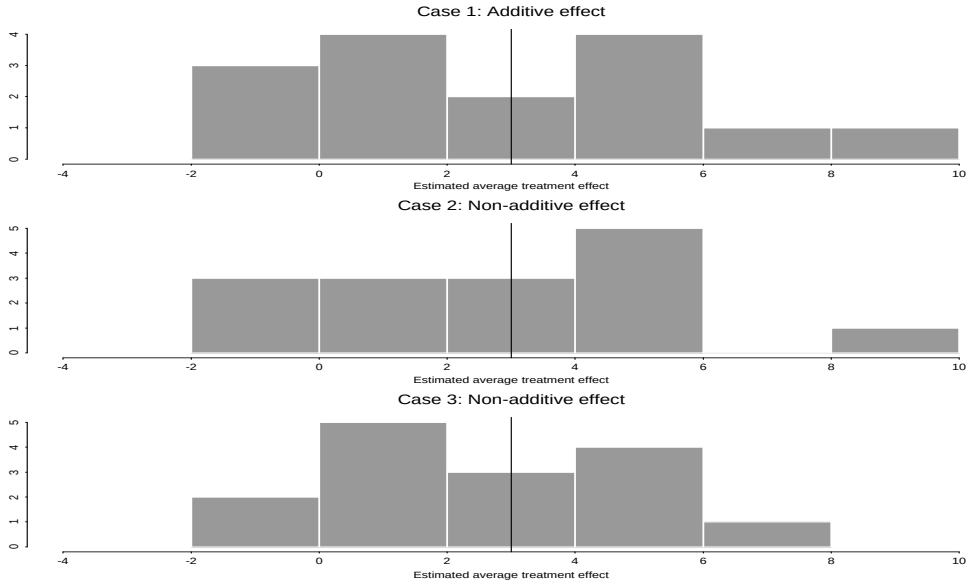
The table below again lists all possible randomizations and the corresponding observed estimates of the treatment effect.

All Possible Assignments

W	$P(\mathbf{W})$	$\overline{y(1)} - \overline{y(0)}$			median(y(1)) − median(y(0))
1 1 1 1 0 0	1/15	3.5			3.5
1 1 1 0 1 0	1/15	5			5
1 1 1 0 0 1	1/15	1.25	=	$\frac{10+3+5+9}{4} - \frac{3+8}{2}$	1.5
1 1 0 1 1 0	1/15	5.75			6.5
1 1 0 1 0 1	1/15	2			3
1 1 0 0 1 1	1/15	3.5			4.5
1 0 1 1 1 0	1/15	7.25	=	$\frac{10+5+12+8}{4} - \frac{3+0}{2}$	7.5
1 0 1 1 0 1	1/15	3.5			4
1 0 1 0 1 1	1/15	5			5.5
1 0 0 1 1 1	1/15	5.75			5.5
0 1 1 1 1 0	1/15	2.0			1.5
0 1 1 1 0 1	1/15	-1.75			-2
0 1 1 0 1 1	1/15	-.25	=	$\frac{3+5+8+9}{4} - \frac{10+3}{2}$	0
0 1 0 1 1 1	1/15	0.5			1
0 0 1 1 1 1	1/15	2			2
Average		3.0			3.3

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 ($4.5 = 8.5 - 4$).

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



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.

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 1 (cont.): The all-knowing, but imperfect doctor

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.

All Possible Assignments

W	$P(\mathbf{W})$	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)	0.6	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	2.6	(-4.0, 3.5)	.21	.92
1 1 0 1 1 0	1/15	14.9	8	2.8	(.2, 8.3)	.65	.13
1 1 0 1 0 1	**1/15**	**15**	**4.5**	**2.4**	**(1.5, 8.5)**	**.41**	**.04**
1 1 0 0 1 1	1/15	9.7	24.5	3.8	(-3.5, 7.5)	.79	.60
1 0 1 1 1 0	1/15	8.9	18	3.4	(0.9, 10.6)	.41	.09
1 0 1 1 0 1	1/15	8.7	12.5	2.9	(2.3, 10.7)	.23	.03
1 0 1 0 1 1	1/15	4.7	40.5	4.6	(-3.2, 10.2)	.91	.45
1 0 0 1 1 1	1/15	2.9	2	1.3	(6.9, 10.6)	.00001	.00001
0 1 1 1 1 0	1/15	15.3	0.5	2.0	(-2.4, 3.4)	.22	.80
0 1 1 1 0 1	1/15	16.3	2	2.3	(-2.0, 4.5)	.44	.58
0 1 1 0 1 1	1/15	7.6	2	1.7	(-4.2, 0.7)	.005	.30
0 1 0 1 1 1	1/15	14	12.5	3.1	(-1.0, 8.0)	.87	.26
0 0 1 1 1 1	1/15	8.3	24.5	3.8	(-0.5, 10.5)	.60	.19

[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 13 out of the 15 intervals contain the true treatment effect (3): $\frac{13}{15} = .87$.

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

86% Confidence 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	1	2
3	5	1	4
4	12	7	5
5	8	7	1
6	9	6	3
Average	7.83	4.83	3
Variance	11.0	9.0	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	$P(\mathbf{W})$	Var(y(1)) $= s_1^2$	Var(y(0)) $= s_0^2$	$\widehat{\text{SD(ATE)}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$	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	(-2.1, 4.1)	.35	.64
1 1 1 0 1 0	1/15	9.7	0.5	1.6	(-2.4, 2.4)	.07	1.0
1 1 1 0 0 1	1/15	10.9	0	1.7	(-2.6, 2.1)	.05	.88
1 1 0 1 1 0	1/15	14.9	12.5	3.2	(0.2, 9.3)	.58	.13
1 1 0 1 0 1	1/15	15	18	3.6	(-0.6, 9.6)	.67	.21
1 1 0 0 1 1	1/15	9.7	18	3.4	(-1.4, 8.4)	.88	.30
1 0 1 1 1 0	1/15	8.9	12.5	2.9	(1.1, 9.4)	.44	.07
1 0 1 1 0 1	1/15	8.7	18	3.3	(0.2, 9.8)	.55	.13
1 0 1 0 1 1	1/15	4.7	18	3.2	(-0.6, 8.6)	.75	.21
1 0 0 1 1 1	1/15	2.9	0	0.9	(7.5, 10.0)	.00001	.00001
0 1 1 1 1 0	1/15	15.3	0.5	2.0	(-2.4, 3.4)	.22	.80
0 1 1 1 0 1	**1/15**	**16.3**	**0**	**2.0**	**(-2.7, 3.2)**	**.17**	**.90**
0 1 1 0 1 1	1/15	7.6	0	1.4	(-2.7, 1.2)	.006	.59
0 1 0 1 1 1	1/15	14	18	3.5	(-1.1, 9.1)	.78	.26
0 0 1 1 1 1	1/15	8.3	18	3.3	(-0.3, 9.3)	.65	.18

[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 11 out of the 15 intervals contain the true treatment effect (3): $\frac{11}{15} = .73$.

