Field Experiments: Design, Analysis and Interpretation Solution Sets

Alan S. Gerber and Donald P. Green*
DO NOT DISTRIBUTE

January 20, 2016

Follow these links to jump to a specific chapter:

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 1 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Core concepts: [25 points]

a) What is an experiment, and how does it differ from an observational study?
Answer:
A randomized experiment is a study in which observations are allocated by chance to receive some type of treatment; in an observational (or non-experimental) study, treatments are not assigned randomly.

b) What is "unobserved heterogeneity," and what are its consequences for the interpretation of correlations?
Answer:
Unobserved heterogeneity refers to the set of unmeasured factors that cause outcomes to vary from one subject to the next. Unobserved heterogeneity complicates the task of drawing causal inferences from correlations between treatments and outcomes because treatments that are not randomly assigned may be correlated with unmeasured factors that predict outcomes.

Question 2

Would you classify the study described in the following abstract as a field experiment, a natural experiment, or a quasi-experiment? Why? [25 points]

> "This study seeks to estimate the health effects of sanitary drinking water among low-income villages in Guatemala. A random sample of all villages with fewer than 2,000 inhabitants were selected for analysis. Of the 250 villages sampled, 110 were found to have unsanitary drinking water. In these 110 villages, infant mortality rates were, on average, 25 deaths per 1,000 live births, as compared to 5 deaths per 1,000 live births in the 140 villages with sanitary drinking water. Unsanitary drinking water appears to be a major contributor to infant mortality."

Answer:
This study is a quasi-experiment. Although villages are sampled randomly, random assignment is

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

not used to determine which villages receive sanitary drinking water (the treatment in this study). The lack of random assignment means that this study does not qualify as either an experiment or natural experiment, the latter being a special kind of experiment in which governments or other non-academic entity allocates treatments randomly.

Question 3

Based on what you are able to infer from the following abstract, to what extent does the study described seem to fulfill the criteria for a field experiment? [25 points]

> "We study the demand for household water connections in urban Morocco, and the effect of such connections on household welfare. In the northern city of Tangiers, among homeowners without a private connection to the city's water grid, a random subset was offered a simplified procedure to purchase a household connection on credit (at a zero percent interest rate). Take-up was high, at 69%. Because all households in our sample had access to the water grid through free public taps ... household connections did not lead to any improvement in the quality of the water households consumed; and despite a significant increase in the quantity of water consumed, we find no change in the incidence of waterborne illnesses. Nevertheless, we find that households are willing to pay a substantial amount of money to have a private tap at home. Being connected generates important time gains, which are used for leisure and social activities, rather than productive activities."[1]

Answer:
This study is an experiment because subjects (those without a private connection to the water grid) were randomly offered an opportunity to purchase a connection. The study satisfies many of the criteria for classification as a field experiment: it was conducted in a naturalistic setting, involved actual consumers, tested the effects of a real intervention (an opportunity to purchase a private water connection on favorable financial terms), and measured meaningful real-world outcomes, such as time use (although we cannot tell from this description whether the measurement of outcomes was unobtrusive).

Question 4

A parody appearing in the British Medical Journal questioned whether parachutes are in fact effective in preventing death when skydivers are presented with severe "gravitational challenge."[2] The authors point out that no randomized trials have assigned parachutes to skydivers. Why is it reasonable to believe that parachutes are effective even in the absence of randomized experiments that establish their efficacy? [25 points]
Answer:
Although randomized experiments could in principle answer the question of whether parachutes are effective against "gravitational challenge," it is unnecessary to conduct a randomized experiment in this case because the threats to inference posed by self-selection into treatment or unobserved heterogeneity seem far-fetched. the laws of physics strongly shape our prior beliefs about what happens to people if they fall from several thousand feet without a parachute. Observational data

[1] Devoto et al. 2011.
[2] Smith and Pell 2003.

corroborate these intuitions – chances of survival are infinitesimal when parachutes malfunction and very good when parachutes work properly – and it is hard to think of a scenario under which this correlation could be spurious, as this relationship holds not only for humans but also for animals, who were used to test parachutes during their development.

Finally, the "treatment effect" of parachutes is extremely large. Nearly all those that fall out of airplanes without them die and nearly all those that have a parachute survive. Any unobserved heterogeneity to bias the estimate of effectiveness would have to be extremely powerful – those who would die regardless of having a parachute would all have to select into the "no parachute" condition. This is a type of informal sensitivity analysis – we reject the notion that the correlation is spurious because the size of the treatment effect is overwhelming.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 2 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Potential outcomes notation:[5 points]

a) Explain the notation "$Y_i(0)$."
Answer:
The potential outcome for subject i if this subject were untreated. Another way to put it: the untreated potential outcome for subject i. Note that the argument in parentheses refers to the case in which d (the treatment indicator) equals zero (lack of treatment).

b) Explain the notation "$Y_i(0)|D_i = 1$" and contrast it with the notation "$Y_i(0)|d_i = 1$"
Answer:
$Y_i(0)|D_i = 1$ The untreated potential outcome for subject i who hypothetically receives the treatment, whereas $Y_i(0)|d_i = 1$ is the untreated potential outcome for subject i if i is actually treated.

c) Contrast the meaning of "$Y_i(0)$" with the meaning of "$Y_i(0)|D_i = 0$."
Answer:
The first is the untreated potential outcome for subject i; the second is the untreated potential outcome for a subject who is untreated under some hypothetical assignment.

d) Contrast the meaning of "$Y_i(0)|D_i = 1$" with the meaning of "$Y_i(0)|D_i = 0$."
Answer:
The first is the untreated potential outcome for a subject in the treatment group under a hypothetical treatment allocation; the second is the untreated potential outcome for a subject who is in the control group under a hypothetical allocation.

e) Contrast the meaning of $E[Y_i(0)]$ with the meaning of $E[Y_i(0)|D_i = 1]$.
Answer: The first is the expectation of the untreated potential outcome for the entire subject pool, whereas the second is the expected untreated potential outcome for a randomly selected subject who would receive the treatment in a hypothetical allocation.

f) Explain why the "selection bias" term in equation (2.15), $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$, is zero when D_i is randomly assigned.
Answer:

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

This equality states that when treatments are allocated randomly, the untreated potential outcome for a subject who actually receives the treatment is, in expectation, the same as the untreated outcome for a subject who goes untreated. This equality follows from the fact that under random assignment, $E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$, since both the treatment and control groups are random samples of the entire set of potential outcomes.

Question 2

Use the values depicted in Table 2.1 to illustrate that $E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$.[5 points]
Answer:
Using the values in the table, we obtain:

$$E(Y_i(0)) = \frac{\sum_{i=1}^{7} Y_i(0))}{7} = \frac{(10+15+20+20+10+15+15)}{7} = \frac{105}{7} = 15$$
$$E(Y_i(1)) = \frac{\sum_{i=1}^{7} Y_i(1))}{7} = \frac{(15+15+30+15+20+15+30)}{7} = \frac{140}{7} = 20$$

And therefore: $E[Y_i(0)] - E[Y_i(1)] = -5$
Alternatively, we may calculate the expectation of each of the differences:

$$\begin{aligned} E[Y_i(0) - Y_i(1)] &= \frac{\sum_{i=1}^{7} Y_i(0) - Y_i(1)}{7} \\ &= \frac{(10-15)+(15-15)+(20-30)+(20-15)+(10-20)+(15-15)+(15-30)}{7} \\ &= \frac{-35}{7} \\ &= -5 \end{aligned}$$

Question 3

Use the values depicted in Table 2.1 to complete the following table.[5 points]

a) Fill in the number of observations in each of the nine cells.
see below.

b) Indicate the percentage of all subjects that fall into each of the nine cells. (These cells represent what is known as the joint distribution of $Y_i(0)$ and $Y_i(1)$, or $p(Y_i(0), Y_i(1))$.
see below.

c) At the bottom of the table, indicate the proportion of subjects falling into each category of $Y_i(1)$ (These cells represent what is known as the marginal distribution of $Y_i(1)$, or $p(Y_i(1))$.
see below.

d) At the right of the table, indicate the proportion of subjects falling into each category of $Y_i(0)$ (i.e., the marginal distribution of $Y_i(0)$, or $p(Y_i(0))$.

Table 1: Table for Question 3

$Y_i(0)$ \ $Y_i(1)$	15	20	30	
10	1: 1/7	1: 1/7	0: 0/7	2/7
15	2: 2/7	0: 0/7	1: 1/7	3/7
20	1: 1/7	0: 0/7	1: 1/7	2/7
	4/7	1/7	2/7	1

e) Use the table to calculate the conditional expectation that $E[Y_i(0)|Y_i(1) > 15]$. (Hint: this expression refers to the expected value of $Y_i(0)$ given that $Y_i(1)$ is greater than 15.)

$$\begin{aligned} E[Y_i(0)|Y_i(1) > 15] &= \sum_i Y_i(0) \frac{pr(Y(0) = Y_i(0), Y_i(1) > 15)}{pr(Y_i(1) > 15)} \\ &= 10 * \frac{(1/7)}{(3/7)} + 15 * \frac{(1/7)}{(3/7)} + 20 * \frac{(1/7)}{(3/7)} \\ &= 15 \end{aligned}$$

f) Use the table to calculate the conditional expectation that $E[Y_i(1)|Y_i(0) > 15]$.

$$\begin{aligned} E[Y_i(1)|Y_i(0) > 15] &= \sum_i Y_i(1) \frac{pr(Y(1) = Y_i(1), Y_i(0) > 15)}{pr(Y_i(0) > 15)} \\ &= 15 * \frac{(1/7)}{(2/7)} + 20 * \frac{0}{(2/7)} + 30 * \frac{(1/7)}{(2/7)} \\ &= 22.5 \end{aligned}$$

Question 4

Define the average treatment effect among the treated, or ATT for short, as $E[\tau_i|D_i = 1]$. Using the equations in this chapter, prove the following claim: "When subjects are randomly assigned to treatment, the ATT is, in expectation, equal to the ATE. In other words, taking expectations over all possible random assignments, $E[\tau_i|D_i = 1] = E[\tau_i]$." [5 points]

Answer:

Because the units assigned to the control group are a random sample of all units, the average of the control group outcomes $Y_i(0)|(D_i = 0)$ is an unbiased estimator of the average value of $Y_i(0)$ among all units. The same goes for the treatment group: the average outcome among units that receive the treatment is an unbiased estimator of the average value of $Y_i(1)$ among all units. Formally, if we order the villages such that the first m observations are from the randomly assigned treatment group and the remaining N-m observations from the control group, we can analyze the expected, or average, outcome over all possible random assignments:

$$
\begin{aligned}
E\left[\frac{\sum_1^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}\right] &= E\left[\frac{\sum_1^m Y_i}{m}\right] - E\left[\frac{\sum_{m+1}^N Y_i}{N-m}\right] \\
&= \frac{E[Y_1] + E[Y_2] + \ldots + E[Y_m]}{m} - \frac{E[Y_{m+1}] + E[Y_{m+2}] + \ldots + E[Y_N]}{N-m} \\
&= E[Y_i(1)D_i = 1] - E[Y_i(0)D_i = 0] \\
&= E[Y_i(1)] - E[Y_i(0)] \\
&= E[\tau_i] \\
&= ATE
\end{aligned}
$$

When treatments are allocated randomly, the expected outcomes in the treated group are the same as for the untreated group and for the subject pool as a whole. Therefore, when treatment is random, $ATT = ATE$.

Question 5

A researcher plans to ask six subjects to donate time to an adult literacy program. Each subject will be asked to donate either 30 or 60 minutes. The researcher is considering three methods for randomizing the treatment. One method is to flip a coin before talking to each person and to ask for a 30-minute donation if the coin comes up heads or a 60-minute donation if it comes up tails. The second method is to write "30" and "60" on three playing cards each, and then shuffle the six cards. The first subject would be assigned the number on the first card, the second subject would be assigned the number on the second card, and so on. A third method is to write each number on three different slips of paper, seal the six slips into envelopes, and shuffle the six envelopes before talking to the first subject. The first subject would be assigned the first envelope, the second subject would be assigned the second envelope, and so on. [10 points]

a) Discuss the strengths and weaknesses of each approach.
Answer:
All three physical methods of random assignment require that the person or persons in charge of implementing the randomization follow the intended protocol: dice must be rolled once per subject, and cards or envelopes must be shuffled thoroughly. Assuming that the mechanics of each physical method of randomization are carried out, the limitation of the dice method is that possibility that the allocation of treatments could wind up being imbalanced; in principle, one could flip a coin 6 times and come up with 6 heads, in which case the treatments would not vary. The card method overcomes this problem and ensures that exactly half of the subjects will receive each treatment. The advantage of the sealed envelope method over the card method is the fact that envelopes help prevent the person who is allocating subjects from deliberately or unconsciously exercising discretion over who receives which treatment, thereby subverting the randomization. It also prevents the implementer from anticipating the next treatment assignment (until the last few envelopes).

b) In what ways would your answer to (a) change if the number of subjects were 600 instead of 6?
Answer:
As the N increases, the dice method becomes more likely to produce a 50-50 division in treatments. For example, with 600 subjects, the probability of obtaining an assignment as imbalanced as 250-350 is less than 1-in-10,000.

c) What is the expected value of D if the coin toss method is used? What is the expected value of D if the sealed envelope method is used?
Answer:
The methods produce identical results, in expectation.
The expected value of X if the dice is used: $E[x_{dice}] = \frac{1}{2}30 + \frac{1}{2}60 = 45$.
The expected value of X if the envelope method is used: $E[x_{envelope}] = \frac{30+30+30+60+60+60}{6} = 45$

Question 6

Many programs strive to help students prepare for college entrance exams, such as the SAT. In an effort to study the effectiveness of these preparatory programs, a researcher draws a random sample of students attending public high school in the United States, and compares the SAT scores of those who took a preparatory class to those who did not. Is this an experiment or an observational study? Why? [10 points]
Answer:
This is an observational study. Subjects are not randomly assigned to the treatment, which in this case is taking the preparatory class. Instead, they self-select into the treatment for unknown reasons. The fact that the students were sampled randomly from the large population is immaterial; the key issue is whether students in the sample were randomly allocated to the treatment or control group. Note that this research method is prone to bias. If students with higher potential outcomes tend to take the prep class, this research design will tend to produce upwardly biased estimates of the ATE; if students with low potential outcomes tend to take the class in order to improve what they expect to be a sub-par score, this research design will tend to produce downwardly biased estimates of the ATE.

Question 7

Suppose that an experiment were performed on the villages in Table 2.1, such that two villages are allocated to the treatment group and the other five villages to the control group. Suppose that an experimenter randomly selects villages 3 and 7 from the set of seven villages and places them into the treatment group. Table 2.1 shows that these villages have unusually high potential outcomes. [10 points]

a) Define the term *unbiased estimator*.
Answer:
An unbiased estimator is a formula that, on average over hypothetical replications of the study, generates estimates that equal the true parameter. Any given estimate may be too high or too low, but on average over hypothetical replications of the study, an unbiased estimator recovers the estimand.

b) Does this allocation procedure produce upwardly biased estimates? Why or why not?
Answer:
No. The procedure is unbiased because the two villages selected for treatment as drawn randomly from the list of villages; therefore their potential outcomes are, in expectation, identical to the average potential outcomes for the entire set of villages. Although in this instance the random allocation procedure produced an estimate that was not equal to the true ATE, the procedure

remains unbiased because across all possible random allocations, the average estimate equals the true ATE.

c) Suppose that instead of using random assignment, the researcher placed Villages 3 and 7 into the treatment group because the treatment could be administered inexpensively in those villages. Explain why this procedure is prone to bias.
Answer:
Unlike random assignment, inexpensiveness is not a criterion the ensures that the treatment group and control group have potential outcomes that are identical in expectation. For example, it may be that villages are inexpensive to treat because they are near transportation networks, which may in turn mean that their potential outcomes are unusual due to increased access to or demand for water sanitation.

Question 8

An experiment by Peisakhin and Pinto[1] reports the results of an experiment in India designed to test the effectiveness of a policy called the Right to Information Act, which allows citizens to inquire about the status of a pending request from government officials. In their study, the researchers hired confederates, slum dwellers who sought to obtain ration cards (which permit the purchase of food at low cost). Applicants for such cards must fill out a form and have their residence and income verified by a government agent. Slum dwellers widely believe that the only way to obtain a ration card is to pay a bribe. The researchers instructed the confederates to apply for ration cards in one of four ways, specified by the researchers. The control group submitted an application form at a government office; the RTIA group submitted a form and followed it up with an official Right to Information request; the NGO group submitted a letter of support from a local nongovernmental organization (NGO) along with the application form; and finally, a bribe group submitted an application and paid a small fee to a person who is known to facilitate the processing of forms. Slum dwellers widely believe that the only way to obtain a ration card is to pay a bribe. The researchers instructed the confederates to apply for ration cards in one of four ways, specified by the researchers. The control group submitted an application form at a government office; the RTIA group submitted a form and followed it up with an official Right to Information request; the NGO group submitted a letter of support from a local nongovernmental organization (NGO) along with the application form; and finally, a bribe group submitted an application and paid a small fee to a person who is known to facilitate the processing of forms. [10 points]

Table 2: Table for Question 8

	Bribe	RTIA	NGO	Control
Number of confederates in the study	24	23	18	21
Number of confederates who had residence verification	24	23	18	20
Median number of days to residence verification	17	37	37	37
Number of confederates who received a ration card within one year	24	20	3	5

a) Interpret the apparent effects of the treatments on the proportion of applicants who have their residence verified and the speed with which verification occurred.
Answer:

[1] Peisakhin and Pinto 2010.

Each of the treatments had a slight effect on the first outcome, the probability of residence verification. In the control group, this rate was 20/21 or approximately 95%. In the three treatment groups, the rate is 100%, implying an average treatment effect of approximately 100 - 95 = 5 percentage points. In terms of the median number of days until residence verification, the RTIA and NGO treatments were the same as the control group, implying an estimated ATE of 37 - 37 = 0. However, the Bribe group received their verification in only 17 days, which is 37 - 17 = 20 days faster than the control group.

b) Interpret the apparent effects of the treatments on the proportion of applicants who actually received a ration card.
Answer:
In the control group, the rate was 5/21 or 24%. The NGO group fared slightly worse 3/18 = 17%. When a right to information request was filed, this rate jumped to 20/23 = 87%, which approaches the 24/24 = 100% success rate among those who paid a bribe.

c) What do these results seem to suggest about the effectiveness of the Right to Information Act as a way of helping slum dwellers obtain ration cards?
Answer:
Although the RTIA treatment does not appear to speed the process of residency verification, it does seem to increase the probability of receiving a card by 20/23 - 5/21 = 63 percentage points over the control group, which seems like a large effect, especially for a treatment that may be implemented inexpensively by applicants.

Question 9

A researcher wants to know how winning large sums of money in a national lottery affects people's views about the estate tax. The researcher interviews a random sample of adults and compares the attitudes of those who report winning more than $10,000 in the lottery to those who claim to have won little or nothing. The researcher reasons that the lottery chooses winners at random, and therefore the amount that people report having won is random. [10 points]

a) Critically evaluate this assumption. (Hint: are the potential outcomes of those who report winning more than $10,000 identical, in expectation, to those who report winning little or nothing?)
Answer:
This assumption may not be plausible in this application. Although lottery winners are chosen at random from the pool of players in a given lottery, this study does not compare (randomly assigned) winners and losers from a pool of lottery players. Instead, winners are compared to non-winners, where the latter group may include non-players. Winning is therefore not randomly assigned. If frequent players are more likely to win than non-players and the two groups have different potential outcomes, the comparison of the two groups may be prone to bias.

b) Suppose the researcher were to restrict the sample to people who had played the lottery at least once during the past year. Is it now safe to assume that the potential outcomes of those who report winning more than $10,000 are identical, in expectation, to those who report winning little or nothing?
Answer:
The assumption is not rooted in a randomization procedure because frequent players are still more likely to be winners than infrequent players. Unfortunately, without detailed information

about how many tickets were purchased for each lottery, we don't know the exact probability that each subject would win. If frequent and infrequent players have different potential outcomes, the comparison is prone to bias (although, arguably, less bias than a comparison of winners to non-players).

Question 10

Suppose researchers seek to assess the effect of receiving a free newspaper subscription on students' interest in politics. A list of student dorm rooms is drawn up and sorted randomly. Dorm rooms in the first half of the randomly sorted list receive a newspaper at their door each morning for two months; dorm rooms in the second half of the list do not receive a paper. [10 points]

a) University researchers are sometimes required to disclose to subjects that they are participating in an experiment. Suppose that prior to the experiment, researchers distributed a letter informing students in the treatment group that they would be receiving a newspaper as part of a study to see if newspapers make students more interested in politics. Explain (in words and using potential outcomes notation) how this disclosure may jeopardize the excludability assumption.
Answer:
The letter is distributed to the treatment group only, so the random assignment is now related to two potential treatments: the newspaper and the letter. In order to use the treatment versus control comparison to identify the ATE of the newspaper, one must assume that the letter has no effect. Formally, this excludability condition states that potential outcomes $Y_i(z, d)$ are affected solely by the treatment ($D_i = d$, whether one receives the newspaper), not by the random assignment and its other consequences ($Z_i = z$, the assigned condition, which determines whether one receives the letter): $Y_i(1, d) = Y_i(0, d)$.

b) Suppose that students in the treatment group carry their newspapers to the cafeteria where they may be read by others. Explain (in words and using potential outcomes notation) how this may jeopardize the non-interference assumption.
Answer:
If the treatment effect is defined as the difference between receipt of the newspaper and no treatment whatsoever, the fact that the control group is exposed to the treatment in the cafeteria is a possible source of bias. In an extreme case where everyone in both treatment and control groups reads the paper (either because they receive it or find it in the cafeteria), a comparison of the treatment and control group may suggest no effect, even if the ATE is large. In a less extreme case, where cafeteria exposure increases with the number of treated friends one has, potential outcomes depend on how the random assignment happens to allocate papers.

Question 11

Several randomized experiments have assessed the effects of drivers' training classes on the likelihood that a student will be involved in a traffic accident or receive a ticket for a moving violation. A complication arises because students who take drivers' training courses typically obtain their licenses faster than students who do not take a course. (The reason is unknown but may reflect the fact that those who take the training are better prepared for the licensing examination.) If students in the control group on average start driving much later, the proportion of students who have an accident or receive a ticket could well turn out to be higher in the treatment group. Suppose a

researcher were to compare the treatment and control group in terms of the number of accidents that occur within 3 years of obtaining a license.[10 points]

a) Does this measurement approach maintain symmetry between treatment and control groups?
Answer:
No, because the measurement procedure differs for treatment and control groups. If control subjects tend to receive their licenses later, the apparent treatment effect may be biased by the fact that the control group is on average older than the treatment group during the period of study. If the groups have different ages, their potential outcomes may differ as well.

b) Would symmetry be maintained if the outcome measure were the number of accidents per mile of driving?
Answer:
No, the problem of asymmetry remains. The control group tends to be older, so their driving patterns may differ, which in turn implies different potential outcomes.

c) Suppose researchers were to measure outcomes over a period of three years starting the moment at which students were randomly assigned to be trained or not. Would this measurement strategy maintain symmetry? Are there drawbacks to this approach?
Answer:
Yes, this approach maintains symmetry, since the clock starts at the same moment for both treatment and control. However, the estimand is now the combined effect of the program on the amount of driving and the quality of the drivers. The program might improve driver quality yet produce more accidents due to increased driving. Some of the uncertainty of interpretation would be eliminated if the driving program were to focus solely on those who already have their licenses, so that eligibility to drive were held constant.

