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

Donald B. Rubin
Department of Statistics
Harvard University

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

Last update: 22 August 2005

The perspective on causal inference taken in this course is often referred to as the "Rubin Causal Model" (e.g., Holland, 1986) to distinguish it from other commonly used perspectives such as those based on regression or relative risk models. Three primary features distinguish the Rubin Causal Model:

1. Potential outcomes define causal effects in all cases: randomized experiments and observational studies
   - Break from the tradition before the 1970's
   - Key assumptions, such as stability (SUTVA) can be stated formally
2. Model for the assignment mechanism must be explicated for all forms of inference
   - Assignment mechanism process for creating missing data in the potential outcomes
   - Allows possible dependence of process on potential outcomes, i.e., confounded designs
   - Randomized experiments a special case whose benefits for causal inference can be formally stated
3. The framework allows specification of a joint distribution of the potential outcomes
   - Framework can thus accommodate both assignment-mechanism-based (randomization-based or design-based) methods and predictive (model-based or Bayesian) methods of causal inference
   - One unified perspective for distinct methods of causal inference instead of two separate perspectives, one traditionally used for randomized experiment, the other traditionally used for observational studies
   - Creates a firm foundation for methods for dealing with complications such as noncompliance and dropout, which are especially flexible from a Bayesian perspective

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

Part I: Framework

Subsection 1: Basic Concepts: Units, Treatments, And Potential Outcomes

Definition of Basic Concepts

Unit: The person, place, or thing upon which a treatment will operate, at a particular time
Note: A single person, place, or thing at two different times comprises two different units.

Treatment: An intervention, the effects of which (on some particular measurement on the units) the investigator wishes to assess relative to no intervention (i.e., the "control")

Potential Outcomes: The values of a unit's measurement of interest after (a) application of the treatment *and* (b) non-application of the treatment (i.e., under control)

Causal Effect: For each unit, the comparison of the potential outcome under treatment and the potential outcome under control

The Fundamental Problem of Causal Inference: We can observe at most one of the potential outcomes for each unit.

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

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). Note we do not use the column "X" in this example, but we will in later ones.

Unit	Initial Headache	Potential	Outcomes	Causal Effect
	X	Y(Asp)	Y(Not)	Y(Asp) - Y(Not)
you	80	25	75	-50

Example I-2: Potential Outcomes and Causal Effect with One Unit: Gain Scores

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

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

Example I-3: Potential Outcomes and Causal Effect with One Unit: Percent Change

Potential Outcomes and Causal Effect with One Unit: In this 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 $\mathrm{Y}^* = 1 - \frac{\mathrm{Y}}{\mathrm{X}}$, where X = intensity of initial headache ($\frac{0}{0}$ is defined to be 1 here).

Unit	Initial Headache	Potential	Outcomes, Y	Causal Effect
	X	Y*(Asp)	Y*(Not)	Y*(Asp) - Y*(Not)
you	80	1 - $\frac{25}{80} =$ = 69%	1 - $\frac{75}{80}$ = 6%	69% - 6% = 63%

A key point here is that the causal effect does not involve probability, nor is it a change over time.

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

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

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

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

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

Subsection 2: Learning about Causal Effects: Replication, Stability, And the Assignment Mechanism

Definition of Basic Concepts

Replication: At least one unit receives treatment and at least one unit receives control

Stable Unit-Treatment-Value Assumption (“SUTVA): Two parts: (a) there is only one form of the treatment and one form of the control, and (b) there is no interference among units

Assignment Mechanism: The process for deciding which units receive treatment and which receive control

We resume with the aspirin example, and we assume only that all aspirin tablets are equally effective.

Example I-5: Potential Outcomes with Two Units Allowing
Interference Between Units (Part (b) of SUTVA Does Not Hold)

Potential Outcomes and Values in Example

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

The causal effect of Asp versus Not for me is 100. We might say that the causal effect for me is “well-defined.” 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. Thus, my outcome does not depend on whether you take aspirin.

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 for you is Y_1([Asp, Not]) - Y_1([Not, Not]) = 50 - 100 = -50, a smaller effect. Perhaps when I have headaches, I complain a great deal to you, inducing whatever head pain you have to increase.

The fact that the causal effect for you depends on what treatment I take makes analyzing the situation difficult. That is why the Stable Unit-Treatment-Value Assumption (“SUTVA”) is so important. We try to hard to construct or find situations in which SUTVA holds.

Note that we have not yet considered the possibility that there may be more effective and less effective aspirin tablets. If such were the case, we would need to expand the above notation to include “Asp+”, for a more effective tablet, and “Asp-”, for a less effective tablet. Now imagine a full bottle of aspirin, with each pill varying in effectiveness; the situation becomes exponentially more complicated even with only two units. With more than two units, SUTVA is even more critical, another reason why it is so commonly assumed.

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

Unit	X	Y(Asp)	Y(Not)	Unit Level 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)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	X_i	Y_i(Asp)	Y_i(Not)	Y_i(Asp) - Y_i(Not)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_N	Y_N(Asp)	Y_N(Not)	Y_N(Asp) - Y_N(Not)

This array of values of X, Y(1), and Y(0) represents the science, about which we want to learn. This is the common sense definition, and is used in common culture (“It’s a Wonderful Life,” ”A Christmas Carol,” law, etc.).

Various Population Level Causal Effects

Comparison of Y_i(Asp) and Y_i(Not)
on a common subset of units

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

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)$\}$

If X_i includes male/female for each unit:
Average causal effect of “Asp” vs. “Not” for males =
$\mathrm{Ave}_{X_i=male}$ $\{Y_i$(Asp) - Y_i(Not)$\}$

- The unit level causal effects cannot be observed; remember the fundamental problem of causal inference. That means that population level causal effects also cannot be observed, even under SUTVA.

- To learn about causal effects, we must have replication. In the example above, we require some units with Y_i(Asp) observed and some with Y_i(Not) observed.

- The assignment mechanism determines how to choose which potential outcome we will observe for each unit. Formally, the assignment mechanism is a probabilistic or deterministic rule for selecting some units to receive control and other units to receive treatment. It describes what we do (or what was done) to learn about the science: X, Y(1), Y(0).

- The assignment mechanism is critical, even if SUTVA holds. We must know or posit a rule for how each unit received treatment or control.

Subsection 3: Transition to Statistical Inference: The Perfect Doctor And Lord's Paradox

Definition of Basic Concepts

Ignorable Assignment Mechanism: The assignment of treatment or control for all units is independent of the unobserved potential outcomes ("**nonignorable**" means not ignorable)

Unconfounded Assignment Mechanism: The assignment of treatment or control for all units is independent of all potential outcomes, observed or unobserved ("**confounded**" means not unconfounded)

Example I-7: Perfect Doctor

This example illustrates that we must consider the assignment mechanism to reach reliable causal inferences.

The hypothetical 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.

The "Truth":

Potential Outcomes

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

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

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

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

What we would actually observe under the Perfect Doctor's assignment mechanism?

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

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

W indicates which treatment each unit received.

In the perfect doctor example, the treatment each unit receives depends on that unit's potential outcomes. The assignment mechanism is nonignorable (and thus confounded). It is difficult to analyze correctly an experiment with a nonignorable assignment mechanism.

For example, if we were to draw an inference based on the observed difference in sample means, we would conclude that the treatment, on average, adds over five years of life. But we "know" that the treatment, on average, subtracts two years from life. Also, from looking at the observed sample means, we would conclude that if everyone received the new operation, people would live an average of eleven years. But from the previous table, we "know" that if everyone received the new operation, people would live an average of five years. This is another incorrect inference. Moreover, from looking at all the observed values, we note that the years lived under the new operation (9, 10, 14) are much greater than the years lived under the old operation (4, 5, 6, 6, 6), an observation that could very easily lead to another incorrect causal inference.

What's wrong with what we did? Where exactly was our mistake? To get a better idea of what's going on, let's take a look at what would be observed in ALL POSSIBLE assignments in this situation.

The Perfect Doctor, continued: All Possible Assignments

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

As the chart on the previous page shows, there are 56 possible assignments in which three of eight units receive treatment.

Observed Outcomes under Assignment 1

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

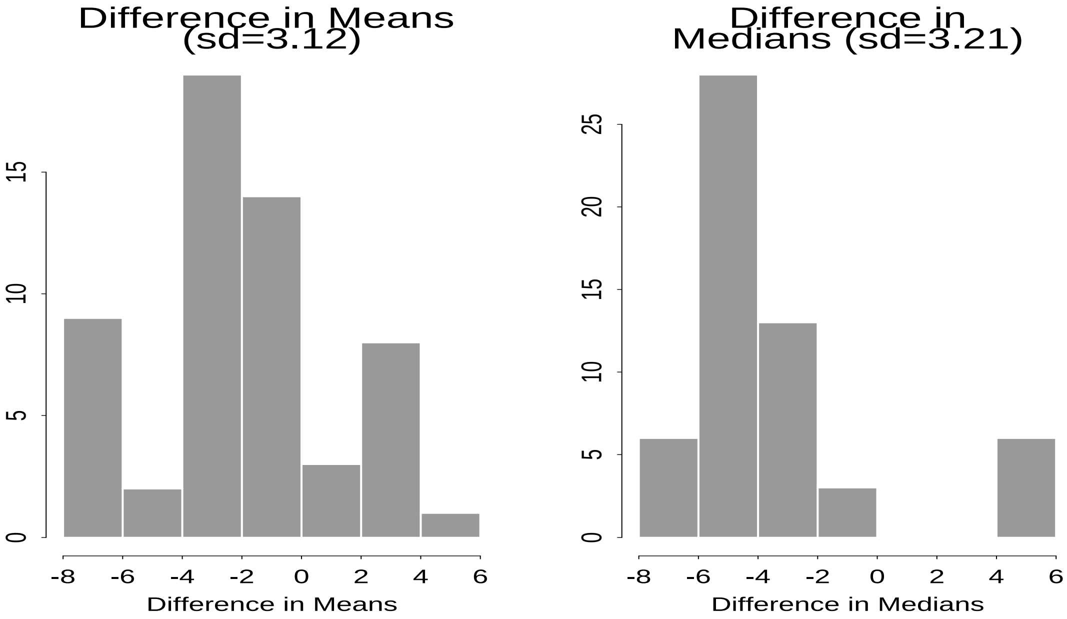
**Observed $\overline{y_1} - \overline{y_0} = -1.6$.

Observed Outcomes under Assignment 56

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

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

Summary of All 56 Assignments



Our statistic was the difference in observed means, and on page I-3.3, we calculated the value of that statistic for all 56 possible assignments. The average of all of these possible values was -2, which equals exactly the “known” truth. This equality suggests that had our observed assignment been, say, a random draw of one of the 56 possible assignments, we would have been OK on average (we will quantify exactly what we mean by “OK on average” in subsequent sections). Random draws do not depend on the potential outcomes, and thus the associated assignment mechanism is unconfounded.

In contrast, the Perfect Doctor’s assignment mechanism depended on the potential outcomes and was therefore nonignorable (implying that it was confounded). We observed with certainty the most extreme assignment possible, and the value of our statistic, the difference of observed sample means, was far from the truth; it even had the wrong sign. The difference of observed sample means is an OK estimate of the average causal effect, in general, only if the assignment mechanism is random.

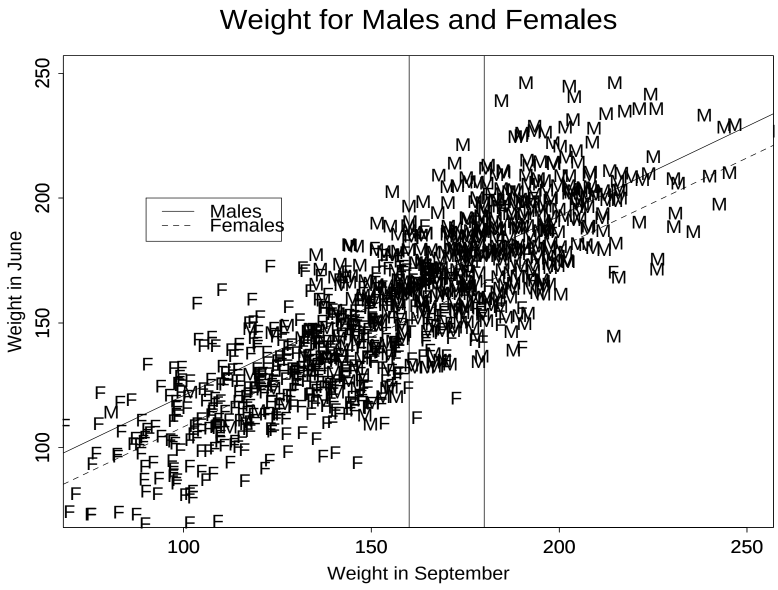
The takeaway message: The observed difference in means is entirely misleading in this situation. The biggest problem when using the difference of sample means here is that we have effectively pretended that we had an unconfounded treatment assignment when in fact we did not. This example demonstrates the importance of finding a statistic that is appropriate for the actual assignment mechanism.

Example I-8: 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 [or her] arrival in September and his [or her] weight the following June are recorded."

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



The average weight for Males was 180 in both September and June. Thus, the average weight gain for Males was zero.
The average weight for Females was 130 in both September and June. Thus, the average weight gain for Females was zero.

Question: What is the differential causal effect of the diet on male weights and on female weights?

Statistician 1: Look at gain scores: No effect of diet on weight for either males or females, and no evidence of differential effect of the two sexes, because no group shows any systematic change.

Statistician 2: Compare June weight (see Figure 1) for males and females with the same weight in September: On average, for a given September weight, men weigh more in June than women. Thus, the new diet leads to more weight gain for men.

Is Statistician 1 correct? Statistician 2? Neither? Both?

Lord's Paradox: Analysis under the Rubin Causal Model

1. What are the units? The students, both male and female, in September.

2. What are the treatments? The university dining hall diet, and the control diet, which is what the students would have had otherwise.

3. What is the assignment mechanism? Because all students in the study were exposed to the dining hall diet, the assignment mechanism is that all units receive treatment with probability one, and all units receive control with probability zero.

4. Is the assignment mechanism unconfounded? Yes; for each unit, the probability of receiving treatment is unrelated to the potential outcomes.

   We can represent the Lord's Paradox situation with the following table. The X's, the covariates, are Sex and Sep. Wt.

Unit #	Sex	Sep. Wt.	W	Y(0)	Y(1)
1	F	90	1	?	92
2	F	92	1	?	95
...					
n$_F$	F	210	1	?	185
n$_F$+1	M	98	1	?	102
n$_F$+2	M	118	1	?	117
...					
N	M	248	1	?	240

   Here, Y(0), the outcome under control, represents what we would observe for a student had that student NOT eaten the dining hall diet. Because all students ate the dining hall diet, we observe this outcome for none of the units. Symbolically, $p(W_i = 1|X_i) = 1$ for all i = 1 to n. That is, the probability that we observe Y(1) is one for each unit.

5. Is this assignment mechanism useful for causal inference? No; we observe the treated outcome for all units and the control outcome for none. To be useful for causal inference, the assignment mechanism must involve some replication of each treatment.

6. Would it have helped if all males received the dining hall diet and all females received the control diet? Not really. We want replication at each value of X. That is, for each possible value of X we would like to see some treated and some control units. The only way to achieve this result, even just in expectation, is to have random assignment at each value of X.

So, is Statistician 1 or Statistician 2 correct? In fact, we could make either one correct, depending on how we filled in Y(0) for each unit. Suppose we wanted to make correct Statistician 1's assertion that there is no effect of treatment for men or for women. We could fill in each unit's Y(0) value with its September weight. (Actually, we could also fill in each unit's Y(0) value with its June weight.) How implausible would these filled in Y(0) values be? We have no idea from the observed data, because we did not observe Y(0) for any of the units.

Now suppose we wanted to make Statistician 2's assertion that the new diet leads to more weight gain for men correct. For each male unit, we could predict Y(0) to be a constant plus that unit's September weight times another constant. We would have to choose the constants correctly (technically, they must come from the linear regression of Y(1) on a vector of ones and September weight). We could do the same thing for each female unit. If we followed these steps, which would result in filling in a Y(0) value for each unit, Statistician 2 would be correct. How implausible would these Y(0) values be? We have no idea based on the observed data, because we did not observe Y(0) for any of the units.

The takeaway message: Both statisticians can be made correct even though they make contradictory assertions, and thus neither is correct, as the positions are stated. The key point is that everyone in this dataset received the treatment (the diet in the university dining halls). We observed Y(1) for all units and Y(0) for none. Would you want to assess the causal effect of a pill by giving everyone the pill? *This dataset has no information in it about the effect of the dining hall diet on weight gain.* To draw causal, rather than descriptive, inferences without making heroic assumptions, we need $0 < p(W_i = 1|X_i) < 1$ for each unit! That is, we need not only an unconfounded assignment mechanism but also a stochastic (meaning probabilistic) assignment mechanism. A stochastic unconfounded assignment mechanism is the most general form of a classical randomized experiment.

Subsection 4: Examples of Unconfounded Assignment Mechanisms, Simple

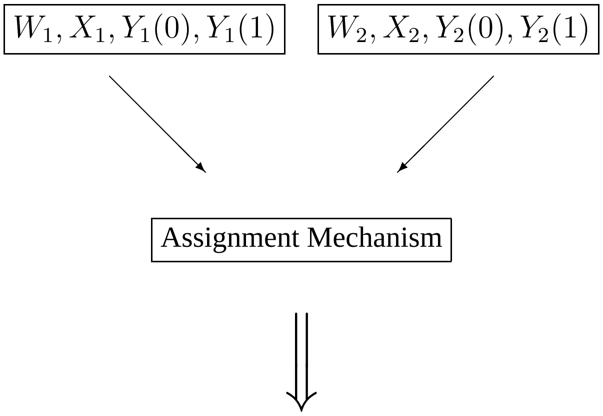
Definition of Basic Concepts

Propensity Score: For each unit, the probability of being assigned treatment: $p(W_i = 1|\boldsymbol{X}, Y(0), Y(1))$

Classical Randomized Assignment Mechanism: A special case of an unconfounded assignment mechanism where each unit's assignment is probabilistic, i.e., the propensity score for each unit is strictly between 0 and 1

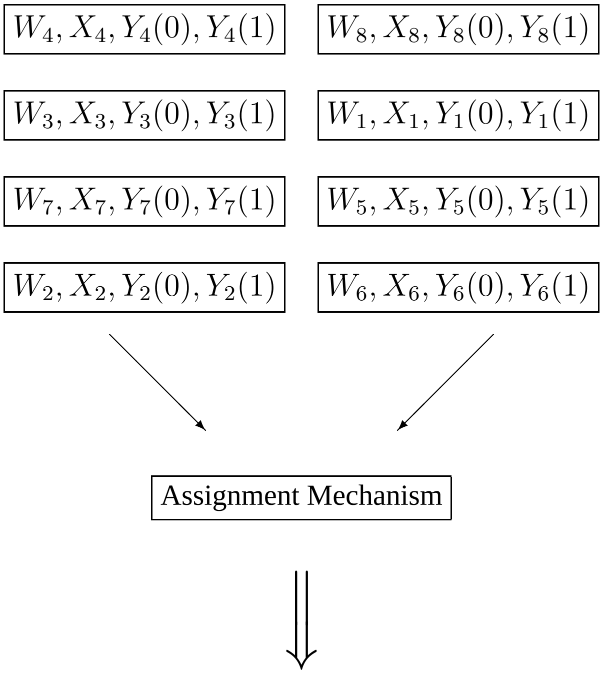
Completely Randomized Assignment Mechanism: A special case of a randomized assignment mechanism where n of N units receive treatment, N - n (i.e., the rest) receive control, with each such assignment equally likely

Example I-9: Completely Randomized Design with $N = 2$ units, 1 Assigned Treatment



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

Example I-10: Completely Randomized Design with $N = 8$ Units, 3 Assigned Treatment


$W_4, X_4, Y_4(0), Y_4(1)$
$W_8, X_8, Y_8(0), Y_8(1)$
$W_3, X_3, Y_3(0), Y_3(1)$
$W_1, X_1, Y_1(0), Y_1(1)$
$W_7, X_7, Y_7(0), Y_7(1)$
$W_5, X_5, Y_5(0), Y_5(1)$
$W_2, X_2, Y_2(0), Y_2(1)$
$W_6, X_6, Y_6(0), Y_6(1)$
Assignment Mechanism

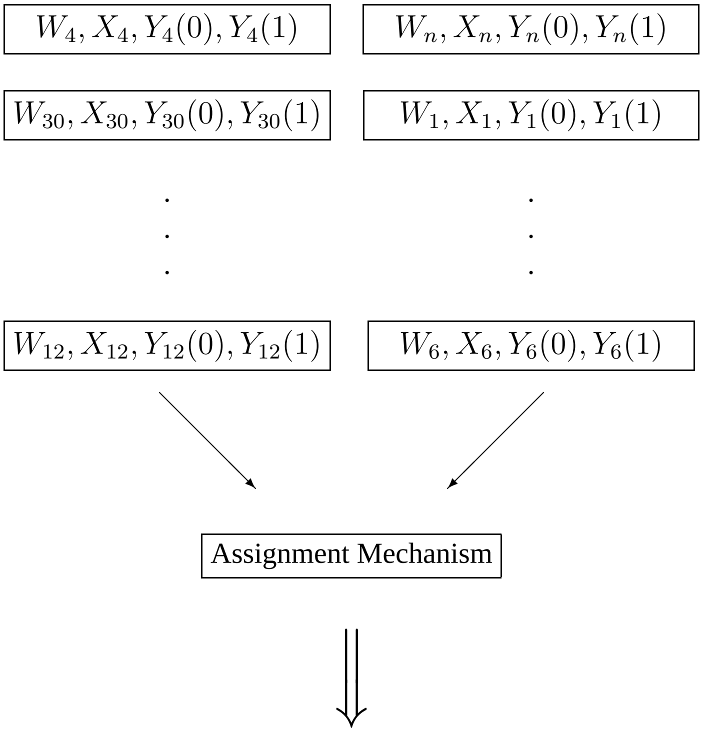
W	Prob of **W**
If $\sum_{i=1}^{8} W_i = 3$ [a]	$\frac{1}{56}$ [b]
If $\sum_{i=1}^{8} W_i \neq 3$	0

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

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

In this example, each person has probability $\frac{3}{8}$ of receiving treatment (and probability $\frac{5}{8}$ of receiving control). Thus, each person's propensity score is $\frac{3}{8}$. Note: Subscripting of units is arbitrary (i.e, random), that is, nothing would change if we called the "first" unit Unit 3, and the "second" unit Unit 7, etc. (so long as we kept Unit 3's W value, its X value, its Y(0) value, and its Y(1) value together). Such a rearrangement is called a "permutation."

Example I-11: Completely Randomized Design with N units, n Assigned Treatment


$W_4, X_4, Y_4(0), Y_4(1)$
$W_n, X_n, Y_n(0), Y_n(1)$
$W_{30}, X_{30}, Y_{30}(0), Y_{30}(1)$
$W_1, X_1, Y_1(0), Y_1(1)$
$W_{12}, X_{12}, Y_{12}(0), Y_{12}(1)$
$W_6, X_6, Y_6(0), Y_6(1)$
Assignment Mechanism

W	Prob of **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

In this example, each person has probability $\frac{n}{N}$ of receiving treatment (and thus probability $1 - \frac{n}{N}$ of receiving control). Thus, each person's propensity score is $\frac{n}{N}$.

Example I-12: “Bernoulli” (fair coin-tossing) Assignment, 4 units

Assignment is random and independent across units; moreover, 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. Thus, each person’s propensity score is .5. Remember that the overall assignment is the collection of all of the individuals’ assignments: $\mathbf{W} = (W_1, W_2, W_3, W_4)$.

Because each individual’s treatment status is assigned independently of the other individuals, the overall assignment probability is the product of the individual probabilities, that is, the product of the propensity scores.

All Possible Assignments

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

Question: Are there some randomized assignments that are less effective than others with respect to learning about the causal effect of treatment versus control? Hint: Remember Lord’s Paradox.

Question: Are there some randomized assignment mechanisms that are less effective than others with respect to learning about the causal effect of treatment versus control?

Example I-13: “Bernoulli” (biased coin-tossing) Assignment, 4 Units

Same as Example I-12, except that now the probability of receiving treatment 1 for each individual is $.4$. Again, treatment is assigned independently for each individual. As with the previous example, each person has the same propensity score, but here, that score is .4.