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

86% Confidence Interval:	[-4,6]
p-value for null of no treatment effect:	$\frac{8}{15} = .53$
p-value for null of treatment effect of 3:	$\frac{5}{15} = .27$

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	10	0
2	3	3	0
3	5	5	0
4	12	3	9
5	8	8	0
6	9	0	9
Average	7.83	4.83	3
Variance	11.0	13.4	21.6

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	$P(\mathbf{W})$	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	32	4.5	(-3.0, 10.0)	.91	.44
1 1 1 0 1 0	1/15	9.7	4.5	2.2	(1.9, 8.1)	.35	.02
1 1 1 0 0 1	1/15	10.9	12.5	3.0	(-3.1, 5.6)	.56	.68
1 1 0 1 1 0	1/15	14.9	12.5	3.2	(1.2, 10.3)	.38	.07
1 1 0 1 0 1	1/15	15	4.5	2.4	(-1.5, 5.5)	.68	.41
1 1 0 0 1 1	1/15	9.7	2	1.8	(0.8, 6.2)	.79	.06
1 0 1 1 1 0	**1/15**	**8.9**	**4.5**	**2.1**	**(4.2, 10.3)**	**.04**	**.0006**
1 0 1 1 0 1	1/15	8.7	12.5	2.9	(-0.7, 7.7)	.86	.23
1 0 1 0 1 1	1/15	4.7	0	1.1	(3.4, 6.6)	.06	.000003
1 0 0 1 1 1	1/15	2.9	2	1.3	(3.9, 7.6)	.04	.00001
0 1 1 1 1 0	1/15	15.3	50	5.4	(-5.7, 9.7)	.85	.71
0 1 1 1 0 1	1/15	16.3	2	2.3	(-5.0, 1.5)	.03	.44
0 1 1 0 1 1	1/15	7.6	24.5	3.8	(-5.7, 5.2)	.39	.95
0 1 0 1 1 1	1/15	14	12.5	3.1	(-4.0, 5.0)	.42	.87
0 0 1 1 1 1	1/15	8.3	24.5	3.8	(-3.5, 7.5)	.79	.60

[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 11 out of the 15 intervals contain the true treatment effect (3): $\frac{11}{15} = .73$.

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

86% Confidence Interval:	[2,12]
p-value for null of no treatment effect:	$\frac{1}{15} = .07$
p-value for null of treatment effect of 3:	$\frac{2}{15} = .13$

Fisher Tests with Unequal Assignment Probabilities

Example 1: Unknown assignment probabilities

Treatment ($W = 1$) is going to a step aerobics class twice a week for six months. Control ($W = 0$) is never attending step class. The observed outcome is cholesterol after one year. For each unit, we observe W and cholesterol, as well as gender. We do not, however, know the probability of W for each individual.

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

We want to determine whether step aerobics leads to decreased cholesterol levels. However, in order to do a Fisher randomization test we need to be able to calculate the assignment probabilities, and we are not given these probabilities. Furthermore, it appears that the assignment probabilities may be different for males and females: females appear more likely than males to receive treatment.

From the observed data, we see that one male received treatment and two males received control, and that three females received treatment and one female received control. We assume Bernoulli treatment assignment with different assignment probabilities for males and females. Based on the observed data, we guess that the probability of being assigned treatment is $\frac{1}{3}$ for males and $\frac{3}{4}$ for females. The implicit assumption is that treatment assignment is unconfounded given each individual's sex. In other words, we assume the treatment probabilities are as follows:

Unit	Gender	$P(W_i = 1 \mid \text{Gender})$	$P(W_i = 0 \mid \text{Gender})$
1	M	.33	.67
2	M	.33	.67
3	M	.33	.67
4	F	.75	.25
5	F	.75	.25
6	F	.75	.25
7	F	.75	.25

We can now carry out a Fisher test based on these estimated assignment probabilities. We will restrict ourselves to only the $\binom{3}{1}\binom{4}{3} = 12$ assignments in which one male and three females receive treatment and two males and one female receive control.

W	$P(\mathbf{W})$	$P^*(\mathbf{W})$
1001110	.016	.083
1001101	.016	.083
1001011	.016	.083
1000111	.016	.083
0101110	.016	.083
0101101	.016	.083
0101011	.016	.083
0100111	.016	.083
0011110	.016	.083
0011101	**.016**	**.083**
0011011	.016	.083
0010111	.016	.083

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, two 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 them combine the two estimates weighted by the sample sizes.

Example: Calculating the observed treatment effect:

1. Calculate the difference in means for males: $\overline{y(1)}_M - \overline{y(0)}_M = \frac{250}{1} - \frac{241+310}{2} = -25$.
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, weighting by the sample sizes: Average treatment effect $= \frac{(-25)*3+(-50)*4}{7} = -39.29$.

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

W	$P^*(\mathbf{W})$	Estimated Treatment Effect
1001110	.083	-7.62
*1001101	.083	-45.7
1001011	.083	-22.86
1000111	.083	7.62
0101110	.083	37.38
0101101	.083	-0.71
0101011	.083	22.14
0100111	.083	52.62
0011110	.083	-1.19
***0011101**	**.083**	**-39.29**
0011011	.083	-16.43
0010111	.083	14.04

The observed test statistic is -39.29. There are two assignments (the starred ones) that give test statistics as extreme or more extreme than -39.29. We calculate the p-value by summing the probabilities of these four assignments:

$$\text{p-value} = .083 + .083 = .16.$$

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

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:

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

Horwitz and Thompson showed that this quantity is an unbiased estimator of the difference in means under the treatment and control conditions.