Question 12

A researcher studying 1,000 prison inmates noticed that prisoners who spend at least 3 hours per day reading are less likely to have violent encounters with prison staff. The researcher therefore recommends that all prisoners be required to spend at least 3 hours reading each day. Let D_i be 0 when prisoners read less than 3 hours each day and 1 when prisoners read more than 3 hours each day. Let $Y_i(0)$ be each prisoner's potential number of violent encounters with prison staff when reading less than 3 hours per day, and let $Y_i(1)$ be each prisoner's potential number of violent encounters when reading more than 3 hours per day. [10 points]

a) In this study, nature has assigned a particular realization of d_i to each subject. When assessing this study, why might one be hesitant to assume that $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ and $E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$?
Answer:
In this case, those who self-select into the treatment may have distinctive potential outcomes – bookish inmates may be less prone to violence. In that case, $E[Y_i(0)|D_i = 0] \neq E[Y_i(0)|D_i = 1]$. Thus, a comparison of readers and non-readers will not tend to produce unbiased estimates of the ATE.

b) Suppose that researchers were to test this researcher's hypothesis by randomly assigning 10 prisoners to a treatment group. Prisoners in this group are required to go to the prison library

and read in specially designated carrels for 3 hours each day for one week; the other prisoners, who make up the control group, go about their usual routines. Suppose, for the sake of argument, that all prisoners in the treatment group in fact read for 3 hours each day and that none of the prisoners in the control group read at all during the week of the study. Critically evaluate the excludability assumption as it applies to this experiment.

Answer:

The excludability assumption implies that potential outcomes respond only to the specified treatment (reading) and not to the random assignment (and other factors it may set in motion). Before attributing the apparent contrast in outcomes between the treatment and control groups to reading per se, we might want to find out what other activities the reading period replaced in the treatment group's schedule. For example, if reading took the place of some activity that often provoked violent encounters with guards (e.g., weightroom exercise), the effect of reading might actually be due to a substitution effect, not the effect of reading per se.

c) State the assumption of non-interference as it applies to this experiment.

Answer:

The requirement that $Y_i(d) = Y_i(\mathbf{d})$ implies that each subject's potential outcomes respond only to the treatment they personally receive, not the treatments received by others. In this case, each prisoner's potential outcomes might depend on which other prisoners are assigned to the reading group (if it's an unruly bunch, reading might not be a quiet, contemplative activity).

d) Suppose that the results of this experiment were to indicate that the reading treatment sharply reduces violent confrontations with prison staff. How does the non-interference assumption come into play if the aim is to evaluate the effects of a policy whereby all prisoners are required to read for 3 hours?

Answer:

The fact that only 10 prisoners were assigned to the reading period means that one must be cautious about generalizing to a policy whereby all prisoners are treated simultaneously. Potential outcomes might be different if a very large proportion of prisoners were sent to reading period, perhaps because a universal reading period would have to be closely monitored by guards in order to maintain control over the entire prison population, which changes the nature of the treatment as well as the likelihood of a violent confrontation.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 3 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts: [10 points]

a) What is a standard error? What is the difference between a standard error and a standard deviation?
Answer:
The standard error is a measure of the statistical uncertainty surrounding a parameter estimate. The standard error is a measure of dispersion in a sampling distribution; the standard deviation is the measure of dispersion of any distribution but is most often used to describe the dispersion in an observed variable. The standard error is the standard deviation of the sampling distribution, or the set of estimates that could have arisen under all possible random assignments.

b) How is randomization inference used to test the sharp null hypothesis of no effect for any subject?
Answer:
The sharp null hypothesis of no effect is a case in which $Y_i(1) = Y_i(0)$; under this assumption, all potential outcomes are observed because treated and untreated potential outcomes are identical. In order to form the sampling distribution under the sharp null hypothesis of no effect, we simulate a random assignment and calculate the test statistic (for example, the difference-in-means between the assigned treatment and control groups). This simulation is repeated a large number of times in order to form the sampling distribution under the null hypothesis. The p-value of the test statistic that is observed in the actual experiment is calculated by finding its location in the sampling distribution under the null hypothesis. For example, if the observed test statistic is as large or larger than 9,000 of 10,000 simulated experiments, the one-tailed p-value is 0.10.

c) What is a 95% confidence interval?
Answer:
A confidence interval consists of two estimates, a lower number and an upper number, that are intended to bracket the true parameter of interest with a specified probability. An estimated confidence interval is a random variable that varies from one experiment to the next due to random variability in how units are allocated to treatment and control. A 95% interval is designed to bracket the true parameter with a 0.95 probability across hypothetical replications of a given experiment. In other words, across hypothetical replications, 95% of the estimated 95% confidence intervals will bracket the true parameter.

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

d) How does complete random assignment differ from block random assignment and clustered random assignment? Answer:
Under complete random assignment, each subject is assigned separately to treatment or control groups such that m of N subjects end up in the treatment condition. Under block random assignment, complete random assignment occurs within each block or subgroup. Under clustered assignment, groups of subjects are assigned jointly to treatment or control; the assignment procedure requires that if one member of the group is assigned to the treatment group, all others in the same group are also assigned to treatment.

e) Experiments that assign the same number of subjects to the treatment group and control group are said to have a "balanced design." What are some desirable statistical properties of balanced designs?
Answer:
One desirable property of a balanced design is that under certain conditions, it generates less sampling variability than unbalanced designs; this property of balanced designs holds when the variance of $Y_i(0)$ is approximately the same as the variance of $Y_i(1)$. Another attractive property is that estimated confidence intervals are, on average, conservative (they tend to overestimate the true amount of sampling variability) under balanced designs. (A final attractive property, which comes up in Chapter 4, is that regression is less prone to bias under balanced designs.)

Question 2

Rewrite equation (3.4) substituting for $Y_i(1)$ using the equation $Y_i(1) = Y_i(0) + \tau_i$. Assume that $N = 2m$, and interpret the implications of the resulting formula for experimental design. [5 points]
Answer:
Substituting $N = 2m$ and $Y_i(1) = Y_i(0) + \tau_i$ gives:

$$
\begin{aligned}
SE(\widehat{ATE}) &= \sqrt{\frac{1}{(N-1)}\left\{\frac{mVar(Y_i(0))}{2m-m} + \frac{mVar(Y_i(0)+\tau_i)}{2m-m}\right\} + 2cov(Y_i(0), Y_i(0)+\tau_i)} \\
&= \sqrt{\frac{1}{(N-1)}\left\{Var(Y_i(0)) + Var(Y_i(0)+\tau_i) + 2Var(Y_i(0)) + 2cov(Y_i(0),\tau_i)\right\}} \\
&= \sqrt{\frac{1}{(N-1)}\left\{3Var(Y_i(0)) + [Var(Y_i(0)) + Var(\tau_i) + 2cov(Y_i(0),\tau_i)] + 2cov(Y_i(0),\tau_i)\right\}} \\
&= \sqrt{\frac{1}{(2m-1)}\left\{4Var(Y_i(0)) + Var(\tau_i) + 4cov(Y_i(0),\tau_i)\right\}}
\end{aligned}
$$

All else being equal, the true standard error is smaller when the variance of the treatment effect is smaller, the variance of $Y_i(0)$ is smaller, and the covariance of the treatment effect and $Y_i(0)$ is smaller.

Question 3

Using the equation $Y_i(1) = Y_i(0) + \tau_i$, show that when we assume that treatment effects are the same for all subjects, $Var(Y_i(0)) = Var(Y_i(1))$ and the correlation between $Y_i(0)$ and $Y_i(1)$ is 1.0.[5 points]

Under constant treatment effects, $Var(Y_i(1) = Var(Y_i(0) + \tau) = Var(Y_i(0))$, and the correlation between $Y_i(1)$ and $Y_i(0)$ is:

$$\begin{aligned} cor(Y_i(1), Y_i(0)) &= \frac{Cov(Y_i(1), Y_i(0))}{\sqrt{Var(Y_i(1)) * Var(Y_i(0))}} \\ &= \frac{Cov(Y_i(0) + \tau, Y_i(0))}{\sqrt{Var(Y_i(0)) * Var(Y_i(0))}} \\ &= \frac{Var(Y_i(0))}{Var(Y_i(0))} \\ &= 1 \end{aligned}$$

Question 4

Consider the schedule of outcomes in the table below. If treatment A is administered, the potential outcome is $Y_i(A)$, and if treatment B is administered, the potential outcome is $Y_i(B)$. If no treatment is administered, the potential outcome is $Y_i(0)$. The treatment effects are defined as $Y_i(A) - Y_i(0)$ or $Y_i(B) - Y_i(0)$. [5 points]

Table 1: Question 4 Table

Subject			
Miriam	1	2	3
Benjamin	2	3	3
Helen	3	4	3
Eya	4	5	3
Billie	5	6	3

Suppose a researcher plans to assign two observations to the control group and the remaining three observations to just one of the two treatment conditions. The researcher is unsure which treatment to use.

a) Applying equation (3.4), determine which treatment, A or B, will generate a sampling distribution with a smaller standard error.
Answer:
First, notice that $Y_i(A) = Y_i(0) + 1$. Then using the results developed in the previous exercise:

$$\begin{aligned} SE(\widehat{ATE_A}) &= \sqrt{\frac{1}{5-1}\left\{\frac{3*2}{2} + \frac{2*2}{3} + 2*2\right\}} \\ &= 1.44 \\ SE(\widehat{ATE_B}) &= \sqrt{\frac{1}{5-1}\left\{\frac{3*2}{2} + \frac{2*0}{3} + 2*0\right\}} \\ &= 0.86 \end{aligned}$$

The standard error for the B vs. control comparison is smaller than the standard error for the A vs. control comparison. Thus, administering treatment B gives rise to a narrower sampling distribution.

b) What does the result in part (a) imply about the feasibility of studying interventions that attempt to close an existing "achievement gap"?
Answer:
When treatment B is administered, the achievement gap between the best and worst student narrows, leaving no variance in $Y_i(B)$. Two of the three terms in equation (3.4) therefore become zero, and the resulting standard error is much lower than it would be under treatment A, which has a constant effect across all subjects. The basic principle here is that it helps to study treatments that reduce the covariance between untreated and treated potential outcomes.

Question 5

Using Table 2.1, imagine that your experiment allocates one village to treatment. [10 points]

a) Calculate the estimated difference-in-means for all seven possible randomizations.
Answer:
There are 7 subjects, 1 of which is assigned to treatment, and thus the number of randomizations is $\frac{7!}{1!(7-1)!} = 7$. Now let's define $\widehat{ATE_i}$ as the difference in means constructed when assuming village i is assigned to treatment.

Table 2: Question 5 Table

Village	$Y_i(0)$	$Y_i(1)$	τ_i	$\widehat{ATE_i}$
1	10	15	5	$15 - \frac{15+20+20+10+15+15}{6} = -\frac{5}{6}$
2	15	15	0	$15 - \frac{10+20+20+10+15+15}{6} = 0$
3	20	30	10	$30 - \frac{10+15+20+10+15+15}{6} = \frac{95}{6}$
4	20	15	-5	$15 - \frac{10+15+20+10+15+15}{6} = \frac{5}{6}$
5	10	20	10	$20 - \frac{10+15+20+20+15+15}{6} = \frac{25}{6}$
6	15	15	0	$15 - \frac{10+15+20+20+10+15}{6} = 0$
7	15	39	15	$30 - \frac{10+15+20+20+10+15}{6} = 15$
Mean	15	20	5	$\frac{-\frac{5}{6}+0+\frac{95}{6}+\frac{5}{6}+\frac{25}{6}+0+15}{7} = 5$
SD	$\sqrt{\frac{2(10-15)^2+2(20-15)^2}{7}} = \sqrt{\frac{100}{7}}$	$\sqrt{\frac{4(15-20)^2+2(30-20)^2}{7}} = \sqrt{\frac{300}{7}}$		$\sqrt{\frac{(-\frac{5}{6}-5)^2+2(-5)^2+(\frac{95}{6}-5)^2+(\frac{5}{6}-5)^2)+(\frac{25}{6}-5)^2+(15-5)^2}{7}}$ $= 6.755$

b) Show that the average of these estimates is the true ATE.
Answer:
The table shows that the average across all randomizations is 5, which is the true ATE.

c) Show that the standard deviation of the seven estimates is identical to the standard error implied by equation (3.4).

Beginning with Equation 3.4:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{(N-1)}\left\{\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)*Var(Y_i(1))}{m} + 2cov(Y_i(0),Y_i(1))\right\}}$$
$$= \sqrt{\frac{1}{6}\left\{\frac{Var(Y_i(0))}{6} + 6Var(Y_i(1)) + 2cov(Y_i(0),Y_i(1))\right\}}$$
$$cov(Y_i(0),Y_i(1)) = \frac{(10-15)(15-20)+(20-15)(30-20)+(20-15)(15-20)}{7} = \frac{50}{7}$$
$$= \sqrt{\frac{1}{6}\left\{\frac{\frac{100}{7}}{6} + 6\frac{300}{7} + 2\frac{50}{7}\right\}}$$
$$= 6.755$$

This is identical to the standard deviation calculated in the table above.

d) Referring to equation (3.4), explain why this experimental design has more sampling variability than the design in which two villages out of seven are assigned to treatment.
Answer:
The covariance term is unaffected, but the first two variance terms are multiplied by different numbers. The first term is multiplied by 1/6 in this example as opposed to 2/5 in the 2-of-7 example. The second term is multiplied by 6/1 in this example as opposed to 5/2 in the 2-of-7 example. Because the second variance term is larger than the first, allocating more sample to the treatment group reduces sampling variance.

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{(N-1)}\left\{\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)*Var(Y_i(1))}{m} + 2cov(Y_i(0),Y_i(1))\right\}}$$
$$= \sqrt{\frac{1}{6}\left\{\frac{1}{6}\frac{100}{7} + \frac{6}{1}\frac{300}{7} + 2\frac{50}{7}\right\}} = 6.755, \text{ if } m = 1$$
$$= \sqrt{\frac{1}{6}\left\{\frac{2}{5}\frac{100}{7} + \frac{5}{2}\frac{300}{7} + 2\frac{50}{7}\right\}} = 4.603, \text{ if } m = 2$$

e) Explain why, in this example, a design in which one of seven observations is assigned to treatment has more[1] sampling variability than a design in which six villages out of seven are assigned to treatment.

[1] Text mistakenly printed "less"

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{(N-1)}\left\{\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)*Var(Y_i(1))}{m} + 2cov(Y_i(0), Y_i(1))\right\}}$$
$$= \sqrt{\frac{1}{6}\left\{\frac{1}{6}\frac{100}{7} + \frac{6}{1}\frac{300}{7} + 2\frac{50}{7}\right\}} = 6.755, \text{ if } m = 1$$
$$= \sqrt{\frac{1}{6}\left\{\frac{6}{1}\frac{100}{7} + \frac{1}{6}\frac{300}{7} + 2\frac{50}{7}\right\}} = 4.23, \text{ if } m = 6$$

By the same logic as above – allocating more units to the condition in which potential outcomes are more variable can reduce sampling variability.

Question 6

The Clingingsmith, Khwaja, and Kremer study discussed in section 3.5 may be used to test the sharp null hypothesis that winning the visa lottery for the pilgrimage to Mecca had no effect on the views of Pakistani Muslims toward people from other countries.Assume that the visa authorities conducted a complete random assignment; generate 10,000 simulated random assignments under the sharp null hypothesis. How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE? What is the implied one-tailed p-value? How many of the simulated random assignments generate an estimated ATE that is at least as large in absolute value as the actual estimate of the ATE? What is the implied two-tailed p-value? [10 points]


```
set.seed(1234567)
D <- as.numeric(hajj$success =="treatment")
Y <- hajj$views

probs <- genprobexact(D)
ate <- estate(Y,D,prob=probs)
perms <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 8.45030476380969e+285 to perform exact estimation.

Ys <- genouts(Y,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)

ate

## [1] 0.4748337

sum(distout>=ate)

## [1] 17
```

```r
sum(abs(distout)>=ate)

## [1] 34

p.value.onesided <- mean(distout>=ate)
p.value.twosided <- mean(abs(distout)>=ate)
p.value.onesided

## [1] 0.0017

p.value.twosided

## [1] 0.0034
```

The estimated ATE is 0.4748337. The number of simulated ATEs under the sharp null hypothesis of no effect that were as large was 15, corresponding to a p-value of 0.0017. The number of simulated ATEs under the sharp null hypothesis of no effect that were as large in absolute value was was 32, corresponding to a p-value of 0.0034.

Question 7

A diet and exercise program advertises that it causes everyone who is currently dieting to lose at least seven pounds more than they otherwise would have during the first two weeks. Use randomization inference (the procedure described in section 3.4) to test the hypothesis that $\tau_i = 7$ for all i. The treatment group's weight losses after two weeks are (2, 11, 14, 0, 3) and the control group's weight losses are (1, 0, 0, 4, 3). In order to test the hypothesis $\tau_i = 7$ for all i using the randomization inference methods discussed in this chapter, subtract 7 from each outcome in the treatment group so that the exercise turns into the more familiar test of the sharp null hypothesis that $\tau_i = 0$ for all i. When describing your results, remember to state the null hypothesis clearly, and explain why you chose to use a one-sided or two-sided test. [10 points]

Table 3: Question 7 Table

Subject	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - 7$
1	?	2	-5
2	?	11	4
3	?	14	7
4	?	0	-7
5	?	3	-4
6	1	?	?
7	0	?	?
8	0	?	?
9	4	?	?
10	3	?	?

```
set.seed(1234567)
D <- c(rep(0,5), rep(1, 5))
Y <- c(1,0,0,4,3,2,11,14,0,3)
Y_star <- Y + D*(-7)    # Subtracts 7 from "treatment" group

probs <- genprobexact(D)
ate <- estate(Y_star,D,prob=probs)
perms <- genperms(D,maxiter=10000)
Ys <- genouts(Y_star,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)
p.value.onesided <- mean(distout<=ate)

ate

## [1] -2.6

p.value.onesided

## [1] 0.2063492
```

There are 10 subjects, 5 of which are assigned to treatment, and thus the number of randomizations is $\frac{10!}{5!5!} = 252$. The null hypothesis is that the true ATE is a 7 pound loss; the alternative hypothesis is that the weight loss ATE is less than 7 pounds. A one-sided hypothesis test is used because we only want to reject the weight loss program's claims if the observed weight loss is less than what they claimed; if they understated the degree of weight loss, their program would be even more effective than claimed, and one would hardly fault them for that. Using the code for randomization inference posted on the website, we find that the observed difference in weight loss between the treatment and control groups (6 - 1.6 = 4.4) is smaller than 79% of all simulated experiments under the null hypothesis of a 7 pound effect for everyone. Thus, the p-value is 0.21, meaning we cannot reject the null hypothesis of a 7-pound effect at the conventional 0.05 significance threshold.

1 Question 8

Natural experiments sometimes involve what is, in effect, block random assignment. For example, Titiunik studies the effect of lotteries that determine whether state senators in Texas and Arkansas serve two-year or four-year terms in the aftermath of decennial redistricting.[2] These lotteries are conducted within each state, and so there are effectively two distinct experiments on the effects of term length. An interesting outcome variable is the number of bills (legislative proposals) that each senator introduces during a legislative session. The table below lists the number of bills introduced by senators in each state during 2003. [10 points]

a) For each state, estimate of the effect of having a two-year term on the number of bills introduced.

```
D <-  titiunik$term2year
Y <- titiunik$bills_introduced
```

[2] Titiunik 2010.

Table 4: Question 8 Table

Texas		Arkansas	
Term Length: 0 = four-year term; 1 = two-year term	# of bills introduced	Term Length: 0 = four-year term; 1 = two-year term	# of bills introduced
0	18	0	11
0	29	0	15
0	41	0	17
0	53	0	23
0	60	0	24
0	67	0	25
0	75	0	26
0	79	0	28
0	79	0	31
0	88	0	33
0	93	0	34
0	101	0	35
0	103	0	35
0	106	0	36
0	107	0	38
0	131	0	52
1	29	0	59
1	37	1	9
1	42	1	10
1	45	1	14
1	45	1	15
1	54	1	15
1	54	1	17
1	58	1	18
1	61	1	19
1	64	1	19
1	69	1	20
1	73	1	21
1	75	1	23
1	92	1	23
1	104	1	24
		1	28
		1	30
		1	32
		1	34

```
block <- titiunik$texas0_arkansas1

ate_texas <- mean(Y[D==1 & block==0]) - mean(Y[D==0 & block==0])
ate_arkansas <- mean(Y[D==1 & block==1]) - mean(Y[D==0 & block==1])
ate_texas

## [1] -16.74167

ate_arkansas

## [1] -10.09477
```

The estimated ATE in Texas is -16.742. In Arkansas, the estimated ATE is -10.095.

b) For each state, estimate the standard error of the estimated ATE.

```
se_texas = sqrt(var(Y[D==0 & block==0])/length(Y[D==0 & block==0]) +
                var(Y[D==1 & block==0])/length(Y[D==1 & block==0]))

se_arkansas = sqrt(var(Y[D==0 & block==1])/length(Y[D==0 & block==1]) +
                   var(Y[D==1 & block==1])/length(Y[D==1 & block==1]))
se_texas

## [1] 9.345871

se_arkansas

## [1] 3.395979
```

The estimated se in Texas is 9.346. In Arkansas, the estimated se is 3.396.

c) Use equation (3.10) to estimate the overall ATE for both states combined.

```
ate_overall <- length(Y[block==0])/length(Y) *ate_texas +
               length(Y[block==1])/length(Y) *ate_arkansas
ate_overall

## [1] -13.2168
```

The overall ATE, -13.217 is the weighted average of the two separate ATEs, where the weights are the shares of overall N in each state.

d) Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimates of the overall ATE.
Answer:

The two states differ in terms of the probability that a given legislator will be assigned to the treatment. Therefore, we cannot pool the data without introducing a correlation between treatment assignment and the potential outcomes associated with the two states. In this study, the experiments take place within each state, and the analyst should pool the state-level results in order to obtain an overall result.

e) Insert the estimated standard errors into equation (3.12) to estimate the standard error for the overall ATE.

```
se_overall= sqrt((length(Y[block==0])/length(Y))^2 *se_texas^2 +
                 (length(Y[block==1])/length(Y))^2 *se_arkansas^2)
se_overall

## [1] 4.74478
```

The overall standard error is (4.745).

f) Use randomization inference to test the sharp null hypothesis that the treatment effect is zero for senators in both states.

```
probs <- genprobexact(D,blockvar=block) # Note differential probabilities
ate <- estate(Y,D,prob=probs)
perms <- genperms(D,maxiter=10000,blockvar=block) # Note blocked randomization

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 1363721466356691712 to perform exact estimation.

Ys <- genouts(Y,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)
p.value.twosided <- mean(abs(distout) >= abs(ate))
ate

## [1] -13.2168

p.value.twosided

## [1] 0.0071
```

Here, we use a two-tailed test because it is not clear theoretically whether longer or shorter terms should make legislators more responsive. Comparing the observed difference-in-means to the distribution of 10,000 simulated randomizations under the sharp null hypothesis reveals a two-tailed p-value of 0.0071, leading us to reject the null hypothesis.