All Possible Assignments

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

Question: Again, are there some assignments that are less effective than others with respect to learning about the causal effect of treatment versus control? Are these assignments more or less likely than in the previous example? Do you have any advice concerning the choice of assignment mechanism for a researcher based on the simple examples of assignment mechanisms given thus far?

Subsection 5: Examples of Unconfounded Assignment Mechanisms, with Covariates – Blocking

Definition of Basic Concepts

Covariate: A characteristic of a unit unaffected by treatment, such as baseline headache in the Perfect Doctor example, or September weight in the Lord's Paradox example

Block: A set of individuals grouped together based on some covariate

Example I-14: Randomization within Blocks

In this case we consider two blocks: one comprising four males and another comprising four females. Two males and two females are chosen completely at random to receive treatment 1. The other two males and two females receive treatment 0.

Males: $X_i = M$

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

Females: $X_i = F$

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

Overall

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

Notice that the propensity score for each unit is .5, but the assignment mechanism is not the same as the Bernoulli(.5) example, nor is it a completely randomized assignment mechanism.

Question: What is the relative merit of this design compared to these other two? Hint: With a completely randomized design or a Bernoulli(.5), is it possible for all four males to be assigned treatment and all four females to be assigned control?

Example I-15: “Bernoulli” (biased coin-tossing) Assignment within Blocks

There are four men and four women. In the notation below, the covariate **X** equals (M, M, M, M, F, F, F, F). For males, the probability of receiving treatment 1 is .2. For females, the probability of receiving treatment 1 is .7. Treatments are assigned independently across units.

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

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

Possible Assignments

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

Example I-16: Probability of Treatment Depends on Age

In this example, the propensity score for unit i (i.e., the probability that $W_i = 1$) is $\frac{\text{age}_i}{\text{age}_i+10}$. In the notation below, **Age** $= (15, 22, 18, 54, 34, 77, 38, 91)$. Treatments are assigned independently across units.

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

Because each individual's treatment status is assigned independently of the other individuals, the probability of $\mathbf{W}$ is the product of the eight propensity scores.

Possible Assignments

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

Again, are some of these assignments more or less useful for causal inference? Any advice on the choice among competing assignment mechanisms?

Example I-17: “Bernoulli” (coin-tossing) Assignment, 4 Units

(Mini) School choice example. This example is based on a real experiment involving a program to give vouchers to students to attend private schools. The 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. Students 1 and 4 come from bad schools.

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

We consider three assignment mechanisms consistent with the above table:

a. We use independent assignment (Bernoulli).

b. We have money only for two vouchers, so two students are given vouchers and two are not; if our initial effort to assign vouchers using the above probabilities does not result in two persons getting vouchers, we redraw.

c. We have money only for three vouchers, and we proceed in the same way as in b, except we redraw until we have three students assigned vouchers.

All Possible Assignments

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

[a] Independent assignment
[b] Constrained so that two given vouchers
[c] Constrained so that three given vouchers

Question: How are the probabilities of the various assignments calculated under the three assignment mechanisms?

Example I-18: Randomized within Matched Pairs

In this design, each block comprises two units.

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, one 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 (0), and so the propensity score for each unit is .5. We consider three 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.

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

Example I-19: Bacterial Growth: Before And After Studies

Supposed we wish to assess the effect of an antibiotic on the growth of bacteria in petri dishes. One reasonable way to accomplish this task is to make N = 2n petri dishes, pick n of them completely at random, administer the drug to the chosen n, do nothing to the other half, and observe the results. Suppose, however, that we have only one petri dish available, so we cannot use this method. Instead, we do the following randomized "before-after" study. First, we randomly pick a number D between 1 and 20. We begin measuring the number of bacteria in the single petri dish on day one. On the Dth day, *after* taking a measurement for that day, we administer the drug to the petri dish. We continue to measure the amount of bacterial on each day thereafter, until we reach day 21.

Note: Few researchers would use this method in an actual experiment. But it may be a useful template with which to analyze a "before-after" observational study where some intervention takes place at some point in time.

In this setting, each unit is the petri dish on a particular day, so we have 21 units. A unit received treatment if the drug was administered before that day's measurement, otherwise the unit received control. What are the possible assignments? Supposed we randomly picked D = 2, and recall that we take day D's measurement *before* administering the drug. So the assignment for D = 2 would be (0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1).

The table below summarizes all possible assignments for this setting.

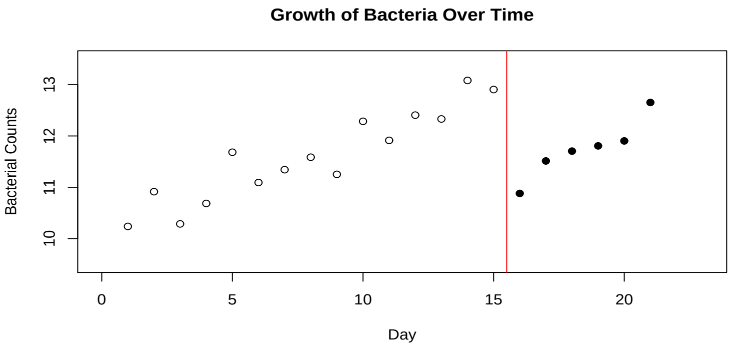
Possible Assignments for Before And After

Assignment **W**	p(**W**)
(0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	.05
(0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	.05
(0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	.05
...	
(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)	.05

Now suppose we randomly choose D = 15. A potential outcomes table for such an experiment under SUTVA, with some hypothetical data, appears below.

Growth of Bacteria: Potential Outcomes

Unit Index	Day	W	Y(0)	Y(1)
1	1	0	10.237	?
2	2	0	10.914	?
3	3	0	10.286	?
4	4	0	10.684	?
5	5	0	11.682	?
6	6	0	11.092	?
7	7	0	11.343	?
8	8	0	11.585	?
9	9	0	11.252	?
10	10	0	12.285	?
11	11	0	11.913	?
12	12	0	12.406	?
13	13	0	12.331	?
14	14	0	13.082	?
15	15	0	12.904	?
16	16	1	?	10.880
17	17	1	?	11.513
18	18	1	?	11.704
19	19	1	?	11.807
20	20	1	?	11.903
21	21	1	?	12.652



Question: Is SUTVA a reasonable assumption in this setting? Suppose that Unit 15 had been treated (along with Units 16-21) – is it reasonable to assume that the observed Y(1) value for units 16-21 would remain the same as it was when Unit 15 received control? What if Unit 10 (and thus Units 11-21) had been treated? (More on this later.)

From the above table it is not clear how one might analyze the observed data from an assignment mechanism such as that of the before and after study. We deal with analysis issues in Part II, but to provide some idea of how one might begin, the table below shows the observed data along with a potentially useful set of statistics. Specifically, the third column represents the difference between the measurements of a particular day and the day before.

Growth of Bacteria: Observed Data And a Statistic Estimating the Causal Effect

Day	W	Y(obs)$_{Day}$	Y(obs)$_{Day}$ - Y(obs)$_{Previous\ Day}$
1	*	10.237	NA
2	0	10.914	0.677
3	0	10.286	-0.628
4	0	10.684	0.399
5	0	11.682	0.998
6	0	11.092	-0.590
7	0	11.343	0.250
8	0	11.585	0.243
9	0	11.252	-0.334
10	0	12.285	1.033
11	0	11.913	-0.372
12	0	12.406	0.492
13	0	12.331	-0.075
14	0	13.082	0.751
15	1	12.904	-0.178
16	1	10.880	-2.024
17	1	11.513	0.632
18	1	11.704	0.192
19	1	11.807	0.102
20	1	11.903	0.096
21	1	12.652	0.749

Question: Is this design unconfounded? Is it ignorable? Why?

Question: Why not use the difference in the means of the treated and control units, i.e., $\overline{y(1)} - \overline{y(0)}$, as the statistic to measure the causal effect in this example? Hint: Look at the previous figure or table.

Earlier, in the context of the example concerning a completely randomized design with N = 8 units, we made the point that the labelling (or ordering) of the units is arbitrary and that the assignment mechanism must be invariant under random permutations. To make this condition true, we sometimes must record information about the order the data occurred in time. Thus, in the table entitled "Growth of Bacteria: Potential Outcomes" a few pages earlier, we recorded the covariate "Day" to keep track of time. Because we did so, that table is identical to the one below in terms of the assignment mechanism.

Growth of Bacteria under a Random Permutation: Potential Outcomes

Unit Index	Day	W	Y(0)	Y(1)
1	11	0	11.913	?
2	21	1	?	12.652
3	3	0	10.286	?
4	4	0	10.684	?
5	19	1	?	11.807
6	6	0	11.092	?
7	7	0	11.343	?
8	8	0	11.585	?
9	18	1	?	11.704
10	10	0	12.285	?
11	17	1	?	11.513
12	9	0	11.252	?
13	13	0	12.331	?
14	1	0	10.237	?
15	15	0	12.904	?
16	16	1	?	10.880
17	2	0	10.914	?
18	20	1	?	11.903
19	12	0	12.406	?
20	14	0	13.082	?
21	5	0	11.682	?

Summary: With an unconfounded assignment mechanism, we know the probability of each possible assignment, and thus the table of **W**'s (the vectors of assignments) and their corresponding probabilities can be written down. The ability to do this for the assignment mechanism ($p(W|X, Y(0), Y(1)$) does not, however, imply anything about the plausibility of SUTVA, which is a property of the science ($X, Y(1), Y(0)$), and thus must always be argued on substantive grounds.

Subsection 6: Examples of Confounded Assignment Mechanisms, Both Ignorable And Nonignorable

Definition of Basic Concepts

Confounded But Ignorable Assignment Mechanism: The probability of assignment depends on the values of some of the observed potential outcomes (repeated from I-3.1)

Nonignorable Assignment Mechanism: The probability of assignment depends on the values of some of the unobserved potential outcomes (repeated from I-3.1)

Example I-20: Bernoulli Assignment, But Probability Depends
on Unobserved Covariate (Nonignorable)

A teacher randomly assigns children in her class to a new reading program (treatment 1). Because she wants motivated children in this new program, she judges each student's motivation on a scale from 1 to 10 and assigns children to the program such that students with higher motivation are more likely to be put into the new program. To ensure confidentiality, she does not write down or disclose to anyone the students' motivation levels (and she promptly forgets them!). Like the perfect doctor, the teacher has great insight, and the motivation score is highly predictive of the children's Y values.

For each student, the assignment to treatment 1 is done using the following rule: the probability that $W_i = 1$ is $.1 * U_i$, where U_i is the student's (unobserved) motivation level. ($\mathbf{U}$ is the vector of the motivation levels of all of the students); for students with motivation level 10, the probability of assignment to treatment 1 is .95, and for students with motivation level 0, the probability of assignment to treatment 1 is .05:

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

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

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 when the researcher either does not observe or ignores the U_i's.

If treatment is assigned independently to each unit, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's, i.e., the product of the unit level propensity scores, but because these propensity scores depend on something we do not observe, and we would not know what they are in an actual experiment.

Possible Assignments

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

Example I-21: ”Play the Winner” Treatment Assignment (Confounded But Ignorable)

The following is based on Ware (1989) and follow-up studies.

- Persistent pulmonary hypertension of the newborn (PPHN) is an acute lung disease that results in newborns' 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% in practice
- Extracorporeal membrane oxygenation (ECMO) treatment mortality rate: less than 20% in experimental settings.
  - ECMO is an extreme therapy that routes the blood out of the jugular vein, oxygenates the blood outside the body, heats it, and then replaces the blood in the body through the carotid artery. It is essentially a simplified heart-lung machine.
- Three randomized studies of ECMO have been done in the treatment of PPHN.
  1. Randomized “play-the-winner” (confounded but ignorable)
     - The probability of each newborn receiving ECMO depends on the previous outcomes.
     - 12 infants are enrolled sequentially (one after another in time).
     - Assignment: Think of an urn that contains two balls: one representing ECMO, one representing CMT. The first infant was randomly given ECMO and future assignment was as follows: “When a treatment was selected and the infant survived, a ball representing that treatment was added to the urn. When the infant died, a ball representing the other treatment was added.” To determine the assignment of the next infant, a ball was drawn out of the urn.
     - $Y = 1$ if the patient died and $Y = 0$ otherwise.

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

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

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

- 11 infants received ECMO and all survived. One infant received CMT and died.
- Note that we can calculate the Prob of $W_i = \mathrm{ECMO}$ for the observed assignment for each individual given the previous assignments and observed outcomes. Let 1 represent the ECMO treatment and 0 represent the CMT treatment. Then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Randomized phase, four deaths in the CMT group (out of ten). By that point nine patients had received ECMO and all survived.
- In non-randomized phase, only ECMO was given.
- By the end of the study, 19 of 20 (97%) ECMO patients survived, whereas six of ten (60%) CMT patients survived.

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

How consistent are these three studies of ECMO? Which do you believe most?

The issue of sequential randomized designs is an important one in statistics (e.g., Chernoff, 1959).

Example I-22: Perfect Doctor, Revisited (Nonignorable)

This example repeats the previous Perfect Doctor example except that there is a unit that lives equally long under either treatment.

The outcome of interest is years lived after surgery. The 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

$\downarrow$

Assignment Mechanism

$\Downarrow$

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

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

$\downarrow$

Assignment Mechanism

$\Downarrow$

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

Note that in this example the only thing that has changed between the left and right sides is the ordering (i.e., labelling) of the units. The probabilities of assignment do not change between the left and right sides because the assignment mechanism cannot depend on the labelling of the units.

The propensity scores, given both potential outcomes, are 0, 1, or .5 (for the unit that lives equally long under either assignment). But, again, in a real-life situation, we would never observe both potential outcomes.

Example I-23: Almost-Perfect Doctor (nonignorable)

We alter Example I-22 slightly here. The doctor now tosses a biased coin for each individual, where the bias is based on $Y(0)$ and $Y(1)$: $Y(1)$ is number of years lived past surgery if given new surgery (treatment 1), and $Y(0)$ is number of years lived past surgery if given traditional surgery (treatment 0).

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

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

Again, because each individual's treatment status is independent of that of the other individuals, the probability of $\mathbf{W}$ is the product of the probabilities of the eight W_i's, the propensity scores given the potential outcomes.

Possible Assignments

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

The take-away message: Some involvement of randomization in the process may not be enough. Here, treatment assignment is nonignorable (and thus the data will be difficult to analyze), even though it has some elements of randomization. Data from nonignorable designs are inherently difficult to analyze correctly for causal effects.

Part II: Causal Inference Based on the Assignment Mechanism – Design before Outcome Data

Subsection 7: “Fisherian” Significance Levels And Intervals for Additive Effects

Fundamental idea: Inference based solely on the assignment mechanism

Definition of Basic Concepts

Null Hypothesis: An initial supposition regarding the nature of the treatment effect, usually specified to test its consistency against observed data

Sharp Null Hypothesis: A null hypothesis articulated with specificity sufficient to allow the researcher to fill in a hypothetical value for each unit’s missing potential outcomes

Test Statistic: A function of the data that the researcher uses to determine the consistency of the null hypothesis with observed data

Randomization Distribution: All possible values of the test statistic and the probabilities associated with each under the (randomized) assignment mechanism

p-value: Under the assumption that the null hypothesis is correct, the probability of observing a value of the test statistic as extreme or more extreme than the one actually observed

Fisher Interval: The set of possible values of the causal quantity of interest corresponding to test statistics with p-values that fall within some range set by the researcher; usually interpreted to be a set of plausible values of the average causal effect

A Primer on Proof by Contradiction

Steps

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

Example II-1: Word Problem

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

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

1. Start by assuming Jane is 30.
2. This means Jane's mother must be 53 (because Jane is 23 years younger than her mother), which means Jane's father must be 5 (because her parents' ages sum to 58). However, Jane's mother is then 48 years older than her father. We've reached a contradiction, because 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, because 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 (there may be more than one solution; this proof does not eliminate the possibility of multiple solutions).

Question: How would you prove that $\sqrt{6}$ is irrational?

Fisher Test in a Completely Randomized Experiment: Six Steps

Before looking at the observed Y data:

1. Specify a sharp null hypothesis (hypothesis regarding the size of the treatment effect on each unit).
   Usually, use the hypothesis of absolutely no effect of treatment ($Y_i(0) = Y_i(1)$ for all units).
2. Specify a test statistic for estimating the treatment effect and evaluating the null hypothesis.
   Often, use the difference in observed sample means of the treated and control groups ($\overline{y(1)} - \overline{y(0)}$).

Using the Y_{obs} data

3. Calculate the value of the test statistic, and specify values that are more extreme (i.e., unusual).
4. Fill in the missing potential outcomes using the sharp null hypothesis and the observed Y values.

Obtaining a p-value

5. 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).
6. Determine how extreme the value observed in Step 3 is. This is called the significance level or probability value (p-value). One calculates it by adding the probabilities of all assignments that lead to a test statistic value as or more extreme than the value observed.

Question: Why isn't $\overline{y(1)} - \overline{y(0)}$ always an appropriate test statistic?

Example II-2: 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 six units, with three assigned treatment and three assigned control in a completely randomized design. The treatment is watching the programs, control is not watching them. Post-program test scores of the children are given below, where the missing potential outcomes are in parentheses and are filled in using the sharp null hypothesis of no treatment effect on anyone.

1. Specify a sharp null hypothesis: There is no effect of the treatment ($Y_i(0) = Y_i(1)$ for any individual).
2. Specify a test statistic: Here, we choose the difference in means: $\overline{y(1)} - \overline{y(0)}$.
3. Calculate the observed value of the test statistic: $\overline{y(1)} - \overline{y(0)} = 5.1$, where bigger is more extreme.
4. Fill in the missing potential outcomes under a sharp null hypothesis of no effect:

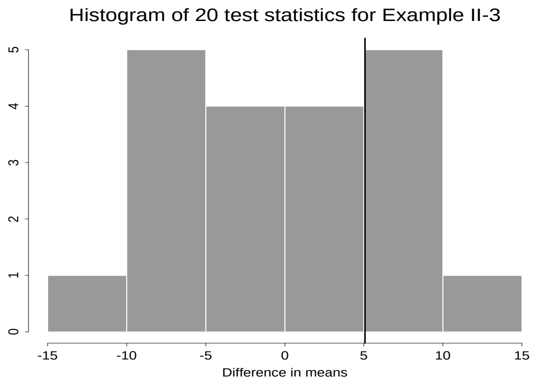
	Actual	Observed	Potential	Outcomes
Unit	Treatment (W)	Outcome	$Y_i(0)$	$Y_i(1)$
1	0	55.0	55.0	(55.0)
2	0	72.0	72.0	(72.0)
3	0	72.7	72.7	(72.7)
4	1	70.0	(70.0)	70.0
5	1	66.0	(66.0)	66.0
6	1	78.9	(78.9)	78.9

5. For each possible assignment, calculate the value of the test statistic: The following table lists all possible randomizations of these data, just like before, except that we have added a third column giving the value of the test statistic that would have been observed under each assignment if the sharp null hypothesis were true. The observed randomization and outcome are in bold.

All Possible Assignments

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

6. Calculate the p-value: We assess the plausibility of the null hypothesis against the observed data: If the null hypothesis were true, the probability of observing the value that we did (5.1) or something more extreme is $6/20 = .3$ (i.e., the p-value or "significance level" is 0.3).



Example II-3: Children’s Television Workshop, Part II

We have the same set-up as in Example II-2, however, we now use a null hypothesis of a treatment effect of 5 points in the table below ($Y_i(1) - Y_i(0) = 5$ for all individuals). This null hypothesis assumes additive treatment effects, i.e., the treatment adds a fixed amount to each control value.

1. Specify a (sharp) null hypothesis: There is an additive treatment effect of 5 points: $Y_i(1) - Y_i(0) = 5$ for all individuals.
2. Specify a test statistic: We again choose the difference in means: $\overline{y(1)} - \overline{y(0)}$.
3. Calculate the observed value of the test statistic: $\overline{y(1)} - \overline{y(0)} = 5.1$, where bigger is more extreme.
4. Fill in missing potential outcomes:

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

5. For each possible assignment, calculate the value of the test statistic: The following table is just like the corresponding one for the previous example, except the third column has values under a different null hypothesis.

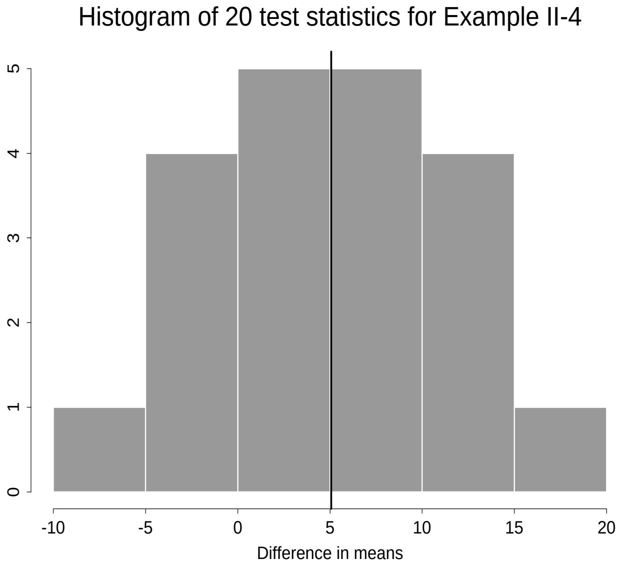
All Possible Assignments

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

6. Calculate the p-value: We assess the plausibility of the null hypothesis against the observed data: If the null hypothesis (of an additive treatment effect of five points) were true, then ten randomizations would give test statistics as extreme or more extreme than what we observed. These randomizations are listed below.

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

The probability of observing the value that we did (5.1) or something more extreme is thus $10/20 = 0.5$ (i.e., the p-value or significance level associated with the null hypothesis of +5 points is 0.5).



Constructing Fisher Intervals in Completely Randomized Experiments

Example II-4: Children's Television Workshop (continued)

We continue example II-3, an experiment regarding the effect of Children's Television Workshop programming on children's reading ability. We are now interested in determining a range of plausible values of the treatment effect.

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

The data are shown below. We now fill in the missing potential outcomes according to a null hypothesis of an additive 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.

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

A few specific examples of this are shown below.

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

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

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

All Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$	$\overline{y(1)} - \overline{y(0)}$	
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) under the null hypothesis of an additive treatment effect of -6 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)

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

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

All Possible Assignments

$\mathbf{W}$	Prob of $\mathbf{W}$	$\overline{y(1)} - \overline{y(0)}$	
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 (here defined as larger than 5.1) under the null hypothesis of an additive causal effect of +12 is $18/20 = .90$ (i.e. the p-value or significance level is 0.90).

Formation of a Fisher Interval

Up until now, we have been careful to define what we meant by “something more extreme” before calculating the p-value associated with the null hypothesis and the observed data. In the current example, for instance, we defined a more extreme value of the test statistic to mean a number larger than the one we observed. Notice by defining “more extreme” in terms of a larger number, we imply that we do not care if the number is smaller, even vastly smaller, than the one we observed. In some scientific settings, a focus on only larger-than-observed numbers might make sense. In others, it may not.

If we do care about departures from the observed value of a statistic in either direction (larger or smaller), we need to extend our p-value calculation procedure. Previously, each time we defined what we meant by “more extreme,” we calculated a p-value, and we interpreted low p-values (say, below .05) as evidence against the null hypothesis. We can see, however, that p-values that are “too high” (say, above .95) may also constitute evidence against the null hypothesis. In effect, by taking p-values that are “too high” as evidence against the null, we have reversed the direction of our definition of “more extreme.”

Why bother with all this? Because if we do, we can use p-values to determine an interval of plausible numbers for the treatment effect, something we call a Fisher interval. Such intervals can be useful to decision makers. To construct a Fisher interval, we systematically go through hypothesized values for the treatment effect. The values that are "plausible," in the sense that their corresponding p-values are between .05 and .95, we keep. The others, we discard. The ones we keep form a 90% Fisher interval.

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

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

Note that -6 is the smallest value considered that has a p-value bigger than .05.

We now consider hypothesized treatment effects bigger than 5:

Treatment Effect	p-value	Treatment Effect	p-value
6	0.50	16	0.90
7	0.55	17	0.90
8	0.60	18	0.90
9	0.65	19	0.90
10	0.70	20	0.90
11	0.75	21	0.90
12	0.85	22	0.90
13	0.85	23	0.90
14	0.85	24	0.90
15	0.85	25	0.95

Note that +24 is the largest value considered that has a p-value smaller than .95. A plausible range of values for the treatment effect (a 90% interval) is thus [-6, 24]: the set of values whose p-value is greater than $.05$ and less than $.95$.