Example 2: Bernoulli assignment where probability of treatment is known and depends on age:

$$P(W_i = 1|\text{Age}_i) = \frac{\text{age}_i}{\text{age}_i+10}$$

$$\mathbf{Age} = (15,22,18,54,34,77).$$

Unit	Age	$P(W_i = 1 \mid \text{Age}_i)$	$P(W_i = 0 \mid \text{Age}_i)$
1	15	.60	.40
2	22	.69	.31
3	18	.64	.36
4	54	.84	.16
5	34	.77	.23
6	77	.86	.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	$P(\mathbf{W})$	$P^*(\mathbf{W})$
111100	.004	.01
111010	.004	.01
111001	.01	.03
110110	.01	.03
110101	.02	.08
110011	.02	.08
101110	.01	.02
101101	.02	.05
101011	.02	.05
100111	**.04**	**.13**
011110	.01	.03
011101	.02	.08
011011	.02	.08
010111	.06	.20
001111	.04	.13

$P(W)$ and $P(W^*)$ are calculated as in previous examples. The bold row is the observed assignment.

Fisher Test:

Example: Calculating the observed treatment effect.

$$\frac{1}{6} * (\frac{9}{.60} + \frac{6}{.84} + \frac{10}{.44} + \frac{15}{.86}) - \frac{1}{6} * (\frac{11}{.31} + \frac{2}{.36}) = 1.92.$$

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	$P^*(\mathbf{W})$	Estimated Treatment Effect
111100	.01	-18.24
111010	.01	-16.26
111001	.03	4.91
110110	.03	-10.27
110101	.08	1.08
* 110011	.08	3.05
101110	.02	-17.40
101101	.05	-6.04
101011	.05	4.07
*** 100111**	**.13**	**1.92**
011110	.03	-15.07
011101	.08	-3.72
011011	.08	-1.75
* 010111	.20	4.24
001111	.13	-2.88

The observed test statistic is 1.92. There are three assignments (the starred ones) that give test statistics as large or larger than 3.5. We calculate the p-value by summing the probabilities of these four assignments:

p-value $= .08 + .13 + .20 = .41.$

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

A 74% confidence 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% ($1 - .13 \times 2 = .74$).

Case Study on Propensity Scores and Matching: 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.

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 a randomized experiment, 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 are in large, prestigious 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 suffer from "selection bias" in the sense that their patients and doctors are not a random sample from 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.
   - Compare like with like: choose patients from SEER who are comparable to those in randomized experiments (similar based on year of treatment, geographic area, tumor size, age, marital status, race or ethnicity).
   - Use "propensity scores" to 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.

Part III: Predictive Inference

Example 1: Age of Death

A doctor is conducting an experiment (again) to determine which of two types of surgery is better in terms of leading to increased life. The control is the standard surgery and treatment is a new surgery he just developed. He recruits twenty study participants and assigns half to treatment and half to control. The design is completely randomized. The outcome measured is age of death. The doctor also measures patients' age at the time of surgery. He observes the following data (the covariate X_{age} is age at the time of surgery):

X_{age}	W	$Y(0)$	$Y(1)$
54	0	62	
46	0	53	
38	0	42	
53	0	60	
48	0	55	
34	0	37	
27	0	28	
44	0	51	
27	0	33	
43	0	51	
48	1		58
59	1		71
44	1		51
44	1		55
47	1		60
27	1		40
61	1		75
39	1		48
45	1		54
44	1		58

The doctor computes a Neyman confidence interval for the observed average treatment effect: $\hat{ATE} = \overline{y_1} - \overline{y_0} = 57 - 47.2 = 9.8$. $s_1^2 = 105.55$ and $s_0^2 = 134.18$. $\hat{VAR}(\hat{ATE}) = \frac{108.26}{10} + \frac{134.18}{10} = 23.97$. The 95% confidence interval $\hat{ATE} \pm 1.96 * \sqrt{\hat{VAR}}$=(.2, 19.4). He also conducts a Fisher randomization test of the sharp null hypothesis of zero treatment effect, and the p-value is .03. The Fisher confidence interval, obtained by inverting the Fisher test, is (2, 18). Neither of the confidence intervals contain zero, so the doctor concludes that his new surgery does increase length of life. However, both intervals are fairly wide, so the doctor doesn't have a very good idea of how much longer patients can expect to live under the new surgery.

Imagine that the doctor is magically able to know each patient's outcome under both treatment and control. This data is shown below. The covariate X_{age} is age at the time of surgery. We also show the treatment effect for each individual, calculated once using the raw outcomes $(y_i(1) - y_i(0))$ and once using the gain scores $((y_i(1) - x_i) - (y_i(0) - x_i))$.

X_{age}	W	$Y(0)$	$Y(1)$	$Y(0) - X$	$Y(1) - X$	$Y(1) - Y(0)$	$(Y(1) - X) - (Y(0) - X)$
54	0	62	69	8	15	7	7
46	0	53	60	7	14	7	7
38	0	42	49	4	11	7	7
53	0	60	67	7	14	7	7
48	0	55	62	7	14	7	7
34	0	37	44	3	10	7	7
27	0	28	35	1	8	7	7
44	0	51	58	7	14	7	7
27	0	33	40	6	13	7	7
43	0	51	58	8	15	7	7
48	1	51	58	3	10	7	7
59	1	64	71	5	12	7	7
44	1	44	51	0	7	7	7
44	1	48	55	4	11	7	7
47	1	53	60	6	13	7	7
27	1	33	40	6	13	7	7
61	1	68	75	7	14	7	7
39	1	41	48	2	9	7	7
45	1	47	54	2	9	7	7
44	1	51	58	7	14	7	7

We see that using the raw difference in outcomes or using the gain scores leads to the same value of the treatment effect.

Example 2 (Example 1 Continued): Years Lived after Surgery

The doctor then realizes that he did the above analysis ignoring the information on age at the time of surgery. He realizes he is actually more interested in how long people live after the treatment than in how long they live overall. He can use the covariate information to do another analysis: now instead of using the outcome age of death, he subtracts the age at the time of surgery from age of death to get years lived after surgery, and does the analysis on the new outcome, Y^*. In other words, $Y^* = Y - X$. The "new" data are given below.