Question 9

Camerer reports the results of an experiment in which he tests whether large, early bets placed at horse tracks affect the betting behavior of other bettors.[3] Selecting pairs of long-shot horses running in the same race whose betting odds were approximately the same when betting opened, he placed two $500 bets on one of the two horses approximately 15 minutes before the start of the race. Because odds are determined based on the proportion of total bets placed on each horse, this intervention causes the betting odds for the treatment horse to decline and the betting odds of the control horse to rise. Because Camerer's bets were placed early, when the total betting pool was small, his bets caused marked changes in the odds presented to other bettors. (A few minutes before each race started, Camerer canceled his bets.) While the experimental bets were still "live," were other bettors attracted to the treatment horse (because other bettors seemed to believe in the horse) or repelled by it (because the diminished odds meant a lower return for each wager)? Seventeen pairs of horses in this study are listed below. The outcome measure is the number of dollars that were placed on each horse (not counting Camerer's own wagers on the treatment horses) during the test period, which begins 16 minutes before each race (roughly 2 minutes before Camerer began placing his bets) and ends 5 minutes before each race (roughly 2 minutes before Camerer withdrew his bets). [10 points]

Table 5: Question 9 Table

	Treatment Horse in Pair			Control Horse in Pair			
	Total bets $T-16$ min	Total bets $T-5$ min	Change	Total bets $T-16$ min	Total bets $T-5$ min	Change	Difference in changes
Pair 1	533	1503	970	587	2617	2030	-1060
Pair 2	376	1186	810	345	1106	761	49
Pair 3	576	1366	790	653	2413	1760	-970
Pair 4	1135	1666	531	1296	2260	964	-433
Pair 5	158	367	209	201	574	373	-164
Pair 6	282	542	260	269	489	220	40
Pair 7	909	1597	688	775	1825	1050	-362
Pair 8	566	933	367	629	1178	549	-182
Pair 9	0	555	555	0	355	355	200
Pair 10	330	786	456	233	842	609	-153
Pair 11	74	959	885	130	256	126	759
Pair 12	138	319	181	179	356	177	4
Pair 13	347	812	465	382	604	222	243
Pair 14	169	329	160	165	355	190	-30
Pair 15	41	297	256	33	75	42	214
Pair 16	37	71	34	33	121	88	-54
Pair 17	261	485	224	282	480	198	26

a) One interesting feature of this study is that each pair of horses ran in the same race. Does this design feature violate the non-interference assumption, or can potential outcomes be defined so that the non-interference assumption is satisfied?

[3] Camerer 1998. This example draws on the second of Camerer's studies and restricts the sample to cases in which a treatment horse is compared to a single control horse.

Answer:
This design feature violates non-interference if the estimand is defined as the difference between the following two potential outcomes: total bets on a given horse when experimental bets are placed on that horse versus no experimental bets on any horse in the race. One could avoid violating non-interference by redefining the estimand as the difference between the following two potential outcomes: total bets on a horse when experimental bets are placed on that horse versus experimental bets are placed on a competing horse in the same race.

b) A researcher interested in conducting a randomization check might assess whether, as expected, treatment and control horses attract similarly sized bets prior to the experimental intervention. Use randomization inference to test the sharp null hypothesis that the bets had no effect prior to being placed.

```
D <-  camerer$treatment
block <- camerer$pair
covs <- as.matrix(camerer$preexperimentbets)

probs <- genprobexact(D,blockvar=block)
perms <- genperms(D,maxiter=10000,blockvar=block)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 131072 to perform exact estimation.

numiter <- ncol(perms)

Fstat <- summary(lm(D~covs))$fstatistic[1]
Fstatstore <- rep(NA,numiter)

for (i in 1:numiter) {
  Fstatstore[i] <- summary(lm(perms[,i]~covs))$fstatistic[1]
        }

p.value <- mean(Fstatstore >= Fstat)
p.value

## [1] 0.3696
```

We conducted 10,000 random assignments, and for each we calculated the F-statistic of a regression of treatment assignment on pre-experimental bets (controlling for blocks). The observed F-statistic for the actual experiment is larger than 3696 of the simulated experiments, implying a p-value of 0.37.

c) Calculate the average increase in bets during the experimental period for treatment horses and control horses. Compare treatment and control means, and interpret the estimated ATE.

```
change <- camerer$change
change_treatment <- mean(change[D==1])
change_control <- mean(change[D==0])
ATE <- change_treatment - change_control
ATE

## [1] -110.1765
```

The average treatment group change was \$461.24, as opposed to an average change of \$571.41 in the control group. Therefore, the estimated ATE is \$−110.18.

d) Show that the estimated ATE is the same when you subtract the control group outcome from the treatment group outcome for each pair and calculate the average difference for the 17 pairs. Answer:

```
pair_diffs <- rep(NA, 17)

for (i in 1:17){
pair_diffs[i] <- diff(change[block==i])
}

mean(pair_diffs)

## [1] 110.1765
```

The average difference between treatment and control outcomes for each pair is also 110.18.

e) Use randomization inference to test the sharp null hypothesis of no treatment effect for any subject. When setting up the test, remember to construct the simulation to account for the fact that random assignment takes place within each pair. Interpret the results of your hypothesis test and explain why a two-tailed test is appropriate in this application.

```
set.seed(1234567)
probs <- genprobexact(D,blockvar=block) # Notice the blocks
ate <- estate(change,D,prob=probs)
perms <- genperms(D,maxiter=10000,blockvar=block)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 131072 to perform exact estimation.

Ys <- genouts(change,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)

ate
```

```
## [1] -110.1765

p.value <- mean(abs(distout) >= abs(ate))
p.value

## [1] 0.3092
```

A two-tailed test generates a p-value of 0.3092, indicating that one cannot reject the sharp null of no effect for any unit. A two-tailed test is appropriate because some theories predict a positive effect while others predict a negative effect: "were other bettors attracted to the treatment horse (because other bettors seemed to believe in the horse) or repelled by it (because the diminished odds meant a lower return for each wager)?" The appropriate null hypothesis in this case is no effect, which would be rejected if we observed either strongly positive or strongly negative differences between treatment and control horses.

Question 10

Suppose that 800 individual students were randomly assigned to classrooms of 25 students apiece, and these classrooms were then randomly assigned as clusters to treatment and control. Assume the non-interference assumption holds. Use equations (3.4) and (3.22) to explain why this clustered design has the same standard error as complete random assignment of individual students to treatment and control. [10 points] Answer:

The equation for the standard error under individual assignment:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{(N-1)}\left\{\frac{mVar(Y_i(0))}{N-m} + \frac{mVar(Y_i(1))}{N-m} + 2cov(Y_i(0), Y_i(1))\right\}}$$

The equation for the standard error under clustered assignment with equal-size clusters:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{(k-1)}\left\{\frac{mVar(\bar{Y}_j(0))}{N-m} + \frac{mVar(\bar{Y}_j(1))}{N-m} + 2cov(\bar{Y}_j(0), \bar{Y}_j(1))\right\}}$$

When the clusters are formed randomly (i.e., individuals are randomly allocated to clusters prior to assignment), the two formulas give approximately the same answer. In order to see the correspondence, notice that the variance of the average treated outcome from random draw of 25 students is $Var(\bar{Y}_j(0)) = \frac{Var(Y_i(0))}{25}$, and similarly, $Var(\bar{Y}_j(1)) = \frac{Var(Y_i(1))}{25}$, and $cov(\bar{Y}_j(0), \bar{Y}_j(1)) = \frac{cov(Y_i(0), Y_i(1))}{25}$. Thus, the quantity inside the braces in both equations differs by a factor of 25, which is approximately $\frac{N-1}{k-1}$.

Question 11

Use the data in Table 3.3 to simulate cluster randomized assignment. [10 points]

a) Suppose that clusters are formed by grouping observations $\{1,2\}, \{3,4\}, \{5,6\} \ldots \{13,14\}$. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to the treatment.

```r
Y0 <- c(0,1,2,4,4,6,6,9,14,15,16,16,17,18)
Y1 <- c(0,0,1,2,0,0,2,3,12,9,8,15,5,17)
cluster <- rep(1:7, each=2)
Ybar0 <- tapply(X=Y0, INDEX=cluster, FUN=mean)
Ybar1 <- tapply(X=Y1, INDEX=cluster, FUN=mean)

var.pop <- function(x){sum((x-mean(x))^2)/(length(x))}
cov.pop <- function(x,y){sum((x-mean(x))*(y-mean(y)))/(length(x))}

var_Ybar0 <- var.pop(Ybar0)
var_Ybar1 <- var.pop(Ybar1)
cov_Ybar0 <- cov.pop(Ybar0,Ybar1)

se_ate <- sqrt((1/6) * ((4/3)*var_Ybar0 + (3/4)*var_Ybar1 + 2*cov_Ybar0))
se_ate

## [1] 4.706192
```

Assuming that 4 out of 7 clusters are assigned to treatment, the standard error of the ATE will be 4.71.

b) Suppose that clusters are instead formed by grouping observations $\{1, 14\}, \{2, 13\}, \{3, 12\} \ldots \{7, 8\}$. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to the treatment.


```r
cluster <- c(1,2,3,4,5,6,7,7,6,5,4,3,2,1)
Ybar0 <- tapply(X=Y0, INDEX=cluster, FUN=mean)
Ybar1 <- tapply(X=Y1, INDEX=cluster, FUN=mean)

var_Ybar0 <- var.pop(Ybar0)
var_Ybar1 <- var.pop(Ybar1)
cov_Ybar0 <- cov.pop(Ybar0,Ybar1)

se_ate <- sqrt((1/6) * ((4/3)*var_Ybar0 + (3/4)*var_Ybar1 + 2*cov_Ybar0))
se_ate

## [1] 0.9766259
```

Assuming that 4 out of 7 clusters are assigned to treatment, the standard error of the ATE will be 0.98.

c) Why do the two methods of forming clusters lead to different standard errors? What are the implications for the design of cluster randomized experiments?
Answer:
The first method clusters the most similar villages together, and the second method clusters the most dissimilar villages together. As a result, the variances of the average within-cluster potential outcomes are much larger in the first method and smaller in the second. As a result, the

second method produces a much narrower standard error of the ATE estimate. The implication for clustered design is that the more similar the observations with a cluster, the less precise the estimates we can produce. When possible, cluster heterogeneous observations together.

Question 12

Below is a schedule of potential outcomes for six classrooms, which are located in three schools. Using a cluster randomized design, researchers will assign one of the three schools (and all the classrooms it contains) to the treatment group. [5 points]

Table 6: Question 12 Table

School	Classroom	$Y_i(0)$	$Y_i(1)$
A	A-1	0	0
B	B-1	0	1
B	B-2	0	1
C	C-1	0	2
C	C-2	0	2
C	C-3	0	2

a) What is the average treatment effect among the six classrooms?

$$\frac{2+2+2+1+1+0}{6} = 1.333$$

b) There are three possible randomizations. Is the difference-in-means estimator unbiased?
Answer:
The estimated ATE is 0 if school A is assigned to treatment, 1 if school B is assigned to treatment, and 2 if school C is assigned to treatment. So if we take the average of three estimates the ATE is $\frac{0+1+2}{3} = 1 \neq 1.33$ and is therefore biased. When potential outcomes are related to cluster size, cluster randomization is prone to bias in small samples, as in this case. This condition holds in this case: the biggest cluster, Cluster C, has larger than average $Y(1)$ values.

c) In general, cluster random assignment generates biased results when (i) clusters vary in size, (ii) potential outcomes vary by cluster, and (iii) the number of clusters is too small to ensure that m of N units are placed into the treatment condition in each randomization. Show what happens in this example when School A and School B are combined for purposes of random assignment, so that there is a 0.5 probability that either School C is placed in treatment or Schools A and B are placed in treatment. Does this design yield unbiased estimates? What are the implications of this exercise for the design of cluster randomized experiments?
Answer:
If A and B are combined and put into treatment, the estimated ATE is 2/3; if C is treated, the estimated ATE is 2. Therefore, the average estimated ATE is $\frac{2/3+2}{2} = 1.33$, which is the true ATE. Therefore, combining clusters to make cluster size constant eliminates bias. The implication is that bias can be avoided by constructing clusters of equal size.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 4 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts:

a) Define "covariate." Explain why covariates are (at least in principle) measured prior to the random allocation of subjects to treatment and control.
Answer:
A covariate is a variable that is (1) unaffected by the treatment and (2) used to predict outcomes. In order to increase the credibility of the claim that a given covariate is unaffected by the treatment, researchers typically restrict the set of covariates to those variables that are measured (or are measurable) prior to the random allocation of treatments.

b) Define "disturbance term."
Answer:
The disturbance term comprises all sources of variation in potential outcomes other than the average treatment effect. For example, in equation (4.7), the disturbance term is $u_i = Y_i(0) - \mu_{Y(0)} + [(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i$. The disturbance term comprises the idiosyncratic variation in untreated responses $Y_i(0) - \mu_{Y(0)}$, plus the idiosyncratic variation in treatment effects $[(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i$.

c) In equation (4.2), we demonstrated that rescaling the outcome by subtracting a pre-test leads to unbiased estimates of the ATE. Suppose that instead of subtracting the pre-test X_i, we subtracted a rescaled pretest cX_i, where c is some positive constant. Show that this procedure produces unbiased estimates of the ATE.
Answer:
The proof is similar to equation (4.2) and again makes use of the fact that the expected value of X_i is the same in the treatment and control groups when treatments are allocated randomly:

$$\begin{aligned}
E[\widehat{ATE}] &= E[Y_i - cX_i|D_i = 1] - E[Y_i - cX_i|D_i = 0)] \\
&= E[Y_i|D_i = 1] - E[cX_i|D_i = 1] - E[Y_i|D_i = 0] + E[cX_i|D_i = 0] \\
&= E[Y_i|D_i = 1] - cE[X_i|D_i = 1] - E[Y_i|D_i = 0] + cE[X_i|D_i = 0] \\
&= E[Y_i(1)] - E[Y_i(0)]
\end{aligned}$$

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

d) Show that the parameter b in equation (4.7) is identical to the ATE.
Answer:
Recall from Equation (4.7) that:

$$\begin{aligned} Y_i &= Y_i(0)(1 - D_i) + Y_i(1)D_i \\ &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\ &= \mu_{Y(0)} + [\mu_{Y(1)} - \mu_{Y(0)}]D_i + Y_i(0) - \mu_{Y(0)} + [(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i \\ &= a + bD_i + u_i \end{aligned}$$

This equation implies that $b = \mu_{Y(1)} - \mu_{Y(0)}$, which is the ATE because the expected value of $Y_i(1)$ is $\mu_{Y(1)}$, and the expected value of $Y_i(0)$ is $\mu_{Y(0)}$.

Question 2

A researcher working with Israeli elementary school students sought to improve students' ability to solve logic puzzles.[1] Students in the treatment and control group initially took a computer-administered test, and the number of correctly solved puzzles was recorded. A few days later, students assigned to the control group were then given 30 minutes to improve their puzzle-solving skills by playing on a computer. During the same allotment of time, students in the treatment group listened to an instructor describe some rules of thumb to keep in mind when solving logic puzzles. All subjects then took a computer-administered post-test, and the number of correctly solved puzzles was recorded. The table below shows the results for each subject.

Table 1: Question 2 Table

Subject	D	Pre-test	Post-test	Improvement
1	1	10	10	0
2	1	9	11	2
3	1	5	6	1
4	1	3	6	3
5	1	3	6	3
6	1	6	7	1
7	1	6	7	1
8	1	5	6	1
9	1	6	7	1
10	0	9	9	0
11	0	6	7	1
12	0	11	10	-1
13	0	4	5	1
14	0	3	3	0
15	0	10	10	0
16	0	7	8	1
17	0	7	7	0
18	0	8	10	2

[1] Dan Gendelman conducted this study in 2004 and shared it with us via personal communication.

a) As a randomization check, use randomization inference to test the null hypothesis that the pre-test scores are unaffected by treatment assignment.

```
D <- rush$treat
Y <- rush$posttest
X <- rush$pretest

# RI: randomization check, testing the effect of the covariate on the treatment

perms <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 48620 to perform exact estimation.

numiter <- ncol(perms)
Fstat <- summary(lm(D~X))$fstatistic[1]

Fstatstore <- rep(NA,numiter)

for (i in 1:numiter) {
  Fstatstore[i] <- summary(lm(perms[,i]~X))$fstatistic[1]
        }

p.value <- mean(Fstatstore >= Fstat)
p.value

## [1] 0.2627
```

We calculated the F-statistic of a regression of treatment assignment on the pretest score for all possible randomizations, and found that the observed F-statistic was larger than 26.27% of the simulated statistics, implying a p-value of 0.263. As expected,we fail to reject the null hypothesis that the treatment assignment is unrelated to the pretreatment covariate, pretest.

b) Use difference-in-means estimation to estimate the effect of the treatment on the post-test score. Form a 95% confidence interval.

```
probs <- genprobexact(D)
ate <- estate(Y,D,prob=probs)
ate

## [1] -0.3333

Ys <- genouts(Y,D,ate=ate)
distout <- gendist(Ys,perms,prob=probs)
ci.95 <- quantile(distout,probs=c(0.025, 0.975))
ci.95
```

```
##   2.5%  97.5%
## -2.333  1.593
```

We obtained a difference-in-means estimate of the ATE of -0.3333333 and a 95% confidence interval of $[-2.33, 1.59]$. This confidence interval is wide enough to include much larger and much smaller treatment effects – even crossing zero.

c) Use difference-in-differences estimation to estimate the effect of the treatment on the post-test score. Form a 95% confidence interval, and compare it to the interval in part (b).

```
Y.improve <- rush$improvement
ate.improve <- estate(Y.improve,D,prob=probs)
ate.improve

## [1] 1

Ys <- genouts(Y.improve,D,ate=ate.improve)
distout <- gendist(Ys,perms,prob=probs)
ci.95.improve <- quantile(distout,probs=c(0.025, 0.975))
ci.95.improve

##   2.5%  97.5%
## 0.1111 1.8889
```

By subtracting a pre-test, we have sharpened our estimates. The difference-in-differences estimate of the ATE is 1 and the 95% confidence interval is [0.11, 1.89]. No longer does the 95% confidence interval cross zero, meaning we can be confidence at the 95% level that the estimated ATE is larger than zero. This contrasts with part b) where the background variability in test scores made the estimation of a small treatment effect more difficult.

Question 3

The table below illustrates the problems that may arise when researchers exercise discretion over what results to report to readers. Suppose the true ATE associated with a given treatment were 1.0. The table reports the estimated ATE from nine experiments, each of which involves approximately 200 subjects. Each study produces two estimates, one based on a difference-in-means and another using regression to control for covariates. In principle, both estimators generate unbiased estimates, and covariate adjustment has a slight edge in terms of precision. Suppose the researchers conducting each study use the following decision rule: "Estimate the ATE using both estimators and report whichever estimate is larger." Under this reporting policy, are the reported estimates unbiased? Why or why not?

Answer:

This procedure leads to biased estimates. Although each estimator is unbiased, the greater of two unbiased estimates is not unbiased. One can think of this procedure as "Report the no-covariates estimate unless the with-covariates estimate is larger, in which case report the with-covariates estimate." On its own, the no-covariates estimate is unbiased, but it tends to be corrected when it

generates a lower-than-average estimate. In this example, the average estimate generated by this reporting procedure is $12/9 = 1.33$, which is greater than the true ATE of 1.0.

Table 2: Question 3 table

Study	No covariates	With covariates	Greater of two estimates
1	5	4	5
2	3	3	3
3	2	2	2
4	6	5	6
5	1	1	1
6	0	0	0
7	-3	-1	-1
8	-5	-4	-4
9	0	-1	0
Average	1	1	1.33
Standard Deviation	3.54	2.83	3.08

Question 4

Table 4.1 contains a column of treatment assignments, D_i, that reflects a complete random assignment of 20 schools to treatment and 20 schools to control.

a) Use equation (2.2) to generate observed outcomes based on these assigned treatments. Regress Y_i on D_i and interpret the slope and intercept. Is the estimated slope the same as the estimated ATE based on a difference-in-means?

```
D  <- teach$D
Y1 <- teach$y1
Y0 <- teach$y0
X <- teach$x

Y <- Y0*(1-D) + Y1*(D)     # Equation 2.2

fit <- lm(Y~D)
arm::display(fit)

## lm(formula = Y ~ D)
##             coef.est coef.se
## (Intercept) 26.85     3.33
## D           10.70     4.71
## ---
## n = 40, k = 2
## residual sd = 14.89, R-Squared = 0.12

diff_means <- mean(Y[D==1])-mean(Y[D==0])
diff_means
```

```
## [1] 10.7
```

The estimate obtained with OLS regression (10.7) is identical to the estimate obtained with difference-in-means (10.7).

b) Regress treated and untreated outcomes on X_i to see whether the condition in equation (4.6) appears to hold. What do you infer about the advisability of rescaling the dependent variable so that the outcome is a change (i.e., $Y_i - X_i$)?

```
fit.1 <- lm(Y~X,subset=D==1)
fit.0 <- lm(Y~X,subset=D==0)

sum_of_coefficients <- fit.1$coefficients[2] + fit.0$coefficients[2]
sum_of_coefficients

##     X
## 1.846

Ydiff <- Y-X
fit.diff <- lm(Ydiff~D)
arm::display(fit.diff)

## lm(formula = Ydiff ~ D)
##             coef.est coef.se
## (Intercept) -1.00     1.15
## D            4.85     1.62
## ---
## n = 40, k = 2
## residual sd = 5.13, R-Squared = 0.19
```

Substituting regression estimates for the true ratio of covariances to variances satisfies the inequality, suggesting that the use of this covariate will improve precision.

$$\frac{Cov\widehat{(Y_i(0), X_i)}}{Var(X_i)} + \frac{Cov\widehat{(Y_i(1), X_i)}}{Var(X_i)} = 0.8995221 + 0.9465386 = 1.8460607$$

We also see that the standard errors have shrunk substantially – the standard error for a regression of Y on D is 4.7093133, whereas the standard error for the regression of the change in Y on D is 1.6226603

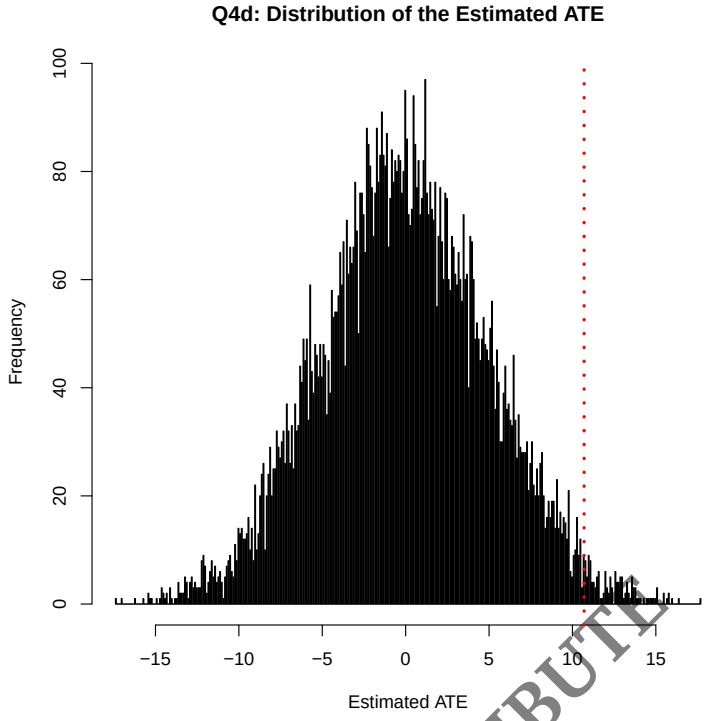
c) Regress Y_i on D_i and X_i. Interpret the regression coefficients, contrasting these results with those obtained from a regression of Y_i on D_i alone.

```
fit.cov <- lm(Y~D+X)
arm::display(fit.cov)

## lm(formula = Y ~ D + X)
##             coef.est coef.se
## (Intercept) 1.22     1.88
## D           5.32     1.63
## X           0.92     0.05
## ---
## n = 40, k = 3
## residual sd = 5.05, R-Squared = 0.90
```