Example II-5: Children's Television Workshop Using Randomized Blocks

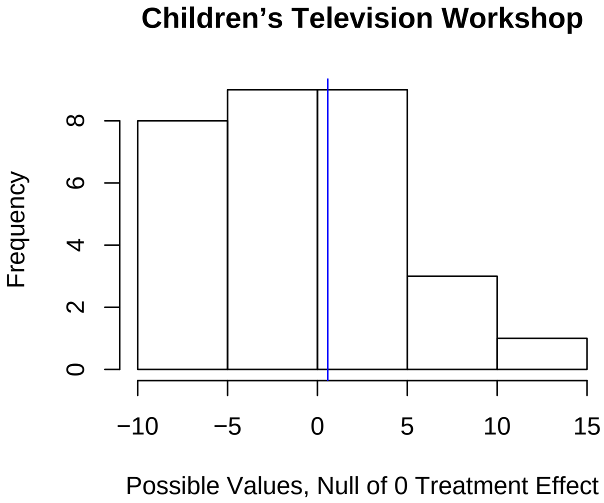
We continue with the Children's Television Workshop example (which we now expand to eight units), but now we block on the covariate indicating female/male. For the purposes of this example, Units 1-5 are girls, 6-8 boys. We assume an assignment mechanism in which three girls and one boy are randomly chosen to receive treatment. We consider two sharp null hypotheses, one of no treatment effect for either boys or girls, and one of a different treatment effect for girls and boys.

a. Sharp null of no treatment effect

1. Specify a sharp null hypothesis: There is no effect of the treatment ($Y_i(0) = Y_i(1)$ for any individual).
2. Specify a test statistic: We again choose the difference in sample means: $\overline{y(1)} - \overline{y(0)}$.
3. Calculate the observed value of the test statistic: $\overline{y(1)} - \overline{y(0)} = \frac{72+64.2+70+66}{4} - \frac{55+72.7+78.9+63.3}{4} = .575$, where more extreme means bigger.
4. Fill in the missing potential outcomes: The sharp null hypothesis allows us to fill in the missing potential outcomes, as follows.

		Actual	Observed	Potential	Outcomes
Unit	Gender	Treatment (W)	Outcome	$Y_i(0)$	$Y_i(1)$
1	F	0	55.0	55.0	(55.0)
2	F	1	72.0	(72.0)	72.0
3	F	0	72.7	72.7	(72.7)
4	F	1	64.2	(64.2)	64.2
5	F	1	70.0	(70.0)	70.0
6	M	1	66.0	(66.0)	66.0
7	M	0	78.9	78.9	(78.9)
8	M	0	63.3	63.3	(63.3)

5. For each possible assignment, calculate the value of the test statistic: There are $\binom{5}{3} = 10$ ways to choose three of five girls to treat and $\binom{3}{1} = 3$ ways to choose one of three boys, and thus the total number of possible assignment vectors is 30. We do not bother to list them all, but a couple of examples include (1, 1, 1, 0, 0, 0, 1, 0, 0), (1, 1, 0, 1, 0, 1, 0, 0), and (0, 0, 1, 1, 1 , 0, 1, 0). Note that (1, 0, 1, 0, 0, 1, 0, 1) is NOT a permissible assignment vector because it has only two treated girls (and two treated boys). Each permissible assignment vector is equally likely, and we can calculate, for each, the value of the test statistic that would be observed if the null hypothesis were true. The thirty values range from -9.28 to 11.3.
6. Calculate the p-value: Assess the plausibility of the null hypothesis against the observed statistic, .575. The observed value of the test statistic is the 18th largest of the thirty equally likely assignments, so the p-value is .6. We summarize this information in a histogram, where the observed value of the test statistic is indicated by a vertical line.



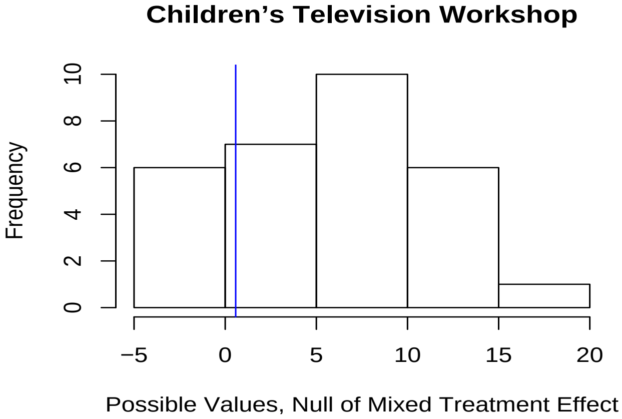
b. We now repeat the analysis of part (a) with a different null hypothesis, this time that the treatment effect for girls is an additive 11 points whereas the treatment effect for boys is an additive -1 point.

1. Specify a (sharp) null hypothesis: $Y_i(1) - Y_i(0) = -1$ for boys, $Y_i(1) - Y_i(0) = 11$ for girls.
2. Specify a test statistic: Here, we again choose the difference in means: $\overline{y(1)} - \overline{y(0)}$.
3. Calculate the observed value of the test statistic: $\overline{y(1)} - \overline{y(0)} = .575$.
4. Fill in the missing potential outcomes: Under the sharp null hypothesis, we can fill in the entire potential outcomes table as follows.

		Actual	Observed	Potential	Outcomes
Unit	Gender	Treatment (W)	Outcome	$Y_i(0)$	$Y_i(1)$
1	F	0	55.0	55.0	(66.0)
2	F	1	72.0	(61.0)	72.0
3	F	0	72.7	72.7	(83.7)
4	F	1	64.2	(53.2)	64.2
5	F	1	70.0	(59.0)	70.0
6	M	1	66.0	(67.0)	66
7	M	0	78.9	78.9	(77.9)
8	M	0	63.3	63.3	(62.3)

5. For each possible assignment, calculate the value of the test statistic. For the reasons articulated in part (a), there are 30 possible treatment assignment vectors, and each is equally likely. For each possible assignment, we calculate a difference that would have been observed if the null hypothesis were true. The values range from -4.28 to 16.3.

6. Calculate the p-value: Assess the plausibility of the null hypothesis against the observed value of the test statistic: .575, the value of the test statistic that we observed, is the 7th largest of 30 equally likely values, so our p-value is $\frac{7}{30} = .23$. Again, we present a histogram.



Questions: Could we construct a Fisher intervals for the females in this example? For the males? Could we construct an overall Fisher interval? What key additional assumption would we need to make to do the latter? Does this assumption seem reasonable?

Example II-6: Bacterial Growth: Before And After, Revisited

We return to the Bacterial Growth example from Part I. The setting, again, is that we wish to assess the effect of a potential antibiotic on the growth of bacteria in a single petri dish. First, we randomly pick a number D between 1 and 20. We begin measuring the number of bacteria in the petri dish on day one. On the Dth day, after taking a measurement for that day, we administer the drug to the dish. We continue to measure the amount of bacteria on each day thereafter, until we reach day 21. In this setting, a unit is the petri dish on a particular day, so we have 21 units. A unit has received treatment if the drug was administered before that day's measurement, otherwise the unit received control.

We supposed that we randomly chose D = 15. We begin with a sharp null hypothesis of no treatment effect. Note that under a sharp null hypothesis, SUTVA is automatically satisfied. If, however, SUTVA makes little scientific sense, then this null hypothesis may make very little scientific sense.

1. Specify a (sharp) null hypothesis: There is no effect of the treatment ($Y_i(0) = Y_i(1)$ for each unit).

2. Specify a test statistic: Here, we do NOT use the difference in means. (Why not?) We use the number of bacteria just after the treatment was administered minus the number of bacteria just before it was administered.

3. Calculate the observed value of the test statistic: Because our randomly chosen D equaled 15, the observed value of this test statistic is $10.880 - 12.904 = -2.024$. We define "more extreme" to mean a number smaller than -2.024.

4. Fill in the missing potential outcomes: The sharp null hypothesis allows us to fill in the following potential outcomes table.

Growth of Bacteria: Potential Outcomes under Sharp Null of No Treatment Effect

Index	Day	W	Y(0)	Y(1)
1	1	0	10.237	(10.237)
2	2	0	10.914	(10.914)
3	3	0	10.286	(10.286)
4	4	0	10.684	(10.684)
5	5	0	11.682	(11.682)
6	6	0	11.092	(11.092)
7	7	0	11.343	(11.343)
8	8	0	11.585	(11.585)
9	9	0	11.252	(11.252)
10	10	0	12.285	(12.285)
11	11	0	11.913	(11.913)
12	12	0	12.406	(12.406)
13	13	0	12.331	(12.331)
14	14	0	13.082	(13.082)
15	15	0	12.904	(12.904)
16	16	1	(10.880)	10.880
17	17	1	(11.513)	11.513
18	18	1	(11.704)	11.704
19	19	1	(11.807)	11.807
20	20	1	(11.903)	11.903
21	21	1	(12.752)	12.752

5. For each possible assignment, calculate the value of the test statistic: For D = 1, if the null hypothesis were true, the test statistic would equal $10.914 - 10.237 = .707$. For D = 2, if the null hypothesis were true, the test statistic would equal $10.286 - 10.914 = -.628$. We continue in this manner to calculate the value of the test statistic over all possible randomizations, assuming that the null hypothesis were true. The table below shows all possible values.

Growth of Bacteria: Values of Test Statistic
under Sharp Null of No Treatment Effect

Randomized Choice: D	Prob. of Selection	Test Stat Value
2	$\frac{1}{20}$	.707
3	$\frac{1}{20}$	-.628
4	$\frac{1}{20}$	.398
5	$\frac{1}{20}$	.998
6	$\frac{1}{20}$	-.59
7	$\frac{1}{20}$	.251
8	$\frac{1}{20}$	.242
9	$\frac{1}{20}$	-.333
10	$\frac{1}{20}$	1.033
11	$\frac{1}{20}$	-.372
12	$\frac{1}{20}$	.493
13	$\frac{1}{20}$	-.075
14	$\frac{1}{20}$	.751
15	$\frac{1}{20}$	-.178
16	$\frac{1}{20}$	-2.024
17	$\frac{1}{20}$	.633
18	$\frac{1}{20}$	.191
19	$\frac{1}{20}$	.103
20	$\frac{1}{20}$	.096
21	$\frac{1}{20}$	.849

6. Of the values corresponding to all 20 possible equally likely randomizations, the observed value of the test statistic, -2.024, is the smallest. The p-value is thus $\frac{1}{20} = .05$.

Now we construct a Fisher interval for an additive causal effect. Because we previously defined “more extreme” to mean a statistic smaller than -2.024, we pick a new hypothesized value for the additive treatment effect smaller than -2.024 (say -2.1), fill in the potential outcomes table, calculate the value of the test statistic over all possible randomizations, and see if the p-value is **bigger than** .05. If the p-value is bigger than .05, then we know that our original number (-2.1) is a plausible one for the additive treatment effect, and we include it in our interval. We repeat this procedure until we find a value leading to a p-value of less than or equal to .05. It turns out that this value is -3.058, so we claim that anything less than -3.058 is an implausible value for the additive causal effect, and so does not belong in our 90% interval. The potential outcomes table under the sharp null hypothesis of a treatment effect of -3.058 appears below.

Growth of Bacteria: Potential Outcomes Table under Sharp
Null Hypothesis of Treatment Effect of -3.058

Randomized Choice: D	Prob. of Selection	$Y(0)$	$Y(1)$	Test Stat Value
*	*	10.237	7.180	*
2	$\frac{1}{20}$	10.914	7.857	-2.381
3	$\frac{1}{20}$	10.286	7.229	-3.686
4	$\frac{1}{20}$	10.684	7.627	-2.660
5	$\frac{1}{20}$	11.682	8.625	-2.060
6	$\frac{1}{20}$	11.092	8.035	-3.648
7	$\frac{1}{20}$	11.343	8.286	-2.807
8	$\frac{1}{20}$	11.585	8.528	-2.816
9	$\frac{1}{20}$	11.252	8.195	-3.391
10	$\frac{1}{20}$	12.285	9.228	-2.025
11	$\frac{1}{20}$	11.913	8.856	-3.430
12	$\frac{1}{20}$	12.406	9.349	-2.565
13	$\frac{1}{20}$	12.331	9.274	-3.133
14	$\frac{1}{20}$	13.082	10.025	-2.307
15	$\frac{1}{20}$	12.904	9.847	-3.236
16	$\frac{1}{20}$	13.937	10.880	-2.024
17	$\frac{1}{20}$	14.57	11.513	-2.425
18	$\frac{1}{20}$	14.761	11.704	-2.867
19	$\frac{1}{20}$	14.864	11.807	-2.955
20	$\frac{1}{20}$	14.96	11.903	-2.962
21	$\frac{1}{20}$	15.809	12.752	-2.209

Next, we start picking value bigger than -2.024 (say, -1.9), and we check to see whether this value generates a p-value **below** .95. If it does, then we know that the value is a plausible additive treatment effect and that it belongs in our 90% interval. We repeat this procedure with progressively bigger values until we find one leading to a p-value greater than or equal to .95. It turns out that this value is -1.395, so we know that anything bigger than than -1.395 is an implausible value for the additive treatment effect. The potential outcomes table under the sharp null hypothesis of a treatment effect of -1.395 appears below.

Growth of Bacteria: Potential Outcomes Table under Sharp
Null Hypothesis of Treatment Effect of -1.395

Randomized Choice: D	Prob. of Selection	$Y(0)$	$Y(1)$	Test Stat Value
*	*	10.237	8.841	*
2	$\frac{1}{20}$	10.914	9.518	-0.718
3	$\frac{1}{20}$	10.286	8.890	-2.023
4	$\frac{1}{20}$	10.684	9.288	-0.997
5	$\frac{1}{20}$	11.682	10.286	-0.396
6	$\frac{1}{20}$	11.092	9.696	-1.985
7	$\frac{1}{20}$	11.343	9.947	-1.144
8	$\frac{1}{20}$	11.585	10.189	-1.153
9	$\frac{1}{20}$	11.252	9.856	-1.728
10	$\frac{1}{20}$	12.285	10.889	-0.362
11	$\frac{1}{20}$	11.913	10.517	-1.767
12	$\frac{1}{20}$	12.406	11.010	-0.902
13	$\frac{1}{20}$	12.331	10.935	-1.470
14	$\frac{1}{20}$	13.082	11.686	-0.644
15	$\frac{1}{20}$	12.904	11.508	-1.573
16	$\frac{1}{20}$	12.276	10.880	-2.024
17	$\frac{1}{20}$	12.909	11.513	-0.762
18	$\frac{1}{20}$	13.100	11.704	-1.204
19	$\frac{1}{20}$	13.203	11.807	-1.292
20	$\frac{1}{20}$	13.299	11.903	-1.299
21	$\frac{1}{20}$	14.148	12.752	-0.546

Our 90% Fisher interval for the additive causal effect is thus $(-3.057, -1.396)$.

Two final aspects of the Before And After/Growth of Bacteria example deserve special attention. First, is SUTVA, specifically the part of SUTVA that specifies no interference among units, a reasonable assumption here? Remember that before making any hypotheses or assumptions, for Unit i, there are 20 different values of $\mathrm{Y}_i(1)$ that could have been observed, each corresponding to a day that might have been selected for the addition of the drug to the petri dish. For the same reason, before making any hypotheses or assumptions, there are 20 different values of $\mathrm{Y}_i(0)$ for Unit i. A sharp null hypothesis of no treatment effect allows us to fill in **all 40** of these potential outcome for unit i, 39 of which are missing, with the same value (the observed one). Under such an assumption, the treatment one unit receives does not affect another unit's potential outcomes, and so SUTVA holds. For similar reasons, SUTVA is satisfied under any sharp null hypothesis that implicitly assumes that there is exactly one value of $Y_i(1)$ and one value of $Y_i(0)$ because by definition such a hypothesis allows us to fill in all of the missing potential outcomes. Once these outcomes are "known," they cannot depend on another unit's treatment assignment.

The difficulty here is that our scientific inferences should also depend on what we could observe if a sharp null hypothesis were not true. The formation of a Fisher interval requires the assessment of the plausibility of several null hypotheses. In other words, it is one thing to assume SUTVA under a sharp null hypothesis of no treatment effect. It is another to assume that SUTVA is scientifically reasonable under a whole range of possible treatment effects. In terms of the specific example above, do you believe that Unit 15's Y(1) value would be the same if Unit 14 had received treatment instead of control? If you do not, then you do not believe that SUTVA holds, and the techniques introduced thus far must be modified to account for this fact, a topic not pursued here.

Second, in this example, we chose not to use the difference in observed means as the test statistic. We made this choice because we expected the bacteria to multiply (or die out) in the petri dish as time progresses, so the difference in means might well have varied systematically in time before and after application of the antibiotic even if the treatment had no effect. In addition, the difference in means might have very little chance of rejecting any null hypothesis; see the figure on page I.5-7. Choosing a test statistic appropriate to the situation can be a difficult task, but it can also be critical for drawing the best inference from the data.

Example II-7: Efficiency Benefits from Using Covariates
Comparing Test Statistics in a Completely Randomized Experiment

In general, we would like our point estimates to be close to the truth. In addition, we would like our intervals to be narrow, because narrow intervals imply a greater knowledge about the quantity of interest. Speaking generally, we use the term ”precision” to refer to interval width. A more “precise” estimate implies a narrower interval.

Proper use of covariates can increase the precision of estimation. For example, suppose we are interested in estimating the effect of an SAT prep class on SAT test scores. There are four students in the experiment, where two will be randomly chosen to receive treatment and the other two will receive control (no class). Consider a hypothetical situation where we know all students' potential outcomes under both treatment and control. We also observe students' SAT scores before treatment assignment (X).

Data

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

Fisher Test Results for All Possible Assignments

W	$\overline{y_1} - \overline{y_0}$	(post) p-value	Interval	$\overline{y_1 - X} - \overline{y_0 - X}$	(gain) p-value	Interval
1100	$475 - 600 = -125$	1	(-151, 1)	$125 - 50 = 75$	0.17	(49, 101)
1010	$475 - 550 = -75$	0.83	(-101, 101)	$75 - 50 = 25$	0.5	(-1, 51)
1001	$550 - 500 = 50$	0.5	(-51, 251)	$100 - 50 = 50$	0.17	(49, 51)
0110	$550 - 500 = 50$	0.5	(-101, 201)	$100 - 50 = 50$	0.5	(-1, 101)
0101	$625 - 450 = 175$	0.17	(149, 351)	$125 - 50 = 75$	0.17	(49, 101)
0011	$625 - 400 = 225$	0.17	(199, 351)	$75 - 50 = 25$	0.5	(-1, 51)

First, look at the table entitled “Data.” Two causal effects are defined, one (in the far right column) that uses the Pre-test SAT covariate to define gain scores, one (in the middle column) uses just the post-test scores. Notice that each unit has the same causal effect under either definition.

Now look at the table entitled ‘Fisher Test Results for All Possible Assignments.” Each row is one possible randomization in a completely randomized experiment (the “W” column specifies the assignment vectors). The second, third, and fourth columns show the observed differences in means, p-values, and Fisher intervals for a comparison of post-test SAT scores with and without treatment; no use is made of the pre-test covariate. The fifth, sixth, and seventh columns show observed differences in means, p-values, and Fisher intervals for a comparison of gain scores; thus, these columns do use the pre-test covariate. You can see that the two different definitions of causal effect produce different p-values and intervals; the p-values and intervals in the right column are formed using the covariate information, whereas those in middle column use post-test scores. Notice that the intervals in the right column are narrower than those in the middle column. Thus, by using the information in the Pre-test SAT covariate, we achieve greater precision in our estimation.

Subsection 8: "Neymanian" Unbiased Estimation And Confidence Intervals

Definition of Basic Concepts

Unbiased estimator of the treatment effect: A statistic whose average value over all possible randomizations equals the true treatment effect. We have already seen in Example I-7 that $\overline{y(1)} - \overline{y(0)}$ is unbiased for $\overline{Y(1)} - \overline{Y(0)}$ in completely randomized experiments.

Confidence (Neyman) Interval: An interval which, over all possible randomizations, includes the true value of the quantity of interest at least as often as advertised. For example, a 95% confidence interval includes the true value 95% of the time *or more*.

Variance: Essentially, the average of the squared deviations from the mean; mathematically, $\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$ *Example of calculating a variance: Variance of the set (4, 5, 9). The mean of this set is* $\frac{4+5+9}{3} = 6$. *The squared deviations are* $(4-6)^2 = 9$, $(5-6)^2 = 1$, *and* $(9-6)^2 = 9$, *so the variance is* $\frac{1}{3-1} * (4+1+9) = 7$. *Note: The critical point about a variance is that it constitutes a measure of how "spread out" the data are.*

Sample variance of Y in the observed treated group, s_1^2: The variance of the set of the observed treated values

Sample variance of Y in the observed control group, s_0^2: The variance of the set of the observed control values

True variance of $\overline{y(1)} - \overline{y(0)}$: the variance of $\overline{y(1)} - \overline{y(0)}$ over all possible randomizations

Usual estimator for the variance of $\overline{y(1)} - \overline{y(0)}$: $\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}$

In 1923, a statistician named Jerzey Neyman proved certain properties about the difference in sample means, $\overline{y(1)} - \overline{y(0)}$. Specifically, he proved that in a completely randomized experiment, this statistic was an unbiased estimate for the true average treatment effect. He also showed certain things about the variance of $\overline{y(1)} - \overline{y(0)}$ in a completely randomized experiment. Specifically, he showed that

1. if the treatment effect is additive, the usual estimator of the variance of $\overline{y(1)} - \overline{y(0)}$ is exactly unbiased for the true variance of $\overline{y(1)} - \overline{y(0)}$; and
2. this usual estimator is positively biased for the true variance if the treatment effect is not additive.

Therefore, assuming additivity when estimating variance tends to lead to overestimates of variance, i.e., is conservative. Note that "conservative" does not necessarily mean "good." Imagine a confidence interval formed by stating that a random 95% of the time, the interval is any positive or negative number, and that 5% of the time, the interval is the number 0. Such an interval would cover the true value of any quantity of interest at least 95% of the time, and thus would also be a "conservative" interval. It would not, however, be of any use.

The following three "Cases" illustrate what Neyman proved (see Neyman (1923)).

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

Because $\overline{Y_i(1)} - \overline{Y_i(0)} = 3$, the true average treatment effect is 3.

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

All Possible Assignments of Four Units to Treatment

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

The average of all possible values of $\overline{y(1)} - \overline{y(0)}$ that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect. However, we see that the difference in medians (which may be of clinical interest) is not unbiased for either the true difference in means or the true difference in medians (which is $3 = 8.5 - 5.5$). Note that the variance of the set of possible values of $\overline{y(1)} - \overline{y(0)}$ is 8.225.

Case 2: There is an average treatment effect of 3 years, but the effect is not additive.

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

Once again, $\overline{Y_i(1)} - \overline{Y_i(0)} = 3$, so the true average treatment effect is 3. The table below again lists all possible randomizations and the corresponding observed estimates of the treatment effect.

All Possible Assignments

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

The average of all possible differences in means that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect, even when there is not an additive treatment effect. However, we see that again, the difference in medians is not unbiased for either the true difference in means or the true difference in medians ($3.0 = 8.5 - 5.5$). Note also that the variance of the set of possible values of $\overline{y(1)} - \overline{y(0)}$ is 7.9, which is smaller than the corresponding value for Case 1, when additivity held.

Case 3: There is an average treatment effect of 3 years, but the effect is far from additive.