W	$Y^*(0)$	$Y^*(1)$
0	8	
0	7	
0	4	
0	7	
0	7	
0	3	
0	1	
0	7	
0	6	
0	8	
1		10
1		12
1		7
1		11
1		13
1		13
1		14
1		9
1		9
1		14

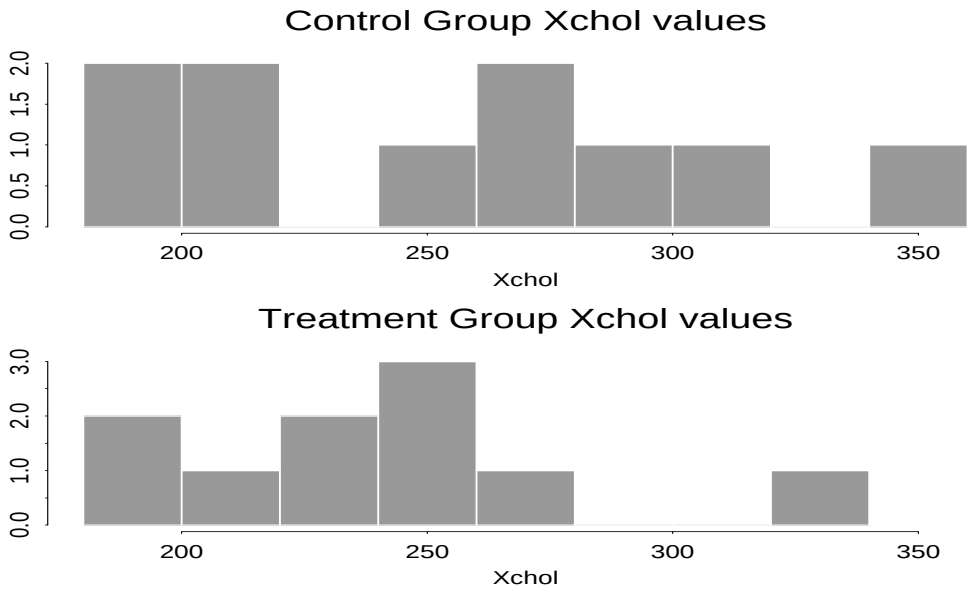
Analyzing these data, the doctor calculates that $\hat{ATE} = 5.4$ and $\hat{VAR}(\hat{ATE}) = 1.12$. A 95% Neymanian confidence interval for the treatment effect is then (3.3, 7.5). The Fisher p-value for the sharp null hypothesis of zero treatment effect is 0.00001 and the Fisher confidence interval is (4,7). These intervals are much more narrow than the original intervals he calculated using age of death as the outcome, so he now has a much sharper idea of how much longer people live on average after receiving the new surgery versus the old surgery.

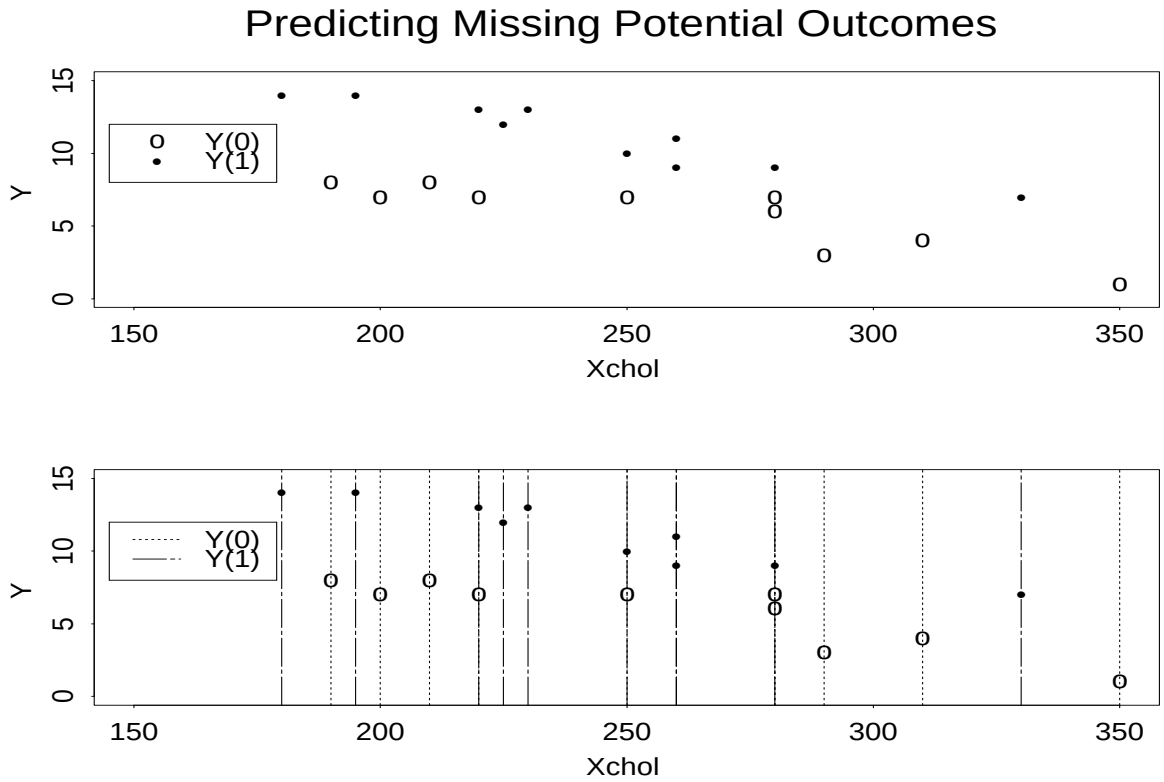
Example 3 (Examples 1, 2 continued): Cholesterol and donor pools for matching

The doctor begins looking through his records, and realizes that he has another possibly relevant piece of information. He recorded the cholesterol level of each patient prior to the surgery (the covariate X_{chol} represents cholesterol prior to surgery).

X_{chol}	W	$Y^*(0)$	$Y^*(1)$
210	0	8	
200	0	7	
310	0	4	
220	0	7	
280	0	7	
290	0	3	
350	0	1	
250	0	7	
280	0	6	
190	0	8	
250	1		10
225	1		12
330	1		7
260	1		11
230	1		13
220	1		13
180	1		14
280	1		9
260	1		9
195	1		14

The doctor took account of age by subtracting age at surgery from age at death. How can he "subtract off" cholesterol from years lived after surgery?