The estimated ATE (5.32) is now roughly half the size as the original difference-in-means. (This estimate also happens to be much closer to the true ATE of 4.0.) Comparing the estimated standard errors from both regressions suggests that the inclusion of a covariate has greatly improved precision.

d) With the estimates obtained in part (a), use randomization inference (as described in Chapter 3) to evaluate the sharp null hypothesis of no effect for any school. To obtain the sampling distribution under the sharp null hypothesis, simulate 100,000 random assignments, and for each simulated sample, estimate the ATE using a regression of Y_i on d_i. Interpret the results.


```
perms <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 137846528820 to perform exact estimation.

probs <- genprobexact(D)
ate <- estate(Y,D,prob=probs)
Ys <- genouts(Y,D,ate=0)

distout <- gendist(Ys,perms,prob=probs)
p.value <- mean(abs(distout)>abs(ate))

ate

## [1] 10.7

p.value

## [1] 0.0291

hist(distout, breaks=1000,
     main="Q4d: Distribution of the Estimated ATE",
     xlab="Estimated ATE")
abline(v=ate, col="red", lty=3, lwd=3)
```



We use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a two-tailed p-value of 0.029, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some positive effect.

e) Using the estimator in part (c), use randomization inference to evaluate the sharp null hypothesis of no effect for any school. To obtain the sampling distribution under the sharp null hypothesis, simulate 100,000 random assignments, and for each simulated sample, estimate the ATE using a regression of Y_i on D_i and X_i. Interpret the results.

```
ate_cov <- estate(Y,D,X,prob=probs)
distout_cov <- gendist(Ys,perms,X,prob=probs)
p.value_cov <- mean(abs(distout_cov)>abs(ate_cov))
ate_cov

##     Z
## 5.316

p.value_cov

## [1] 0.0026
```

We again use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a two-tailed p-value of 0.003, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

f) Use the estimated ATE in part (a) to construct a full schedule of potential outcomes for all schools, assuming that every school has the same treatment effect. Using this simulated schedule

of potential outcomes, construct a 95% confidence interval for the sample average treatment effect in the following way. First, assign each subject to treatment or control, and estimate the ATE by a regression of Y_i on D_i. Repeat this procedure until you have 100,000 estimates of the ATE. Order the estimates from smallest to largest. The 2,501st estimate marks the 2.5th percentile, and the 97,500th estimate marks the 97.5th percentile. Interpret the results.

```
Ys <- genouts(Y,D,ate=ate)
distout <- gendist(Ys,perms,prob=probs)
ci.95 <- quantile(distout, probs=c(0.025, .975))
ci.95

##  2.5% 97.5%
##  1.53 19.84
```

The confidence interval stretches from [1.53, 19.84] implying that the ATE is positive but its location is subject to a great deal of statistical uncertainty. Our best guess is 10.7, but the interval ranges from a small positive value to a truly massive effect.

g) Use the estimated ATE in part (c) to construct a full schedule of potential outcomes for all schools, assuming that every school has the same treatment effect. Using this simulated schedule of potential outcomes, simulate the 95% confidence interval for the sample average treatment effect estimated by a regression of Y_i on D_i and X_i. Interpret the results. Is this confidence interval narrower than one you generated in question (f)?

```
Ys_cov <- genouts(Y,D,ate=ate_cov)
distout_cov <- gendist(Ys_cov,perms,X,prob=probs)
ci.95_cov <- quantile(distout_cov, probs=c(0.025, .975))
ci.95_cov

##  2.5% 97.5%
## 2.225 8.442
```

The confidence interval now stretches from [2.23, 8.44]. Interestingly, this interval no longer contains the estimate obtained without controls for covariates. Our best guess is now 5.32, and our 95% interval is now roughly one-third as wide as before.

Question 5

Randomizations are said to be "restricted" when the set of all possible random allocations is narrowed to exclude allocations that have inadequate covariate balance. Suppose, for example, that the assignment of treatments (D_i) in Table 4.1 was conducted subject to the restriction that a regression of D_i on X_i (the pretest) does not allow the researcher to reject the sharp null hypothesis of no effect of X_i on D_i at the 0.05 significance level) produces a p-value on that is greater than 0.05. In other words, had the researcher found that the assigned D_i were significantly predicted by X_i, the random allocation would have been conducted again, until the D_i met this criterion.

a) Conduct a series of random assignments in order to calculate the weighting variable w_i; for units in the treatment group, this weight is defined as the inverse of the probability of being assigned

to treatment, and for units in the control group, this weight is defined as the inverse of the probability of being assigned to control. See Table 4.2 for an example. Does w_i appear to vary within the treatment group or within the control group?

```
D  <- teach$D
Y1 <- teach$y1
Y0 <- teach$y0
X <- teach$x

Y <- Y0*(1-D) + Y1*(D)
N <- length(D)


randfun <- function() {
  teststat <- -1
  while (teststat < 0.05) {
                Zri <- sample(D)
                teststat <- summary(lm(Zri~X))$coefficients[2,4]
        }
        return(Zri)
}

# notice the use of the restricted randomization function.
# restricted randomization often generates unequal probabilities of assignment.
# if so, inverse probability weighting is required.

perms <- genperms.custom(numiter=10000,randfun=randfun)
probs <- genprob(perms)
weights <- (1/probs) *D + (1/(1-probs))*(1-D)
var.weights.treat <- var(weights[D==1])
var.weights.control <- var(weights[D==0])
```

The variance of the weights is 4×10^{-4} in the treatment condition and 6×10^{-4} in the control condition. Indeed, units do have different probabilities of assignments as a result of the restriction scheme, but the differences are small.

b) Use randomization inference to test the sharp null hypothesis that D_i has no effect on Y_i by regressing Y_i on D_i and comparing the estimate to the sampling distribution under the null hypothesis. Make sure that your sampling distribution includes only random allocations that satisfy the restriction mentioned above. Be sure to weight units by inverse probability weights as produced by the random allocation procedure. Estimate the ATE, calculate the p-value, and interpret the results.

```
ate <- estate(Y,D,prob=probs)
Ys <- genouts(Y,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)
p.value <- mean(abs(distout) > abs(ate))
ate
```

```
## [1] 10.73

p.value

## [1] 0.0054
```

The IPW estimate of the ATE is 10.73, which is close to the unweighted estimate above. Using a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject, we find a p-value of 0.005, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

c) Use randomization inference to test the sharp null hypothesis that D_i has no effect on Y_i by regressing Y_i on D_i and X_i and comparing the estimate to the sampling distribution under the null hypothesis. Estimate the ATE, calculate the p-value, and interpret the results.

```
perms <- genperms.custom(numiter=10000,randfun=randfun)
probs <- genprob(perms)
ate_cov <- estate(Y,D,X,prob=probs)
Ys <- genouts(Y,D,ate=0)
distout_cov <- gendist(Ys,perms,X,prob=probs)
p.value_cov <- mean(abs(distout_cov) > abs(ate_cov))
ate_cov

##     Z
## 5.346

p.value_cov

## [1] 0.0017
```

The IPW estimate of the ATE is 5.35, which is close to the unweighted estimate above. We again use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a p-value of 0.002, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

d) Compare the sampling distributions under the null hypothesis in parts (a) and (b) to the sampling distributions obtained in exercises 4(d) and 4(e), which assumed that the randomization was unrestricted.

```
## Sampling Distributions from 4(d) and 4(e)
perms_complete_RA <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 137846528820 to perform exact estimation.
```

```
probs_complete_RA <- genprobexact(D)

ate_complete_RA <- estate(Y,D,prob=probs_complete_RA)
Ys_complete_RA <- genouts(Y,D,ate=ate_complete_RA)
distout_complete_RA <- gendist(Ys_complete_RA,perms_complete_RA,
                               prob=probs_complete_RA)
se_complete_RA <- sd(distout_complete_RA)
se_complete_RA

## [1] 4.601

ate_cov_complete_RA <- estate(Y,D,X,prob=probs_complete_RA)
Ys_cov_complete_RA <- genouts(Y,D,ate=ate_cov_complete_RA)
distout_cov_complete_RA <- gendist(Ys_cov_complete_RA,perms_complete_RA,X,
                                   prob=probs_complete_RA)
se_cov_complete_RA <- sd(distout_cov_complete_RA)
se_cov_complete_RA

## [1] 1.593

## Sampling Distributions from 5(a) and 5(b)
perms_restricted_RA <- genperms.custom(numiter=10000,randfun=randfun)
probs_restricted_RA <- genprob(perms_restricted_RA)

ate_restricted_RA <- estate(Y,D,prob=probs_restricted_RA)
Ys_restricted_RA <- genouts(Y,D,ate=ate_restricted_RA)
distout_restricted_RA <- gendist(Ys_restricted_RA,perms_restricted_RA,
                                 prob=probs_restricted_RA)
se_restricted_RA <- sd(distout_restricted_RA)
se_restricted_RA

## [1] 4.199

ate_cov_restricted_RA <- estate(Y,D,X,prob=probs_restricted_RA)
Ys_cov_restricted_RA <- genouts(Y,D,ate=ate_cov_restricted_RA)
distout_cov_restricted_RA <- gendist(Ys_cov_restricted_RA,perms_restricted_RA,X,
                                     prob=probs_restricted_RA)
se_cov_restricted_RA <- sd(distout_cov_restricted_RA)
se_cov_restricted_RA

## [1] 1.607
```

Table 3: Summary of Estimated Standard Errors

	Without Covariates	With Covariates
Complete Random Assignment	4.601	1.593
Restricted Random Assignment	4.199	1.607

Without covariates and assuming complete randomization, we obtain a standard error of 4.601. Under restricted randomization, the standard error declines to 4.199. Including a covariate and assuming complete randomization, we obtain a standard error of 1.593. Under restricted randomization, the standard error remains essentially unchanged at 1.607. Restricted randomization is akin to blocking, in that it rules out random allocations that result in imbalance; however, its advantages in terms of precision are limited when the researcher controls for a strongly prognostic covariate, which achieves most of the precision gains associated with blocking.

Question 6

One way to practice your experimental design skills is to undertake a mock randomization of an existing non-experimental dataset. In this exercise, the existing dataset is treated as though it were a baseline data collection effort that an experimental researcher gathered in preparation for a random intervention. The actual data in question come from a panel study of Russian villagers. Villagers from randomly selected rural areas of Russia were interviewed in 1995 and re-interviewed in 1996 and 1997. Our attention focuses on the 462 respondents who were interviewed in all three waves and provided answers to questions about their income, church membership, and evaluation of national conditions (i.e., how well are things going in Russia?). Imagine that an experimental intervention occurred after the 1996 survey and that national evaluations in the 1997 survey were the experimental outcome of interest. The dataset provided at isps.research.yale.due/FEDAI contains the following pre-treatment covariates that may be used for blocking: sex, church membership, social class, and evaluations of national conditions in 1995 and 1996. As you design your experiment, imagine that "post-intervention" evaluations of national conditions in 1997 were unknown.

a) One way to develop a sense of which variables are likely to predict post-intervention evaluations of national conditions in 1997 is to regress evaluations of national conditions in 1996 on sex, church membership, social class, and evaluations in 1995. Which of these variables seem to most strongly predict evaluations of national conditions in 1996? What is the R^2 from this regression?

```
russia <- within(russia,{
  female <- as.numeric(sexresp6 == "woman")
  class <- relevel(group6,ref="very poor")
  church_member <- as.numeric(memberc6=="yes")
  id <- 1:nrow(russia)
  class_verypoor <- as.numeric(class=="very poor")
  class_poor <- as.numeric(class=="poor")
  class_middle <- as.numeric(class=="middle")
  class_morethanmiddle <- as.numeric(class=="more than middle")
})
```

```
fit <- lm(index96 ~ index95 + female + church_member + class, data=russia)
summary(fit)$r.squared

## [1] 0.3937

fit.nolag <- lm(index96 ~ female + church_member + class, data=russia)
summary(fit.nolag)$r.squared

## [1] 0.02828
```

The regression treats "index95" as a continuous variable and all others as categorical. the R-squared is 0.394, which implies that the regressors predict about 40% of the variance in "index96". The strongest predictor is 95, the lagged dependent variable. Had we omitted this variable from the model, the R-squared would have fallen to 0.028.

b) Suppose you were to design a block random assignment in order to predict evaluations in 1997. Use the R package `blockTools` (for example code, see isps.research.yale.due/FEDAI) to perform a block random assignment, blocking on sex, church membership, social class, and evaluations in 1996. Decide for yourself how many subjects to include in each block. Compare the treatment and control groups to verify that blocking produced groups that have the same profile of sex, church membership, social class, and evaluations in 1996.


```
block.out <- block(data = russia, n.tr = 2,
                   id.vars = "id",algorithm="randGreedy",
                    block.vars = c("female", "church_member",
                                   "index96","class_verypoor",
                                   "class_poor", "class_middle"))
assign.out <- assignment(block.out)

# extracting the treatment assignment from blockTools takes some work
# The commands below check to see which ID numbers appear on the
# list of assign.out's assignment to Treatment 1

russia$Z_blocked <- as.numeric(russia$id %in%
                         as.numeric(as.character(
                           unlist(assign.out$assg[[1]]["Treatment 1"]))))

arm::display(lm(Z_blocked ~ female + church_member + class + index96,
                data=russia))

## lm(formula = Z_blocked ~ female + church_member + class + index96,
##     data = russia)
##                         coef.est coef.se
## (Intercept)              0.53     0.17
## female                   0.00     0.06
## church_member            0.01     0.08
## classpoor               -0.04     0.16
```

```
## classmiddle            -0.05     0.16
## classmore than middle -0.05     0.22
## index96                 0.00     0.01
## ---
## n = 462, k = 7
## residual sd = 0.50, R-Squared = 0.00
```

Using the package `blockTools`, we created blocks of size 2 based on gender, church membership, evaluations in 1996, and social class. The package also conducts complete random assignment – with some work, this assignment can be extracted. Regressing this treatment assignment on the set of pretreatment covariates reveals that the groups are well balanced.

c) Suppose you wanted to assess how well your blocking design performed in terms of increasing the precision with which treatment effects are estimated. Of course, there was no actual treatment in this case, but imagine that shortly after the survey in 1996, a treatment were administered to a randomly selected treatment group. (Here is an instance in which the sharp null hypothesis of no effect is known to be true!) The outcome from this imaginary experiment is evaluations of national conditions in 1997. Compare the sampling distribution of the estimated treatment effect (which should be centered on zero) under balanced complete random assignment to the sampling distribution of the estimated treatment effect under block random assignment.
Answer:
See below

d) Calculate the sampling distribution of the estimated treatment effect under balanced complete random assignment using regression to control for the variables that would have otherwise been used to form blocks. Compare the resulting distribution to the sampling distribution of the estimated treatment effect under block random assignment. Does blocking produce an appreciable gain in precision over what is achieved by covariate adjustment?

```
sims <- 10000
results <- matrix(NA,sims,3)
colnames(results) <- c("complete","adjusted","blocked")
N <- nrow(russia)

for(i in 1:sims) {
    # Complete RA, with and without adjustment
    russia$Z_complete <- ifelse(1:N %in% sample(N, N/2), 1, 0)
    results[i,1] <- lm(index97 ~ Z_complete, data=russia)$coefficients[2]
    results[i,2] <- lm(index97 ~ Z_complete + female + church_member + class + index96,
                       data=russia)$coefficients[2]

    # Blocked RA, without adjustment
    assign.out <- assignment(block.out)
    russia$Z_blocked <- as.numeric(russia$id %in%
                         as.numeric(as.character(
                           unlist(assign.out$assg[[1]]["Treatment 1"]))))
    results[i,3] <- lm(index97 ~ Z_blocked, data=russia)$coefficients[2]
    }
```

```r
# use apply() to extract means and SDs for each column (2 refers to columns)
results_table <- rbind(apply(results,2,mean),apply(results,2,sd))
rownames(results_table) <- c("Average Estimate", "Standard Error")
results_forxtable <- xtable(results_table,caption="Comparison of 3 estimators")
```

```r
print.xtable(results_forxtable,caption.placement="top",table.placement="H")
```

Table 4: Comparison of 3 estimators

	complete	adjusted	blocked
Average Estimate	0.00	0.00	-0.00
Standard Error	0.17	0.13	0.13

The table above shows a comparison of three estimators of the ATE: difference-in-means under complete random assignment, OLS with covariate adjustment under complete random assignment, and difference-in-means under blocked random assignment. All three estimators are centered on the true ATE of zero. The least precise method is complete random assignment with the difference-in-means estimator, which produces a standard error of 0.169. The most precise approach is blocking, which produces a standard error of 0.131. Slightly inferior to blocking is covariate adjustment, which produces a standard error of 0.133. Blocking's slight superiority stems from the fact that, under blocking, there is no incidental correlation between the covariates and random assignments and therefore no "collinearity penalty."

Question 7

Researchers may be concerned about using block randomization when they are unsure whether the variable used to form the blocks actually predicts the outcome. Consider the case in which blocks are formed randomly – in other words, the variable used to form the blocks has no prognostic value whatsoever. Below is a schedule of potential outcomes for four observations.

Table 5: Question 7 Table

Subject	Y(0)	Y(1)
A	1	2
B	0	3
C	2	2
D	5	5

a) Suppose you were to use complete random assignment such that $m = 2$ units are assigned to treatment. What is the sampling variance of the difference-in-means estimator across all six possible random assignments?

The average estimated ATE is 1.0, which is the true ATE. The variance of the estimated ATEs over all 6 possible randomizations is 2.833.

Table 6: Question 7a table

Treated Units	$\bar{Y(1)}$	$\bar{Y(0)}$	$\widehat{ATE}$
A and B	2.5	3.5	-1
A and C	2	2.5	-0.5
A and D	3.5	1	2.5
B and C	2.5	3	-0.5
B and D	4	1.5	2.5
C and D	3.5	0.5	3

b) Suppose you were to form blocks by randomly pairing the observations. Within each pair, you randomly allocate one subject to treatment and the other to control so that $m = 2$ units are assigned to treatment.There are three possible blocking schemes; for each blocking scheme, there are four possible random assignments. What is the sampling variance of the difference-in-means estimator across all twelve possible random assignments?

Table 7: Question 7b table

	Treated Units	$\bar{Y(1)}$	$\bar{Y(0)}$	$\widehat{ATE}$
AB and CD blocked	A,C	2	2.5	-0.5
	A,D	3.5	1	2.5
	B,D	4	1.5	2.5
	B,C	2.5	3	-0.5
AC and BD blocked	A,B	2.5	3.5	-1
	A,D	3.5	1	2.5
	C,B	2.5	3	-0.5
	C,D	3.5	0.5	3
AD and BC blocked	A,B	2.5	3.5	-1
	A,C	2	2.5	-0.5
	D,B	4	1.5	2.5
	D,C	3.5	0.5	3

Across the 12 possible random assignments, the variance of the estimated ATE is again 2.833. Notice that every estimate in the previous table appears in this table twice.

c) From this example, what do you infer about the risks of blocking on a non-prognostic covariate?
Answer:
There is no risk of increasing variance with a useless blocking variable; at worst, the variable will be random noise, in which case the sampling variance will be the same as a design without blocking.

Question 8

Sometimes researchers randomly assign subjects from lists that are later discovered to have duplicate entries. Suppose, for example, that a fund-raising experiment randomly assigns 500 of 1,000 names to a treatment that consists of an invitation to contribute to a charitable cause. However, it is later discovered that 600 names appear once and 200 names appear twice. Before the invitations are mailed, duplicate invitations are discarded, so that no one receives more than one invitation.

a) What is the probability of assignment to the treatment group among those whose names appeared once in the original list? What is the probability of assignment to the treatment group among those whose names appeared twice in the original list?
Answer:
The probability of being assigned to treatment if your name appears once is 0.5. The probability of being assigned to treatment if your name is a duplicate is $0.5 + (0.5)(0.5) = 0.75$, where the first term is the probability you were assigned to treatment the first time your name came up and the second term is the probability you were assigned to control the first time multiplied by the probability you were assigned to treatment the second time.

b) Of the 800 unique names in the original list, how many would you expect to be assigned to treatment and control?
Answer:
Of the 600 unique names that appear once, 300 are, in expectation, allocated to treatment. Of the 200 unique names that appear twice, 150 are, in expectation, allocated to treatment. Thus, we expect a total of 450 unique names in the treatment group.

c) What estimation procedure should one use in order to obtain unbiased estimates of the ATE?
Answer:
One should analyze the experiment as though it were randomized in two blocks: the names that appear once and the names that appear twice. Use an estimator like equation (4.11).

Question 9

Gerber and Green conducted a mobilization experiment in which calls from a large commercial phone bank urged voters in Iowa and Michigan to vote in the November 2002 election.[2] The randomization was conducted within four blocks: uncompetitive congressional districts in Iowa, competitive congressional districts in Iowa, uncompetitive congressional districts in Michigan, and competitive congressional districts in Michigan. Table 4.3 presents results only for one-voter households in order to sidestep the complications of cluster assignment.

a) Within each of the four blocks, what was the apparent effect of being called by a phone bank on voter turnout?
Answer:
From the "Estimated ATE" Row: Block 1: .0096, Block 2: -.0078, Block 3: -.0136, Block 4: .0083. Substantively, these results suggest that calls encouraging voter turnout had effects ranging from -1.4 percentage points to +1.0 percentage point.

b) When all of the subjects in this experiment are combined (see the rightmost column of the table), turnout seems substantially higher in the treatment group than the control group. Explain why

[2] Gerber and Green 2005.

this comparison gives a biased estimate of the ATE.
Answer:
This estimator is biased because individuals in each stratum had different propensities to enter into treatment. The uncompetitive Michigan block has the lowest rate of treatment and also has the lowest rate of voting in the control group. Overall, blocks with higher rates of treatment tend to have higher rates of voting in the control group, which accounts for the upward bias.

c) Using the weighted estimator described in Chapter 3, show the calculations used to generate an unbiased estimate of the overall ATE.

```
ests <- c(.00964, -.007829, -.01362,.008271)
shareoftotalN <- c(0.049487, 0.1520981, 0.626616, 0.171799)
overall_ate <-sum(ests*shareoftotalN)
overall_ate

## [1] -0.007827
```

d) When analyzing block randomized experiments, researchers frequently use regression to estimate the ATE by regressing the outcome on the treatment and indicator variables for each of the blocks (omitting one block if the regression includes an intercept.) This regression estimator places extra weight on blocks that allocate approximately half of the subjects to the treatment condition (i.e., $P_j = 0.5$) because these blocks tend to estimate the within-block ATE with less sampling variability. Compare the four OLS weights to the weights W_j used in part (c).
Answer:
The weights used in part (c) are based on the share of the subject pool that is in each block. This weighting scheme places a great deal of weight on the relatively large Michigan block. By contrast, the OLS weights are a blend of the number of subjects in each block and each block's balance between treatment and control allocations. Because the blocks do not differ very much in terms of their allocation rates, the OLS weights tend to be similar across blocks.

e) Regression provides an easy way to calculate the weighted estimate of the ATE in part (c) above. For each treatment subject i, compute the proportion of subjects in the same block who were assigned to the treatment group. For control subjects, compute the proportion of subjects in the same block who were assigned to the control group. Call this variable q_i. Regress outcomes on treatment, weighting each observation by $1/q_i$, and show that this type of weighted regression produces the same estimate as weighting the estimated ATEs for each block.

```
Y <- phones$vote02
block <- phones$strata
Z <- phones$treat2

## Proportion of subjects in each block assigned to treatment
block.pr <- tapply(Z, block, mean)

q <- rep(NA, length(Y))

for(i in 1:4){
  q[block==i] <- block.pr[i]*Z[block==i] + (1-block.pr[i])*(1-Z[block==i])
```

```
}

fit <- lm(Y ~ Z, weights=1/q)
summary(fit)

##
## Call:
## lm(formula = Y ~ Z, weights = 1/q)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -4.051 -0.469 -0.469  0.537  4.786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.466198   0.000727  641.31  < 2e-16 ***
## Z           -0.007828   0.001028   -7.61  2.7e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.705 on 940713 degrees of freedom
## Multiple R-squared:  6.16e-05,Adjusted R-squared:  6.06e-05
## F-statistic:   58 on 1 and 940713 DF,  p-value: 2.65e-14
```

The coefficient on the treatment indicator is −0.0078, which is the same as was found in part c.