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

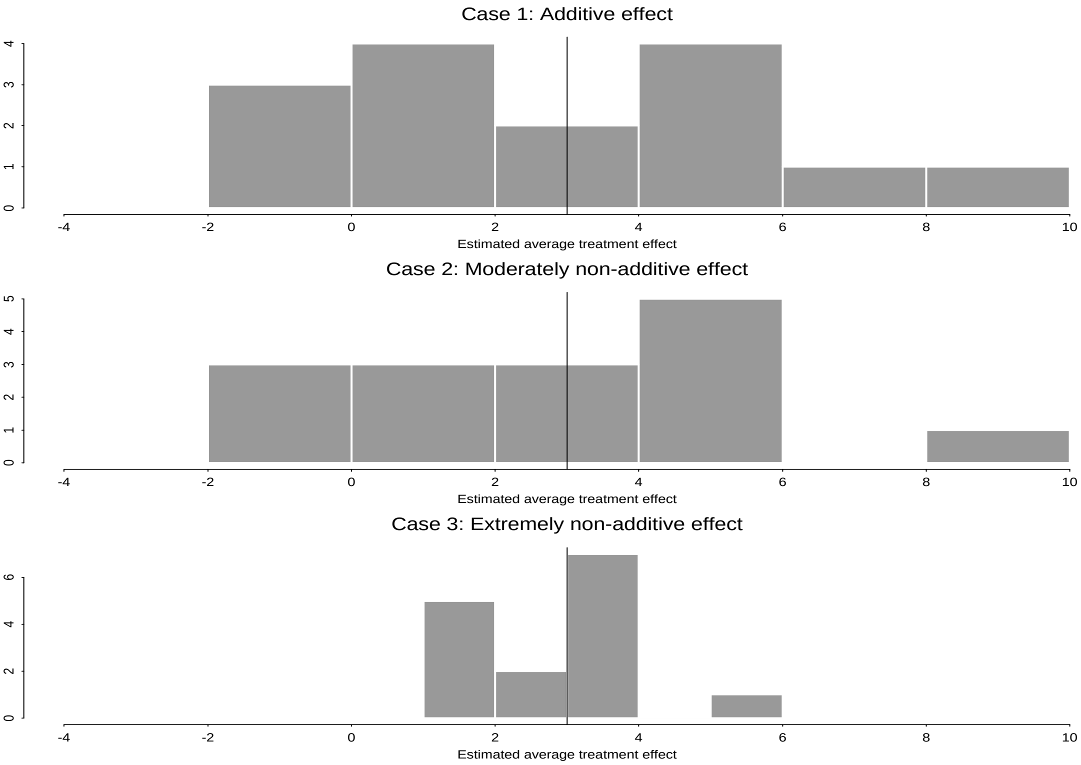
Again, the true average treatment effect is 3. The table below again lists all possible randomizations and the corresponding observed estimates of the treatment effect.

All Possible Assignments

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

The average of all possible test statistics that the doctor could see is 3. Thus, the difference in means is an unbiased estimate of the true average treatment effect, even when there is not an additive treatment effect. We see that again, the difference in medians is not unbiased for either the true difference in means or the true difference in medians ($3.0 = 8.5 - 5.5$). Note also that the variance of the set of possible values of $\overline{y(1)} - \overline{y(0)}$ is 1.09, which is much less than the corresponding values for Case 1 and Case 2.

The following histograms show the estimates of the average treatment effect under all possible randomizations for each of the three cases above. Note that greater spread occurs as the causal effect becomes more additive.


Case 1: Additive effect

Case 2: Moderately non-additive effect

Case 3: Extremely non-additive effect

The take-away message from Cases 1-3: In a completely randomized experiment, the observed difference in means, $\overline{y(1)} - \overline{y(0)}$, is an unbiased estimator of the true average treatment effect, $\overline{Y(1)} - \overline{Y(0)}$. The true variance of the the estimator $\overline{y(1)} - \overline{y(0)}$ decreases as the treatment effect becomes less additive. These principles hold for any completely randomized experiment, as well as for simple generalizations such as randomized blocks.

Neyman Confidence Intervals in Completely Randomized Experiments

In completely randomized experiments with a large number of observations, we can construct an approximate upper bound and lower bound of a Fisher interval for an additive treatment effect by using the following formulas.

$$\left(\left(\overline{y(1)}-\overline{y(0)}\right)-2*\sqrt{\frac{s_1^2}{n_1}+\frac{s_0^2}{n_0}},\ \left(\overline{y(1)}-\overline{y(0)}\right)+2*\sqrt{\frac{s_1^2}{n_1}+\frac{s_0^2}{n_0}}\right)$$

Neyman showed that this interval is typically wider than needed to obtain 95% coverage of the true average treatment effect, except when the true effect is additive and sample sizes are large. This method of obtaining confidence intervals can be extended to settings other than completely randomized designs, but we do not delve into that.

We illustrate the formation of such an interval with Case 3, above. We use the data from Case 3, which we reproduce below for convenience, and we suppose that we observe random assignment (1, 0, 1, 1, 1, 0).

Patient	$Y_i(1)$	$Y_i(0)$
1	10	?
2	?	9
3	5	?
4	12	?
5	8	?
6	?	5

$\overline{y(1)} = \frac{10+5+12+8}{4} = 8.75$

$\overline{y(0)} = \frac{9+5}{2} = 7$

$\overline{y(1)} - \overline{y(0)} = 8.75 - 7 = 1.75$

$s_1^2 = \frac{1}{4-1} * ((10-8.75)^2 + (5-8.75)^2 + (12-8.75)^2 + (4-8.75)^2) = 16.25$

$s_0^2 = \frac{1}{2-1} * ((9-7)^2 + (5-7)^2) = 8$

$2 * \sqrt{\frac{s_1^2}{n_1}+\frac{s_0^2}{n_0}} = 2 * \sqrt{\frac{16.25}{4}+\frac{8}{2}} = 5.7$

The Neyman Interval: $(1.75 - 5.7, 1.75 + 5.7) = (-3.95, 7.45)$

Although this interval certainly contains the true average treatment effect (3), it is quite wide. With the widespread availability of computers, when doing assignment-based inference, there is little need to prefer Neyman-type formulas for creating confidence intervals over simulation methods. Two portions of Neyman's work remain important, however: (1) the test statistic used should be unbiased or approximately unbiased for the true treatment effect (remember the Perfect Doctor example), and (2) assuming additive treatment effects is often conservative, at least in large samples.

Example II-7, Continued: Comparing Test Statistics in a Completely Randomized Experiment: Efficiency Benefits from Using Covariates

Review page II-7.20-.21. That example demonstrated how proper incorporation of a covariate into the analysis could lead to more significant p-values and smaller Fisher intervals for a chosen unbiased test statistic. This principle remains true with the Neyman interval, as the table below demonstrates (using the same data).

Estimates for All Possible Assignments

W	$\overline{y_1} - \overline{y_0}$	(post) Std. Dev.	$\overline{y_1 - X} - \overline{y_0 - X}$	(gain) Std. Dev.
1100	$475 - 600 = -125$	90	$125 - 50 = 75$	25
1010	$475 - 550 = -75$	125	$75 - 50 = 25$	25
1001	$550 - 500 = 50$	158	$100 - 50 = 50$	0
0110	$550 - 500 = 50$	150	$100 - 50 = 50$	50
0101	$625 - 450 = 175$	125	$125 - 50 = 75$	25
0011	$625 - 400 = 225$	90	$75 - 50 = 25$	25
Average	50	125.8*	50	28.9**

* 125.8 = $\sqrt{\frac{90^2+125^2+\cdots+90^2}{6}}$

** 28.9 = $\sqrt{\frac{25^2+25^2+\cdots+25^2}{6}}$

Subsection 9: Extension to Studies with Variable But Known Propensities – Blocking

Definition of Basic Concepts

Propensity Score: For each unit, the probability of being assigned treatment; formally, $p(W_i = 1|\boldsymbol{X}, Y(0), Y(1))$ (definition repeated from Subsection 4)

In Example II-5, with the Children's Television Workshop data, we introduced the idea of using covariates to divide the units into blocks and doing randomized experiments within these blocks. In Example II-5, the covariate we used was male/female, and we implemented one experiment for the boys and another one for the girls. This structure allowed us to assign a probability of treatment for girls ($\frac{3}{5}$) that is different from the probability of treatment for boys ($\frac{1}{3}$), something we might want to do because of the scientific setting involved.

What should we do, however, if we want to use more covariates than just sex? Supposed we also want to use race (say African-American, Caucasian, Asian-American, Other), ethnicity (Hispanic versus non-Hispanic), and height (short, medium, tall)? Now we have 2 x 4 x 2 x 3 = 48 potential blocks. We may not have enough experimental units to go around. Moreover, suppose we wanted to "block" in some way on a continuous covariate, such as blood pressure. We could divide the continuous variable blood pressure into Low, Medium, and High, much as we did with height, but still there are many possible categories – how many should we use?

In these circumstances, we can still have the probability that units receive treated or control depend on the covariates. Once we have these individual probabilities, then the collection of the propensity scores become a single covariate to be used for blocking. In fact, as long as we have a reasonable number of units, we can get approximately unbiased estimates of the average treatment effect (and variance estimates) by grouping units with similar propensity scores together, then treating each block as a mini randomized experiment. When designing the study, each unit's propensity score can depend in almost any way we want on the unit's covariates; as long as the propensity score does not depend on the unit's potential outcomes (i.e., as long as treatment assignment is unconfounded), knowing the propensity score is enough to allow us to form meaningful blocks. If we know the set of propensity scores, we can obtain approximately unbiased estimates by focusing on it while ignoring the specific covariates. The theorems that prove why this works are beyond the scope of this introduction. The method, however, is easy to understand, and we illustrate it with examples.

Example II-8: Blocking with Known Propensity Scores

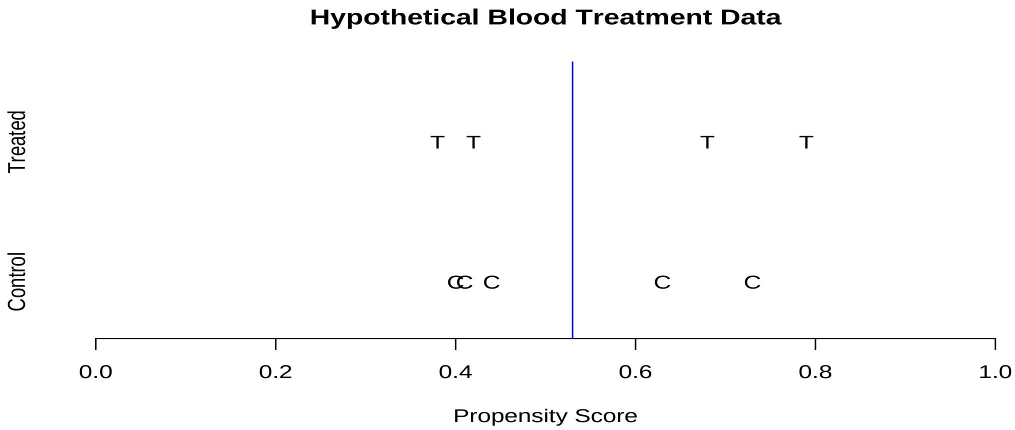
A (hypothetical) company wants to study the effect of a new kind of pill designed to treat a temporary but nontrivial illness of the blood. Both the company and the general public suspect (perhaps based on animal trials) that the pill works, but the FDA will not allow the company to begin sales until the medication's effectiveness is assessed in a randomized trial. The pill is designed for persons between the ages of 20 and 45, and its side effects include temporary hair loss and sensitivity to sunlight. To protect the integrity of the experimental trial, the company has developed a placebo that mimics these side effects. No previous evidence suggests that the drug's effectiveness varies with either gender or age, but it is possible. The outcome to be measured is the blood level of Substance A.

The side effects place the company in a difficult position. The company believes that women are more likely than men to object to the hair loss side effect, and that twenty-somethings are more likely than thirty-somethings to object to a prohibition on sunbathing. The company wants to encourage women and the youthful to volunteer for the study by offering them a higher probability of receiving the active treatment. It tells potential study volunteers that it will decide who receives treatment as follows: For female volunteers, the probability of receiving treatment is $\frac{19}{Age}$. For male volunteers, the probability of receiving treatment is $\frac{14}{Age}$. Thus, for a specific age, women are more likely to receive the active treatment. For example, the probability that a 22-year-old woman receives treatment (i.e., her propensity score) is $\frac{19}{22} = .86$, whereas the propensity score for a 38-year-old man is $\frac{14}{38} = .36$.

The company's strategy attracts nine volunteers who all agree to comply during the experiment. The resultant data appear below.

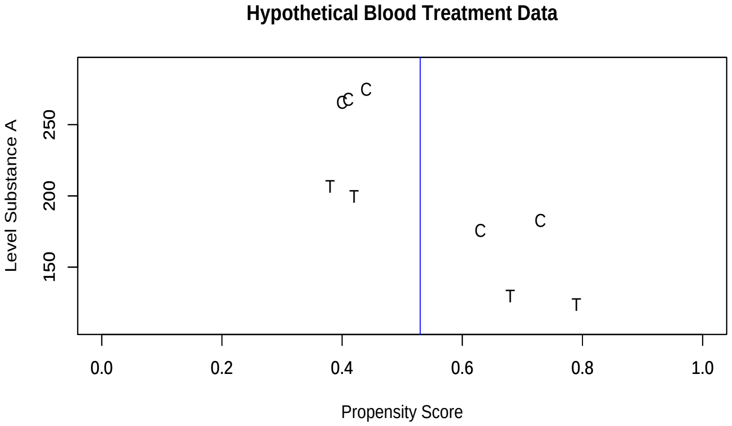
Unit	Sex	Age	Prob of W_i = 1	W	Y(0)	Y(1)
1	F	28	$\frac{14+5}{28} = .68$	1	?	130
2	F	45	$\frac{14+5}{40} = .42$	1	?	200
3	F	26	$\frac{14+5}{26} = .73$	0	183	?
4	F	24	$\frac{14+5}{24} = .79$	1	?	124
5	M	22	$\frac{14}{22} = .63$	0	176	?
6	M	35	$\frac{14}{35} = .40$	0	266	?
7	M	37	$\frac{14}{37} = .38$	1	?	207
8	M	32	$\frac{14}{32} = .44$	0	275	?
9	M	34	$\frac{14}{34} = .41$	0	268	?

The theorems tell us that once we have the propensity score, we can obtain an approximately unbiased estimate for the average treatment effect by focusing on it alone (ignoring sex and age). Thus, the display below contains the information we need to form propensity score blocks.



The propensity scores suggest forming two blocks, one consisting of the five units to the left of the vertical line with low propensity scores, the other consisting of the four units to the right with high propensity scores. Block One will consist of Units 2, 6, 7, 8, and 9, whereas Block Two will consist of Units 1, 3, 4, and 5. Each block can be considered to be its own completely randomized experiment. Notice that we did **not** use the outcome variable in deciding on these blocks.

Having formed the blocks, we collect the data and compare the outcomes for the units. A plot of the data including the outcomes appears below; “C” signifies control and “T” signifies treated.



We now demonstrate the use of the Fisher and Neyman techniques on the two blocks. In both methods, we treat each block as its own mini-randomized experiment. In the Fisher analysis, we hypothesize a constant additive treatment effect that applies to the units in both blocks. The number of randomizations is a little tricky. In Block One, there are $\binom{5}{2} = 10$ possible ways to choose two treated from five total units. In Block Two, there are $\binom{4}{2} = 6$ possible ways to choose two treated from four units. We are considering the two blocks as independent experiments, so any randomization from Block One can be combined with any randomization from Block Two. Therefore, there are $10 \times 6 = 60$ possible randomizations. We will not display them all here, but we show two possible randomizations in the table below. In both tables, the units from Block One are in **bold** type.

Assignments under
Randomization 1

Unit	W_i
1	1
2	**0**
3	0
4	0
5	1
6	**1**
7	**1**
8	**0**
9	**0**

Assignments under
Randomization 2

Unit	W_i
1	0
2	**0**
3	1
4	1
5	0
6	**0**
7	**0**
8	**1**
9	**1**

We use the value of the constant additive treatment effect specified under the null hypothesis to fill in the missing values of the potential outcomes tables in the two blocks. Then, for each randomization, we calculate a difference in means for the first block and a difference in means for the second block. To combine the two differences into a single number, we weight the differences according the number of units in each block, meaning we multiply the first difference by $\frac{5}{9}$, multiply the second difference by $\frac{4}{9}$, and add the two resulting numbers together. Doing this procedure for each randomization gives us 60 possible weighted average treatment effects that could have been observed. Ranking them allows us to assess how plausible the observed statistic (which is -60.09) is under the null hypothesis; for a null hypothesis of no treatment effect, the observed statistic is the most extreme, so the p-value $\frac{1}{60} = .0167$. Repeating this whole procedure over and over again for different additive null hypotheses allows us to construct a 93.3% $(1 - (\frac{1}{60} * 4))$ Fisher interval, which equals (-68, -52).

For the Neyman analysis, we do the following. For Block One, the difference in observed means is $\frac{200+207}{2} - \frac{266+275+268}{3} = 203.5 - 269.7 = -66.2$. The estimated sampling variance of the difference in observed means for Block One, using the formula on page II.8-1, is 19.7. For Block Two, the difference in observed means is $\frac{130+124}{2} - \frac{183+176}{2} = 127 - 179.5 = -52.5$, and the estimated sampling variance for the difference in observed means is 21.25. To combine the two means, we again weight according to the fraction of observations in each block, and the result is $(\frac{5}{9} * -66.2) + (\frac{4}{9} * -52.5) = -60.1$. For the variance, we weight according to the squared fraction of observations in each block, and the result is $[(\frac{5}{9})^2 * 19.7] + [(\frac{4}{9})^2 * 21.25] = 10.3$, which corresponds to a standard error of $\sqrt{10.3} = 3.21$. Thus, the large-sample Neyman 95% interval is (-66.5, -53.7), which is quite similar to the 93.3% Fisher interval.

Subsection 10: Extension to Studies with Unknown Propensities – Blocking on Estimated Propensities

Definition of Basic Concepts

Observational study: An attempt to draw inferences about the causal effect of an active treatment versus a control treatment based on data in which the investigator did not decide which units would receive treatment and which would receive control, but rather observed the assignments that the units received, according to an unknown assignment mechanism

Logistic regression: One method of using the units' covariates to estimate their propensity scores

Thus far, nearly all of our discussion has focused on experiments, i.e., situations in which someone decides who will receive treatment and who will receive control. In many cases, however, complications prevent us from doing an experiment. For example, suppose we wish to assess the effect of smoking in teenage populations on lung cancer for the purpose of making policy next year. Could we take a group of teenagers, randomly assign half to smoke and half not to smoke, then wait to see how many of each group develop lung cancer? This is unrealistic for at least two reasons. The first is ethics. The second is that lung cancer takes 30-40 years to develop – too long to wait to give advice about next year's policy. In some situations, then, we must perform observational studies on existing data.

A fundamental precept of the understanding of causal inference advanced in this course is that to obtain valid and reliable inferences from an observational study, one must imagine a corresponding "template" randomized experiment and pretend that each unit's propensity score was lost or discarded, and therefore must be reconstructed. To implement an observational study, we first estimate each unit's propensity score, then proceed as we did in the previous section, that is, by blocking on it. We must assume our propensity score estimates are reasonably close to the true probabilities of being treated (given the covariates), and we must hope (assume) that each unit's probability of being treated does not depend on its unobserved potential outcomes (unconfounded treatment assignment – ignorability). The first assumption is less difficult to assess; the second is needed to make progress, and its plausibility depends on the science of the situation.

To proceed in this manner, we need a way to estimate each unit's probability of being treated as a function of covariates. The following example illustrates how to do this by creating blocks of observations based on the covariates when only a couple of covariates are available. If we have measurements on more covariates, we must use more sophisticated techniques to obtain propensity score estimates.

Example II-9: Estimating Unknown Assignment Probabilities

We are interested in determining which of two types of surgery is better with respect to post-operative life expectancy. The control is a standard surgery and treatment is a recently developed surgery. Ethically it is difficult to do a randomized experiment because the new surgery is thought to be better, so we decide to do an observational study using data on patient outcomes that hospitals have already collected. We have data on 2000 patients, where half received the standard surgery (treatment 0) and half received the new surgery (treatment 1). The hospitals also recorded each patient's age and cholesterol level before surgery.

Just as when we are designing a randomized experiment, when designing an observational study, we must not consider the outcome data: for objective and valid results, we cannot look at the implied estimated causal effect of treatment versus control when choosing our design.

As a first step, we need to estimate each individual's propensity score. We could do this in a number of ways. First, we might estimate the probability of receiving treatment 1 as $\frac{1}{2}$ for every unit, because we observe that 1000 of the 2000 patients received the new treatment. If we did so, we would essentially be assuming that the units' probabilities of being treated had nothing to do with their ages or pretreatment cholesterol levels. Is such an assumption reasonable? The following two tables suggest otherwise for this dataset.

Age Block	Total number	Number Treated	Number Control	Block's Estimated Treatment Probability
0-20	137	94	43	$\frac{94}{137} = 0.69$
20-40	455	276	179	0.61
40-60	790	393	397	0.50
60-80	479	193	286	$\frac{193}{479} = 0.28$
80-100	118	31	87	0.26

The probability of receiving treatment certainly seems to be lower for older patients. Perhaps the new surgery is more invasive, and thus doctors are more reluctant to recommend it to older patients.

Cholesterol Block	Total Number	Number Treated	Number Control	Block's Estimated Treatment Probability
0-200	175	155	20	0.89
200-250	475	354	121	$\frac{354}{475}$ =0.75
250-300	704	343	361	0.49
300-350	464	130	334	0.28
350-400	162	16	146	0.10

This table shows that people with low cholesterol are more likely to receive the new treatment. Again, the nature of the new surgery may explain this difference.

Because treatment assignment appears to vary with age and pretreatment cholesterol, we should include these covariates when estimating the propensity scores. We can do so by looking at a two-way table with both age and cholesterol level. Each cell shows the number of treated individuals divided by the total number in that cell, and then the simple estimate of the propensity score based on this ratio.

Age

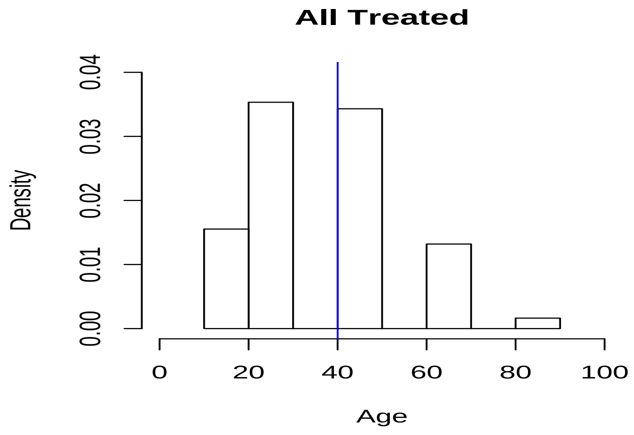
		0-20	20-40	40-60	60-80	80-100
Chol. Level	0-200	11/11 1.00	32/38 0.84	32/49 0.65	17/29 0.59	2/7 0.29
	200-250	57/61 0.93	100/119 0.84	75/141 0.53	40/103 0.39	4/25 0.16
	250-300	48/57 0.84	145/191 0.76	148/293 0.51	43/177 0.24	7/67 0.10
	300-350	28/33 0.85	63/98 0.64	72/172 0.42	28/125 0.22	2/46 0.04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

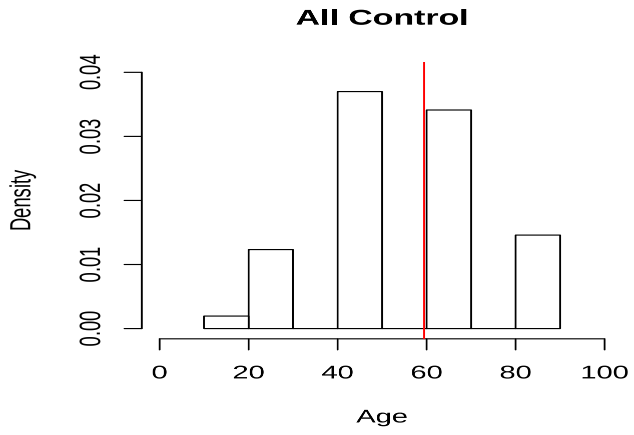
We could try to treat each cell of the table as a mini randomized experiment and compare outcomes for each. But what would we do with the top left cell, in which no individuals received control? What would we do with the bottom right cell, in which only one unit received treatment? These estimates of the propensity scores are regarded as "noisy" because of the small sample sizes.

Instead, we form four blocks based on the following estimated propensity score ranges: 0-.24, .25-.49, *.50-.74*, and **.75-1.00**, where we identify the four blocks with regular, underlined, *italics*, and **bold** type. The table below displays the four blocks.