Control Group Xchol values
0.0
0.5
1.0
1.5
2.0
200
250
300
350
Xchol
Treatment Group Xchol values
0.0
1.0
2.0
3.0
200
250
300
350
Xchol


Predicting Missing Potential Outcomes
Y
0
5
10
15
Y(0)
Y(1)
150
200
250
300
350
Xchol
Y
0
5
10
15
Y(0)
Y(1)
150
200
250
300
350
Xchol

Unit	X_{chol}	W	$Y^*(0)$	$Y^*(1)$	Units for Donor Pool	Donor $Y(0)$	Donor $Y(1)$
1	210	0	8		20,16,12,15		14,13,12,13
2	200	0	7		17,20,16,12		14,14,13,12
3	310	0	4		14,19,18,13		11,9,9,7
4	220	0	7		20,16,12,15		14,13,12,13
5	280	0	7		11,14,19,18		10,11,9,9
6	290	0	3		11,14,19,18,13*		10,11,9,9,7
7	350	0	1		14,19,18,13		11,9,9,7
8	250	0	7		15,11,14,19		13,10,11,9
9	280	0	6		11,14,19,18		10,11,9,9
10	190	0	8		17,20,16,12		14,14,13,12
11	250	1		10	4,8,5,9	7,7,7,6	
12	225	1		12	2,1,4,8	7,8,7,7	
13	330	1		7	5,9,6,3,7*	7,6,3,4,1	
14	260	1		11	8,5,9,6	7,7,6,3	
15	230	1		13	2,1,4,8	7,8,7,7	
16	220	1		13	10,2,1,4,8*	8,7,8,7,7	
17	180	1		14	10,2,1,4	8,7,8,7	
18	280	1		9	8,5,9,6,3*	7,7,6,3,4	
19	260	1		9	8,5,9,6	7,7,6,3	
20	195	1		14	10,2,1,4	8,7,8,7	

Noticing that the values of the covariate (cholesterol level) overlap between the treatment and control groups, we wish to obtain an estimate of the treatment effect by using matching on this covariate. 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).

We define the donor pool as the units in the other treatment group with the four closest values of the covariate. This is shown on the right hand side of the above table. The values shown are the cholesterol level and the values in parentheses are the unit numbers. In pools marked by a star, five values were chosen because it was impossible to pick the four that were the closest (there was a "tie" in closeness).

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. The imputed values are in parentheses. We can then calculate each individual's treatment effect and obtain an estimate of the average treatment effect. This is done repeatedly to examine the variability in the estimate.

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	4	(9)	5
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(9)	6
7	350	0	1	(11)	10
8	250	0	7	(13)	6
9	280	0	6	(11)	5
10	190	0	8	(14)	6
11	250	1	(7)	10	3
12	225	1	(8)	12	4
13	330	1	(3)	7	4
14	260	1	(6)	11	5
15	230	1	(8)	13	5
16	220	1	(7)	13	6
17	180	1	(7)	14	7
18	280	1	(6)	9	3
19	260	1	(7)	9	2
20	195	1	(8)	14	6
Average					**5.15**
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	4	(11)	7
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(10)	7
7	350	0	1	(7)	6
8	250	0	7	(11)	5
9	280	0	6	(9)	3
10	190	0	8	(14)	6
11	250	1	(7)	10	3
12	225	1	(7)	12	5
13	330	1	(4)	7	3
14	260	1	(3)	11	8
15	230	1	(7)	13	6
16	220	1	(7)	13	6
17	180	1	(8)	14	6
18	280	1	(3)	9	6
19	260	1	(3)	9	6
20	195	1	(8)	14	6
Average					**5.45**
Median					**6**

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	4	(7)	3
4	220	0	7	(13)	6
5	280	0	7	(9)	2
6	290	0	3	(11)	8
7	350	0	1	(9)	8
8	250	0	7	(13)	6
9	280	0	6	(9)	3
10	190	0	8	(13)	5
11	250	1	(7)	10	3
12	225	1	(8)	12	4
13	330	1	(4)	7	3
14	260	1	(7)	11	4
15	230	1	(7)	13	6
16	220	1	(7)	13	6
17	180	1	(8)	14	6
18	280	1	(3)	9	6
19	260	1	(7)	9	2
20	195	1	(8)	14	6
Average					**4.9**
Median					**5.5**

Imputation 4:

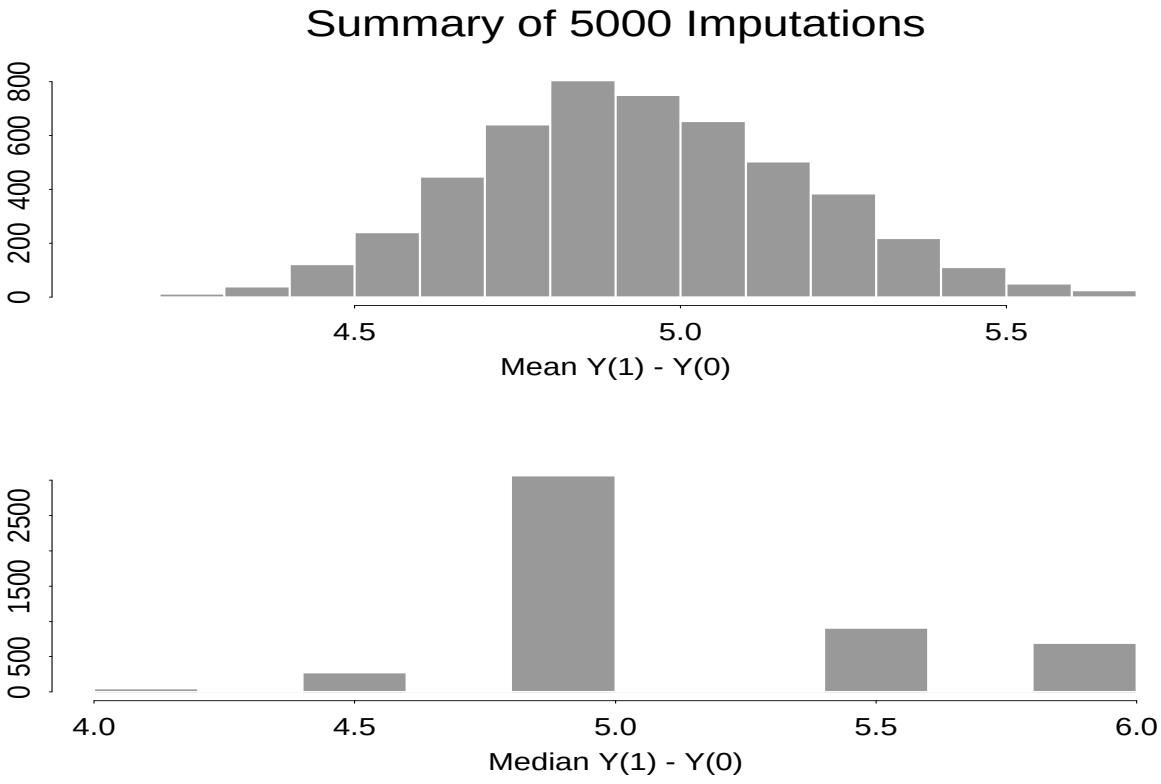
Unit	X_{chol}	W	$Y^*(0)$	$Y^*(1)$	$Y^*(1) - Y^*(0)$
1	210	0	8	(13)	5
2	200	0	7	(14)	7
3	310	0	4	(9)	5
4	220	0	7	(12)	5
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	350	0	1	(9)	8
8	250	0	7	(13)	6
9	280	0	6	(9)	3
10	190	0	8	(14)	6
11	250	1	(7)	10	3
12	225	1	(7)	12	5
13	330	1	(1)	7	6
14	260	1	(7)	11	4
15	230	1	(8)	13	5
16	220	1	(7)	13	6
17	180	1	(7)	14	7
18	280	1	(4)	9	5
19	260	1	(6)	9	3
20	195	1	(7)	14	7
Average					**5.3**
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	4	(11)	7
4	220	0	7	(12)	5
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	350	0	1	(9)	8
8	250	0	7	(9)	2
9	280	0	6	(10)	4
10	190	0	8	(13)	5
11	250	1	(7)	10	3
12	225	1	(7)	12	5
13	330	1	(1)	7	6
14	260	1	(6)	11	5
15	230	1	(7)	13	6
16	220	1	(8)	13	5
17	180	1	(8)	14	6
18	280	1	(7)	9	2
19	260	1	(3)	9	6
20	195	1	(8)	14	6
Average					**5.1**
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	4	(7)	3
4	220	0	7	(13)	6
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	350	0	1	(11)	10
8	250	0	7	(9)	2
9	280	0	6	(11)	5
10	190	0	8	(12)	4
11	250	1	(7)	10	3
12	225	1	(7)	12	5
13	330	1	(3)	7	4
14	260	1	(3)	11	8
15	230	1	(7)	13	6
16	220	1	(8)	13	5
17	180	1	(8)	14	6
18	280	1	(7)	9	2
19	260	1	(6)	9	3
20	195	1	(8)	14	6
Average					**4.9**
Median					**5**



Example 1: Need for Covariate Overlap

In this example, we observe people after receiving either a new surgery ($W = 1$) or the standard surgery ($W = 0$). The outcome, Y, is years lived after surgery, and age (at the time of surgery) is a covariate. We observe the following data. Treatments were assigned as some stochastic function of age; i.e., treatment is ignorable.

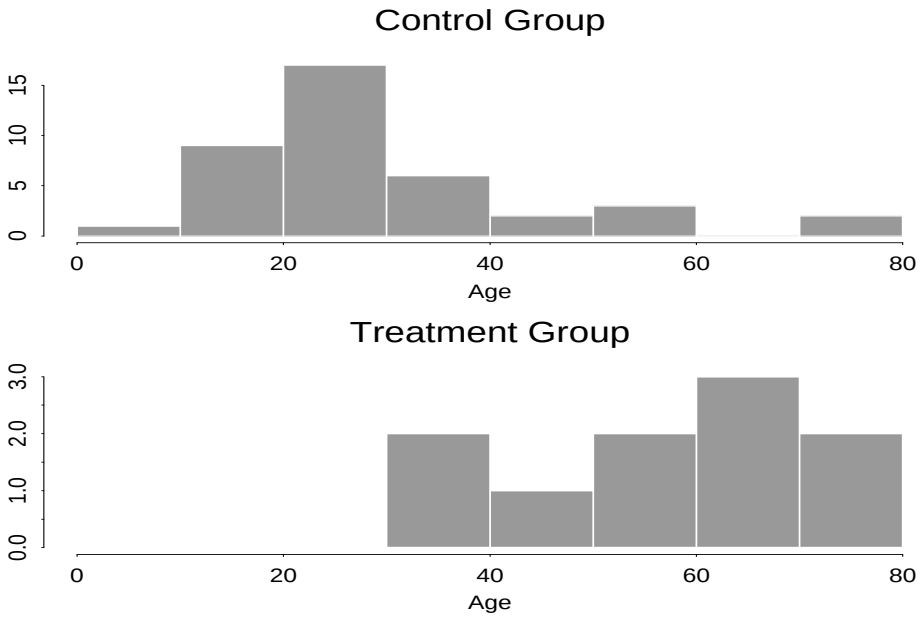
W	Age	$Y(0)$	$Y(1)$	W	Age	$Y(0)$	$Y(1)$
0	10	64		0	30	50	
0	12	66		0	30	46	
0	14	62		0	31	45	
0	15	60		0	31	43	
0	16	63		0	31	43	
0	17	57		1	32		48
0	18	57		0	33	42	
0	18	56		0	36	40	
0	19	55		0	40	36	
0	20	56		1	40		39
0	21	52		0	42	33	
0	22	54		0	48	20	
0	22	50		1	49		32
0	23	48		0	52	22	
0	25	48		1	53		27
0	25	51		1	55		28
0	25	49		0	55	20	
0	26	48		0	60	17	
0	27	45		1	62		18
0	28	45		1	65		12
0	29	47		1	68		15
0	29	49		0	72	3	
0	29	48		1	73		9
0	29	44		1	79		2
0	30	44		0	80	0	

Now consider only the treatment group:

W	Age	$Y(1)$
1	32	48
1	40	39
1	49	32
1	53	27
1	55	28
1	62	18
1	65	12
1	68	15
1	73	9
1	79	2