Question 10

The 2003 Kansas City voter mobilization experiment described in Chapter 3 is a cluster randomized design in which 28 precincts comprising 9,712 voters were randomly assigned to treatment and control.[3] The study contains a wealth of covariates: the registrar recorded whether each voter participated in elections dating back to 1996. The dataset may be obtained at isps.research.yale.edu/FEDAI.

a) Test the balance of the treatment and control groups by looking at whether past turnout predicts treatment assignment. Regress treatment assignment on the entire set of past votes, and calculate the F-statistic. Use randomization inference to test the null hypothesis that none of the past turnout variables predict treatment assignment. Remember that to simulate the distribution of the F-statistic, you must generate 1,000 random cluster assignments and calculate the F-statistic for each simulated assignment. Judging from the p-value of this test, what does the F-statistic seem to suggest about whether subjects in the treatment and control groups have comparable background characteristics?

```
Z <-  kansas$treatmen
Y <- kansas$vote03
clust <- kansas$unit
```

[3] Arceneaux 2005.

```
covs <- as.matrix(kansas[,2:21])  # covariates are past voter turnout

probs <- genprobexact(Z,clustvar=clust)  # subjects are clustered by precinct
perms <- genperms(Z,maxiter=1000,clustvar=clust)    # clustered assignment

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 40116600 to perform exact estimation.

numiter <- ncol(perms)

Fstat <- summary(lm(Z~covs))$fstatistic[1]   # F-statistic from actual data

Fstatstore <- rep(NA,numiter)
for (i in 1:numiter) {
  Fstatstore[i] <- summary(lm(perms[,i]~covs))$fstatistic[1]
        }

p.value <- mean(Fstatstore >= Fstat)
p.value

## [1] 0.936
```

Using randomization inference, we recover a p-value of 0.936; we therefore cannot reject the null hypothesis of random assignment.

b) Regress turnout in 2003 (after the treatment was administered) on the experimental assignment and the full set of covariates. Interpret the estimated ATE. Use randomization inference to test the sharp null hypothesis that experimental assignment had no effect on any subject's decision to vote.


```
ate <- estate(Y,Z,X=covs,prob=probs)
Ys <- genouts(Y,Z,ate=0)
distout <- gendist(Ys,perms,X=covs,prob=probs)
p.value.onetailed <- mean(distout >= ate)

ate

##       Z
## 0.05596

p.value.onetailed

## [1] 0.005
```

The estimate of the treatment effect is 0.056, implying that treatment increased turnout by 5.6

percentage points. This finding is statistically significant. Under the sharp null, estimates as large or larger only occur 0.5% of the time.

c) When analyzing cluster randomized experiments with clusters of varying size, one concern is that difference-in-means estimation is prone to bias. This concern also applies to regression. In order to sidestep this problem, researchers may choose to use the difference-in-totals estimator in equation (3.24) to estimate the ATE. Estimate the ATE using this estimator.

```
ateHT <- estate(Y,Z,prob=probs,HT=TRUE)
ateHT

## [1] 0.05395
```

The difference-in-totals estimate of the treatment effect is that treatment increased turnout by 5.4 percentage points.

d) Use randomization inference to test the sharp null hypothesis that treatment assignment had no effect, using the difference-in-totals estimator.

```
distoutHT <- gendist(Ys,perms,prob=probs,HT=TRUE)
p.value.onesidedHT <- mean(distoutHT >= ateHT)
p.value.onesidedHT

## [1] 0.198
```

Estimates generated under the sharp null equaled or exceeded the observed difference-in-totals 19.8% of the time, meaning we cannot reject the null.

e) The difference-in-totals estimator can generate imprecise estimates, but its precision can be improved by incorporating information about covariates. Create a new outcome variable that is the difference between a subject's turnout (1 = vote, 0 = abstain) and the average rate of turnout in all past elections. Now, using this "differenced" outcome variable, estimate the ATE using the difference-in-totals estimator, and test the sharp null hypothesis of no effect.

```
Y_diff <- Y - rowMeans(covs)

ateHT2 <- estate(Y_diff,Z,prob=probs,HT=TRUE)  # difference-in-differenced totals
Ys <- genouts(Y_diff,Z,ate=0)
distoutHT2 <- gendist(Ys,perms,prob=probs,HT=TRUE)
p.value.onesidedHT2 <- mean(distoutHT2 >= ateHT2)

ateHT2

## [1] 0.04874

p.value.onesidedHT2

## [1] 0.012
```

Using the differenced outcome variable tightened our estimates – the p-value under the sharp null is now 0.012, meaning we can reject the sharp null of no effect for any unit.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 5 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Using the data in Table 5.2:

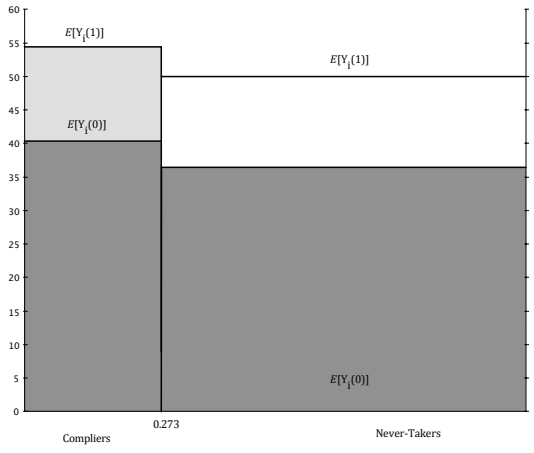
a) Estimate the following quantities: $E[d_i(1)]$, $E[Y_i(0)|d_i(1) = 0]$, $E[Y_i(0)|d_i(1) = 1]$,and $E[Y_i(1)|d_i(1) = 1]$.

$$E[d_i(1)] = \frac{395}{1445} = 0.273$$
$$E[(Y_i(0)|d_i(1) = 0)] = 36.48$$
$$E[(Y_i(0)|d_i(1) = 1)] = \frac{37.54 - (0.727 * 36.48)}{0.273} = 40.36$$
$$E[(Y_i(1)|d_i(1) = 1)] = 54.43$$

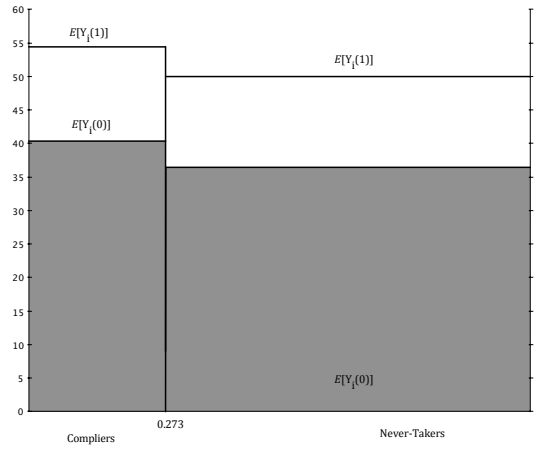
b) Using these estimates and assuming that $E[Y_i(1)|d_i(1) = 0] = 0.5$, construct a figure that follows the format of Figure 5.1. Show the apparent proportion of Compliers, the ITT, and the CACE.

Figure 1: Question 2 Figure

(a) Treatment Group



(b) Control Group



*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

Question 2

Make up a hypothetical schedule of potential outcomes for three Compliers and three Never-Takers in which the ATE is positive but the CACE is negative. Suppose that an experiment were conducted on your pool of subjects. In what ways would the estimated CACE be informative or misleading?
Answer:

Table 1: Hypothetical schedule of potential outcomes

subject	$Y(0)$	$Y(1)$	$D(1)$	$Y(Z=0)$	$Y(Z=1)$
1	15	5	1	15	5
2	10	5	1	10	5
3	5	5	1	5	5
4	5	25	0	5	5
5	10	20	0	10	10
6	15	30	0	15	15

$$ATE = \frac{90 - 60}{6} = 5$$
$$CACE = \frac{15 - 30}{3} = -5$$

Question 3

Explain whether each of the following statements is true or false for the case of one-sided noncompliance, assuming that an experiment satisfies non-interference and excludability.

a) If the ITT is negative, the $CACE$ must be negative.
Answer:
True. The ITT can be written as $E[D(1)] * CACE$. If this quantity is negative, then since $E[D(1)]$ must be non-negative, the $CACE$ must be negative.

b) The smaller the ITT_D, the larger the $CACE$.
Answer:
False. There is no necessary relationship between the ITT_D, the proportion of the subjects that are compliers, and the CACE, the average response of the compliers to the treatment. This confusion sometimes arises due to the algebra of calculating the CACE from the ITT and ITT_D. Because the ITT can be written as $ITT_D * CACE$, the $CACE$ can be calculated by ITT/ITT_D. From this ratio it might appear that when the ITT_D is smaller we are dividing the ITT by a smaller number, leading to a larger CACE. However, changing the rate of compliance may change the ATE among those who now comply.

c) One cannot identify the CACE if no one in the experiment receives the treatment.
Answer:
True. If no one receives the treatment, it is impossible to estimate the effect of the treatment. Algebraically, the ITT estimate is divided by zero, leading to an undefined CACE estimate.

Question 4

Explain whether each of the following equalities follows as a consequence of the excludability assumption.

a) $E[Y_i(z=1)] = E[Y_i(z=1)|d_i(1)=1]$
Answer:
No. The exclusion restriction (ER) says ignore Z_i. However, the ER does not imply that $E[Y(D(1))]$ for the entire subject pool is equal to the average $Y(D(1))$ for the compliers.

b) $E[Y_i(z=0,d=0)|d_i(1)=0] = E[Y_i(z=1,d=0)|d_i(1)=0]$
Answer:
Yes. According to the ER, $Y(z=1,d) = Y(z=0,d)$ for all d.

c) $E[Y_i(z=1,d(1))] = E[Yi(z=1),d(0))]$
Answer:
No. The ER does not imply that $D(1) = D(0)$ for all i.

d) $E[Y_i(z=0,d(0))] = E[Yi(z=1),d(0))]$
Answer:
Yes. The ER implies that $Y_i(Z=1,D) = Y_i(Z=0,D)$.

Question 5

Critically evaluate the following statement: "If you are conducting an experiment that encounters one-sided noncompliance, you will never know which of your subjects are Compliers and which of your subjects are Never-Takers."
Answer:
Subjects are assigned to treatment or control. For those subjects assigned to the control group, all subjects are untreated so there is no way to distinguish Compliers from Never-takers. For those subject assigned to the treatment group, the Compliers are treated and the Never-Takers are not. This is observable, and so you can tell which subjects are of each type for those assigned to the treatment group. Using the subjects assigned to the treatment group, you can contrast the compliers and never-takers based on pretreatment variables. However, as suggested by the statement, there is a limit to what you can know about individual subjects assigned to the control group. Since both types remain untreated in the control group, you cannot partition the entire subject pool into compliers and never takers.

Question 6

Suppose that a researcher hires a group of canvassers to contact a set of 1,000 voters randomly assigned to a treatment group. When the canvassing effort concludes, the canvassers report that they successfully contacted 500 voters in the treatment group, but the truth is that they only contacted 250. When voter turnout rates are tabulated for the treatment and control groups, it turns out that 400 of the 1,000 subjects in the treatment group voted, as compared to 700 of the 2,000 subjects in the control group (none of whom were contacted).

a) If you believed that 500 subjects were actually contacted, what would your estimate of the CACE be?

Answer:
ITT estimate is $0.40 - 0.35 = 0.05$. The ITT_D estimate is $= 500/1000 = 0.5$.

$$\widehat{CACE} = \frac{\widehat{ITT}}{\widehat{ITT_D}} = \frac{0.05}{0.5} = 0.10$$

b) Suppose you learned that only 250 subjects were actually treated. What would your estimate of the CACE be?
Answer:
ITT estimate stays the same: $0.40 - 0.35 = 0.05$. The ITT_D estimate is now $= 250/1000 = 0.25$.

$$\widehat{CACE} = \frac{\widehat{ITT}}{\widehat{ITT_D}} = \frac{0.05}{0.25} = 0.20$$

c) Do the canvassers' exaggerated reports make their efforts seem more or less effective? When formulating your answer, you may define effectiveness in terms of either the ITT or the CACE.
Answer:
The exaggerated reports made their efforts look less effective in terms of the CACE. Since the share of the treatment group that actually received the "boost" associated with the treatment was smaller than was claimed, the observed difference was attributed to 500 people being treated rather than 250 being treated. Consequently, the average effect of each treatment seems half as large. This misreport has no effect on the estimate of the number of voters produced by the canvassing effort, which is estimated by the ITT.

Question 7

Make up a schedule of potential outcomes that would generate Figure 5.2, which illustrates the consequences of an exclusion restriction violation. Hint: you will need to allow for potential outcomes that respond to both d and z.
Answer:
Figure 5.2 illustrates a situation in which the potential outcome when untreated among the non-compliers depends on whether the subject is in the treatment versus control group. Note that this is just an example of an exclusion restriction violation, in this case limited to one of the average potential outcomes (untreated non-compliers). Other patterns of ER violations are possible as well. The 6 quantities needed to construct a figure similar to figure 5.2 are:

1. $E[D_i(1)]$
2. $E[(Y(0)|D_i(1) = 0), Z_i = 0]$
3. $E[(Y(0)|D_i(1) = 0), Z_i = 1]$
4. $E[(Y(1)|D_i(1) = 0)]$
5. $E[(Y(0)|D_i(1) = 1)]$
6. $E[(Y(1)|D_i(1) = 1)]$

Further, $E[(Y(0)|D_i(1) = 0), Z_i = 0]$ and $E[(Y(0)|D_i(1) = 0), Z_i = 1]$ must be different – this is the crucial violation of the exclusion restriction. Suppose the subject pool is comprised of only two type subjects. 25 percent are of type 1 and the remainder are of type 2.

Table 2: Question 7 Table

Subject	$Y(D = 1)$	$Y(D = 0, Z = 0)$	$Y(D = 0, Z = 1)$	$D(1)$
Type 1	10	5	5	1
Type 2	8	4	6	0

Question 8

Cotterill et al. report the results of an experiment conducted in an area of the United Kingdom where only half of the local residents recycle their trash.[1] Canvassers visited homes and encouraged residents to recycle. Outcomes were measured by whether the home put out a recycling bin on at least one occasion during the following three weeks. We restrict our attention here to homes that did not recycle trash during a pre-experimental period of observation. When implementing the intervention, researchers encountered one-sided noncompliance: 1,015 of the 1,849 homes assigned to the treatment group were successfully canvassed; none of the 1,430 homes assigned to the control group were canvassed. These researchers found that 591 homes in the treatment group recycled, as opposed to 377 in the control group. The researchers also observed that 429 of 1,015 homes that were successfully canvassed recycled, as opposed to 539 of the 2,264 homes that were not canvassed.

a) Estimate the ITT, and interpret the results.
Answer:

$$\widehat{ITT} = \frac{591}{1849} - \frac{377}{1430} = 0.32 - 0.264 = .056$$

Assignment to being canvassed caused an estimated 5.6 percentage point increase in recycling.

b) Estimate the ITT_D, and interpret the results. Answer:

$$\widehat{ITT_d} = \frac{1015}{1849} = 0.549$$

The estimated probability a subject randomly assigned to the treatment group will be canvassed (is a Complier) is 54.9%

c) Using the equations in Theorem 5.1 as a guide, write down a model of the expected recycling rate among those assigned to the control group. Do the same for the expected recycling rate among those assigned to the treatment group. Show that under the assumptions of Theorem 5.1, the $CACE$ can be identified based on the design of this experiment.

Answer:

Assume that the standard assumptions (Exclusion restriction, non-interference) hold. For each subject, the potential outcomes are a collection of 4 values: $(Y_i(d), D_i(z))$.

[1] Cotterill et al. 2009.

- Define the $CACE = E[Y_i(1) - Y_i(0)|D_i(1) = 1]$.
- Expected value of recycling rate (Y) in the control group = $E[Y_i(0)|D_i(1) = 1] * ITT_d + E[Y_i(0)|D_i(1) = 0] * (1 - ITT_d)$
- Expected value of recycling rate in treatment group = $E[Y_i(1)|D_i(1) = 1] * ITT_d + E[Y_i(0)|D_i(1) = 0] * (1 - ITT_d)$
- Expected value of rate of successful canvassing = $E[D_i(1)] = ITT_d$
- Expected value of recycling rate among treatment group minus recycling rate of control group = $[E[Y_i(1)|D_i(1) = 1] - E[Y_i(0)|D_i(1) = 1)] * (ITT_d) = CACE * ITT_d$.
- $CACE = [E[Y_i(1)|D_i(1) = 1] - E[Y_i(0)|D_i(1) = 1)]]/(ITT_d)$.

The expected value of ITT estimate (difference in treatment and control group recycling rates) is equal to the numerator (it is an unbiased estimator), and the expected value of the observed compliance rate is an unbiased estimate of the denominator. The ratio of these estimators is a consistent estimator of $E[Y_i(1)|D_i(1) = 1] - E[Y_i(0)|D_i(1) = 1)]/(ITT_d)$, therefore the experiment produces enough information to obtain an estimate of the CACE.

d) Estimate the $CACE$, and interpret the results.
Answer:

$$\widehat{CACE} = \frac{0.056}{0.549} = 0.102$$

The estimated average increase in the probability a Complier recycles when treated versus not treated is 10.2 percentage points.

e) Explain why comparing the recycling rates of the treated and untreated subjects tends to produce misleading estimates of the $CACE$ and ATE.
Answer:
Comparisons of the treated and untreated conflate the effect of the treatment and other differences across the groups that might be correlated with being treated. In contrast to random assignment, which produces groups with the same expected potential outcomes through the procedure, claims about the similarity of the potential outcomes for the actually treated and untreated typically rest on assumptions. In the particular example presented here, an unbiased estimate of the treatment effect for compliers requires an estimate of the average outcome for the compliers when untreated, but the untreated households are a mixture of untreated compliers and non-compliers, rather than a collection of untreated compliers. There is substantial evidence that the average recycling rate among the untreated is lower than the recycling rate among the untreated Compliers. The non-Compliers recycle at a rate of 19.4%, which is much lower than the control group (a mixture of the two types) recycling rate of 26.4%. This implies that comparing the treated Compliers with the mixture of untreated Compliers and non-compliers will exaggerate the treatment effect.

Question 9

One way to detect heterogeneous treatment effects across subgroups is to employ a design that randomly manipulates the level of compliance. One such study was conducted in Michigan in 2002.[2] Subjects were randomly allocated to three experimental groups. The first treatment group was targeted for a phone call that encouraged subjects to vote in the upcoming November election. The second treatment group was targeted for the same call using the same script on the same day, but more attempts were made to reach subjects. No attempts were made to contact the control group. The table below shows the contact rates and voting rates for each of the three assigned groups.

Table 3: Question 9 Table

	Control	Treatment group #1 (minimal effort)	Treatment group #2 (maximal effort)
Percent reached by callers	0	29.97	47.31
Percent voting	55.89	55.91	56.53
N	317182	7500	7500

a) Define two types of Compliers: those who respond when called with minimal (or maximal) effort and those who respond only when called with maximal effort. Write down a model expressing the expected voting rate among those assigned to the control group as a weighted average of potential outcomes among Minimal Compliers, Maximal Compliers, and Never-Takers. Do the same for the expected rate of voting among those assigned to each of the treatment groups.
Answer:
Let $Z_i = 0$ (no call), 1 (minimal effort), or 2 (maximal effort). $Y_i(Z_i) = 1$ if subject i votes, 0 otherwise. Assuming monotonicity as outlined in the problem description, there are three types ($D_i(0) = 0$ for all types):

- $D_i(1) = D_i(2) = 0$ [Never-Takers]
- $D_i(2) = 1, D_i(1) = 1$ [Easy to reach subjects, or Minimal Effort Compliers]
- $D_i(2) = 1, D_i(1) = 0$ [Hard to reach subjects, or Maximal Effort Compliers]

Expected Vote rate in Control (EV, Control) =

$$E(Y(0)|\text{ never taker}) * Pr(\text{never taker}) +$$
$$E(Y(0)|\text{ easy to reach}) * Pr(\text{easy to reach}) +$$
$$E(Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})$$

Expected Vote rate in minimal effort (EV, minimal) =

$$E(Y(0)|\text{ never taker}) * Pr(\text{never taker}) +$$
$$E(Y(1)|\text{ easy to reach}) * Pr(\text{easy to reach}) +$$
$$E(Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})$$

[2] Gerber and Green 2005.

Expected Vote rate in maximal effort (EV, maximal) =

$$E(Y(0)|\text{ never taker}) * Pr(\text{never taker}) +$$
$$E(Y(1)|\text{ easy to reach}) * Pr(\text{easy to reach}) +$$
$$E(Y(1)|\text{ hard to reach}) * Pr(\text{hard to reach})$$

b) Show that the CACE for each of the treatments can be identified based on the design of this experiment.

(EV, minimal - EV,control) =

$$E(Y(1)|\text{ easy to reach}) * Pr(\text{easy to reach}) -$$
$$E(Y(0)|\text{ easy to reach}) * Pr(\text{easy to reach})$$
$$= E(Y(1) - (Y(0)|\text{ easy to reach}) * Pr(\text{easy to reach})$$
$$= (\text{ATE}|\text{ easy to reach}) * Pr(\text{easy to reach}).$$
$$\text{ATE}|\text{ easy to reach} = \frac{(\text{EV, minimal - EV,control})}{Pr(\text{easy to reach})}$$

Similarly,

(EV, maximal - EV,minimal) =

$$E(Y(1)|\text{ hard to reach}) * Pr(\text{hard to reach}) -$$
$$E(Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})$$
$$= E(Y(1) - (Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})$$
$$= (\text{ATE}|\text{ hard to reach}) * Pr(\text{hard to reach}).$$
$$\text{ATE}|\text{ hard to reach} = \frac{(\text{EV, maximal - EV,minimal})}{Pr(\text{hard to reach})}$$

To estimate the numerators, which involve EV for the subject pool when untreated, given the minimal treatment or maximal treatment, use the average of the randomly assigned groups, which are unbiased estimators of the respective quantities. To obtain unbiased estimates of the subject pool proportions for the three types, the proportion that complies in the minimal treatment group is an unbiased estimate of the proportion of easy to reach, and the proportion that complies in the maximal effort group is an estimate of the combined proportion of easy and hard to reach. Subtract the estimated proportion that is easy to reach to obtain the proportion that is hard to reach.

c) Estimate the share of the subject pool that Maximal Compliers comprise. Estimate the share of the subject pool that Minimal Compliers comprise.
Answer:
The share of compliers in the minimal treatment group provides an estimate of the share of easy to reach: 29.97%
The share of compliers in the maximal treatment group provides an estimate of the share of easy to reach plus the share of hard to reach, which equals 47.31%. Subtracting the estimated share of the easy to reach, 29.97%, produces an estimate of the share of hard to reach, 47.31 - 29.97 = 17.34%.

d) Estimate the average treatment effect among each type of Complier, and interpret the results.
Answer:
The CACE for the easy to reach: $\frac{0.5591-0.5589}{0.2997} = 0.0007$
The CACE for the hard to reach: $\frac{0.5653-0.5591}{0.1734} = 0.0358$
The treatment effect estimate for the hard to reach is larger than the estimated effect for the easy to reach, although further calculations are needed to determine whether the difference in CACEs is greater than one would expect from random sampling variability.