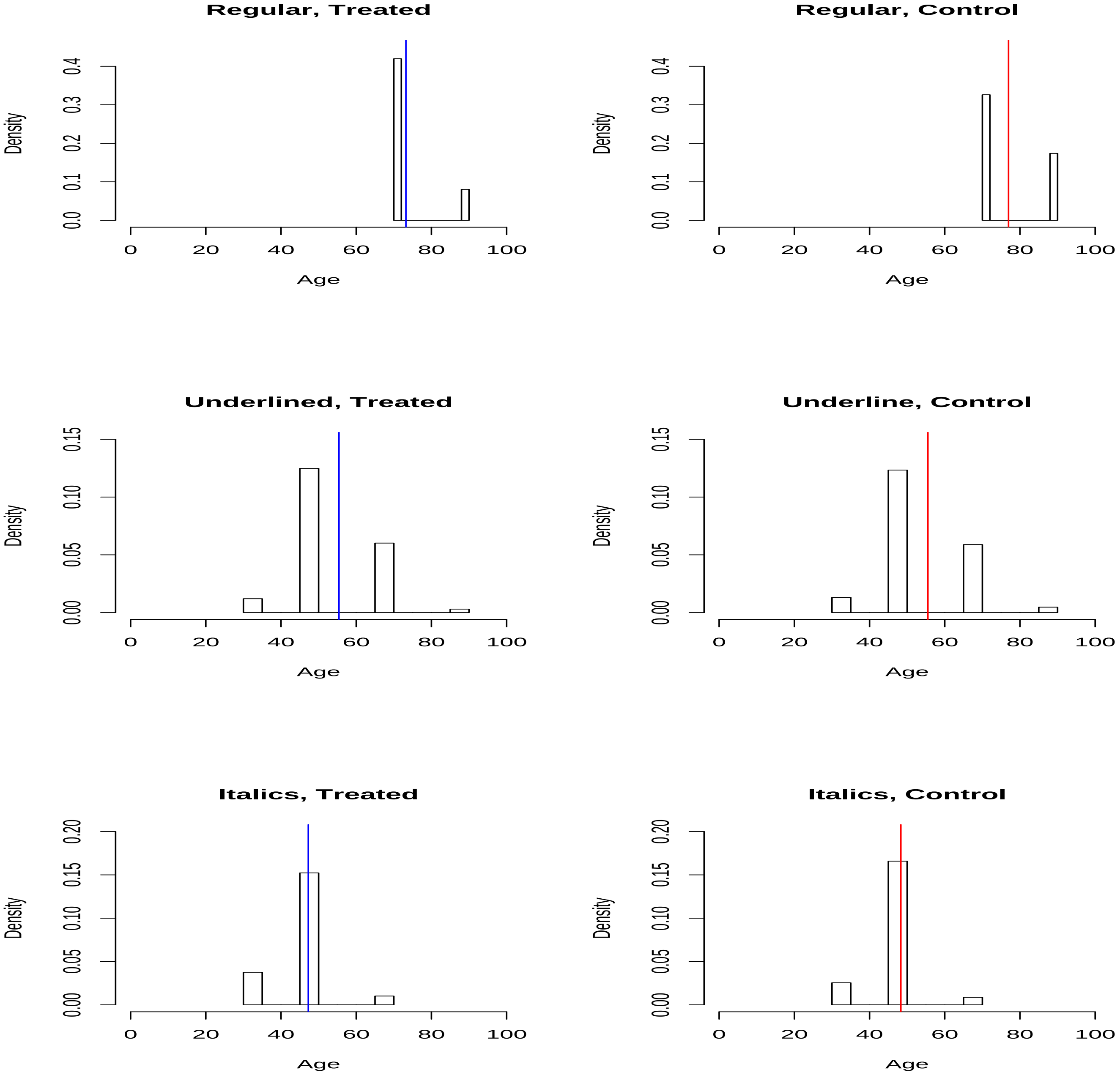
		Age				
		0-20	20-40	40-60	60-80	80-100
Chol. Level	0-200	**11/11** **1.00**	**32/38** **0.84**	*32/49* *0.65*	*17/29* *0.59*	2/7 0.29
	200-250	**57/61** **0.93**	**100/119** **0.84**	*75/141* *0.53*	40/103 0.39	4/25 0.16
	250-300	**48/57** **0.84**	**145/191** **0.76**	*148/293* *0.51*	43/177 0.24	7/67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

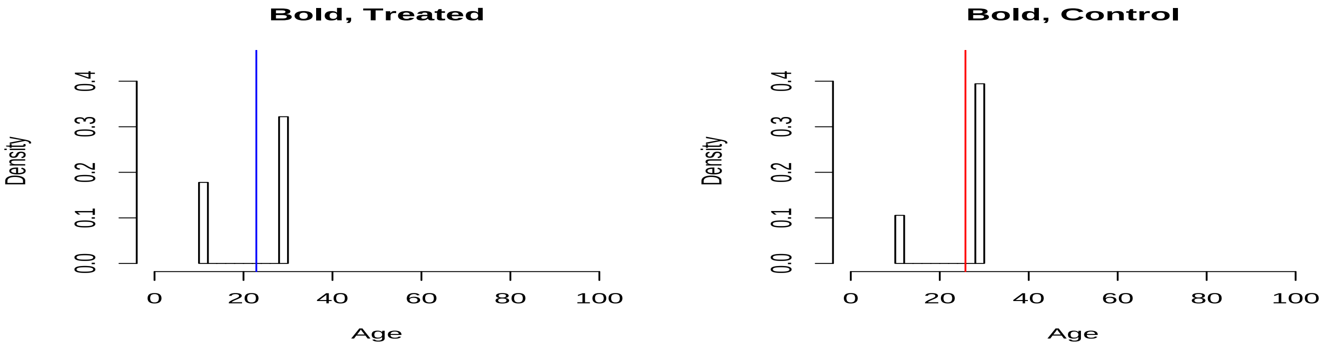
What have we accomplished with this blocking procedure? To show what this has done, we need to do two "before blocking" and "after blocking" comparisons, one for age and one for cholesterol. We start with age. Immediately below are histograms of the ages of the all treated and all control units before blocking. Notice that the two histograms look different. The treated are younger, e.g., the left histogram has a larger fraction of observations in the 20-30 year range.



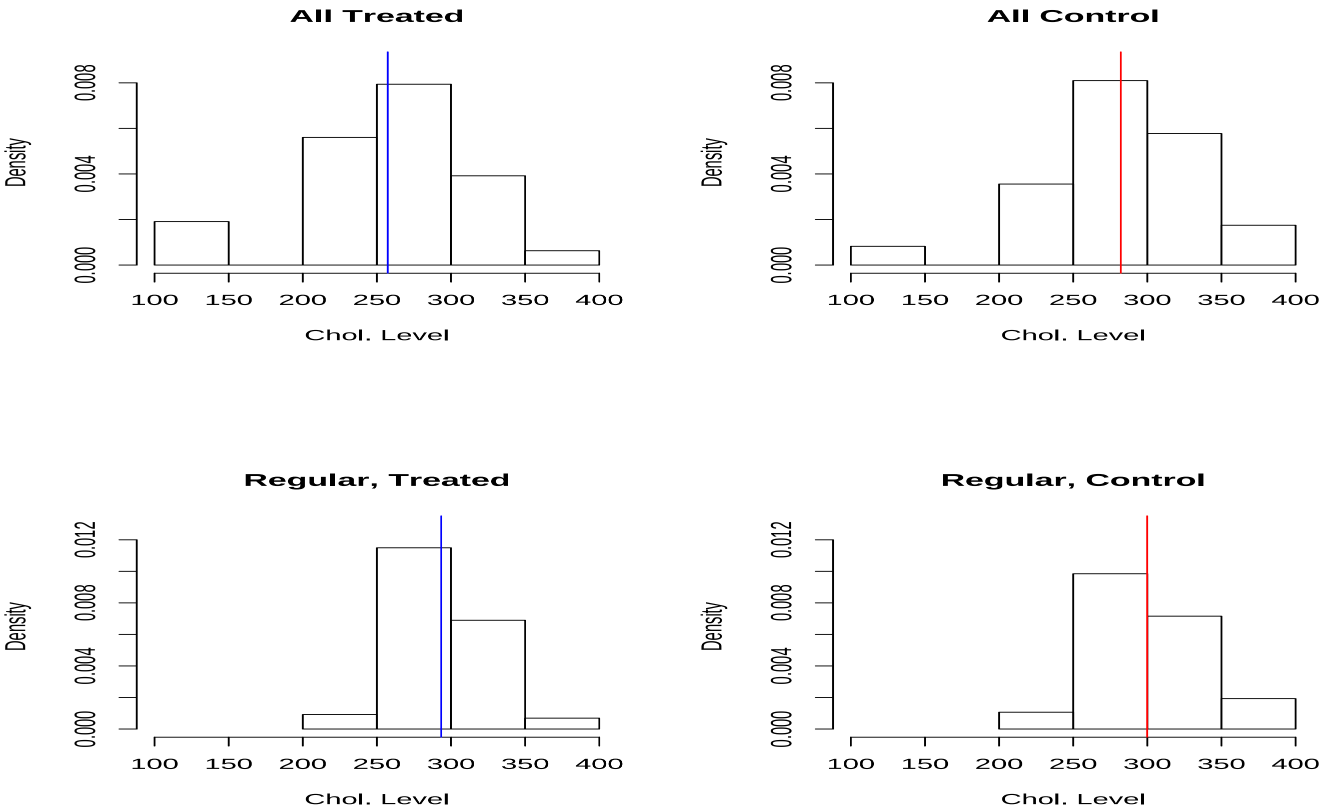


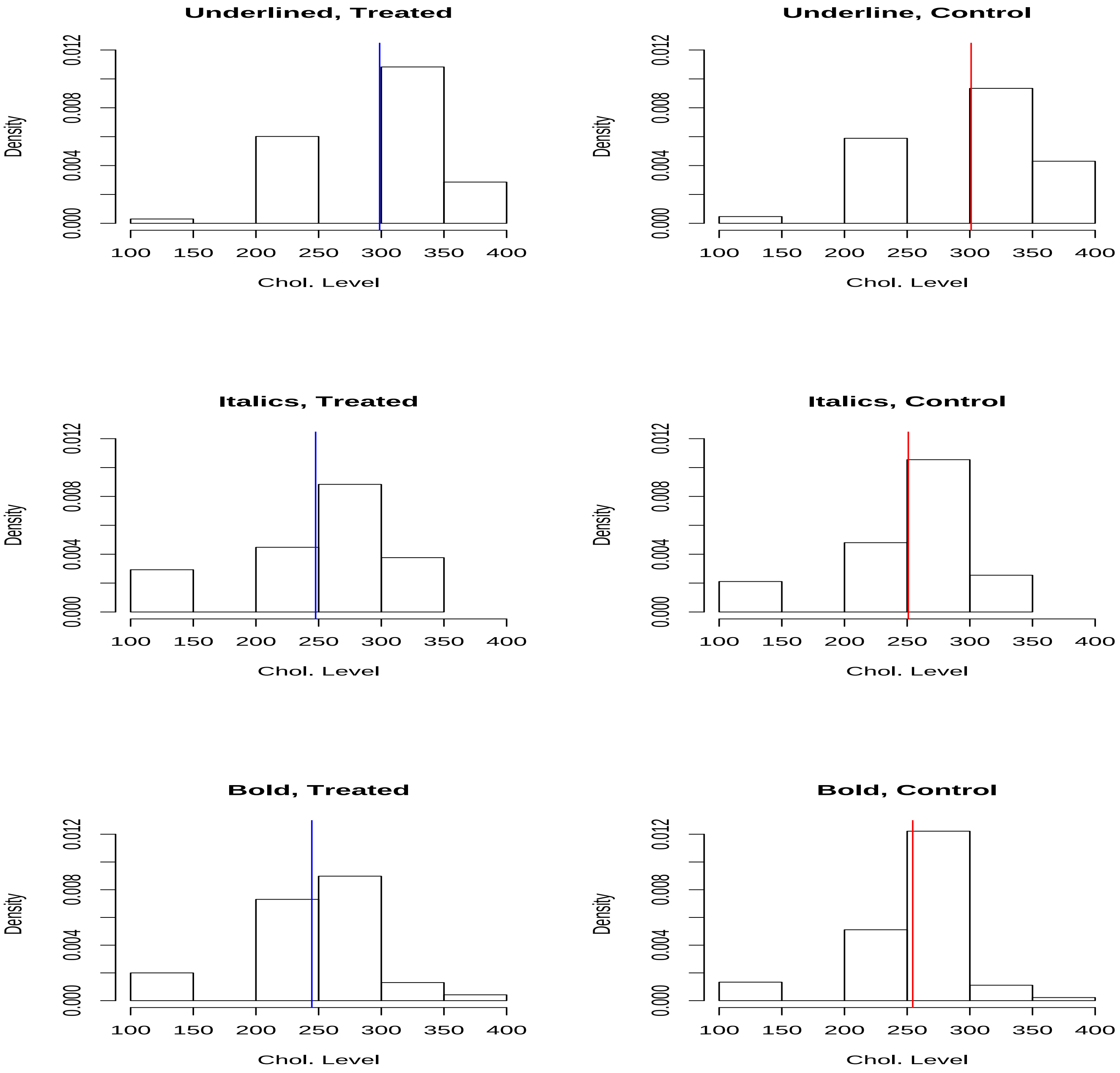
The previous two histograms were “before blocking” in the before and after comparison. Now examine the following four sets of histograms. Each represents a comparison of the distribution of age among the treated and control units within a block. Within each block, the distribution of age is roughly the same. The message: for age, the “after blocking” distributions resemble each other (within each block) far more than the “before blocking” distributions.





Now we do the same comparisons for the pretreatment cholesterol covariate. Again, notice that the first set of histograms, representing the distribution of cholesterol for all treated versus all control observations, look different (although this covariate seems better balanced than age). In particular, the control group has a greater fraction of observations at the 300-350 level. Matters improve after blocking.





Although the estimation of the probabilities is fairly straightforward for this example, it could get very complicated if there were more covariates available. Statistical methods have been developed that model the probability of receiving treatment given the covariates. One common method is called logistic regression, which uses statistical techniques to estimate the propensity score for each unit. With these estimates in hand, we can proceed to form blocks, where within each block, the propensity scores are relatively constant.

One of the key benefits of a randomized experiment is the implied balance of all of the background covariates between the treated and control groups within each block. A well designed observational study will also have this feature with respect to the observed covariates, with causal effects being estimated by comparing treated and control units with the same distribution of observed covariates within blocks. (Of course, randomized experiments are preferable in the sense that this balance is also implied for all unobserved covariates, too.) Here we give an example of how to create a series of hypothetical randomized experiments from observational data.

Example II-10: Reconstructing hypothetical randomized experiments through subclassification

From Rubin, D.B. (1997). Example adapted from Cochran (1968).

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

Comparison of Mortality Rates for Three Smoking Groups in Three Databases in the 1960s

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

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

An explanation for this surprising result can be found by looking at the average age of the people in each of the smoking categories. The table below shows these ages. Nonsmokers, and especially cigarette smokers, tend to be younger than cigar or pipe smokers. Age is thus highly related to the decision to smoke.

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

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

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

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

This table comes from Cochran (1968), who also gave theoretical results for subclassification. He showed that as long as there is reasonable overlap in the distributions of age in the treated and control groups, subclassification using five or six subclasses should remove 90% or more of the initial bias due to a single covariate such as age.

This example is in terms of a single covariate. Theoretical results from Rosenbaum and Rubin (1983, 1984) have shown that balance on the estimated propensity score between the treated and control groups implies balance on all covariates that went into the propensity score estimation. Thus, by subclassifying on the propensity score, we should (on average) obtain these results (balance between the treatment groups in these mini-randomized experiments) for all of the observed covariates. At least in large samples, within a group of individuals with similar values of the propensity score, we can treat the data as arising from a randomized experiment, assuming we accept that the assignment mechanism is unconfounded given the observed covariates. We saw this in the previous artificial example.

As stated before, a key feature of a randomized experiment is that the outcome data are not used in the design (randomization phase). We replicate that feature here by forming subclasses using only covariate values. The subclasses are defined and the analysis organized and planned without even seeing the outcome variable, although for pedagogical purposes we showed the results on the outcome variable here in the Cochran (1968) example.

Example II-11: 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.

The scientists in this study were interested in estimating the survival rates of breast cancer patients who receive breast conservation (lumpectomy, nodal dissection, and radiation) versus mastectomy. We summarize results from two types of studies, randomized experiments and an observational study, both of which were implemented for investigating the treatment of breast cancer.

1. Randomized Experiments
   - “Gold standard” of medical research
     - Assignment mechanism is unconfounded. Randomization assures balance in the distribution of observed *and* unobserved covariates.
     - In some randomized experiments, blinding/double blinding can be used so patients (and possibly doctors) do not know which treatment they are receiving. (This option is not available for studying these breast cancer treatments.)
   - Characteristics of patients and procedures in randomized experiments may be different from those in an “ordinary” setting or from the population of those to be treated
     - Randomized trials typically are in large, university hospitals (not many “community physicians”).
     - Physicians must follow pre-specified procedures in randomized experiments.
     - Patients (and doctors) have to be willing to be randomized before participation.
   - For these breast cancer treatments, six randomized experiments had been done around the world
   - Results:

	5 year survival rates		**Difference in rates**
Study	**Breast Conservation**	**Mastectomy**	**(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%

2. Observational Study: SEER database

   - 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 five years follow-up available on all patients in the early 1990s
   - Choose patients from SEER who could have been in randomized experiments (similar based on year of treatment, geographic area, tumor size, age, marital status, race or ethnicity)
   - Propensity score estimates the probability of each individual's receiving breast conservation based on the covariates available

     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 and 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 and diagnosed in 1985 with a very small tumor would have a relatively high propensity score.

     Two 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
   - Split all eligible patients in SEER (5,326 women) into five (approximately equal-sized) 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
     - Consider the data 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
     - Do analysis within each block and combine across blocks using Neyman/Fisher

- 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 is found by averaging the five (essentially equal-sized) blocks
- Similar results are found as in the 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 are lower than survival rates in the randomized experiments, as would be expected considering that the randomized trials were done in specialized conditions
- Note trend in signs of the estimated effects (i.e., the differences)

Subsection 11: Theory And Practice of Matched Sampling – Using Propensities And Covariates

Suppose we are interested only in the effect of the treatment "on the treated." For example, if we are trying to estimate the effect of smoking on incidence of lung cancer, we may not care about the effect of smoking on persons who do not smoke because we are probably not contemplating an intervention designed to induce them to start. If we are interested in the effect of treatment on the treated, we can use a technique that has become increasingly popular over the past several decades called "matching," introduced in the previous example. Conceptually, matching is similar to blocking, the difference being that each treated observation becomes its own little quasi-block, and we find the control observation that is "closest" to it in some sense. There are a great many ways to define how "close" one observation is to another. Two basic ways are (i) the difference in the values of some covariate thought to be critical, and (ii) the difference in the values of the single most important covariate, the propensity score.

We illustrate this technique using the toy dataset from Example II-8, where we match on the propensity score.

Example II-12: Matching on a Covariate

Recall that in Example II-8, we wanted to estimate the effect of a new pill designed to address a blood ailment. The data from the study, including the (known) propensity scores, are reproduced below.

Unit	Sex	Age	Prob of W_i = 1	W	Y(0)	Y(1)
1	F	28	$.68$	1	?	130
2	F	45	$.42$	1	?	200
3	F	26	$.73$	0	183	?
4	F	24	$.79$	1	?	124
5	M	22	$.63$	0	176	?
6	M	35	$.40$	0	266	?
7	M	37	$.38$	1	?	207
8	M	32	$.44$	0	275	?
9	M	34	$.41$	0	268	?

The treated units are Units 1, 2, 4, and 7, whereas Units 3, 5, 6, 8, and 9 received control. To begin, we pick the unit from the control group that has a propensity score closest to that of Unit 1's .68. Units 3 and 5 both have propensity scores .05 away from .68. In cases of ties like these, we might pick one randomly, or we might include both units. Let us suppose for now that we decided to choose one randomly, and that the random choice was Unit 3. The best match for Unit 2's propensity score of .42 is Unit 9's .41. Unit 4 is matched to Unit 3, and Unit 7 is matched to Unit 6. Thus, we have used matching to create a reduced set of control observations (Units 3, 6, and 9) similar to our set of treated observations (Units 1, 2, 4, and 7). To obtain an inference for a causal effect, we could use Fisher's method, Neyman's large-sample method, or the model-based methods discussed in Part III, as applied to a paired randomized experiment. For example, a 94% Fisher interval for the difference in means assuming an additive treatment effect is (-144, 24). A large-sample Neyman interval is (-143, -4). Which of these two intervals is more intuitive to you in this setting?

The matching technique outlined here can be particularly powerful in an observational study in which (as often happens) the number of control observations far exceeds the number of treated observations. In such instances, we use matching to discard control observations that do not resemble the treated units.

Example II-13: National Supported Work Demonstration

The National Supported Work Demonstration was a program run by the US Government during the 1970s. It was designed to help move disadvantaged workers into the labor market by providing them with work experience and counselling. 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 program.

The results of this program have been analyzed in several ways. Because it was a randomized experiment, a good estimate of the "true" treatment effect is available. However, as a way to illustrate methods for dealing with observational studies, this data set has also been used as a part of an observational study, essentially ignoring the randomized control group data and finding a comparison group using large national data sets already available. For more information on these analyses, see Lalonde (1986) and Dehejia & Wahba (1999).

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

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

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

We see that the treated and control groups are very similar, but that the treated group and the PSID comparison group are very different. (Variables that are marked with a * are “significantly” different at the .05 level from each other in the treated group and the PSID comparison group).

To form a better comparison group, propensity scores were estimated, and then the treated group members were matched to individuals in the PSID on the basis of their estimated propensity scores, effectively looking for blocks of size two with one treated and one control – a matched pair design, also called a “paired comparison.” Some controls, however, had to be used more than once. Thus, only comparison group members who looked like the treated group were kept in the resulting dataset. The following shows the covariate means for the treated group and the new matched comparison group.

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

The treated and matched comparison groups now appear to be similar to each other. None of the variables are “significantly” different between these two groups.

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

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

Estimating Propensity Scores – Some Details of Example II-12

Here we give some details on estimating propensity scores.

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

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

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

The following procedure summarizes an implementation method. Note that this is just one possible method.

1. Start with a model (for example, logistic regression with treatment received as the response variable) with main effects for each of the observed covariates and thereby obtain estimated propensity scores for each unit.

2. Discard control units outside the range of the treated group propensity scores, and/or treated units outside the range of the control group propensity scores.

3. Form one block (with propensity scores in the range 0-1), do a t-test of $\hat{e}$ between the treated and control groups. If significant, split into two blocks at the median. Continue this process, splitting a block if it has a t-statistic greater than 2 and if there are more than two treated and control units in each new block formed.

4. Within each block formed in Step 3, test for equality of means of functions of X (e.g., each covariate, each covariate squared, two-way interactions of covariates). If any t-statistic is greater than 2.5 in any block, include that term in the new propensity score specification.

5. Repeat Steps 1-4 until there are no more (or very few, as few as possible) significant t-statistics. This will imply that within each block, the treated and control groups are well balanced.

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

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

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

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

Lower and Upper Block Boundaries: Specification 1

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

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

t-statistics of propensity score: Specification 1

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

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

Block Results: Specification 1

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

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

The following table shows t-statistics for the covariates, their squares, and their two-way interactions. This table is used as a diagnostic, to determine which terms should be added to the propensity score specification; t-statistics greater than 2.5 imply that the block is poorly balanced on that covariate.

Block T-statistics: Specification 1

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

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

Block Results: Specification 2

	Discard	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Discard
Lower Bound	0.00	0.00	0.02	0.10	0.23	0.40	0.60	0.81	0.99
Upper Bound	0.00	0.02	0.10	0.23	0.39	0.59	0.80	0.99	1.00
# Obs.	1428	702	204	102	67	39	30	100	3
# Controls	1428	696	198	91	44	16	8	9	0
# Trainees	0	6	6	11	23	23	22	91	3
Mean Controls	0.03	0.01	0.05	0.15	0.31	0.47	0.67	0.97	–
Mean Trainees	–	0.01	0.06	0.16	0.31	0.49	0.71	0.92	0.65
t-Stat Diff	–	1.42	1.11	0.28	0.66	1.27	1.89	-0.18	–

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

Block T-statistics: Specification 2

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

There are now no terms with t-statistics greater than 2.5. The blocks are well balanced on the observed covariates. An analysis of the outcomes can now be done because the design phase of the observational study is complete.