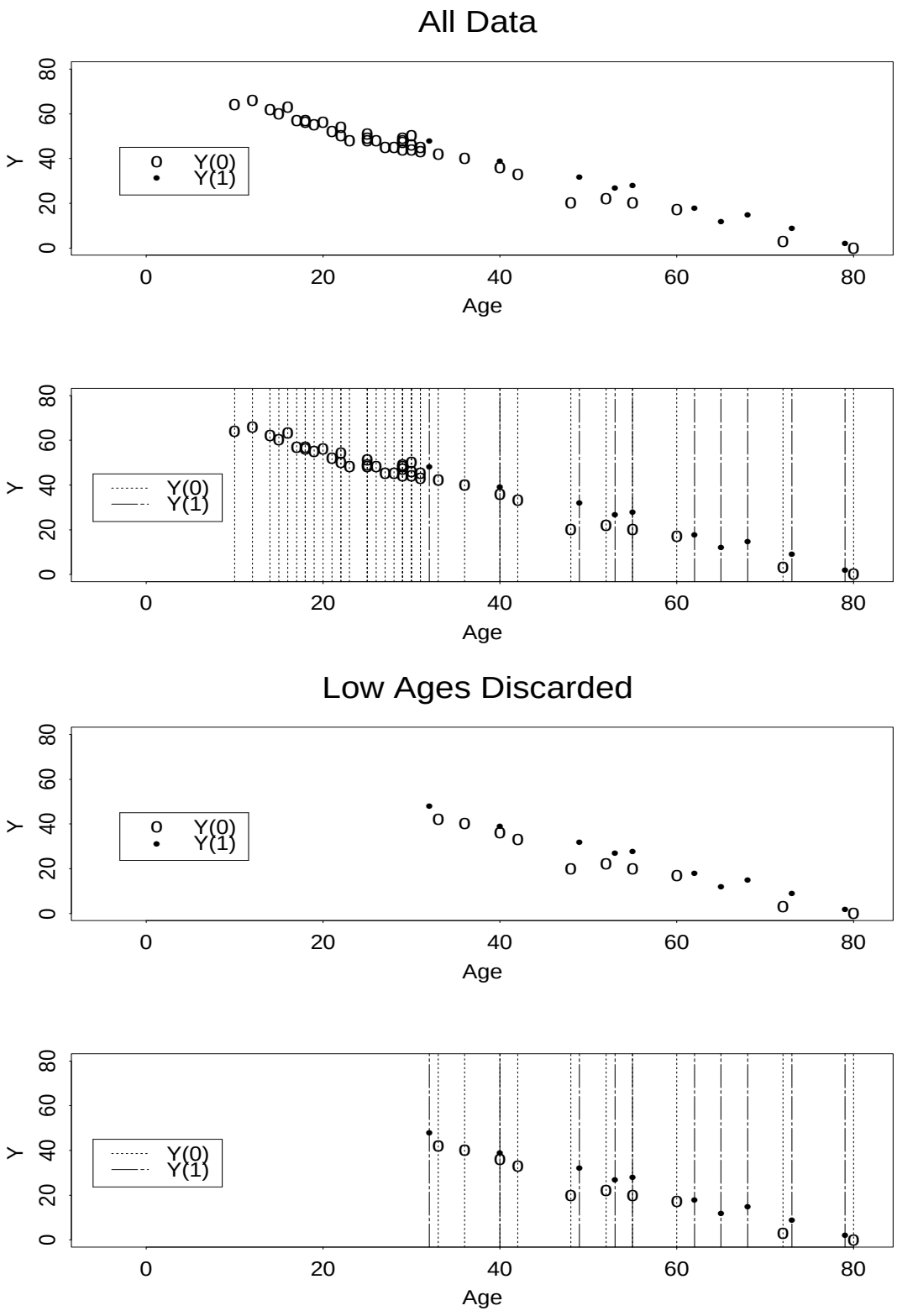
The 32-year-old person is the youngest in the treated group. Notice that most people in the control group were younger than the 32-year-old.

It is not really "fair" to compare the 10- to 31-year-olds in the control group with the people in the treatment group, since 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.



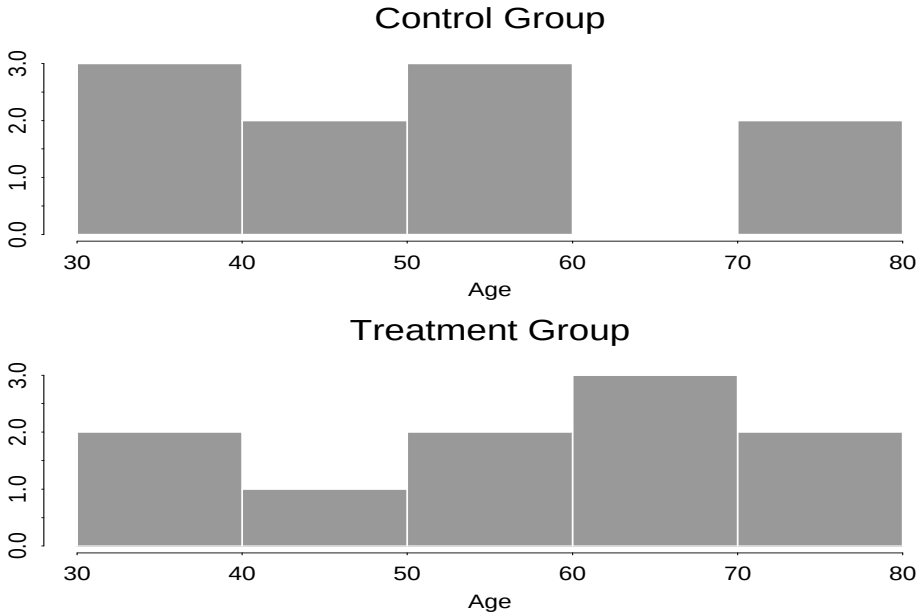
The comparison will be more fair if we only use the controls who are "like" the treated, in the sense that they have similar ages.

We will do this by discarding any control group members whose age is lower than 32, the lowest observed age in the treated group. Discarding these controls gives the following data.


All Data
Y
Age
Y(0)
Y(1)
0
20
40
60
80
Low Ages Discarded

W	Age	$Y(0)$	$Y(1)$
1	32		48
0	33	42	
0	36	40	
0	40	36	
1	40		39
0	42	33	
0	48	20	
1	49		32
0	52	22	
1	53		27
1	55		28
0	55	20	
0	60	17	
1	62		18
1	65		12
1	68		15
0	72	3	
1	73		9
1	79		2
0	80	0	

Note that age is more balanced between the two groups now. However, still the average age in the treated group is 57.6, and the average age in the control group is now 51.8.