Question 10

Guan and Green report the results of a canvassing experiment conducted in Beijing on the eve of a local election.[3] Students on the campus of Peking University were randomly assigned to treatment or control groups. Canvassers attempted to contact students in their dorm rooms and encourage them to vote. No contact with the control group was attempted. Of the 2,688 students assigned to the treatment group, 2,380 were contacted. A total of 2,152 students in the treatment group voted; of the 1,334 students assigned to the control group, 892 voted. One aspect of this experiment threatens to violate the exclusion restriction. At every dorm room they visited, even those where no one answered, canvassers left a leaflet encouraging students to vote.

a) Using the dataset at http://isps.research.yale.edu/FEDAI, estimate the ITT.

```
# get rid of a couple of observations with missing outcome data
beijing <- na.omit(beijing.all)
Z <-     beijing$treat2
Y <-     beijing$turnout
D <-     beijing$contact
clust <- beijing$dormid

ITT <- mean(Y[Z==1]) - mean(Y[Z==0])
ITT

## [1] 0.1319
```

The estimated ITT is 0.132.

b) Use randomization inference to test the sharp null hypothesis that the ITT is zero for all observations, taking into account the fact that random assignment was clustered by dorm room. Interpret your results.

```
probs <- genprobexact(Z,clustvar=clust)  # subjects are clustered by dorm room
itt <- estate(Y,Z,prob=probs)
itt

## [1] 0.1319

numiter <- 10000
perms <- genperms(Z,maxiter=numiter,clustvar=clust)    # clustered assignment
```

[3] Guan and Green 2006.

```
## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 5.83377797524832e+275 to perform exact estimation.

Ys <- genouts(Y,Z,ate=0)
distout <- gendist(Ys,perms,prob=probs)

p.value <- sum(abs(distout) >= abs(itt))        # two-tailed comparison
p.value

## [1] 0
```

c) Assume that the leaflet had no effect on turnout. Estimate the $CACE$.

```
itt <- estate(Y,Z,prob=probs)
itt

## [1] 0.1319

ittd <- estate(D,Z,prob=probs)
ittd

## [1] 0.8858

cace <- itt/ittd
cace

## [1] 0.1489
```

d) Assume that the leaflet raised the probability of voting by one percentage point among both Compliers and Never-Takers. In other words, suppose that the treatment group's turnout rate would have been one percentage point lower had the leaflets not been distributed. Write down a model of the expected turnout rates in the treatment and control groups, incorporating the average effect of the leaflet.
Answer:
Let $Y_i(1)$ be whether i votes when reached for canvassing but not treated with a leaflet, $Y_i(0)$ be whether i votes when not reached for canvassing and not left a leaflet. Let $D_i = 1$ when canvassed when assigned to treatment group, 0 otherwise.
Expected Turnout in Treatment Group (EV, Treatment):

$$E[Y_i(1) + .01)|D_i(1) = 1] * ITT_d + E[Y_i(0) + .01)|D_i(1) = 0] * (1 - ITT_d) =$$
$$E[Y_i(1)|D_i(1) = 1] * ITT_d + E[Y_i(0)|D_i(1) = 0] * (1 - ITT_d) + .01$$

Expected Turnout in Control Group (EV, Control):

$$E[Y_i(0)|D_i(1) = 1] * ITT_d + E[Y_i(0)|D_i(1) = 0] * (1 - ITT_d)$$

(EV, Treatment) - (EV, Control) = ITT:

$$\begin{aligned} ITT &= .01 + E[Y_i(1)|D_i(1) = 1] * ITT_d - E[Y_i(0)|D_i(1) = 1] * ITT_d \\ &= .01 + E[Y_i(1) - Y_i(0)|D_i(1) = 1] * ITT_d \\ ITT/ITTd &= E[Y_i(1) - Y_i(0)|D_i(1) = 1] + \frac{.01}{ITT_d}. \end{aligned}$$

Therefore:

$$CACE = E(Yi(1) - Yi(0)|Di(1) = 1) = ITT/ITT_d - .01/ITT_d$$

e) Given this assumption, estimate the CACE of canvassing.

$$\widehat{CACE} = 0.149 - \frac{0.01}{0.885} = 0.138$$

f) Suppose, instead, that the leaflet had no effect on Compliers (who heard the canvasser's speech and ignored the leaflet) but raised turnout among Never-Takers by 3 percentage points. Given this assumption, estimate the CACE of canvassing.
Answer:
Under this assumption, we write:
Expected Turnout in Treatment Group (EV, Treatment)=

$$E[Y_i(1)|D_i(1) = 1] * ITT_d + E[Y_i(0) + .03)|D_i(1) = 0] * (1 - ITT_d)$$

Therefore, subtracting the expected vote in the control group from the expected vote in the treatment group gives:

$$\begin{aligned} ITT = E[Y_i(1)|D_i(1) = 1] * ITT_d + E[Y_i(0) + .03)|D_i(1) = 0] * (1 - ITT_d) - \\ E[Y_i(0)|D_i(1) = 1] * ITT_d - E[Y_i(0)|D_i(1) = 0] * (1 - ITT_d) \\ = E[Y_i(1) - Y_i(0)|D_i(1) = 1] * ITT_d + 0.03 * (1 - ITT_d) \end{aligned}$$

Therefore:

$$\begin{aligned} \frac{ITT}{ITT_d} &= CACE + \frac{0.03(1 - ITT_d)}{ITT_d} \\ CACE &= \frac{ITT}{ITT_d} - \frac{0.03(1 - ITT_d)}{ITT_d} \\ &= \frac{0.132}{0.885} - \frac{0.03(1 - 0.885)}{0.885} \\ &= 0.145 \end{aligned}$$

Question 11

Nickerson describes a voter mobilization experiment in which subjects were randomly assigned to one of three conditions: a baseline group (no contact was attempted), a treatment group (canvassers attempted to deliver an encouragement to vote), and a placebo group (canvassers attempted to deliver an encouragement to recycle).[4] Based on the results presented below, calculate the following:

Table 4: Question 11 Table

Treatment assignment	Treated?	N	Turnout
Baseline	No	2572	0.3122
Treatment	Yes	486	0.3909
	No	2086	0.3274
Placebo	Yes	470	0.2979
	No	2109	0.3215

a) Estimate the proportion of Compliers based on subjects' responses to the treatment. Estimate the proportion of Compliers based on subjects' responses to the placebo. Assuming that the individuals are assigned randomly to the treatment and placebo groups, are these rates of compliance consistent with the null hypothesis that both groups have the same proportion of Compliers?


```
Z <- c(rep("baseline", 2572), rep("treatment", 486+2086), rep("placebo", 470+2109))
D <- c(rep(0, 2572), rep(1, 486), rep(0,2086), rep(1, 470), rep(0,2109))
Y <- c(rep(1, round(2572*0.3122)), rep(0,round(2572*(1-0.3122))),
       rep(1, round(486*0.3909)), rep(0,round(486*(1-0.3909))),
       rep(1, round(2086*0.3274)), rep(0,round(2086*(1-0.3274))),
       rep(1, round(470*0.2979)), rep(0,round(470*(1-0.2979))),
       rep(1, round(2109*0.3215)), rep(0,round(2109*(1-0.3215))))
pr.c.treatment <- mean(D[Z=="treatment"])
pr.c.treatment

## [1] 0.189

pr.c.placebo <- mean(D[Z=="placebo"])
pr.c.placebo

## [1] 0.1822
```

The estimated proportion of compliers in the vote encouragement group is 0.189. The estimated proportion in the placebo group is 0.182. The difference between these rates is negligible. These rates of compliance are consistent with the null hypothesis that both groups have the same proportion of Compliers.

[4] Nickerson 2005, 2008.

b) Do the data suggest that Never-Takers in the treatment and placebo groups have the same rate of turnout? Is this comparison informative?

```
rate.nt.treatment <- mean(Y[Z=="treatment" & D==0])
rate.nt.placebo <- mean(Y[Z=="placebo" & D==0])
rate.nt.treatment

## [1] 0.3274

rate.nt.placebo

## [1] 0.3215
```

Yes, the turnout rate among the encouragement never takers is 32.7% versus 32.2% for the placebo group. Under random assignment and equivalent compliance to treatment and placebo contact, placebo and the encouragement groups have the same expected average potential outcomes when untreated. If the observed difference in average potential outcomes when untreated is too large, we may reject the maintained hypothesis that the group is formed by random draws from a common pool of subjects.

c) Estimate the CACE of receiving the placebo. Is this estimate consistent with the substantive assumption that the placebo has no effect on turnout?

```
itt.placebo <- mean(Y[Z=="placebo"]) - mean(Y[Z=="baseline"])
cace.placebo <- itt.placebo/pr.c.placebo
```

The CACE is 0.027. The placebo has an unexpectedly positive effect on turnout (although further analysis shows that the effect is not larger than one would expect due to random sampling variability).

d) Estimate the CACE of receiving the treatment using two different methods. First, use the conventional method of dividing the ITT by the ITT_D. Second, compare turnout rates among Compliers in both the treatment and placebo groups. Interpret the results.

```
## Method 1
itt.treatment <- mean(Y[Z=="treatment"]) - mean(Y[Z=="baseline"])
cace.treatment1 <- itt.treatment/pr.c.treatment
cace.treatment1

## [1] 0.1440329

## Method 2
cace.treatment2 <- mean(Y[Z=="treatment" & D==1]) - mean(Y[Z=="placebo" & D==1])
cace.treatment2

## [1] 0.09307416
```

Using the ITT and the compliance rate, the estimated average treatment effect for the compliers is a 14.4 percentage point increase in turnout. Comparing the compliers when treated and when untreated (assuming compliance with the placebo isolates the same group of subjects as compliance with the encouragement and the placebo has no effect of Y(0) for compliers), the estimated CACE is 9.3 percentage points.

The two methods arrive at similar estimates of the CACE. Because Method 1 involves a ratio estimator, it is biased but consistent. Method 2 is both unbiased and consistent. As noted above, the (chance) higher turnout in the placebo group makes Method 2 generate smaller estimates of the CACE in this case.

Question 12

Imagine a math tutoring program that involves two day-long sessions with instructors. Given the likely possibility that some of the students who are randomly assigned to the program will attend only the first of the two sessions, suppose that administrators are primarily interested in finding out whether two sessions improve performance on end-of-year tests but are also interested in assessing the effectiveness of the first session alone.

a) Propose an experimental design that addresses the possibility that some students will only attend the first session.
Answer:
Anticipating that some subjects would drop out after a day if assigned to a two-day treatment, randomize subjects into 3 groups: Those who are assigned to a two day treatment, those who are assigned to a one day treatment, and those who are assigned to control.

b) Show that your experimental design is capable of identifying the causal effects of the full program and the abbreviated program. Answer:
Define 3 types of subjects: 2 day compliers, 1 day compliers, and Never-takers. 2 day compliers fully comply with their treatment assignment. 1 day compliers only attend 1 day if assigned either to 2 or 1 day of treatment. Never-takers attend 0 days of treatment regardless of assignment. Assume that showing up on day one is unaffected by assignment conditional on being a 1 day complier or a 2 day complier.

Expected Perfomance in Control (EP, Control) =

$$E(Y(0)|\text{ never-taker}) * Pr(\text{never-taker}) +$$
$$E(Y(0)|\text{ 1 day compliers}) * Pr(\text{1 day compliers}) +$$
$$E(Y(0)|\text{ 2 day compliers}) * Pr(\text{2 day compliers})$$

Expected Perfomance in 1-day group (EP, 1-day) =

$$E(Y(0)|\text{ never-taker}) * Pr(\text{never-taker}) +$$
$$E(Y(1)|\text{ 1 day compliers}) * Pr(\text{1 day compliers}) +$$
$$E(Y(1)|\text{ 2 day compliers}) * Pr(\text{2 day compliers})$$

Expected Perfomance in 2-day group (EP, 2-day) =

$$E(Y(0)|\text{ never-taker}) * Pr(\text{never-taker}) +$$
$$E(Y(1)|\text{ 1 day compliers}) * Pr(\text{1 day compliers}) +$$
$$E(Y(2)|\text{ 2 day compliers}) * Pr(\text{2 day compliers})$$

$$\text{CACE of 1-day} = \frac{(EP, 1-day) - (EP, Control)}{Pr(\text{1 \& 2 day compliers})}$$
$$\text{CACE of the second day} = \frac{(EP, 2-day) - (EP, 1-day)}{Pr(\text{2 day compliers})}$$

We can identify two quantities: the average effect of the one-day program for all subjects except never-takers, and the average effect of the second day for the subgroup of 2-day compliers. These can be estimated using group average outcomes, and the proportions of the three types. The proportion of 2-day compliers is estimated with the full 2-day compliance rate in the 2-day group. The proportion of 1 day compliers is the compliance rate in the 1-day group minus the proportion of 2-day compliers.

If we are willing to assume that the CACE of 1-day is the same for 1 and 2 day compliers, then the effect of the entire program can be estimated by adding the CACE of the second day to the CACE of the 1-day program. This assumption might be plausible or not, depending on the program.

c) Suppose that one-sided noncompliance occurred in the following way: everyone in the treatment group received the treatment, and some of the control group was inadvertently treated. Show that by modifying the CACE theorem (for example, replacing Never-Takers with Always-Takers) one can still identify the CACE in this case.
Answer:
Define compliers as those who comply with their treatment assignment and always takers as those who always reveal their treated potential outcome, regardless of their treatment assignment.

Expected Perfomance in Control (EP, Control) =

$$E(Y(1)|\text{ always-taker}) * Pr(\text{always-takers}) +$$
$$E(Y(0)|\text{ complier}) * Pr(\text{compliers})$$

Expected Perfomance in Treatment (EP, Treatment) =

$$E(Y(1)|\text{ always-taker}) * Pr(\text{always-takers}) +$$
$$E(Y(1)|\text{ complier}) * Pr(\text{compliers})$$

$$\text{CACE} = \frac{(EP, Treatment) - (EP, Control)}{Pr(\text{compliers})}$$

The proportion of compliers can be estimated as one minus the proportion of those assigned to the control group who take the treatment, i.e., the proportion of always-takers.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 6 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

The following three quantities are similar in appearance but refer to different things. Describe the differences.

- $E[Y_i(d(1))|D_i = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects who would actually receive the treatment ($D_i = 1$).

- $E[Y_i(d(1))|d_i(1) = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $D_i(1)$ for the subgroup of subjects who received the treatment when assigned to it ($d_i(1) = 1$). In the case of one-sided non-compliance, this subgroup is the Compliers.

- $E[Y_i(d(1))|d_i(1) = d_i(0) = 1]$
  Answer:
  This expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $d_i(1)$ for the subgroup of subjects known as Always-Takers, who always receive the treatment regardless of whether they are assigned to the treatment group ($d_i(1) = d_i(0) = 1$).

Question 2

The following expression appears in the proof of the CACE theorem. Interpret the meaning of each term in the expression, and explain why the expression as a whole is equal to zero: $E[Y_i(d(1))|d_i(1) = d_i(0) = 0] - E[Y_i(d(0))|d_i(1) = d_i(0) = 0]$.
Answer:
The first expression refers to the expected potential outcome of Y_i given the treatment received by the assigned treatment group $d_i(1)$ for the subgroup of subjects known as Never-Takers, who never receive the treatment regardless of whether they are assigned to the treatment group ($d_i(1) =$

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

$d_i(0) = 0$). The second expression refers to the expected potential outcome of Y_i given the treatment received by the assigned control group $d_i(0)$ for the subgroup of subjects known as Never-Takers, who never receive the treatment regardless of whether they are assigned to the treatment group ($d_i(1) = d_i(0) = 0$). These are equal because Never-Takers reveal the same potential outcome regardless of treatment assignment.

Question 3

Assuming that the excludability and non-interference assumptions hold, are the following statements true or false? Explain your reasoning.

a) Among Compliers, the ITT equals the ATE.
Answer:
True. For Compliers, treatment assigned equals treatment received, and so ITT = ATE.

b) Among Defiers, the ITT equals the ATE.
Answer:
False: For Defiers, treatment assigned is the opposite of treatment received, and so ITT = -ATE.

c) Among Always-Takers and Never-Takers, the ITT and ATE are zero.
Answer:
False. For Always-takers and Never-takers, the ITT is zero because they respond the same to both experimental assignments. The ATE among these subgroups may not be nonzero; the ATE is not revealed empirically.

Question 4

When analyzing experiments with two-sided noncompliance, why is it incorrect to define Compliers as "those who take the treatment if assigned to treatment"?
Answer:
This definition should be "those who take the treatment if AND ONLY IF assigned to treatment." The definition given in the question would also hold for Always-takers, who take the treatment if assigned to treatment.

Question 5

Suppose that a sample contains 30% Always-Takers, 40% Never-Takers, 15% Compliers, and 15% Defiers. What is the ITT_D?
Answer:
Recall from equation (6.19): $ITT_D = \pi_C + \pi_A T - (\pi_D + \pi_A T) = \pi_C - \pi_D$, which in this case implies that the $ITT_D = 0$.

Question 6

Suppose that, in violation of the monotonicity assumption, a sample contains both Compliers and Defiers. Let π_C be the proportion of subjects who are Compliers, and let π_D be the proportion of subjects who are Defiers. Show that the CACE is nevertheless identified if (i) the ATE among

Defiers equals the ATE among Compliers and (ii) $\pi_C \neq \pi_D$.
Answer:
Let's re-write equation (6.20) assuming that the two ATEs are the same and that the denominator is nonzero:

$$\frac{ITT}{ITT_D} = \frac{(ATECompliers)\pi_C - (ATE|Defiers)\pi_D}{(\pi_C - \pi_D)} = \frac{(ATECompliers)(\pi_C - \pi_D)}{(\pi_C - \pi_D)} = (ATECompliers).$$

Question 7

In experiments with one-sided noncompliance, the ATE among subjects who receive the treatment (sometimes called the average treatment-on-the-treated effect, or ATT) is the same as the CACE, because only Compliers receive the treatment. Explain why the ATT is not the same as the CACE in the context of two-sided noncompliance.
Answer:
Under two-sided noncompliance, both Compliers and Always-takers receive treatment when assigned to the treatment group, and Always-takers receive treatment when assigned to the control group. Therefore, as we move from one-sided to two-sided noncompliance, "the treated" no longer refers to Compliers, and the ATT no longer equals the CACE.

Question 8

In the Milwaukee domestic violence experiment, researchers working in collaboration with police officers randomly assigned one of three treatments when officers responded to an incident involving domestic violence.[1] Officers were instructed to arrest the perpetrator and hold him overnight, arrest the perpetrator but release him after a brief period, or issue a warning. The full breakdown of assigned and actual treatments is presented in Table 6.7, along with observed rates of later arrest in the three treatment conditions.

Table 1: Question 8 Table

		Assigned Treatment		
		Full Arrest	Brief Arrest	Warning
Actual Treatment	Full Arrest	400	13	1
	Brief Arrest	1	384	1
	Warning	3	1	396
	Total N	404	398	398
Subsequent Outcomes	Calls to hotline to report perpetrator	296	301	261
	Perpetrators later arrested	146	157	151

a) Consider a simplified coding of the assigned and actual treatment, dividing subjects into two categories: arrest or non-arrest. Evaluate the plausibility of the non-interference, excludability, and monotonicity assumptions in this application.
Answer:
Non-interference: Potential outcomes reflect only the subject's own treatment status and not the status of other observations. It is possible, though unlikely, that perpetrators discuss their

[1] Sherman et al.1992.

treatments with one another, in which case potential outcomes might be affected not only by one's own treatment but whether it seems severe or lenient vis-a-vis other treatments that other subjects have received.
Excludability: Potential outcomes respond solely to receipt of the treatment and not the random assignment of the treatment or any indirect byproduct of random assignment (e.g., other actions that the responding police officer takes in addition to or instead of issuing a warning or making an arrest; for example, a police officer who is instructed to make no arrest might say/do other threatening things to compensate for the lenient punishment). There is no reason to believe that compensatory actions were taken in this study.
Monotonicity implies that there are no subjects who would be arrested if assigned to non-arrest and not arrested if assigned to arrest. This assumption seems plausible, as it is hard to imagine a scenario by which arrests occur if and only if no arrest is assigned.

b) Assume that the core assumptions hold, and calculate the $\widehat{ITT_D}$, $\widehat{ITT}$, and $\widehat{CACE}$ given the simplified treatment categorization. Interpret the results.

```
pi_at <- 2/398
pi_nt <- (3+1)/(404+398)
pi_c <- 1 - pi_at - pi_nt
itt_d <- pi_c
itt_d

## [1] 0.99

itt_hotline <- (296 + 301)/(404 + 398) - (261/398)
itt_hotline

## [1] 0.08861

cace_hotline <- itt_hotline/itt_d
cace_hotline

## [1] 0.08951

itt_arrest <- (303)/(802) - (151/398)
itt_arrest

## [1] -0.001591

cace_arrest <- itt_arrest/itt_d
cace_arrest

## [1] -0.001608
```

Let's first consider the effects on hotline calls. The estimated CACE of 0.09 means that among Compliers (those who are arrested if and only if assigned to arrest), an actual arrest appears

to increase the probability of subsequent hotline calls by 9 percentage points. The estimated CACE of -0.002 for subsequent arrests means that among Compliers, actual arrest decreases the probability of subsequent arrest by a mere 0.2 percentage points.

c) Suppose monotonicity were not assumed. What do the results of the simplified treatment suggest about the maximum and minimum values of π_{NT}, π_{AT}, and π_C?
Answer:
Without assuming monotonicity, we return to what the observable quantities imply about the latent groups' shares of the subject pool.