This analysis can be done in the full matched groups, or within subclasses (i.e., further blocks). Doing the analysis within subclasses is useful if the overall groups are still not well matched. We see that within subclasses the groups are similar to each other, and Neyman/Fisher methods for randomized experiments can be used within each block. Under the assumption of unconfoundedness given the observed covariates, the subclasses could be defined as above, or by quantiles of the propensity score in the treated group, control group, or overall. An overall effect is estimated using a weighted average of the within subclass estimates, as illustrated earlier.

For reference purposes, some of the technical terms associated with various ways of implementing the methods described in this section appear below, along with non-technical definitions.

- Exact matching on critical covariates:
  - Forming donor pools or comparison groups that have the same value of particular covariates, then using methods that allow approximate matches to deal with the remaining covariates
  - Used when the investigator, for scientific reasons, believes that certain covariates are critically related to the potential outcomes
  - Usually not possible to implement unless the critical covariates have only a few possible values (e.g., sex can be only male or female)
  - Often, within blocks formed by exact matches on critical covariates, investigators attempt to compare observations with similar propensity score values, essentially using the propensity score as a one-dimensional summary of the covariates
- Caliper matching:
  - For each unit that received treatment, identifying the unit's values of certain covariates, then forming "calipers" of control observations whose values of these critical covariates are within a particular range of the treated unit's (e.g., within plus or minus five years of age)
  - Used when the investigator, for scientific reasons, believes that certain covariates are importantly related to the potential outcomes, and differences within these calipers are not important
- Stratified matching: Dividing the range of covariates or of the propensity score into blocks (i.e., strata), then finding the control observations that are exact matches to the treated with respect to the strata
- Metric: Any measure of the distance between the covariates of one observation and the covariates of another
- Mahalanobis metric: A particular kind of metric; in two dimensions, the Mahalanobis metric is akin to the hypotenuse of a right triangle formed by two points, where measurements are scaled so that, for example, it makes no difference if length is stated in feet or inches

Part III: Causal Inference Based on Predictive Distribution of Potential Outcomes

Subsection 12: Prediction Inference – Intuition under Ignorability

Introductory Discussion: Standard, Model-Based Methods Versus Models in the RCM

Much of causal inference in the physical, biomedical, and social science is attempted with a standard set of models. These models take a dizzying variety of forms, and some are complex and mathematical. Typically, an investigator "fits" more than one model, perhaps using variables in different ways or with different mathematical forms.

Almost all of these standard methods for attempting causal inference have two essential characteristics in common. First, they all use the outcome variable Y at every stage of the analysis, particularly while fitting the model. The process of fitting different models **automatically** generates estimates of causal effects during the fitting process, and the investigator sees these estimates immediately. There is thus some risk that the investigator's choice among models will be influenced by the "desirability" of the resulting causal estimate. Second, these standard methods deal solely with Y_{obs}, meaning $Y(1)$ if the unit was assigned treatment and $Y(0)$ if the unit was assigned control. None of these models, in theory or in practice, deals with potential outcomes. Most of these methods share a third characteristic, namely, that they can be accomplished using canned statistical software.

As noted previously, the framework for causal inference introduced in this course is often referred to as the Rubin Causal Model ("RCM"). Disciples of the RCM do not reject the use of models in causal inference. To the contrary, models are often necessary, and they may be especially useful to increase the precision of estimates. Instead, the RCM contrasts with standard model-based methods of causal inference in at least two ways. First, the maxim "design before inference" means that much of the hard work of inference is accomplished **without using** Y. That encourages the investigator to commit to an algorithm for drawing inferences before knowing what the result will be. Second, the backbone of the RCM is the potential outcomes framework, meaning that the investigator must be thinking in terms of missing data. The purpose of Part III is to introduce model-based methods for filling in the missing values.

In your own work, you may encounter some of the dizzying variety of models referred to above. You may see, or may already have seen, something resembling the following equation (for a study involving two covariates): $Y_{obs,i} = \beta_0 + \beta_W * W_i + \beta_1 * X_{1,i} + \beta_2 * X_{2,i} + error_i$. The goal here would be to draw inferences about the β's (the parameters), particularly β_W. Verbally, the investigator relates the observed value of Y for each unit to some function of both the unit's treatment assignment and its vector of covariates. The "error" term makes sure that the left side equals the right side. "Error" is a poor choice of words here; no one has made a mistake, nor has a measurement been taken improperly. "Error" refers simply to the deviation of the observed value from the value the model predicts. You may have heard terms such as regression, ordinary least squares, probit regression, logistic regression, splines, or the like. The above equation represents an ordinary least squares repression. The point of all this is not to have you be dazzled by the mathematics, the notation, or the jargon. Think about basic principles first. Does the model work solely with Y_{obs}? Does it automatically generate estimates of causal effects in the fitting process? For the ordinary least squares regression referred to immediately above, the answer to both of these question is "yes," which should give you pause about the reliability of the technique.

The imputation methods described below are not formally exactly correct, but convey the essential idea of how to impute, or fill in, the missing potential outcomes. Formally, we are using an exchangeable model on the science (X, Y(1), Y(0)) to multiply impute the missing potential outcomes given all observed values.

Example III-1: Using Donor Pools To Fill in Missing Potential Outcomes – Discrete Covariate

Consider again an example of a study attempting to estimate the effect of a new surgery on years lived after the surgery. A randomized experiment (blocked by sex) was done, with ten patients (five male and five female) receiving the new surgery and ten (five male and five female) receiving the old surgery. There is one covariate available for each individual: male versus female.

Instead of running a Fisher test and generating an interval for a hypothetical additive effect, or instead of using Neyman's method of unbiased estimation (and large-sample confidence intervals), we can use the observed values to predict the missing potential outcomes and fill in ("impute") the missing values. Once these missing values are filled in, it is straightforward to calculate, for the units in this study, the average treatment effect (corresponding to that imputation), average treatment effect within subclasses defined by the covariates, or even values such as the average difference in squared potential outcomes, the median causal effect, etc.

We will first impute using "donor pools": a set (or pool) of individuals in the other treatment group with similar covariate values. As noted above, this approach is not formally exactly correct, but it is very intuitive and close to being correct.

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

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

Each unit is missing either Y(0) or Y(1), and we wish to fill in these missing values. Because of the unconfounded treatment assignment, units of the same sex are exchangeable (in an intuitive and a formal sense), and we thus can ignore the treatment assignment indicator. Therefore, we can use the observed values of Y(0) for males to fill in the missing value of Y(0) for males, and similarly for females.

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

Donor Pools Defined by Gender

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

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

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

Sample Imputation

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

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

This process should be repeated multiple times, using the same donor pools but randomly drawing a new donor for each individual each time. Each of these multiple draws (i.e., multiple imputations) provides a separate estimate of the quantity of interest, and this set of estimates allows us to gauge the uncertainty in the estimate and also allows us to create an interval for the effects. For a 95% interval, we need a lower bound and an upper bound. Suppose we have imputed 5000 times. For the lower bound, we would find the .025 * 5000 = 125th largest estimate, whereas for the upper bound, we would find the .975 * 5000 = 4875th largest.

Note that because we are filling in the entire potential outcomes table with each draw, we can calculate any quantity we desire each time and thereby get an idea about that quantity's distribution. For example, suppose a government official wanted to know how many years it would be before 15% of patients died if everyone received the new surgery. We would proceed as before, except that for each draw, we would calculate the .15 * 20 = 3rd largest Y(1) value; 5000 draws would provide 5000 such values, and we could use these to construct, say, a 95% interval for the quantity the government official desired.

Six specific imputations are shown below, and the histograms for the mean causal effect and the median causal effect based on 5000 imputations are also given. The vertical bars in each plot show the bounds for a 95% interval. These 95% intervals can be compared to an approximate 95% Fisher interval (assuming an additive treatment effect) for the difference in means of (.9, 4.02), and a 95% Neyman large-sample interval for the difference in means of (.61, 4.4).

Imputation 1:

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

Imputation 2:

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

Imputation 3:

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

Imputation 4:

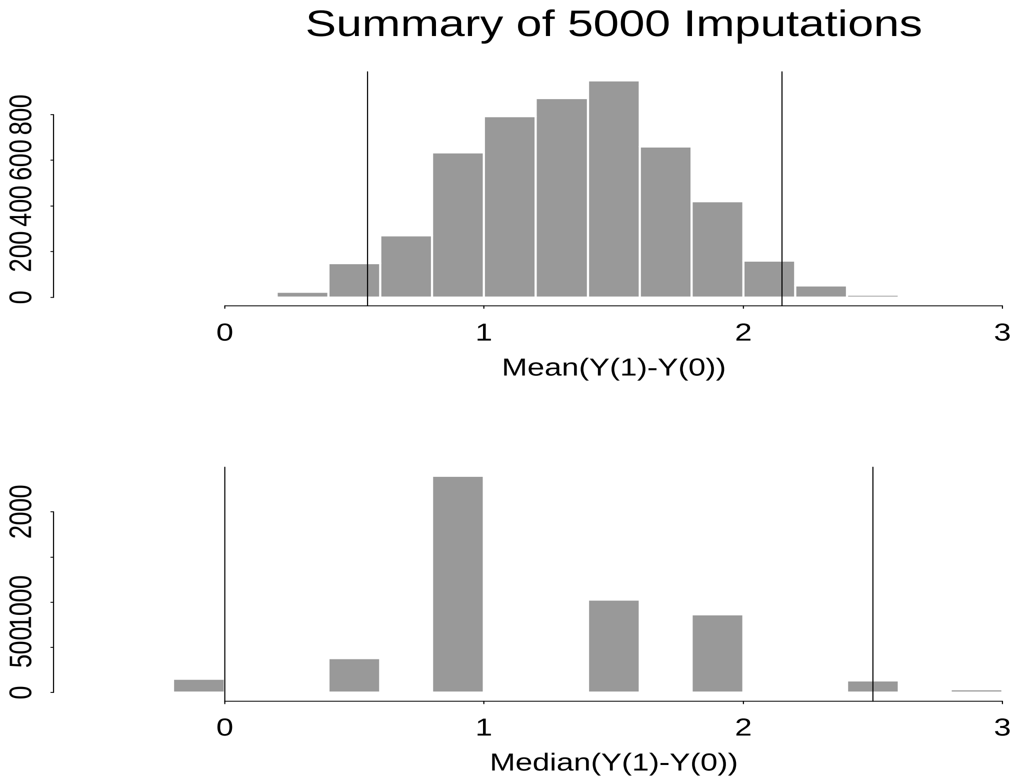
Unit	Gender	W	Y(0)	Y(1)	Y(1)-Y(0)
1	M	1	(8)	12	4
2	M	1	(6)	9	3
3	M	1	(11)	9	-2
4	M	1	(7)	7	0
5	M	1	(7)	8	1
6	F	1	(6)	12	6
7	F	1	(5)	11	6
8	F	1	(6)	10	4
9	F	1	(8)	14	6
10	F	1	(6)	12	6
11	M	0	6	(12)	6
12	M	0	8	(9)	1
13	M	0	7	(9)	2
14	M	0	11	(12)	1
15	M	0	11	(7)	-4
16	F	0	5	(7)	2
17	F	0	7	(7)	0
18	F	0	6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

Imputation 5:

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

Imputation 6:

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

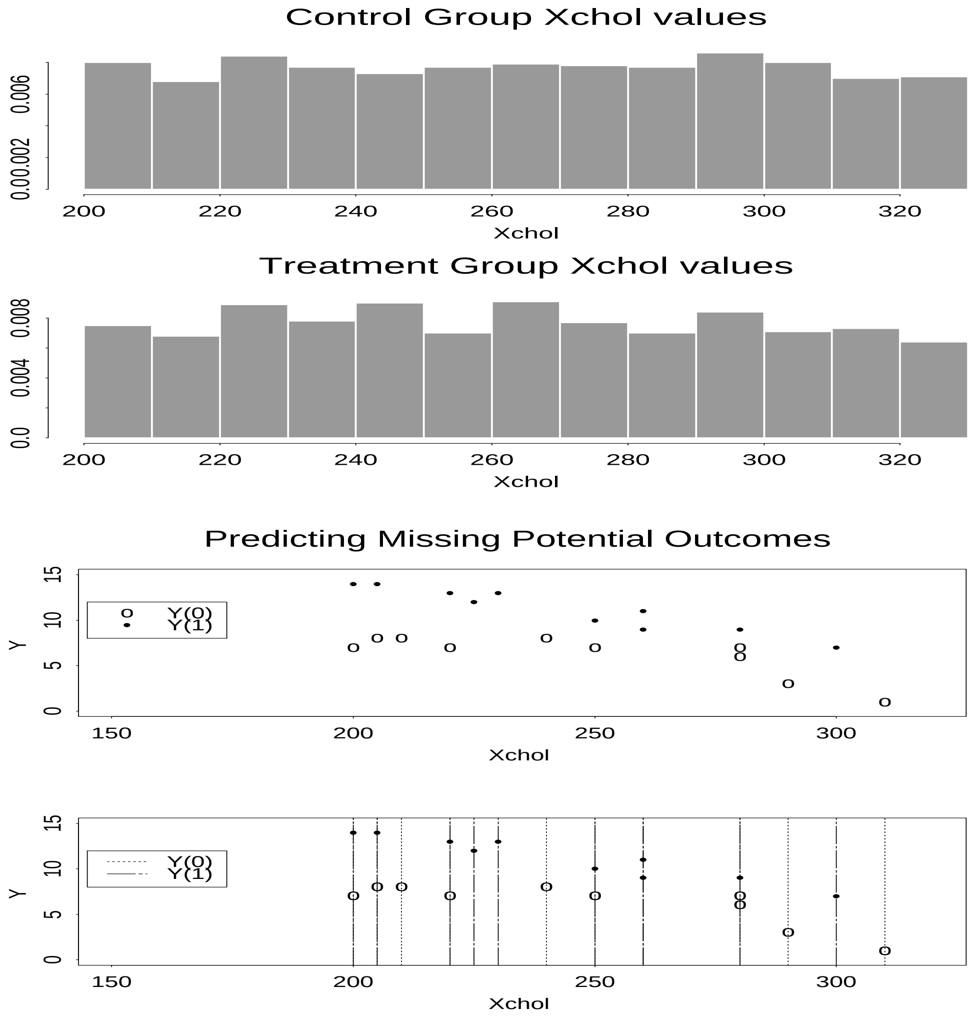


Subsection 13: Matching To Impute Missing Potential Outcomes – Donor Pools

Example III-2: Donor Pools with a Continuous Covariate

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

In addition to Y(0) or Y(1), the doctor also has recorded the cholesterol level of each patient prior to the surgery (the covariate X_{chol} represents cholesterol prior to surgery). The histograms below show all values of cholesterol for the 2,000 patients, and the bottom plots show the observed potential outcome data for the first 10 in each group.



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

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

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

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

Six specific imputations are displayed below, and the histograms for the mean causal effect and the median causal effect are also given. The vertical bars in each plot show the bounds for a 95% interval. For the mean, this 95% interval from simulation is approximately (4.15, 5.05). This simulation interval can be compared to a Neyman 95% confidence interval, which for these data is (4.12, 5.07). Which interval do you find more intuitive?

What would we do here if the assignment mechanism had been that of the perfect doctor? Hint: A good answer requires some thought!

Imputation 1:

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

Imputation 2:

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

Imputation 3:

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

Imputation 4:

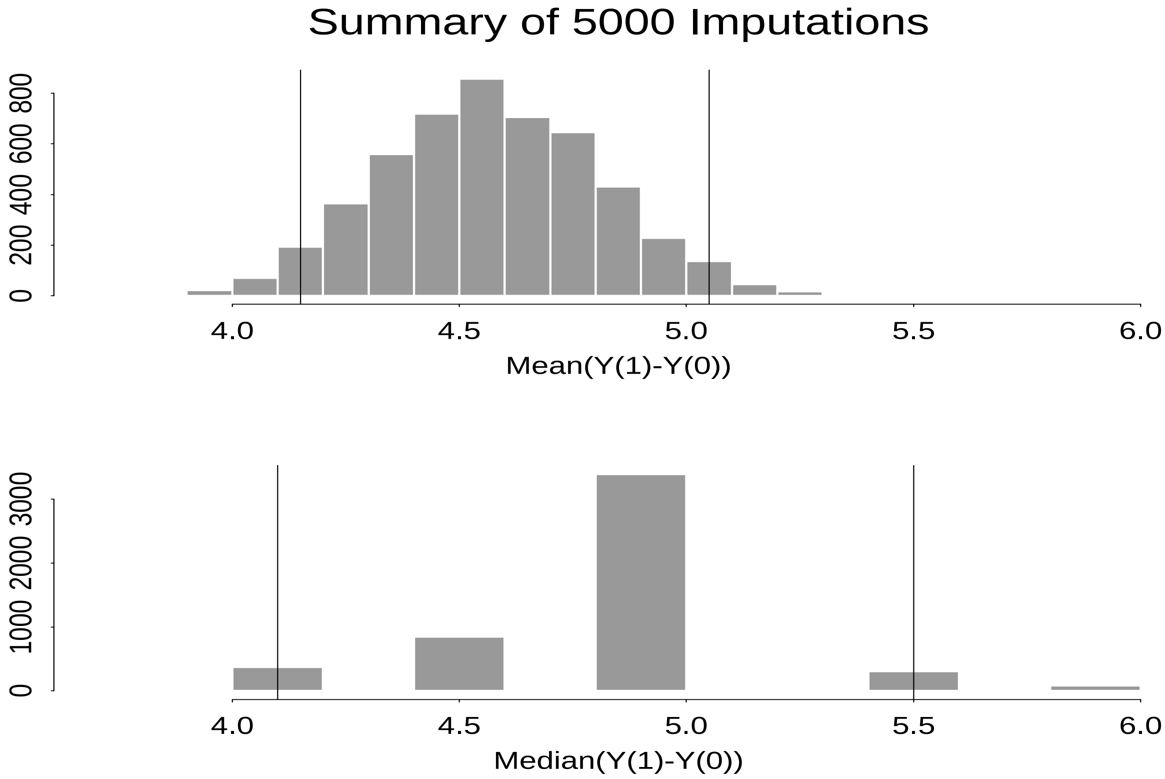
Unit	X_{chol}	W	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	210	0	8	(13)	5
2	200	0	7	(14)	7
3	310	0	1	(9)	8
4	220	0	7	(12)	5
5	280	0	7	(11)	4
6	290	0	3	(9)	6
7	240	0	8	(9)	1
8	250	0	7	(13)	6
9	280	0	6	(9)	3
10	205	0	8	(14)	6
1001	250	1	(7)	10	3
1002	225	1	(7)	12	5
1003	300	1	(4)	7	3
1004	260	1	(7)	11	4
1005	230	1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

Imputation 5:

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

Imputation 6:

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



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

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

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

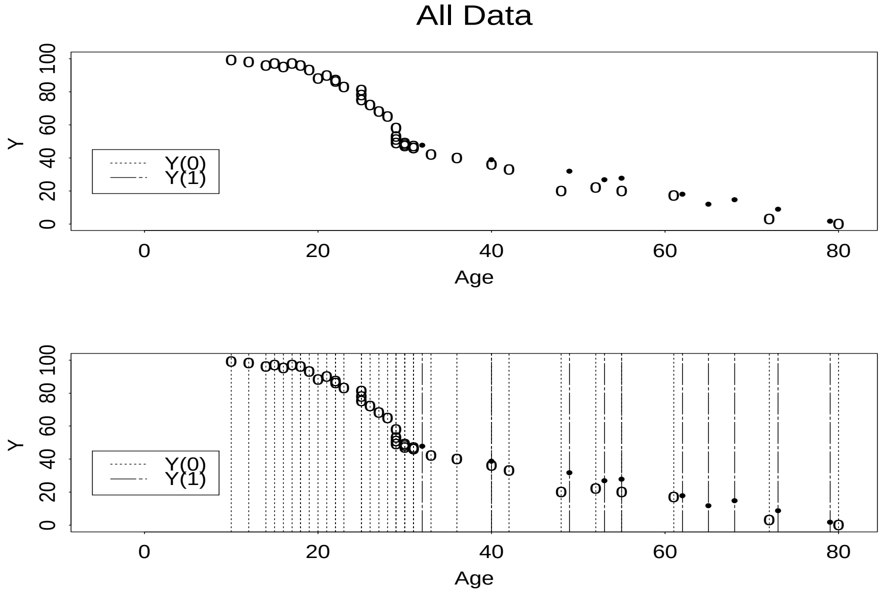
Suppose interest focuses on the effect of the new surgery on those who might receive it. That is, we care only about estimating the effect in the population of people who might receive the new surgery, not on the population in general.

We thus would like to impute the missing potential outcomes only in the treated group; that is predict the missing $Y(0)$ for these units:

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

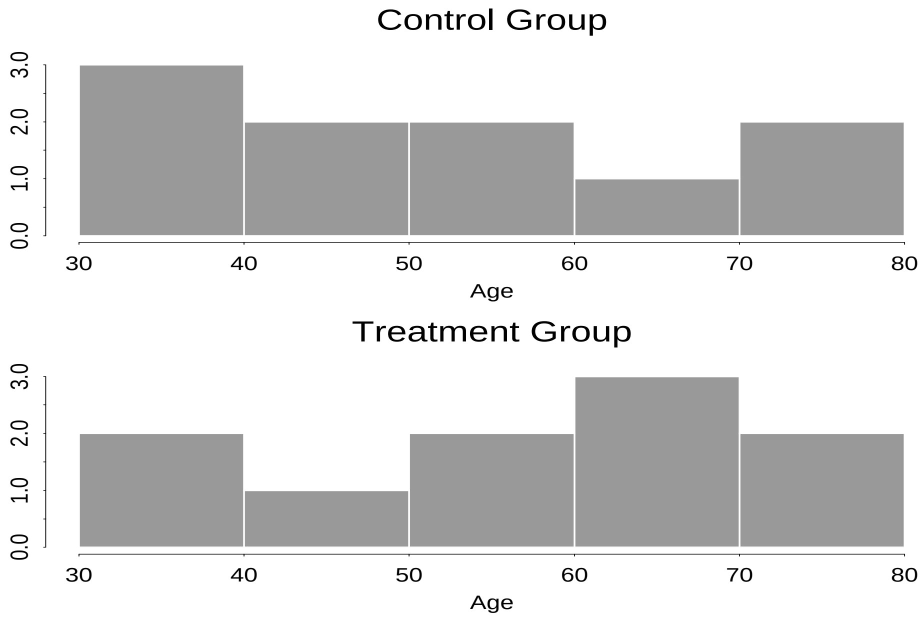
The 32-year-old person is the youngest in the treated group. Notice that most people in the control group were younger than the 32-year-old.

It is not really relevant to use the under those under 30 in the control group to help impute $Y(0)$ for the people in the treatment group because their ages are so different from those in the treatment group.



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

The ideas of matching and donor pools automatically correct for this difference in ages. For each individual in the treated group, we form a donor pool consisting of the four control individuals close in age to the treated individual (here, we choose the four closest, but we make sure in general that all are "close"). We thus only use control members with similar ages to individuals in the treated group. As seen in the plot below, there is good overlap in the age distributions in the range of ages of those in the treated group. We are thus able to form a decent donor pool for each treated individual.