If we use all the data to estimate the average treatment effect, we get an estimate of -44, i.e., the new surgery decreases life span by 44 years on average. Using only the data from those aged 32 or older, the estimated average treatment effect is -0.3, i.e., the new surgery has almost no effect on life span. However, the average age in the treatment group is still higher than the average age in the control group, and younger people tend to live more years after the surgery than older people. The difference in ages between the two groups may be biasing the results.

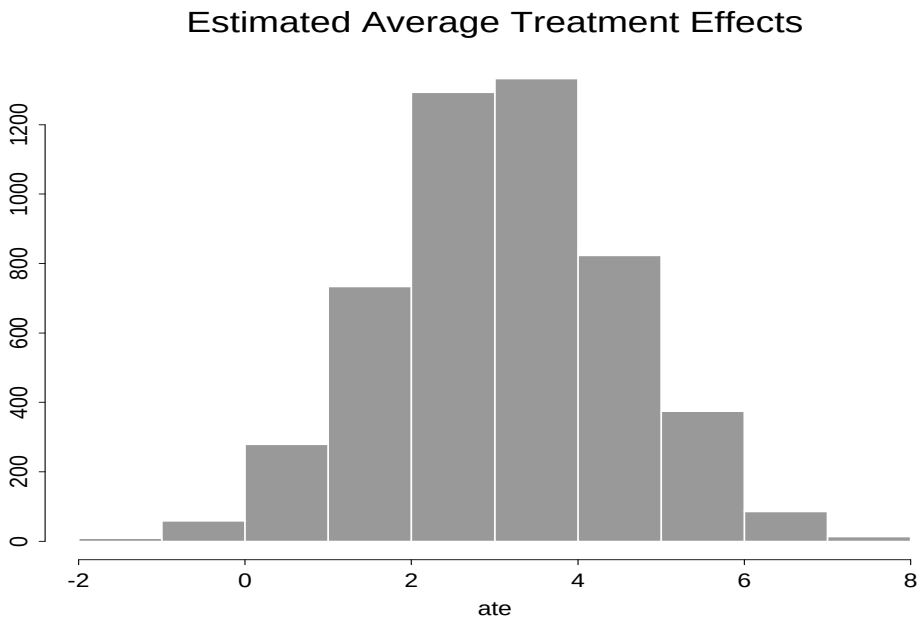
Now we want to estimate the average treatment effect using matching, since age and years lived appear to be correlated within each treatment group (as stated above, years lived tends to decrease as age increases). We form donor pools based on age, using the four closest individuals in the other treatment group to define the pool.

Unit	Age	W	$Y(0)$	$Y(1)$	Units for Donor Pool	Donor $Y(0)$	Donor $Y(1)$
1	33	0	42		11,12,13,14		48,39,32,27
2	36	0	40		11,12,13,14		48,39,32,27
3	40	0	36		11,12,13,14		48,39,32,27
4	42	0	33		11,12,13,14		48,39,32,27
5	48	0	20		12,13,14,15		39,32,27,28
6	52	0	22		13,14,15,16		32,27,28,18
7	55	0	20		13,14,15,16		32,27,28,18
8	60	0	17		14,15,16,17		27,28,18,12
9	72	0	3		17,18,19,20		12,15,9,2
10	80	0	0		17,18,19,20		12,15,9,2
11	32	1		48	1,2,3,4	42,40,36,33	
12	40	1		39	1,2,3,4	42,40,36,33	
13	49	1		32	4,5,6,7	33,20,22,20	
14	53	1		27	5,6,7,8	20,22,20,17	
15	55	1		28	5,6,7,8	20,22,20,17	
16	62	1		18	6,7,8,9	22,20,17,3	
17	65	1		12	6,7,8,9	22,20,17,3	
18	68	1		15	7,8,9,10	20,17,3,0	
19	73	1		9	7,8,9,10	20,17,3,0	
20	79	1		2	7,8,9,10	20,17,3,0	

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 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 obtain an estimate of the average treatment effect.

Age	W	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
33	0	42	(39)	-3
36	0	40	(48)	8
40	0	36	(39)	3
42	0	33	(32)	-1
48	0	20	(28)	8
52	0	22	(27)	5
55	0	20	(32)	12
60	0	17	(18)	1
72	0	3	(15)	12
80	0	0	(9)	9
32	1	(33)	48	15
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				**4.9**

By repeating this process many times, we can get an estimate of the average treatment effect as well as an estimate of the variance of this 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 3.12 and its estimated variance is 1.97. A 95% interval for the treatment effect is (0.4, 5.85). The estimated treatment effect calculated using Neyman's method on the above data, but ignoring the covariate is -0.3, with a 95% confidence interval of (-13, 12.36). A 95% confidence interval calculated by inverting the Fisher test and ignoring covariates is (-11, 11).



The method described here 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 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.

Example 1: Matching in an Observational Study

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 many 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 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).

One of the sources for a comparison group was the Panel Survey of Income Dynamics (PSID), a national survey of income that includes a number of covariates. A subset of the individuals in the PSID was chosen (on the basis of a single covariate) to form a comparison group. The following table shows the means of the covariates for the true treated group and this comparison group from the PSID. The means of the true (randomized) control group are also shown.

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 "significantly" different from each other in these two groups are marked with a *).

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. 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

We see that the treated and the matched comparison group 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

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)

Example 2: 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 the full sample of controls is.
- 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. Treated outcome 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.

Example 1: 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 were 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):

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

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:

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

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 2: 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. 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 will 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.

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 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 3: The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT)

The following study was reported in Goetghebeur and Molenberghs, "Causal Inference in a Placebo-Controlled Clinical Trial with Binary Outcome and Ordered Compliance," *Journal of the American Statistical Association*, September 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 is 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 4: 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 ITT estimate. 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 who didn't:

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

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 one of the assignments. 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 who receives the letter and does not get a flu shot must be a never-taker. Similarly, someone who does not receive the letter but 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.

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 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.

QR33: Causal Inference

Course Packet

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

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

3. 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.

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

5. 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.

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

7. Bickman, L. (1985). Randomized Field Experiments in Education: Implementation Lessons. Chapter 4 (pages 39–53) in *Randomization and Field Experimentation*, ed. Boruch, R.F. and Wothke, W. New Directions for Program Evaluation, Number 28. Publication of the Evaluation Research Society. San Francisco: Jossey-Bass Inc., Publishers.

8. 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.

9. 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.

10. 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.

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

12. 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.

13. 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.