Subjects who were treated when assigned to control: $\hat{\pi_{AT}} + \hat{\pi_D} = \frac{2}{398} = 0.00503$
Subjects who were not treated when assigned to control: $\hat{\pi_{NT}} + \hat{\pi_C} = \frac{396}{398} = 0.99497$
Subjects who were treated when assigned to treatment: $\hat{\pi_{AT}} + \hat{\pi_C} = \frac{401+397}{404+398} = 0.99501$
Subjects who were not treated when assigned to treatment: $\hat{\pi_{NT}} + \hat{\pi_D} = \frac{4}{404+398} = 0.00499$

The lower bounds for shares of Defiers, Always-takers, or Never-takers is 0. The last line enables us to put an upper bound on the share of Defiers and Never-takers: 0.00499, or 0.499%. The first line enables us to put an upper bound on the share of Always-takers: 0.503%. If Never-takers were as high as 0.499%, the share of Compliers cannot be lower than 0.99497 - 0.00499 = 0.98998. If there were no Never-takers, the maximum share of Compliers is 99.497%.

d) More complexity is introduced when we consider the full array of three treatment assignments and three forms of actual treatment. In the case of two assigned treatments and two actual treatments, we have four types of subjects (Compliers, Defiers, Never-Takers, and Always-Takers). How many types of subjects are there with three treatment assignments and three forms of actual treatment?
Answer:
There are 27 possible types. See table below.

e) How many types are there if you make the following "monotonicity" stipulations:

(i) Anyone who is fully arrested if assigned to be warned would also be fully arrested if assigned to be briefly arrested or fully arrested

(ii) Anyone who is fully arrested if assigned to be briefly arrested would also be fully arrested if assigned to be fully arrested

(iii) Anyone who is briefly arrested if assigned to be warned would also be briefly arrested if assigned to be briefly arrested

(iv) Anyone who is warned if assigned to be arrested would also be warned if assigned to be warned.

The table below shows all 27 possible combinations of potential outcomes for treatment. After accounting for all four restrictions, 10 types remain.

Table 2: Question 8 Table

	Z = Warning	Z = Brief Arrest	Z = Full Arrest	Monotonicity Constraints				
Type	D(0)	D(1)	D(2)	i	ii	iii	iv	all four
1	0	0	0					
2	0	0	1					
3	0	0	2					
4	0	1	0					
5	0	1	1					
6	0	1	2					
7	0	2	0		X			X
8	0	2	1		X			X
9	0	2	2					
10	1	0	0			X	X	X
11	1	0	1			X	X	X
12	1	0	2			X	X	X
13	1	1	0				X	X
14	1	1	1					
15	1	1	2					
16	1	2	0		X	X	X	X
17	1	2	1		X	X		X
18	1	2	2			X		X
19	2	0	0	X			X	X
20	2	0	1	X			X	X
21	2	0	2	X			X	X
22	2	1	0	X			X	X
23	2	1	1	X				X
24	2	1	2	X				X
25	2	2	0	X	X		X	X
26	2	2	1	X	X			X
27	2	2	2					
Total Types	27	27	27	19	21	21	17	10

Question 9

In their study of the effects of conscription on criminal activity in Argentina, Galiani, Rossi, and Schargrodsky use official records of draft lottery numbers, military service, and prosecutions for a cohort of men born between 1958 and 1962.[2] Draft eligibility is scored 1 if an individual had a draft lottery number that caused him to be drafted, and 0 otherwise. Draft lottery numbers were selected randomly by drawing balls from an urn. Military service is scored 1 if the individual actually served in the armed services, and 0 otherwise. Subsequent criminal activity is scored 1 if the individual had a judicial record of prosecution for a serious offense. For a sample of 5,000 observations, the authors report an $\widehat{ITT_D}$ of 0.6587 (SE = 0.0012), an $\widehat{ITT}$ of 0.0018 (SE = 0.0006), and a $\widehat{CACE}$ of 0.0026 (SE = 0.0008). The authors note that the $\widehat{CACE}$ implies a 3.75% increase in the probability of criminal prosecution with military service.

[2]Galiani, Rossi, and Schargrodsky 2010.

a) Interpret the $\widehat{ITT_D}$, $\widehat{ITT}$, $\widehat{CACE}$, and their standard errors.
Answer:
The $\widehat{ITT_D}$ refers to the difference in rates of military service between the treatment and control groups. Evidently, the treatment group was 65.87 percentage points more likely to serve in the military than the control group. The $\widehat{ITT}$ refers to the difference in prosecution rates between the assigned treatment and control groups (irrespective of whether a subject actually served). The estimate of 0.0018 implies that the treatment group was 0.18 percentage points more likely to be prosecuted than the assigned control group. The $\widehat{CACE}$ is the estimated ATE among Compliers, those who serve in the military if and only if they have a draft-eligible number. This estimate is 0.0026, which implies that Compliers become 0.26 percentage points more likely to be prosecuted as a result of serving in the military. The standard errors are a measure of statistical uncertainty, and a rule of thumb is that a 95% confidence interval may be formed by adding and subtracting +/- 2SEs. In this case, the 95% interval for $\widehat{ITT_D}$ is 65.87 +/- 0.0024; for $\widehat{ITT}$ is 0.0018 +/- 0.0012; for $\widehat{CACE}$, it is 0.0026 +/- 0.0016. The margin of uncertainty for the $\widehat{ITT}$ and $\widehat{CACE}$ is fairly wide, but the intervals are on the positive side of zero, suggesting that military service (if the exclusion restriction holds) has a criminogenic effect.

b) The authors note that 4.21% of subjects who were not draft eligible nevertheless served in the armed forces. Based on this information and the results shown above, calculate the proportion of Never-Takers, Always-Takers, and Compliers under the assumption of monotonicity.
Answer:
Monotonicity means that the proportion of Defiers is zero. The 4.21% who served without being drafted implies that Always-takers are 4.21% of the subject pool. From the $\widehat{ITT_D}$ of 0.6587 we infer the Compliers are 65.87% of the subject pool. That leaves 1 - 4.21% - 65.87% = 29.9% who are Never-takers.

c) Discuss the plausibility of the monotonicity, non-interference, and excludability assumptions in this application. If an assumption strikes you as implausible, indicate whether you think the $\widehat{CACE}$ is biased upward or downward.
Answer:
Let's analyze each assumption. Monotonicity implies no Defiers. Defiers are those who serve in the military if and only if they are not drafted. Given that one ordinarily think of people who join the military on their own volition as being willing to go if drafted, it is difficult to imagine that many people fit this description, so this assumption seems plausible. Random assignment implies that treatment assignment is independent of the potential outcomes. Although some lotteries are implemented incompetently or corruptly, we are given no reason to suspect that here. Non-interference means that potential outcomes reflect only the treatment or control status of the subject in question and do not depend on the status of other observations. In this case the potential outcome is whether a subject will be prosecuted. It seems possible that one's criminal career could be shaped by whether one's friends are or aren't drafted, but it is not clear how this violation of non-interference would bias the results, since if my friends are drafted it might make me more likely to engage in criminal conduct regardless of whether I am assigned to treatment or control. Excludability means that potential outcomes respond solely to receipt of the treatment (military service) and not the random assignment of the treatment or any indirect byproduct of random assignment (e.g., draft dodging). If citizens who are drafted are more easily monitored (e.g., their finger prints are recorded) then there might be an upward bias in the measurement of the crime committed by those assigned to treatment simply because it is easier to solve a crime committed by them.

Question 10

In her study of election monitoring in Indonesia, Hyde randomly assigned international election observers to monitor certain polling stations.[3] Here, we consider a subset of her experiment where approximately 20% of the villages were assigned to the treatment group. Because of difficult terrain and time constraints, observers monitored 68 of the 409 polling places assigned to treatment. Observers also monitored 21 of the 1,562 stations assigned to the control group. The dependent variable here is the number of ballots that were declared invalid by polling station officials.

a) Is monotonicity a plausible assumption in this application?
Answer:
Monotonicity implies no Defiers. Defiers in this context are polling stations that are monitored if and only if they are assigned to the control group. Monotonicity seems a plausible assumption if one imagines that polling stations are monitored in the control group because monitors are tourists who like to monitor spots that are close to tourist attractions. These attractions would also draw their attention if the polling stations in question were in the treatment group.

b) Under the assumption of monotonicity, what proportion of subjects (polling locations) would you estimate to be Compliers, Never-Takers, and Always-Takers?
Answer:
Monotonicity implies that Defiers make up 0% of the subject pool. Always-takers make up $\hat{\pi_{AT}} = \frac{21}{1562} = 0.013$ or 1.3%. Compliers make up $\hat{\pi_C} = \frac{68}{409} - \frac{21}{1562} = 0.153$ of 15.3%. Therefore, Never-takers make up 83.4%.

c) Explain what the non-interference assumption means in the context of this experiment.
Answer:
Non-interference means that each polling station's potential outcomes respond only to whether the polling station itself is treated. This assumption would be jeopardized if monitors, for example, displace corruption when they monitor nearby polling stations. Under that scenario, the "untreated" potential outcome may rise when neighboring stations are treated, causing bias when treated and untreated stations are compared in order to gauge the ATE of treatment vs. no treatment.

d) Download the sample dataset at http://isps.research.yale.edu/FEDAI and estimate the ITT and the CACE. Interpret the results.

```
Z <- as.integer(hyde$Sample) -1    # monitoring treatment
Y <- hyde$invalidballots
D <- as.numeric(hyde$observed=="yes")
probs <- genprobexact(Z)
ITT <- estate(Y,Z,prob=probs)
ITTd <- estate(D,Z,prob=probs)
CACE <- ITT/ITTd
ITT

## [1] 4.824

CACE
```

[3] Hyde2010.

```
## [1] 31.57
```

The ITT is estimated by comparing means in the assigned control 81.33 and treatment groups 86.16, for a difference of 4.82: assignment to monitoring appears to increase the number of invalid ballots by 4.82 per polling station. The CACE is estimated by dividing the ITT by the ITTd, calculated above: 31.57. Assuming non-interference, excludability, and monotonicity, this estimate implies that an actual visit by observers causes an increase of 31.57 invalid ballots among Compliers (polling stations that are observed if and only if assigned to treatment).

e) Use randomization inference to test the sharp null hypothesis that there is no intent-to-treat effect for any polling location. Interpret the results. Explain why testing the null hypothesis that the ITT is zero for all subjects serves the same purpose as testing the null hypothesis that the ATE is zero for all Compliers.

```
perms <- genperms(Z,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least Inf to perform exact estimation.

Ys <- genouts(Y,Z,ate=0)
distout <- gendist(Ys,perms,prob=probs)
p.value.twosided <- mean(abs(distout) >= abs(ITT))
p.value.twosided

## [1] 0.4973
```


Testing the null hypothesis of zero ITT for all subjects, we generated 10,000 randomizations and compared the observed ITT to the sampling distribution of simulated ITTs. We obtained a two-tailed p-value of 0.5 because the observed value was larger than approximately 50% of the simulated ITTs. We therefore cannot reject the sharp null hypothesis of no effect for any unit. Testing the null that the ITT is zero for all subjects is the same as testing the null that the CACE is zero for all compliers because the ITT is in the numerator of the CACE.

Question 11

A large-scale experiment conducted between 2002 and 2005 assessed the effects of Head Start, a preschool enrichment program designed to improve school readiness.[4] The assigned treatment encouraged a nationally representative sample of eligible (low-income) parents to enroll their four-year-olds in Head Start. Of the 1,253 children assigned to the Head Start treatment, 79.8% actually enrolled in Head Start; 855 of the children assigned to the control group (13.9%) nevertheless enrolled in Head Start. One of the outcomes of interest is pre-academic skills, as manifest at the end of the yearlong intervention. The principal investigators report that scores averaged 365.0 among students assigned to the treatment group and 360.5 among students assigned to the control group, with a two-tailed p-value of .041. Two years later, students completed first grade. Their

[4] Puma et al. 2010. We focus here on one part of the study, the sample of four-year-old subjects.

first grade scores on a test of academic skills averaged 447.7 in the treatment group and 449.0 in the control group, with a two-tailed p-value of 0.380.

a) Estimate the CACE for this experiment, using pre-academic skills scores as the outcome.
Answer:
The estimated CACE is: $\widehat{CACE} = \frac{365-360.5}{0.798-0.139} = 6.28$

b) Estimate the CACE for this experiment, using academic skills in first grade as an outcome.
Answer:
The estimated CACE is: $\widehat{CACE} = \frac{447.7-449.0}{0.798-0.139} = -1.97$

c) Estimate the average downstream effect of pre-academic skills on first grade academic skills. Hint: Divide the estimated ITT (from a regression of first grade academic skills on assigned treatment) by the estimated ITT_D (from a regression of pre-academic skills on assigned treatment). Interpret your results. Are the assumptions required to identify this downstream effect plausible in this application? If not, would you expect the apparent downstream effect to be overestimated or underestimated?
Answer:
The estimated downstream CACE is: $\widehat{CACE} = \frac{447.7-449.0}{365-360.5} = -0.29$
The results suggest, surprisingly, that an improvement in pre-academic skills among Compliers (those whose pre-academic skills change if they are exposed to the treatment) led to a deterioration of academic skills in first grade. For every one-point gain in pre-academic skills, there was a 0.29 drop in first grade skills. Ordinarily, one would expect a positive relationship (building early skills help build skills later on). One possible explanation for this anomalous result is sampling variability. Another is a violation of the exclusion restriction. Suppose, for the sake of argument, that Head Start teachers were coaching students to help them perform better on tests of pre-academic skills. (One could define this sort of teaching-to-the-test as the effect of Head Start, in which case there would be no excludability violation.) Suppose that coaching boosts pre-academic skills scores but lowers first grade scores because the same tricks that are used on the pre-academic skills test lower grades on the first grade test. The excluded factor of coaching boosts the denominator and lowers the numerator, and so the net bias is difficult to predict.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 7 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

a) Equation (7.1) describes the relationship between potential missingness and observed missingness. Explain the notation used in the expression $r_i = r_i(0)(1 - z_i) + r_i(1)z_i$.
Answer:
The variable r_i represents whether a given observation is actually observed ($r_i = 1$) or not ($r_i = 0$). The potential outcomes $r_i(1)$ and $r_i(0)$ refer to whether a given observation would be observed if assigned to the treatment group or the control group, respectively. When $Z_i = 0$, the revealed outcome is $r_i = r_i(0)$, and when $Z_i = 1$, the revealed outcome is $r_i = r_i(1)$. The expression above is analogous to the "switching equation" that maps potential outcomes to revealed outcomes via the realized treatment assignment – depending on the treatment assignment, subjects reveal their $r_i(1)$ or $r_i(0)$.

b) Explain why the assumption that $Y_i(z) = Y_i(z, r(z) = 1) = Y_i(z, r(z) = 0)$ amounts to an "exclusion restriction."
Answer:
An exclusion restriction is an assumption that says that a given input variable has no effect on a potential outcome. In this example, the input variable $r_i(z)$, which indicates whether outcomes will be observed given a treatment assignment, has no effect on the potential outcomes of $Y_i(z)$.

c) What is an "If-Treated-Reporter"?
Answer:
An If-Treated-Reporter is a subject that whose outcomes are observed if and only if they are assigned to the treatment group. For this type of subject $r_i(1) = 1$ and $r_i(0) = 0$.

d) What are extreme value bounds?
Answer:
Extreme value bounds indicate the largest and smallest estimates one would obtain if one were to substitute the largest or smallest possible outcomes in place of missing data in each experimental group.

Question 2

Suppose that $r_i(1) = r_i(0)$ for all subjects in an experiment. In other words, all subjects are either Always-Reporters or Never-Reporters. Show that when the treatment effect is the same for all

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

subjects, the difference-in-means for subjects with observable outcomes shown in equation (7.6) is the same as the overall ATE in equation (7.5).
Answer:
Assuming $r_i(1) = r_i(0)$, we substitute for equation (7.5):

$$E[r_i(1)] * E[Y_i(1)|r_i(1) = 1] + (1 - E[r_i(1)]) * E[Y_i(1)|r_i(1) = 0] -$$
$$E[r_i(1)] * E[Y_i(0)|r_i(1) = 1] - (1 - E[r_i(1)]) * E[Y_i(0)|r_i(1) = 0] =$$
$$E[Y_i(1) - Y_i(0)|(r_i(1) = 1)]$$

which is what we get when we make the same substitution into equation (7.6).

Question 3

Construct a hypothetical schedule of potential outcomes to illustrate each of these cases:

a) The proportion of missing outcomes is expected to be different for the treatment and control groups, yet the difference-in-means estimator is unbiased when applied to observed outcomes in the treatment and control groups.

$Y_i(0)$	$Y_i(1)$	$r_i(0)$	$r_i(1)$
4	0	1	0
5	5	1	1
6	4	1	1
2	5	0	1
3	6	0	1

Using the general formula for the ATE,

$$E[r_i(1)] * E[Y_i(1)|r_i(1) = 1] + (1 - E[r_i(1)]) * E[Y_i(1)|r_i(1) = 0] -$$
$$E[r_i(1)] * E[Y_i(0)|r_i(1) = 1] - (1 - E[r_i(1)]) * E[Y_i(0)|r_i(1) = 0] =$$
$$0.8 * 5 + 0.2 * 0 - 0.6 * 5 + 0.4 * 2.5 = 0$$

In this special case, calculating the ATE among the non-missing did not lead to biased estimates of the ATE among the entire subject pool.

b) The proportion of missing outcomes is expected to be the same for the treatment and control groups, yet the difference-in-means estimator is biased when applied to observed outcomes in the treatment and control groups.

$Y_i(0)$	$Y_i(1)$	$r_i(0)$	$r_i(1)$
4	0	1	0
5	5	1	1
6	4	1	1
2	5	1	1
3	6	0	1

Using the general formula for the ATE,

$$E[r_i(1)] * E[Y_i(1)|r_i(1) = 1] + (1 - E[r_i(1)]) * E[Y_i(1)|r_i(1) = 0] -$$
$$E[r_i(1)] * E[Y_i(0)|r_i(1) = 1] - (1 - E[r_i(1)]) * E[Y_i(0)|r_i(1) = 0] =$$
$$0.8 * 5 + 0.2 * 0 - 0.8 * 4.25 + 0.2 * 3 = 0$$

Focusing solely on the non-missing values gives us $E[Y_i(1)|r_i(1) = 1) - Y_i(0)|(r_i(0) = 1)]$ or $54.25 = 0.75$, which is biased.

Question 4

Construct a hypothetical schedule of potential outcomes for $Y_i(z)$ and $R_i(z)$ to show that under some random assignments, a researcher may estimate extreme value bounds that do not encompass the true ATE.

$Y_i(0)$	$Y_i(1)$	$r_i(0)$	$r_i(1)$
3	4	0	1
3	4	0	1
3	4	1	1
8	9	1	1

From the table we see that the ATE is 1. Now let's assume that subject 1 and 3 are assigned to treatment and subject 2 and 4 to control. Subject 2âĂŹs outcome is missing. Let's assume that the outcome measure can range from 0 to 10, in which case the extreme value bounds substitute 0 or 10.

$$ATE = 1$$
$$ATE_{max} = \frac{4+4}{2} - \frac{0+8}{2} = 0$$
$$ATE_{min} = \frac{4+4}{2} - \frac{10+8}{2} = -5$$

This example reminds us that the extreme value bounds are estimates that vary according to the particular randomization; they are not logical bounds on the minimum and maximum values of the ATE.

Question 5

Suppose you were to encounter missingness in the course of conducting an experiment. You look for clues about the causes and consequences of missingness by conducting three lines of investigation: (1) assessing whether rates of missingness differ between treatment and control groups, (2) assessing whether covariates predict which subjects have missing outcomes, and (3) assessing whether the predictive relationship between missingness and covariates differs between treatment and control groups. In what ways would these three lines of investigation inform the analysis and interpretation of your experiment?
Answer:

The value of each analysis depends in part on the researcher's interpretation of why attrition occurs. If, for example, the researcher's hypothesis is that attrition occurs for reasons that are effectively random (e.g., administrative oversights), the three analyses might be informative. If rates of missingness are similar across experimental groups and covariates that predict the (observed) outcome are weakly related to missingness, the researcher's MIPO interpretation gains credence.(The limitations of these tests should also be kept in mind: the covariates cannot speak definitively to the question of how unobserved potential outcomes are related to missingness.) Alternatively, a researcher might posit that missingness is systematic (and therefore likely to be related to covariates) yet posit that missingness is symmetric across experimental groups in the sense that the sample contains Always-Reporters and Never-Reporters. The researcher aspires to estimate the ATE among Always-Reporters and looks for signs of asymmetry in rates of attrition (test 1) and predictors of attrition (test 3). Although these tests cannot establish that the hypothesis is true, our degree of belief in the hypothesis grows if neither test shows signs of asymmetry.

Question 6

From the online appendix (http://isps.research.yale.edu/FEDAI), download the data used in the Angrist, Bettinger, and Kremer article.[1] Using the voucher treatment and two covariates (sex and valid phone number), develop a linear regression model that predicts nonmissingness. Use the predicted values from this model to generate inverse probability weights, taking care to verify that predicted values are nonnegative and not greater than 1.0. Run a weighted regression of reading test scores on winning the voucher, using inverse probability scores as weights. Interpret the estimates.


```
angrist_s <- subset(angrist, age>=9 & age <= 25 & checkid==1 )
angrist_s <- within(angrist_s,{
  read[is.na(read)] <- 0
  sex <- sex_name
  observed <- 1 - (read == 0)
  probobs <- glm(observed~(vouch0*sex)+(vouch0*phone)+(vouch0*age),
                 family=binomial(link="logit"))$fitted
  weights <- 1/probobs
})

# Verify that all probabilities are less than one and greater than zero
with(angrist_s, {
rbind(summary(probobs[vouch0==0]),
summary(probobs[vouch0==1]))
})

##          Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## [1,] 0.005258 0.09059 0.2953 0.3022  0.4137 0.8876
## [2,] 0.006938 0.23770 0.4494 0.3758  0.5037 0.8721

# Coefficients for unweighted regression (restricting analysis to observed subjects)
lm(read~vouch0, data=subset(angrist_s, observed==1))$coefficients
```

[1] Angrist, Bettinger, and Kremer 2006.

```
## (Intercept)      vouch0
##  46.9208148   0.6827378

# Coefficients for IPW regression (restricting analysis to observed subjects)
lm(read~vouch0, weights=weights, data=subset(angrist_s, observed==1))$coefficients

## (Intercept)      vouch0
##  46.4378182   0.7230303
```

The estimated ATE from the weighted regression is 0.68, which is very similar to the unweighted estimate. This estimate suggests that assignment to the voucher increased reading scores by an average of 0.68 scale points (which is fairly small given a standard deviation of 5.6). None of the probabilities used for weights is outside the 0-1 range; the weights vary from 0.26 to 0.42.

Question 7

Sometimes experimental researchers exclude subjects from their analysis because the subjects (1) appear to understand what hypothesis the experiment is testing, (2) seem not to be taking the experiment seriously, or (3) fail to follow directions. Discuss whether each of these three practices is likely to introduce bias when the researcher compares average outcomes among non-excluded subjects.

Answer:

Each of these practices may produce biased estimates. Subjects who "understand what hypothesis the experiment is testing" may have distinctive potential outcomes; discarding these observation may lead to bias, especially if they are more likely to suspect the hypothesis when assigned to the treatment group. Subjects who seem to not be taking the experiment seriously or fail to follow directions may also have distinctive potential outcomes, and behavior that might cause them to be expelled may differ depending on experimental assignment.

Question 8

Ditlmann and Lagunes report the results of an experiment in which Hispanic and non-Hispanic confederates attempted to use a personal check to purchase $10 gift certificates at 217 retail stores.[2] Confederates, who were trained to behave in a similar manner, were randomly assigned to each store. One of the outcome measures is whether the retail clerk asks to see the confederate's photo identification. A second outcome is whether, for those who were asked to present identification, the identification card (which was supplied by the experimenters) was accepted as valid. Suppose the question of interest were: Are clerks more likely to accept the identification card when it is presented by a white or Hispanic shopper? Because some shoppers were never asked to present identification, their outcomes are missing. Define the treatment as 0 if non-Hispanic and 1 if Hispanic. Define the request for identification as 0 if no request is made and 1 if a request is made. Define the acceptance of identification as 0 if identification is rejected and 1 if it is accepted. The table below shows the number of retailers who requested and/or accepted identification, by experimental condition.

a) The data seem to suggest that Hispanics who presented identification were more likely to have their IDs accepted than whites who presented identification. Explain why this pattern in the

[2] Ditlmann and Lagunes 2010.