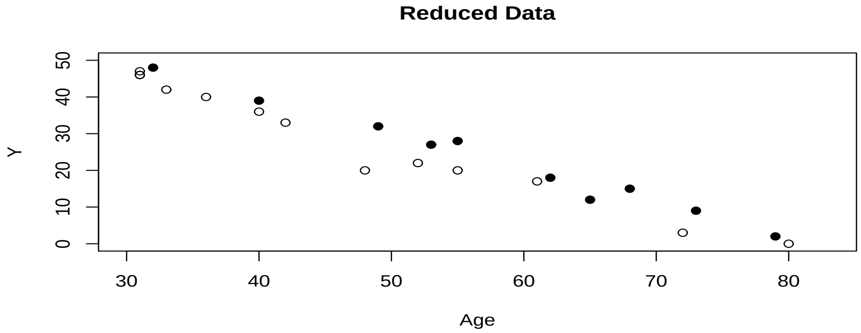

Control Group

Treatment Group

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

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

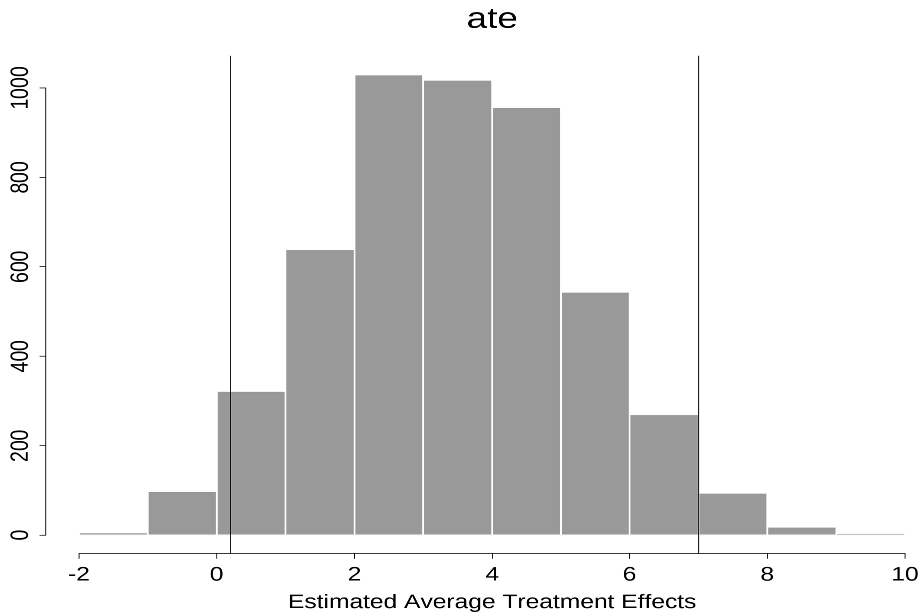
A plot of the age covariate against the observed Y for the treated observations and the control observations in the donor pools makes it easy to see that these data suggest the treatment has some effect. Compare this graph to the one on page III-13.8.



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 calculate the average (or median) treatment effect for those in the treated group.

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

By repeating this process many times, we can estimate the average treatment effect as well as the variance of the estimated treatment effect. We repeated the process 5,000 times, and the histogram below shows the values of the estimated treatment effect. The estimated mean treatment effect is 3.5 and its estimated variance is 3.1. A 95% interval for the treatment effect is (0.2, 7.0), which can be compared to a Neyman large-sample interval of (.19, 7.04).



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

More generally, we can use any of the variety of different matching methods described earlier. Whenever we use covariates for matching, however, it is generally important that they be PROPER covariates. A proper covariate is a characteristic that is not affected by treatment. In the immediately preceeding example, age is a proper covariate because surgery does not affect age at surgery. All pre-treatment assignment variables are also proper covariates: for example, treatment cannot change someone’s pre-treatment cholesterol level. But we should not include a variable such as post-treatment weight in our matching; the cholesterol reducing drug might well affect this variable. We refer to a variable such as post-treatment weight as an “improper” covariate.

Note also that all of this matching is done without knowing the values of the outcome variables! This is a critical design criterion, regardless of whether the outcome variable is included in the same dataset as the covariates.

Subsection 14: Fitting Distinct Predictive Models within Each Treatment Group

In this section, we address what to do after we have applied the methods of previous sections to identify groups of treated and control observations with balanced covariate distributions. We may have matched, or formed subclasses, but we have a group of treated observations that is similar to those receiving control with respect to background covariates. We could, of course, use the methods discussed in section II-11, and often these procedures work well, especially in large samples. For more sophisticated problems, however, it may help to use models. The following example illustrates one method.

Example III-3: 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 examined the effects of in utero exposure in humans.

Two studies with similar designs were conducted. We concentrate on the larger one here. Medical records were used to identify the treated and control groups and to collect covariate data.

- Treated (exposed) group: Men born at the largest hospital in Copenhagen, Denmark between 1959 and 1961 whose mother took phenobarbital while pregnant.
  - 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; see Reinisch (1995)).
  - 101 controls selected.

The following table summarizes the effects of the matching:

Variable	Full Set of Controls	Matched Controls	Exposed Subjects
Proper Covariates			
% 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
Improper Covariates			
Mean gestational length (wks)	38.59	38.63	38.73
Mean birth weight (g)	3233	3260	3219
Mean birth length (cm)	51.28	51.64	51.57
Mean # cigarettes in 3rd trimester	6.40	5.26	5.03
Mean maternal weight gain (kg/m^3)	26.88	28.18	27.65
Mean maternal complaint	1.70	3.97	4.95
Sample size	3308	101	81

- We see that the matched sample of controls is much more similar to the exposed group than is the full sample of controls.

- The following variables are "significantly" different between the full set of controls and the exposed individuals: % unwanted pregnancy, % single mother, mean socioeconomic status, mean mother's age, and mean maternal complaint.

- There are no variables that are "significantly" different between the matched controls and the exposed subjects.

- Note that we did NOT use the improper covariate "mean maternal complaint" in the model used to estimate the propensity score. Actually, we did not use any of the improper covariates in any part of process of forming matched samples. We keep track of what is happening with these variables because they may be of scientific interest, perhaps as potential ways in which the treatment (here, barbiturates) affects the outcome (a test score). Thus, it may be of interest to scientists that the process of forming matched samples to balance the proper covariates also had the effect of balancing the improper covariates for which we had measurements. This fact might suggest that barbiturates do not affect the outcome of interest solely by, for example, reducing birth weight or birth length.

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.
- Use of models: A linear model was used relating outcome Y(0) to covariates using the matched control subjects. One model was fit for the treated group and a second for the units assigned control. Because we have already generated comparable treated and control groups via the propensity score matching, the form of the model used is typically not important for the estimates of causal effects. The main purpose of the model here is to increase the estimated precision of our estimates.
  - Covariates used: family's socioeconomic status (SES) when child one 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 Y(0) for the treated subjects. The observed treated outcome is then compared with the predicted control outcome.
- Also looked within subgroups.

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

Conclusions:

- Effects of exposure to phenobarbital in utero can be seen well into adulthood even in the absence of physical abnormalities.
- Timing of drug exposure affects the size of the effect.
- Social and psychological factors interact with in utero exposure to affect the size of the effect.

Subsection 15: Formal Predictive Inference – Bayesian

Concluding Discussion: Standard, Model-Based Methods Versus Models in the RCM

The examples in Part III illustrate the two critical characteristics of the way the RCM uses models as predictive tools for causal inference. First, this framework recognizes that the investigator's task is to find principled ways to fill in the missing potential outcome for each unit. Second, the hard work is to create comparable sets of treated and control observations. Ideally this step should be accomplished without any examination of the observed potential outcomes, and ideally it should be done separately from the actual process of filling in missing potential outcomes. Thus, as investigators are implementing the critical steps, they do not know whether the analyses will result in large, small, or zero estimated causal effects, nor whether the estimates will be positive or negative. If the outcome data are available, of course, nothing prevents investigators from peeking, but the disciplined investigator will not do so, and sometimes structural safeguards can be put in place (such as committing to certain matched samples before the outcome data are even measured).

When treatment assignment is unconfounded, such a prescription can in principle be followed, at least in large samples. In smaller samples, more restrictive models (e.g., linear models or, better yet, spline models) may be needed, but the general prescription should be followed wherever possible. All of these guidelines can be given a formal Bayesian foundation, as initially outlined in Rubin (1978). When treatment assignment is confounded, following these guidelines is more complex, although as shown in Rubin (1978), the formal Bayesian analysis remains the same as in the unconfounded case if treatment assignment is ignorable. When treatment assignment is nonignorable, life becomes much more difficult, both formally and practically. The concluding section in Part III offers some advice on the nonignorable situation.

Subsection 16: Nonignorable Treatment Assignment – Sensitivity Analysis

Formal Methods To Deal with Nonignorable Treatment Assignment

Once the observed covariates have been controlled satisfactorily (through matching, subclassification, modeling within subclasses, etc.), then attention can be shifted to consider the possible impact of unobserved covariates, when nonignorable treatment assignment is a possibility. Of course, if imbalance in observed covariates cannot be corrected, then either heroic modelling assumptions must be relied on to support causal inferences, or we must search for a more appropriate dataset. Ideally, speculation about nonignorable models should take place before the results are seen, but it is appropriate at any time in an observational study.

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

Rosenbaum and Rubin (1983) extended Cornfield's approach by offering a method to assess sensitivity to an unobserved binary covariate, also taking into account the observed covariates **x**. The main approach involves positing, like Cornfield, an unobserved binary covariate u such that treatment assignment is confounded when only **x** is observed but unconfounded when both u and **x** are observed. The method estimates the treatment effect over ranges of plausible correlations between u and both treatment assignment and the potential outcomes. This method effectively imputes the missing u under each of the hypothetical scenarios specified by the relationship between u and the other quantities, adjusts for the imputed u in addition to **x**, then assesses whether the causal inferences would change in important ways. If the conclusions are insensitive to this range of plausible correlations, then the results from assuming unconfounded treatment assignment given only **x** (but not u) are more believable.

We can also use the ideas of prediction to think about what we would do if assignment were nonignorable. In Example III-2 we assumed that treatment assignment was ignorable given age and pre-treatment cholesterol. What if instead the data had arisen from the perfect doctor? Remember that the perfect doctor could effectively see each individual's potential outcomes under both treatments, and assigned the treatment that was best for each person (or tossed a fair coin when there was no difference in survival outcomes). We would like to match on variables that would make treatment assignment ignorable, but because the doctor did not write down the potential outcomes under both treatments for each individual, we do not have that information. Assignment is no longer ignorable, but we can still think about imputing the missing potential outcomes. Consider Unit 1, who lived for eight years after the old surgery ($W_1 = 0$, $Y_1(0) = 8$). Since the perfect doctor assigned the treatment that would be best for each individual, we know that Unit 1 must have a potential outcome under the new surgery less than or equal to eight years! In other words, $Y_1(1) \leq 8$. We could then think about imputing values less than eight years for the missing $Y_1(1)$. Maybe any value uniformly between zero and eight? Or perhaps specify a different lower bound that involved a donor who received new treatment and lived less than eight years?

And wouldn't we want to use baseline cholesterol level (when available) to help with this imputation? For example, we could find individuals among the treated group who match Unit 1 on baseline cholesterol and lived for less than 8 years. These people could then form a donor pool for Unit 1.

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

Part IV: Principal Stratification: Dealing with Explanatory/Intermediate Post-Treatment Variables

Subsection 17: Simple Noncompliance And Instrumental Variables

Example IV-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 *Statistics in Medicine*. 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, whereas 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($Y = 0$ means alive at one year, $Y = 1$ means the opposite).

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

Row	True Compliance Type	Treatment Assigned	Treatment Received	Y_{obs}	Number of Children
1	?	0	0	0	11514
2	?	0	0	1	74
3	N	1	0	0	2385
4	N	1	0	1	34
5	C	1	1	0	9663
6	C	1	1	1	12
					23682

The standard analysis for randomized studies with noncompliance is called Intention to Treat (ITT). This method ignores observed compliance information and compares those assigned to treatment to those assigned to control. This procedure gives a valid estimate of the effect of treatment assignment on outcome, at least from the randomization-based perspective.

"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, i.e., compares those who appeared to comply with the protocol.

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	Calculation	Row Comparison
ITT	-.0026	$= \frac{12+34}{9663+2385+12+34} - \frac{74}{11514+74}$	3, 4, 5, & 6 vs. 1 & 2
As-treated	-.0065	$= \frac{12}{9663+12} - \frac{34+74}{11514+2385+34+74}$	5 & 6 vs. 1, 2, 3, & 4
Per protocol	-.0052	$= \frac{12}{9663+12} - \frac{74}{11514+74}$	5 & 6 vs. 1 & 2

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 generally do not even estimate true causal effects because they compare groups of people who are fundamentally different. [This difference is evident from the data: note that the death rate for the noncompliers in the treatment group is $\frac{34}{34+2385}$ = .014, much higher than the control group's $\frac{74}{74+11514}$ = .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 true compliers, but those who received control are a mixture of true compliers and true noncompliers.

The per protocol estimate ignores the observed true noncompliers in the treatment group, comparing those who complied in the treatment group with those who "complied" in the control group (in our case the whole control group). But this also compares true compliers with a mixture of true compliers and noncompliers, because the control group contains both true compliers and noncompliers.

The ITT estimate compares two groups, each of which is a mixture of true compliers and noncompliers, and because treatment was assigned randomly, the proportion of true compliers and noncompliers on average will 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 essentially the same as "instrumental variables" in economics (Angrist, Imbens, & Rubin, 1996), we can estimate the effect of the treatment assignment on outcome for true compliers, i.e., the causal effect of receiving treatment on outcome for true compliers.

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

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

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

ACE, p_N, and p_C can all be estimated from the data. The ITT estimate is unbiased for the ACE, and p_N and p_C can be estimated as the proportion of compliers and noncompliers in the treatment group, because 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:

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

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

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

This estimate is called the instrumental variables (IV) estimate of the complier average causal effect (CACE). Note that this does estimate the effect of *treatment* received for the true compliers, because 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; in other words, to estimate the complier average causal effect, there must be some compliers!
4. NACE = 0. We assume that, because behavior cannot be changed by assignment for noncompliers, neither can the outcome. This assumption is called the Exclusion Restriction, and must be considered carefully for each experiment because it can be unreasonable.

Noncompliance can be more complicated than illustrated in the previous example. For instance, some individuals may take the active treatment regardless of whether they were assigned control or treatment; we call such people “always-takers.” Examples with both always-takers and never-takers are especially common in so-called “encouragement designs.”

Example IV-2: Encouragement Designs – "Two-Sided" Noncompliance

From Hirano et al. (2000).

One ethical issue in randomized studies concerns 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, 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 is unlikely to be deleterious to anyone, it is often considered unethical to refuse the new, better treatment to some individuals who may want it.

To avoid this dilemma, encouragement designs are sometimes used. In an encouragement design, one group is particularly encouraged to take the treatment of interest. The encouragement should increase the use of the treatment in one group without affecting the use of the treatment in the other group. An example 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. Randomized encouragement designs can then be analyzed in ways similar to randomized studies with noncompliance because subjects may or may not take the treatment that is being encouraged.

In our example, we are interested in estimating the effect of an influenza vaccine on flu-related hospitalizations for elderly patients. Because the flu vaccine is known to be effective in the laboratory, the experimenter could not ethically assign randomly some elderly patients not to receive this treatment. A randomized encouragement design was therefore implemented. Physicians were randomly selected to receive a computer generated reminder encouraging them to give their at-risk patients flu vaccine shots. The outcome of interest is flu-related hospitalization.

There are also two covariates available: patient's age and whether they have chronic obstructive pulmonary disease (COPD). A summary of the data (in terms of means) 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, because 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 did not, due to noncompliance: The patients who received the flu shot are older and more likely to have COPD. We cannot simply compare the outcomes by flu shot status to obtain a reasonable estimate of the effect of taking the vaccine.

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

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

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

As noted before, 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 received (vaccine) status and health, thus invalidating both an as-treated analysis and a per-protocol analysis.

To estimate the causal effect of the vaccine itself on hospitalizations, it helps 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 units' full compliance statuses. We only observe their behavior under the observed assignment. To simplify the calculations, we make the assumption that there are no defiers, that is, people who would do the opposite of their encouragement. That is, we assume that there are no people who would get a flu shot when not encouraged to do so but who not get a flu shot when encouraged to do so. This assumption seems reasonable in this setting. We are then able to identify some people as specific types. Specifically, someone whose doctor receives the letter and who does not get a flu shot must be a never-taker. Therefore, about 1 - 0.307 = .693 = 69.3% of the subjects randomly assigned treatment are never-takers. Similarly, someone whose doctor does not receive the letter but who does get a flu shot must be an always-taker. Therefore, about 19% of the subjects assigned control are always takers. For individuals who are not identified as a specific type, their compliance status can be imputed using a model for compliance status, but a simple estimate for the percentage is 100% - 19% - 69% = 12%.

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, treatment assignment does not affect their flu related hospitalizations.
2. Exclusion restriction for always-takers: for always-takers, treatment assignment does not affect their flu related hospitalizations.

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 after receiving their letters, their doctors might prompt them to receive other health benefits and be more aware of the risks of the flu. The always-takers tend to be sicker than true compliers and never-takers, and so receiving these extra benefits could affect their outcome. In addition, they may receive the vaccine earlier than they would have otherwise. The never-takers may be unlikely to receive other benefits from their doctors, because they are not even receiving the flu vaccine, and so assignment to receive the letter may be unlikely to affect their outcomes. Under the predictive framework, either or both of these assumptions can be relaxed.

In this setting, we observe each unit's covariates (here, age and COPD), assignment (doctor did/did not receive a reminder letter), and one potential outcome (was/was not hospitalized) under assigned treatment, but not the true compliance type. The assumption that there are no defiers allows us to know with certainty the compliance type of any unit that did the opposite of what was assigned; such a unit is either a never-taker or an always taker. But for units who did what they were assigned to do, we do not know compliance status with certainty. For example, for a unit whose doctor received a letter and who obtained a flu shot, is the unit a complier or an always-taker?

We proceed in two steps. First, we build a model to predict the compliance status for all units for whom that variable is missing (i.e., all units who did what they were assigned to do). Then, conditional on this predicted compliance status, we build a model to predict each unit's missing potential outcome (hospitalization or not).

The specifics of the overall model require fancy statistical techniques. Once implemented, the model allows us, among other things, to estimate ITT effects, i.e., for each compliance type, differences in means under the two assignments. The following table summarizes the results for these ITT effects (the standard error estimates are 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)

The results in the "Excl. for never-takers" column lead to an interesting conclusions: encouragement seems to have a similar beneficial effect on the always-takers as it does on the compliers, and thus encouragement to get the shot rather than the shot itself may be reducing flu related hospitalizations!

A final note: recall the estimates in the table above represent hospitalization rates, so a negative number is "good" in the sense that it represents a decreased rate of hospitalization. Is it reasonable to suppose that sending a doctor an encouragement letter could increase the patient's risk of hospitalization? This setting might be a case in which we have a strong prior belief that any causal estimate should be zero or negative. Note that in the third and fourth columns of the table above, the point estimate for the never-takers is positive, although it is within one standard deviation of zero. If this positive point estimate were more than two or three standard deviations from zero, it might provide a reason to doubt the corresponding model. We must be careful any time we use substantive results of the model to assess its adequacy, but this might have been a case where such use would have been prudent.

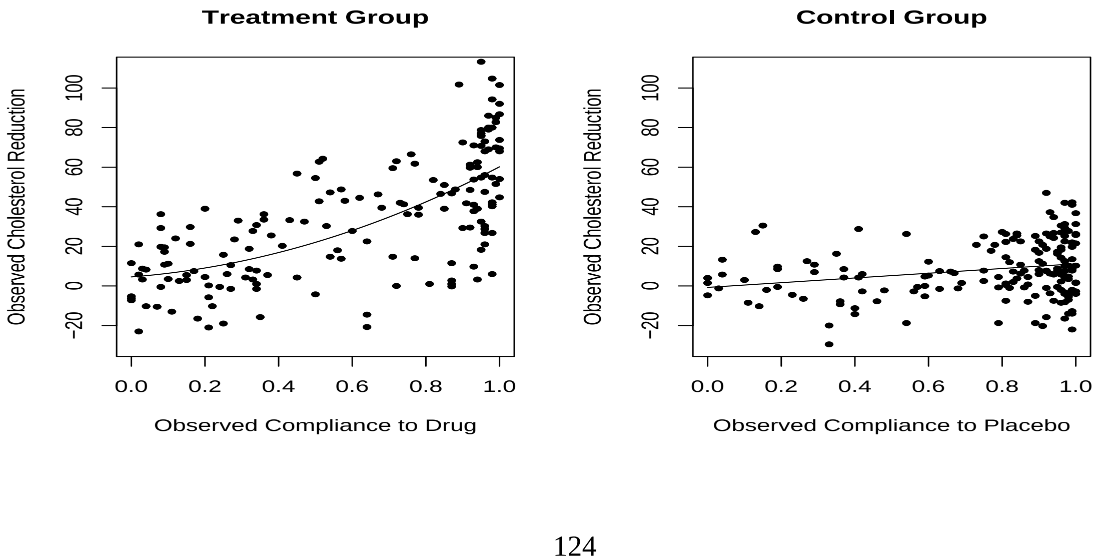
Subsection 18: More Complex Examples of Noncompliance

More complex examples of noncompliance exist, even when active treatment is not available to those not assigned to take it. For example, compliance can be “partial” in the sense that only a fraction of an assigned dose of pills is taken. And we might encounter “extended noncompliance,” where even those assigned “control” may not take it, as when “control” is a placebo pill. The following example from Efron-Feldman (1991) suffers from both complications of noncompliance. The discussion that follows is adopted from Jin and Rubin (2005).

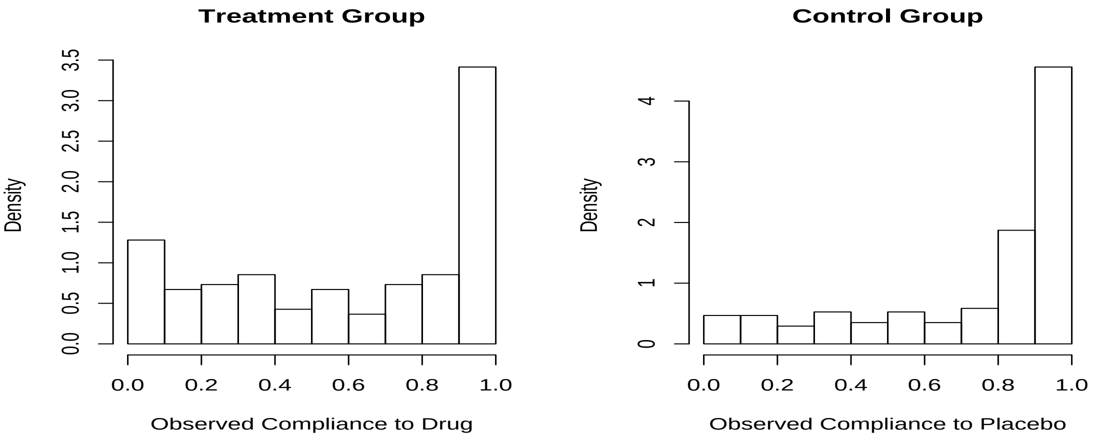
Example IV-??: Extended Partial Compliance in a Cholesterol Drug Study

The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) was a placebo-controlled double blind study of men designed to assess the effect of a drug on cholesterol level. A “double blind” study is a trial in which, in theory, neither the patient nor the administering medical staff know whether the patient is receiving active treatment or placebo. In the LRC-CPPT, treatment assignment was randomized and was thus unconfounded. Each study patient was instructed to take pills provided to them at regular intervals. By careful monitoring, scientists were able to estimate how many pills each patient actually took, i.e., compliance.

If the placebo mimicked the active pill perfectly in all ways except for the effect (if any) of the active ingredient, we might expect compliance in the placebo (control) group to be the same in expectation as compliance in the treated group. In this study, however, the active pill had various side effects, including increased colonic gas, that the placebo did not mimic perfectly. Some question thus existed as to whether a patient assigned active treatment would take the same percentage of pills as he would have taken had he been assigned placebo. Why might this matter? Patients who take a higher percentage of pills (even placebos) may be more conscientious generally, more likely to exercise, more likely to change their diets for the better, etc., and thus these patients might show greater improvement in their cholesterol levels (even if their pills contain no active ingredient). The figures below demonstrate that something like this has occurred in the LRC-CPPT. The Y-axis for both graphs is cholesterol reduction (here higher values are better). The X-axis is the percentage of pills the patients took. As you can see from the left plot, patients who took more pills with the active ingredient tended to have greater cholesterol reduction. But as the right plot demonstrates, patients who took more placebo pills also tended to have greater cholesterol reduction!



Analysis of these data is further complicated by the fact that patients assigned the active treatment appeared to take, on average, fewer pills than those assigned placebo (one might speculate that this result is due to the unpleasant side effects of the active drug that the placebo did not mimic perfectly). The two histograms below reveal this.



This discussion and these figures suggest that we encounter special difficulties when considering (i) treatments more complicated than taking a single pill versus no pill, and (ii) patients who do not comply 100% with their assigned treatment. Stepping back from the specifics of the LRC-CPPT for a moment, in some study situations, it might be possible for patients assigned placebo to obtain the active drug from a source other than the study. Similarly, it might be possible for patients assigned treatment to obtain placebo from an alternative source. For this reason, our potential outcomes table looks like the one below. In this table, observed values are represented by an "x," unobserved by a "?."

					Potential Outcomes					
						If Assigned			If Assigned	
		Covariates				Placebo			Active	
Unit	X_1	...	X_p	W	% Plac.	% Active	Y(0)	% Plac.	% Active	Y(1)
1	x	...	x	1	?	?	?	?	x	x
2	x	...	x	1	?	?	?	?	x	x
...										
n_T	x	...	x	1	?	?	?	?	x	x
$n_T + 1$	x	...	x	0	x	?	x	?	?	?
$n_T + 2$	x	...	x	0	x	?	x	?	?	?
...										
n	x	...	x	0	x	?	x	?	?	?

This table is more complicated than those we have previously encountered because it includes multiple columns under each assignment.

In the specific context of the LRC-CPPT, certain assumptions allowing simplification of this table seemed reasonable. Because the study concerned an experimental drug, it was unlikely that units assigned placebo could obtain the active ingredient from a source other than the trial. Similarly, the placebo pill given was specially created for the LCR-CPPT, so units assigned active treatment probably had no access to placebo. These assumptions, which have been called "access monotonicity," allow us to fill in some columns of the table above, as follows.

					If Assigned			If Assigned		
		Covariates				Placebo			Active	
Unit	X_1	. . .	X_p	W	% Plac.	% Active	Y(0)	% Plac.	% Active	Y(1)
1	x	. . .	x	1	?	0	?	0	x	x
2	x	. . .	x	1	?	0	?	0	x	x
. . .										
n_T	x	. . .	x	1	?	0	?	0	x	x
$n_T + 1$	x	. . .	x	0	x	0	x	0	?	?
$n_T + 2$	x	. . .	x	0	x	0	x	0	?	?
. . .										
n	x	. . .	x	0	x	0	x	0	?	?

Continuing with the specific context of the LRC-CPPT, recall that the purpose of the drug under study was to reduce cholesterol levels in the bloodstream. If the drug did have some effect on blood-level cholesterol, study units were unlikely to perceive this effect on their own (although they might discover their cholesterol levels during visits to doctors not involved in the study, such as their personal physicians). The negative side effects of the active ingredient were easily apparent. It might seem reasonable to assume that for a unit assigned active treatment, the percentage of active pills actually taken is less than (or equal to) the percentage of placebo pills that unit would have taken had he been assigned placebo. Thus, if Unit 2 were assigned active treatment and took 70% of his pills, we might assume that Unit 2 would have taken at least 70% of the placebo pills had he been assigned placebo. Similarly, for a unit assigned placebo, the percentage of pills taken represents an upper bound on the percentage of active pills the unit would have taken had he been assigned the active treatment. Thus, if a unit were assigned control and took 50% of his pills, we might assume that he would have taken at worst 50% of the active pills had he been assigned treatment. This assumption has been called "negative side effect monotonicity," and it allows us to fill in the table above still further in the following manner (the specific numbers used are just for illustration; the decimals represent the fraction of the total number of assigned pills actually taken).

	Covariates				If Assigned Placebo			If Assigned Active		
Unit	X_1	. . .	X_p	W	% Plac.	% Active	Y(0)	% Plac.	% Active	Y(1)
1	x	. . .	x	1	.15 $\leq$	0	?	0	.15	x
2	x	. . .	x	1	.87 $\leq$	0	?	0	.87	x
. . .										
n_T	x	. . .	x	1	.44 $\leq$	0	?	0	.44	x
$n_T + 1$	x	. . .	x	0	.87	0	x	0	$\leq$ .87	?
$n_T + 2$	x	. . .	x	0	.23	0	x	0	$\leq$ .23	?
. . .										
n	x	. . .	x	0	.64	0	x	0	$\leq$ .64	?

The access monotonicity and negative side effect monotonicity assumptions have done much to clarify matters. We can now proceed to analyze the data using a model to fill in the missing potential outcomes and calculate the quantity of interest. The process of doing so is beyond the scope of this course.

Two final questions: first, some aspects of the way patients actually take pills have not been included in the analysis. For example, suppose a unit took 87% of the active pills assigned to him; what in our analysis would change if he took them all at once instead of at the regularly prescribed intervals? What implicit assumption about this point is begin made in the above formulation? Second, in many settings, investigators are interested in something called “dose-response,” or how much of an effect is produced for different levels of the treatment. Without further assumptions, the data above cannot be used for causal inferences about dose-response. Why not? What additional assumptions are necessary to allow dose-response inferences from these data?

Subsection 19: Surrogate Outcomes: "Direct" And "Indirect" Causal Effects

The basic message in this subsection is simple: Avoid using the terminology "direct" causal effect or "indirect" causal effect in the context of a particular study. At least since R.A. Fisher used the terms in 1935, their meaning has been confused and their usage unhelpful, particularly when dealing with post-treatment variables. This situation may clarify in the next few years, but at present, the terms are poorly defined except in simple, artificial situations.

Example IV-3: Race Discrimination in Award of Hiring Bonuses

Plaintiffs filed a lawsuit alleging that a company gives lower hiring bonuses for persons it thinks are non-white. The company's hiring and bonus system are as follows. Applicants fill out forms detailing their backgrounds, education, and experiences. On one of the forms, the applicants check boxes to indicate whether they are "white" or "non-white"; the company's form recites that the information will be used "solely for historical purposes." After it has made its hiring decisions, and for each person hired, the company reviews the application again to assign a "motivation" score M of between 1-6. The M score represents the company's best guess as to the new employee's eagerness to succeed and advance. The company then hands out hiring bonuses.

To make their case, the plaintiffs point to a particular application of the company's hiring and bonus system. Eight applicants applied for positions at the company. In fact, all eight were non-white, but at the request of a discrimination officer, a random half checked the "white" box on the company's forms. All eight were hired. The table below represents the observed data. Let W = 0 if the employee checked the "non-white" box, W = 1 for "white." M$_{obs}$ is the observed value of the motivation score, Y_{obs} is the bonus awarded in thousands of dollars.

Observed Data

Unit #	W	M$_{obs}$	Y$_{obs}$
1	0	2	10
2	1	3	10
3	0	3	12
4	1	4	12
5	0	4	14
6	1	5	18
7	0	5	20
8	1	6	26

The defendant company first argues that there is little observable "direct" effect of race on motivation score. The scores for the four white units (i.e., $M_{obs} = 3, 4, 5$) looks similar to the Ms for the four non-whites (i.e., $M_{obs} = 2, 3, 4)$. Next, the company argues the court should examine whether there is any "direct" effect of race on bonuses controlling for motivation, especially because there is no apparent effect of W (i.e., the race box checked) on motivation score. The company contends that because measuring motivation is a good idea and, here, is apparently non-discriminatory, the court must exclude any "indirect" effect of race on bonuses "through" the motivation score M.

How does one find the “direct” effect of “race” (i.e., the box checked on the application forms) on bonuses while excluding the “indirect” affect race might have through the M-score? The defendant company suggests that one should compare the bonus awarded for non-white and white hires with the same value of M_{obs}. Three such comparisons are possible. For M_{obs} = 3, we have one white person who received a bonus of $10,000.00, and one non-white person who received a bonus of $12,000.00. For M_{obs} = 4, we have one white who received a bonus of $12,000.00, and one non-white who received $14,000.00. For M_{obs} = 5, the white to non-white comparison is $18,000.00 to $20,000.00. (Note that this process is called “conditioning” on the value of M.) The company thus argues that any race discrimination that exists is in favor of non-whites; for each M score, the non-white received a bonus $2000.00 higher than the corresponding white.

The court asks you to assess the company’s interpretation of the data. How do you respond?

You try to clarify the situation. First, the company is thinking only in terms of observed data, not potential outcomes for either the motivation score M (which is a post-treatment variable) or the observed outcome Y (recall that W is randomly assigned before M or Y is observed). Because potential outcomes are not in the picture, none of the conclusions the company has drawn are causal. The (hypothetical) full set of potential outcomes for the units in the table above appears below. M(0) is the motivation score that the unit would have received had that unit checked the non-white box, whereas M(1) is the motivation score that the unit would have received had that unit checked the white box. Analogously, Y(0) is the bonus the unit would have received had that unit checked the non-white box, Y(1) the bonus corresponding to the white box.

	Potential Outcomes			
Unit #	M(0)	M(1)	Y(0)	Y(1)
1	2	3	10	10
2	2	3	10	10
3	3	4	12	12
4	3	4	12	12
5	4	5	14	18
6	4	5	14	18
7	5	6	20	25
8	5	6	22	26

One look at this table verifies that checking the “white” box versus the “non-white” (which was randomly assigned) never results in a lower bonus, and, for units four to eight, substantially increases the bonus awarded. The company is discriminating against persons who check the “non-white” box, but one would never know this from the langauge of “direct” and “indirect” causal effects as used in this example and, we submit, in common use. (Adapted from Rubin (2005)).

Subsection 20: Censoring And/Or Truncation of Outcomes, Such as Due to Death

In some biological and social science settings, we are interested in the effect of a treatment only if some post-treatment condition does not occur. For example, we might be evaluating the effect of hormone replacement therapy (versus placebo) on development of cancer within five years. By definition, we are interested in whether study subjects get cancer within five years post treatment, and we can only observe whether this occurs for study subjects who live that long. What do we do with subjects who die for some reason unrelated to cancer (e.g., a car accident or heart attack) before the stated time period has elapsed? We call the data from such subjects "censored" or "truncated," in this case, censored due to death.

Conceptually, the problem of censoring occurs in a variety of settings. For example, economists define the concept of "wages" as the amount earned by persons who are employed; a person's wages must by definition be above zero. Wages are different from "income," which is the amount of money a person earns even if the person is unemployed (in which case income might be zero). For a person without a job, the concept of wages is undefined. Imagine that we conduct a study to assess the effect of a job training program on wages. The data from study subjects who are unemployed after treatment or control are censored by their failure to obtain jobs.

One (incorrect) way to deal with censoring is to remove censored units from the dataset and compare the treated and control units that remain. One of the difficulties with this approach is that the treatment received may affect the censoring of a unit in a way that also affects the (counterfactual) potential outcome that would have been observed if the unit's data had not been censored. For example, in the job training/wage context, some people who receive training (i.e., are assigned treatment) may be *less* likely to be employed than they would have been had they not received training. Such people might believe that their time is now worth more, or that because they now have new skills, jobs better than the ones they can presently find, and which would have been acceptable in the absence of training, are just around the corner. If there were a sufficient number of such people, removing the censored units from the study before data analysis could seriously bias the results. The now-reduced dataset would include treated units who took only jobs that were "good enough" for them post-training program, whereas the control group would include control units who took the first jobs they could find, but who might have refused those same jobs had they received training. To deal with censoring properly, we need tools similar to those we employed in Subsection 17, where (for example) we divided units into compliers, always-takers, never-takers, and defiers.

Example IV-4: Job Training And Wages

(Adapted from Zhang et al. 2005.) We return to the National Supported Work Demonstration study from Example II-12. As noted in that example, several investigators have attempted to use these data to assess the validity of models to estimate the effect of the job training program at issue on income. Here, we use the data to assess the effect of the program on wages. As noted above, the fact that some units in the study were unemployed several years after receiving either treatment (the training program) or control (no training program) requires us to think carefully and to analyze the resultant data equally carefully.

The fundamental conceptual point is that the training program has no well-defined effect on *wages* except for units who would be employed under *both* treatment and control. There is no well-defined causal effect on wages for persons who would be employed under treatment but unemployed under control, or unemployed under treatment but employed under control, or unemployed under both treatment and control. In each of the latter three cases, the training program cannot affect wages because the concept of wages is not defined for at least one of the potential outcomes. This understanding leads us to define the following table, similar to the table in Subsection 17 dealing with flu shots, which now defines four types of people according to their employment statuses: EE, EU, UE, and UU.

Type	Assigned to	Employment Status
EE	Training	Employed
	No Training	Employed
EU	Training	Employed
	No Training	Unemployed
UE	Training	Unemployed
	No Training	Employed
UU	Training	Unemployed
	No Training	Unemployed

To assess the causal effect of the training program on wages, we need to isolate all individuals who belong to the first group, the EEs. We cannot do so directly because we only get to observe employment status for one of the potential outcomes (under treatment or under control). In other words, we do not observe an individual's status as an EE, EU, UE, or UU. We do know that anyone unemployed is not an EE, and persons who received training and who were employed might be EUs or EEs, whereas persons who received no training and who were employed might be UEs or EEs. Distinguishing the EEs from the EUs assigned treatment from the UEs assigned control is a tricky and technical statistical problem, and we omit the details here. Once done, however, we can proceed to use the techniques discussed in previous sections to estimate the causal effect of training on the EEs.

We summarize some results. Initially, we define some notation. First, we classify the units using the symbols $O(assignment, employment)$, so that $O(1,1)$ refers to the set of units assigned training who were employed after training; $O(1,0)$ refers to those assigned treatment who were unemployed; $O(0,1)$ refers to those assigned no training who were employed; and $O(0,0)$ refers to those assigned control who were unemployed. Next, the covariates measured in this dataset included race (B = black, NB = non-nlack), marital status (M = married, NM = not married), and degree (H = high school degree, NH = no high school degree). Finally, the symbols $\pi_{EE}, \pi_{EU}, \pi_{UE}$, and π_{UU} refer to the probability that a unit is of a particular status. Thus, π_{EE} refers to the probability that a unit is within the class of people who would be employed both with and without training. Other symbols are defined below.

The following table shows summary statistics for the four observed groups in different classes of people. In each cell of the first column, the first row shows the characteristics of a class of people, and the second row shows the number and proportion of that class among the total N=445 participants. In each cell of columns 2-5, the first row shows the number and proportion of the corresponding observed group, and the second row shows the mean and the standard deviation of wage in US dollars for the corresponding observed group that is observed to be employed.

	O(1,1)	O(1,0)	O(0,1)	O(0,0)
B, M, H 15 (3%)	5 (33%) 14621 (13592)	1 (7%)	6 (40%) 5967(4088)	3 (20%)
B, M, NH 49 (11%)	20 (41%) 8108 (6720)	3 (6%)	14 (29%) 7341 (5576)	12 (24%)
B, NM, H 63 (14%)	28 (44%) 10249 (8447)	9 (14%)	15 (24%) 6624 (5988)	11 (18%)
B, NM, NH 244 (55%)	60 (25%) 7250 (8387)	30 (12%)	96 (39%) 6721 (5646)	58 (24%)
NB, M, H 4 (1%)	2 (50%) 11255 (1644)	1 (25%)	1 (25%) 6735 (0)	0 (0%)
NB, M, NH 7 (2%)	3 (43%) 4603 (2026)	0 (0%)	3 (43%) 5569 (1148)	1 (14%)
NB, NM, H 15 (3%)	8 (53%) 6490 (4354)	0 (0%)	6 (40%) 11140 (4471)	1 (7%)
NB, NM, NH 48 (11%)	14 (29%) 9220 (7368)	1 (2%)	27 (56%) 7808 (4684)	6 (13%)

The table below shows estimates of the proportions of number of people of each status type and the average treatment effect on employment (AteE=π_{EU}-π_{UE}) for different classes of people, and their associated 95% interval estimates. In each cell of the first column, the first row shows the characteristic of the corresponding class of people, and the second row shows the number and proportion of that class in the total N=445 participants.

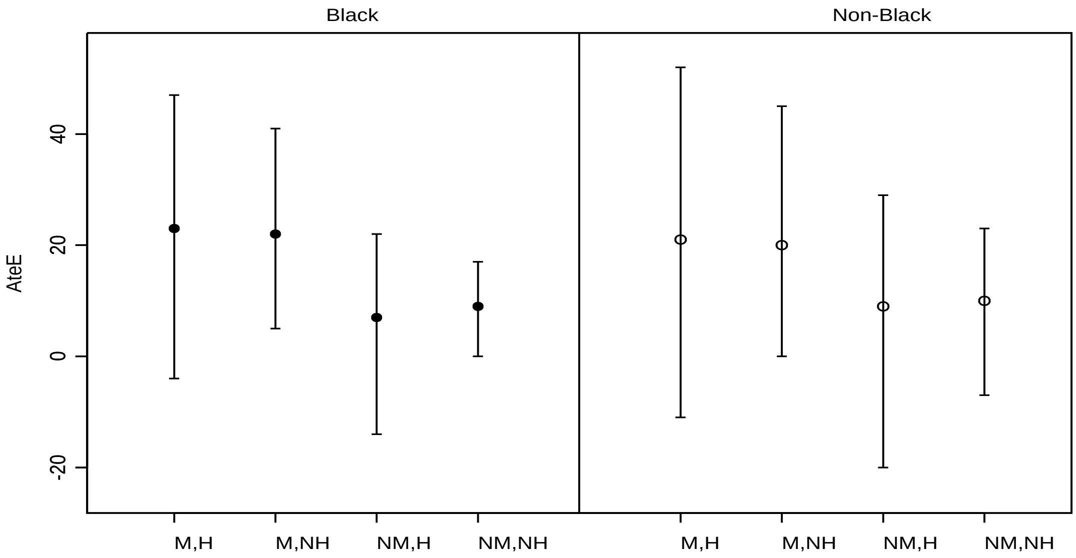
	π_{EE}	π_{EU}	π_{UE}	π_{UU}	$AteE$
B, M, H 15 (3%)	51% (30%,66%)	33% (14%,54%)	10% (2%,29%)	6% (1%,33%)	23% (-4%,47%)
B, M, NH 49 (11%)	57% (38%,68%)	27% (12%,46%)	5% (1%,15%)	11% (2%,39%)	22% (5%,41%)
B, NM, H 63 (14%)	52% (35%,62%)	20% (9%,34%)	13% (4%,32%)	15% (4%,39%)	7% (-14%,22%)
B, NM, NH 244 (55%)	53% (46%,60%)	16% (10%,24%)	7% (3%,15%)	24% (16%,33%)	9% (0%,17%)
NB, M, H 4 (1%)	65% (33%,85%)	28% (10%,57%)	7% (1%,30%)	0%	21% (-11%,52%)
NB, M, NH 7 (2%)	73% (46%,88%)	23% (9%,48%)	4% (1%,20%)	0%	20% (0%,45%)
NB, NM, H 15 (3%)	71% (45%,86%)	19% (8%,37%)	10% (2%,34%)	0%	9% (-20%,29%)
NB, NM, NH 48 (11%)	79% (62%,88%)	15% (8%,28%)	6% (1%,20%)	0%	10% (-7%,23%)

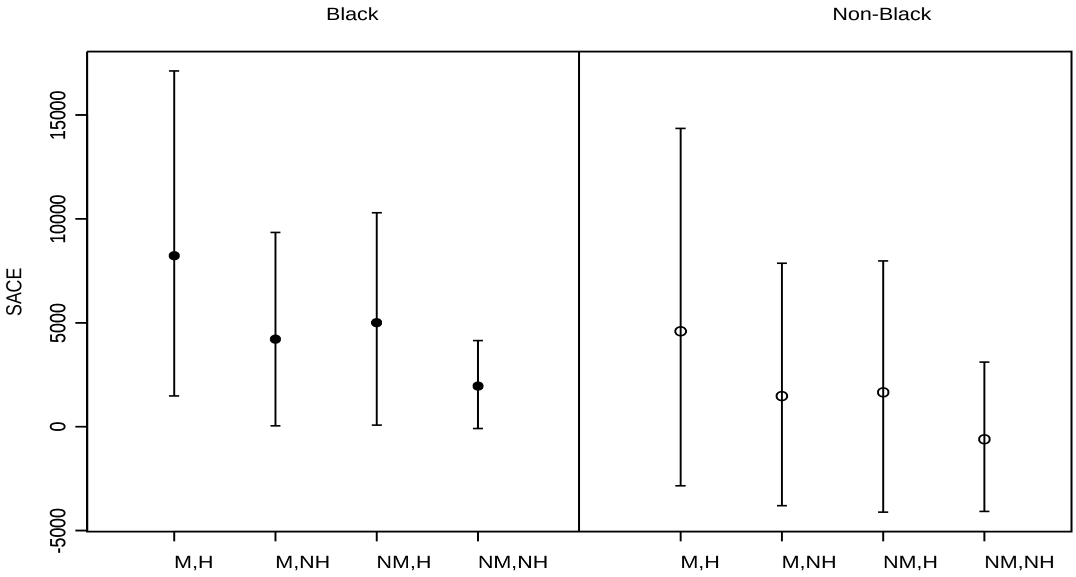
IV-20.5

The table below shows estimates of the treatment effects on wages in US dollars for different classes of people, and their associated 95% interval estimates. "AveEE (T)", "AveEE (C)", "AveEU" and "AveUE" refer to the average wage under treatment for the EE group, the average wage under control for the EE group, the average wage under treatment for the EU group and the average wage under control for the UE group, respectively. SACE refers to the "Survivor Average Causal Effect" – the average effect of the training program on the EE group.

	$SACE$	AveEE (T)	AveEE (C)	AveEU	AveUE
B, M, H	8226 (1482,17124)	16434 (10864,24982)	8208 (5626,11987)	3349 (1393,8022)	345 (127,943)
B, M, NH	4215 (39,9347)	11198 (7839,16069)	6982 (5197,9383)	2161 (1136,4148)	99 (36,266)
B, NM, H	5008 (79,10297)	13615 (10148,18390)	8607 (6014,12298)	1798 (1082,2973)	2309 (1099,4890)
B, NM, NH	1955 (-92,4146)	9277 (7677,11228)	7322 (6236,8642)	1161 (745,1796)	665 (337,1301)
NB, M, H	4587 (-2848,14360)	14353 (8681,23778)	9766 (6472,14759)	4457 (1092,18530)	1623 (253,10327)
NB, M, NH	1472 (-3801,7864)	9780 (6115,15732)	8308 (5700,12250)	2876 (874,9675)	467 (77,2787)
NB, NM, H	1650 (-4111,7978)	11891 (8102,17534)	10241 (7056,14782)	2393 (855,7026)	10853 (2481,48563)
NB, NM, NH	-609 (-4076,3108)	8102 (5805,11379)	8711 (6726,11336)	1545 (626,3908)	3124 (788,12300)

The following figures provide some additional summary. The first shows estimates of the treatment effects on employment (in percentages) among classes of people defined by their covariates (the bars represent 95% intervals). The second shows estimates of the causal effect in the EE group in US dollars, again among classes of people defined by their covariates and again with 95% intervals.





Note that in some scientific settings, we might be able to make some assumptions that simplify the analysis. In Example IV-3, it was hard to assume away any of the four EE, EU, UE, and UU groups. Such assumptions have to be assessed carefully, but if some could reasonably be made, the statistical task would be simpler.

Part V: Conclusion

Subsection 21: Bibliography

The following is a list of suggested readings in this field. Note that they come from a variety of academic disciplines.

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

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

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

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

5. Chernoff, H. (1959). Sequential Design of Experiments. *Annals of Statistics* 30: 755-770.

6. Cochran, W.G. (1968). "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics* 24: 295-313.

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

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

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

10. 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.

11. Efron, B. (1971). Forcing a Sequential Experiment To Be Balanced. *Biometrika* 58:403–417.

12. Efron, B., and Feldman, D. (1991). Compliance as an Explanatory Variable in Clinical Trials *Journal of the American Statistical Association* 86:9–17.

13. 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.

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

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

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

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

18. 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.

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

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

21. 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.

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

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

24. Jin, H. and Rubin, D.B. (2005). Principal Stratification for Causal Inference with Extended Partial Compliance (forthcoming).

25. Lalonde, R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review* 76: 604–620.

26. 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.

27. 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.

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

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

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

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

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

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

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

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

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

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

38. Rubin, D. (1978), Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics* 6:34–58.

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

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

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

42. 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.

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

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

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

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

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

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

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

50. Rubin, D.B. (2005). Causal Inference Using Potential Outcomes: Design, Modelling, Decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* 100: 322-331.

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

52. Ware, J. (1989) Investigating Therapies of Potentially Great Benefit: ECMO. *Statistical Science* 4:298-306.

53. Zhang, J.L. et al. (2005). Evaluating Causal Effects in the Presence of "Truncation by Death" – Likelihood-based Analysis Via Principal Stratification (forthcoming).