Table 1: Question 8 Table

	White Shopper	Hispanic Shopper
No ID Requested	28	17
ID Requested and Accepted	50	68
ID Requested but Rejected	28	26
Total N	106	111

data may give the misleading impression that retailers discriminate in favor of Hispanics.
Answer:
The problem with this interpretation is that whites were less likely to be asked to present their identification. We therefore do not observe how the clerks would have responded to the 28 white shoppers who were not asked to present identification had they done so (nor do we observe the corresponding outcomes for the 17 Latinos whose identification was not requested).

b) Use extreme value bounds to fill in the missing outcomes (acceptance or rejection of identification) for those subjects who never presented identification. Interpret your results.
Answer:
When using extreme value bounds, we impute a value of 1 to the Latino group's missing values and a value of 0 to the White group's missing values in order to generate the upper bound on the effect of being assigned to the Latino group. The lower bound is obtained by imputing a value of 0 in place of the Latino group's missing values and a value of 1 in place of the White group's missing values:

$$ATE_{upper} = \frac{68+17}{111} - \frac{50}{106} = 0.294$$
$$ATE_{lower} = \frac{68}{111} - \frac{50+28}{106} = -0.123$$

The extreme value bounds are estimated to be negative 12.3 percentage points to positive 29.4 percentage points.

c) Is the monotonicity assumption on which trimming bounds rest defensible in this application? Calculate the trimming bounds and interpret the results.
Answer:
In this context, monotonicity implies that and that retail clerks who do not request identification from a Latino customer would not also request it from a white customer. We may express this assumption formally as $r_i(W) \leq r_i(L)$, since a request for identification causes a subject's outcome to be recorded. This restriction excludes clerks whose potential outcomes are $r_i(W) = 1$ and $r_i(L) = 0$ (they request identification from whites but not Latinos). This assumption seems plausible, if we are willing to believe that clerks uniformly believe that whites are a better credit risk than Latinos.

Under monotonicity we can bound the ATE of being assigned to the Latino treatment for Always-Reporters (those who would be asked for identification regardless of their assignment):

$$E[Y_i(L)|r_i(W) = 1; r_i(L) = 1] - E[Y_i(W)|r_i(W) = 1; r_i(L) = 1]$$

The identification problem stems from the fact that our experiment only provides direct evidence about the second quantity, since the only ones who request identification from whites are those with potential outcomes $r_i(W) = 1$ and $r_i(L) = 1$. The first quantity presents more of an empirical challenge because we observe outcomes from two types of retail clerks, those with potential outcomes $r_i(W) = 1$ and $r_i(L) = 1$ and with potential outcomes $r_i(W) = 1$ and $r_i(L) = 1$. However, the first term in the expression above refers only to the potential outcomes among clerks whose potential outcomes are $r_i(W) = 1$ and $r_i(L) = 1$.

Trimming bounds make use of the fact that we can estimate the relative shares of Always-Ask clerks ($r_i(W) = 1$ and $r_i(L) = 1$) and If-Latino-Ask clerks ($r_i(W) = 1$ and $r_i(L) = 1$) in the subject pool. The share of Always-Ask in the entire subject pool can be estimated based on the non-missingness rate in the White treatment group. The difference in non-missingness rates in the two experimental groups estimates the share of subjects who are If-Latino-Ask clerks:

$$Q = \frac{\pi(r_i(L) = 1) - \pi(r_i(W) = 1)}{\pi(r_i(L) = 1)} = \frac{\frac{94}{111} - \frac{78}{106}}{\frac{94}{111}} = 13\%$$

The last step is to place bounds on the average $Y_i(L)$ for Always-Ask clerks. Using formula 7.21:

$$\hat{E}[Y_i(L)|r_i(L) = 1; Y_i(L) > \hat{Y}_i(L, q)] - \hat{E}[Y_i(W)|r_i(W) = 1]$$

where $\hat{Y}_i(L, q)$ refers to the 13^{th} percentile of the distribution of outcomes in the Latino treatment group. We trim (68+26)*13/100 of the 0 outcome (i.e., ID rejected), obtaining an average of 68/(26+68-12)=0.829. From that we subtract the average outcome in the White treatment group in order to obtain the upper bound:

$$ATE_{upper} = 0.829 - \frac{50}{78} = 0.188$$

In order to obtain the lower bound, we instead trim 12 of the 1 outcomes (i.e. ID accepted) to obtain the lower bound estimate:

$$\hat{E}[Y_i(L)|r_i(L) = 1; Y_i(L) < \hat{Y}_i(L, q)] - \hat{E}[Y_i(W)|r_i(W) = 1]$$

The estimate is:

$$ATE_{lower} = \frac{68 - 12}{26 + 68 - 12} - \frac{50}{78} = 0.0419$$

Thus, the trimming bounds range from 4.2 percentage points to 18.8 percentage points.

```r
Z <- c(rep(0, 106), rep(1, 111)) #W =0, H=1
y <- c(rep(NA, 28), rep(1, 50), rep(0, 28),
       rep(NA, 17), rep(1,68), rep(0, 26))

prob.na.treated <- sum(is.na(y[Z==1]))/length(y[Z==1])
prob.na.control <- sum(is.na(y[Z==0]))/length(y[Z==0])

Q <- ((1 - prob.na.treated) - (1- prob.na.control))/(1 - prob.na.treated)

Y.Z1 <- sort(y[Z==1])
Y.Z1.low <- Y.Z1[1:ceiling(length(Y.Z1)*(1-Q))]
Y.Z1.high <- Y.Z1[ceiling(length(Y.Z1)*Q):length(Y.Z1)]

trim <- c(mean(Y.Z1.low) - mean(y[Z==0], na.rm=TRUE),
          mean(Y.Z1.high) - mean(y[Z==0], na.rm=TRUE))
trim

## [1] 0.04190119 0.18824265
```


Question 9

Suppose a researcher studying a developing country plans to conduct an experiment to assess the effects of providing low-income households with cash grants if they agree to keep their children in school and take them for regular visits to health clinics. The primary outcome of interest is whether children in the treatment group are more likely to complete high school. A random sample of 1,000 households throughout the country is allocated to the treatment group (cash grants), and another sample of 1,000 households is allocated to the control group.

a) Suppose that halfway through the project, a civil war breaks out in half of the country. Researchers are prevented from gathering outcomes for 500 treatment and 500 control subjects living in the war zone. What are the implications of this type of attrition for the analysis and interpretation of the experiment?
Answer:
In this case, one might suppose that the source of missingness operates the same on the treatment and control subjects, so that the only two latent types in the subject pool are Always-Reporters and Never-Reporters. One may not be able to estimate the ATE for the entire country without assuming $MIPO|X$ and re-weighting the outcomes in the observed section of the country to reflect the covariate profile in the wartorn region. However, if one is content to estimate the ATE for the observed section of the country on the assumption that experimental assignment is unrelated to potential missingness, this type of attrition does not cause bias.

b) Another identical experiment is performed in a different developing country. This time the attrition problem is as follows: households that were offered cash grants are more likely to live at the same address years later, when researchers return in order to measure outcomes. Of the 1,000 households assigned to the treatment group, 900 are found when researchers return to measure outcomes, as opposed to just 700 of the 1,000 households in the control group. What are the implications of this type of attrition for the analysis and interpretation of the experiment?

Answer:
This type of attrition may be a source of bias. Migration (missingness) may be related to potential education outcomes, and the treatment (or lack thereof) may cause some households to relocate. For example, if students with lower potential education outcomes tend to migrate when their incomes are low, the treatment has the effect of causing some lower-performing students to remain in the non-missing sample, thereby reducing the estimated effect of the treatment based on a comparison of non-missing subjects in treatment and control. In this case, a researcher might turn to trimming bounds on the assumption that those who would have been available for an interview if assigned to the control group would also have been available for an interview if assigned to the treatment group.

Question 10

Table 7.6 summarizes the results of a series of simulations. Recall from the discussion of this table in section 7.6 that each simulation considers the accuracy with which conventional and second-round sampling procedures recover the ATE. Based on the results presented in the table, address the following questions.

a) The first six rows of the table consider scenarios in which missingness is unrelated to potential outcomes. Each scenario varies the rate of missingness and the number of observations gathered in the second round of data collection. Compare the two estimators (one based on initial data collection only, the other based on both rounds of data collection) in terms of bias, precision, and the width of the extreme value bounds.
Answer:
The difference-in-means estimates based on the first round of data collection are the same for each of the simulations; what varies are the results of the second round of data collection. Because missingness is unrelated to potential outcomes, the point estimates are nearly identical across all simulations (all are unbiased), regardless of which estimator we choose. The standard errors increase as we move to the second round of data collection, because the two-round estimator is putting extra weight on imprecisely estimated quantities from the second round. Across all pairs of simulations, the SE for the second round is reduced substantially when more observations are gathered in the second round. Even though the two-round approach does not reduce the standard error, it does markedly reduce the spread of the extreme value bounds, even when relatively few observations are gathered in the second round. Second round sampling does better in this regard because a smaller proportion of the sample has its missing values filled in with extreme values.

b) The next six rows of the table consider scenarios in which missingness is related to potential outcomes. Again, compare the two estimators in each scenario in terms of bias, precision, and the width of the extreme value bounds.
Answer:
The first round estimates are now severely biased; in each case, the average estimate is close to zero when the true ATE is 10. Second round sampling is much less biased, especially when the share of missing in the second round is low. Extreme value bounds are much smaller under second round sampling. For example, under scenario (A,B,100), the extreme value bounds for first round sampling are [-0.224,0.276], as opposed to [0.075,0.105] for second round sampling.

c) Perform the same comparisons for the scenarios in the final six rows of the table, in which missingness is related to potential outcomes in the first round but unrelated to potential outcomes

in the second round.
Answer:
Second round sampling tends to perform better when missingness is unrelated to potential outcomes in the second round. For example, under scenario (A,B,100), the extreme value bounds are [0.075,0.105] for second round sampling, whereas for scenario (A,5,100) the extreme value bounds are [0.084,0.109]. The point estimates also become less biased when second round sampling is independent of potential outcomes: 0.091 vs. 0.098. Precision remains about the same.

d) Overall, under which scenario does estimation based on both rounds of data analysis have the greatest comparative advantage over estimation based only on the initial round of data collection? Under which scenario does estimation based only on the initial round of data collection have the greatest comparative advantage over estimation based on both rounds?
Answer:
If one seeks to point estimate the ATE, first round sampling is best (unbiased and, in comparison to second round sampling, more precise) when missingness is independent of potential outcomes. However, since analysts will not know whether outcomes are missing at random, they may turn to extreme value bounds, in which case second round sampling is clearly superior. The superiority of the second round strategy is most apparent when missingness in the first round is related to potential outcomes, and when missingness in the second round is unrelated to potential outcomes.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 8 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts:

a) Interpret the expression $Y_i(\mathbf{d}) = Y_i(d)$ and explain how it conveys the non-interference assumption.
Answer:
The expression $Y_i(d)$ refers to the potential outcome that would be expressed based on the input d, which refers to the treatment that subject i receives. By contrast, $Y_i(\mathbf{d})$ refers to the potential outcome that subject i would express based on the assignments that all of the subjects receive. The equality means that the only input that matters is the treatment that subject i receives.

b) Why are experiments that involve possible spatial spillover effects (such as the example described in section 8.4) said to involve "implicit" clustered assignment?
Answer:
Because certain units are so closely spaced that if a subject receives spillovers from one of the units, it must receive spillovers from all of the units. In that sense, spillovers are assigned as clusters.

c) In what ways might a within-subjects design violate the non-interference assumption?
Answer:
In a within-subjects design, the unit of observation is the time period. Non-interference presupposes that each unit's potential outcomes are solely a function of the treatments administered in that period. Possible violations include the following: subjects in one period are affected by the treatments that they may have received in a previous period; subjects in one period may be affected because they anticipate treatments that will be administered in a subsequent period.

d) What are the attractive properties of a waitlist (or stepped-wedge) design?
Answer:
If the assumptions for unbiased inference are met, this within-subjects design may provide precise estimates of causal effects even when the number of subjects is limited. In terms of implementation, it may be easier to gain the cooperation of subjects or groups administering the treatment that might otherwise object to the use of a control group because everyone in the study eventually receives the treatment.

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

Question 2

National surveys indicate that college roommates tend to have correlated weights. The more one roommate weighs at the end of the freshman year, the more the other freshman roommate weighs. On the other hand, researchers studying housing arrangements in which roommates are randomly paired together find no correlation between two roommates' weights at the end of their freshman year. Explain how these two facts can be reconciled.

Answer:

A correlation describes the relationship between roommates' weights. A correlation might arise even if subjects have no effect on one another; for example, if subjects from similar regional or ethnic backgrounds tend to room together, one might observe a correlation. Random pairing of roommates means that, in expectation, there would be no correlation between their weights unless they either affected one another's weights or were both affected by similar environmental factors.

Question 3

Sometimes researchers are reluctant to randomly assign individual students in elementary classrooms because they are concerned that treatments administered to some students are likely to spill over to untreated students in the same classroom. In an attempt to get around possible violations of the non-interference assumption, they assign classrooms as clusters to treatment and control, and administer the treatment to all students in a classroom.

a) State the non-interference assumption as it applies to the proposed clustered design.

Answer:

The non-interference assumption depends on the estimand. If the aim is to estimate the causal effect of the intervention on individual students, the non-interference assumption is the same as usual, namely, that each student's potential outcomes are affected only by the treatment administered to that subject. If one is concerned about transmission of treatments between students in the same classroom, that concern would still apply to a clustered design, since potential outcomes may be affected by the treatments that other subjects in the same classroom receive. On the other hand, if one is interested in classroom-level treatment effects (i.e., the difference between a 100% treated classroom and a 0% treated classroom), this design sidesteps concerns about within-classroom interference because they are built into the definition of the estimand. In the latter case, the relevant non-interference assumption holds that classroom outcomes are unaffected by the treatment status of other classrooms (e.g., other classrooms in the same school or grade).

b) What causal estimand does the clustered design identify? Does this causal estimand include or exclude spillovers within classrooms?

Answer:

The causal estimand identified by the clustered design is the average effect of a classroom being 100% treated versus 0% treated. This includes within-classroom spillovers at the individual level, but assumes that across-classroom spillovers do not occur.

Question 4

Recall from Chapter 3 (exercise 9) the field experiment conducted by Camerer in which he selected pairs of similar horses running in the same race and randomly placed large wagers on one of them

to see if his bets affected the amount of money that other bettors placed on both horses.

a) Define the potential outcomes in Camerer's study. What non-interference assumption is invoked?
Answer:
Two definitions for potential outcomes are possible.

The outcome variable in this study was defined as the change in total bets that occur between the pre-intervention period and the post-intervention period. Presumably, the original study assumed that potential outcomes $Y_i(d)$ reflect only the treatment assignment of horse i, not the treatment or non-treatment of the other horse in the experimental pair.

Nevertheless, one could define four potential outcomes for each horse: $Y_i(D_i = 1, D_j = 1)$, $Y_i(D_i = 1, D_j = 0)$, $Y_i(D_i = 0, D_j = 1)$, and $Y_i(D_i = 0, D_j = 0)$, which correspond to the the change in bets for horse i depending on whether i, j, both, or neither are treated with experimental bets. Because of the matched pair randomization, horses only ever revealed the second or third of these potential outcomes.

b) What is the causal parameter that this study identifies?
Answer:
Depending on which non-interference assumption is made, the same difference-in-means estimate refers either to the average difference between $Y_i(D_i = 1)$ and $Y_i(D_i = 0)$ or the average difference between $Y_i(D_i = 1, D_j = 0)$ and $Y_i(D_i = 0, D_j = 1)$. The first is the average effect of treatment on the change in bets. The second is the average effect of being the treated horse in a pair in which one horse receives the experimental bets.

Question 5

In their study of spillover effects, Sinclair, McConnell, and Green sent mailings to randomly selected households encouraging them to vote in an upcoming special election.[1] The mailings used a form of "social pressure," disclosing whether the targeted individual had voted in previous elections. Because this type of mail had proven to increase turnout by approximately 4-5 percentage points in previous experiments, Sinclair, McConnell, and Green used it to study whether treatment effects are transmitted across households. Employing a multi-level design, they randomly assigned all, half, or none of the members of each nine-digit zip code to receive mail. For purposes of this example, we focus only on households with one registered voter. The outcome variable is voter turnout as measured by the registrar of voters. The results are as follows. Among registered voters in untreated zip codes, 1,021 of 6,217 cast ballots. Among untreated voters in zip codes where half of the households received mail, 526 of 3,316 registered voters cast ballots. Among treated voters in zip codes where half of the households received mail, 620 of 2,949 voted. Finally, among treated voters in zip codes where every household received mail, turnout was 1,316 of 6,377.

a) Using potential outcomes, define the treatment effect of receiving mail addressed to subject i.
Answer:
The definition of personally receiving mail could be defined in three ways (given our focus on one-voter households). It could be (a) the effect of mail on those whose zip code neighbors receive no mail, (b) the effect of mail on those for whom half of the neighboring households in the zip code receive mail, or (c) the effect of mail on those whose zip code neighbors all receive mail. Given the design of this study, only (b) can be estimated empirically because no one receives mail in an untreated zip code, and everyone receives mail in a 100% treated zip code.

[1] Sinclair, McConnell, and Green 2010.

b) Define the "spillover" treatment effect of being in a zip code where varying fractions of households are treated.
Answer:
Holding constant one's own treatment status, one may define three potential outcomes depending on whether none, half, or all of the neighboring households are treated. When defining the ATE of spillover, one may compare half to none, full to half, or full to none.

c) Propose an estimator for estimating the firsthand and secondhand treatment effects. Show that the estimator is unbiased, explaining the assumptions required to reach this conclusion.
Answer:
The firsthand effects can be estimated only for those in half-treated zip codes by comparing average outcomes among treated and untreated subjects. One can assess the spillover effect among subjects who receive no mail themselves but reside in either half-treated or untreated zip codes. Similarly, one can assess the spillover effect among subjects who received mail themselves and reside in either 100% or 50% treated zip codes. The three assumptions are random assignment (satisfied by design because direct treatments and rates of treatment among neighbors are randomly assigned), non-interference (satisfied if we believe that potential outcomes are solely a function of firsthand treatment and treatment of others in the same zip code; treatment of those outside the zip code is ignored), and excludability (satisfied if we believe that potential outcomes are affected only by firsthand and second hand receipt of mail and not by other factors that might be correlated with treatment assignment).

d) Based on these data, what do you infer about the magnitude of the mailing's direct and indirect effects?


```
Z_ind <- c(rep(0, 6217), rep(0, 3316), rep(1, 2949), rep(1, 6377))
Z_zip <- c(rep("none", 6217), rep("half", 3316), rep("half", 2949), rep("all", 6377))
Y <- c(rep(1, 1021), rep(0, 6217-1021),
       rep(1, 526), rep(0, 3316-526),
       rep(1, 620), rep(0, 2949-620),
       rep(1, 1316), rep(0, 6377-1316))
ate.firsthand.half <-
  mean(Y[Z_ind==1 & Z_zip=="half"]) -
  mean(Y[Z_ind==0 & Z_zip=="half"])
ate.secondhand.untreated <-
  mean(Y[Z_ind==0 & Z_zip=="half"]) -
  mean(Y[Z_ind==0 & Z_zip=="none"])
ate.secondhand.treated <-
  mean(Y[Z_ind==1 & Z_zip=="all"]) -
  mean(Y[Z_ind==1 & Z_zip=="half"])
ate.firsthand.half

## [1] 0.05161591

ate.secondhand.untreated

## [1] -0.00560227
```

```
ate.secondhand.treated

## [1] -0.00387413
```

Here, the firsthand effects can be estimated only for those in half-treated zip codes: 620/2949 - 526/3316 = 0.052, or 5.2 percentage points. One can assess spillover effect by way of two different comparisons. In order to assess the effects of spillover among subjects who receive no mail themselves, compare voting rates for those living in 50% treated zip code to those living in the 0% treated zip code: 526/3316 - 1021/6217 = -0.006, or negative 0.6 percentage points. In order to assess the effects of spillover among subjects who received mail themselves, compare voting rates for those living in 100% treated zip codes to those living in 50% treated zip codes: 1316/6377 - 620/2949 = -0.004, or negative 0.4 percentage points. Although the estimated firsthand effect is strongly positive, both of the estimated spillover effects are close to zero.

Question 6

Using the potential outcomes from the clinic example in Table 8.2, calculate the following estimates.

a) Estimate $E[Y_{01} - Y_{00}]$ for the random assignment that places the treatment at location A.
Answer:
The treated and untreated averages are $\frac{\frac{0}{0.2}}{\frac{1}{0.2}} = 0$ and $\frac{\frac{0}{0.4}+\frac{6}{0.6}+\frac{6}{0.8}}{\frac{1}{0.4}+\frac{1}{0.6}+\frac{1}{0.8}} = 3.23$, respectively. The estimated ATE is -3.23.

b) Estimate $E[Y_{10} - Y_{00}]$ for the random assignment that places the treatment at location A, restricting the sample to the set of villages that have a non-zero probability of expressing both of these potential outcomes.
Answer:
The treated and untreated averages are $\frac{\frac{2}{0.4}}{\frac{1}{0.4}} = 2$ and $\frac{\frac{0}{0.4}+\frac{6}{0.6}}{\frac{1}{0.4}+\frac{1}{0.6}} = 2.4$, respectively. The estimated ATE is -0.4.

c) In order to make a more direct comparison between these two treatment effects, estimate $E[Y_{01} - Y_{00}]$, restricting the sample to the same set of villages as in part (b).
Answer:
The treated and untreated averages are $\frac{\frac{0}{0.2}}{\frac{1}{0.2}} = 0$ and $\frac{\frac{0}{0.4}+\frac{6}{0.6}}{\frac{1}{0.4}+\frac{1}{0.6}} = 2.4$, respectively. The estimated ATE is -2.4.

Question 7

Lab experiments sometimes pair subjects together and have them play against one another in games where each subject is rewarded financially according to the game's outcome. One such game involves making monetary contributions to a public good (e.g., preserving the environment); the game can be arranged such that each player gains financially if both of them make a contribution, but each player is better off still if they contribute nothing while their partner in the game makes a contribution. The treatment is whether the pair of players is allowed to communicate prior to deciding whether to contribute. Suppose that a lab experimenter recruits four subjects and assigns them randomly as pairs to play this game. The outcome is whether each player makes a contribution: Y_i is 1 if the player contributes and 0 otherwise. Each player has three potential

outcomes: Y_{0i} is the outcome if players are prevented from communicating, Y_{1i} is the outcome if a player communicates with another player who is "persuasive," and Y_{2i} is the outcome if a player communicates with another player who is "unpersuasive." The table below shows the schedule of potential outcomes for four players, two of whom are persuasive and two of whom are unpersuasive.

Table 1: Question 7 Table

Subject	Type	Y_{0i}	Y_{1i}	Y_{2i}
1	Persuasive	0	1	0
2	Persuasive	1	1	0
3	Unpersuasive	0	0	0
4	Unpersuasive	1	1	1

a) Calculate the average treatment effect of $Y_{1i} - Y_{0i}$. Calculate the average treatment effect of $Y_{2i} - Y_{0i}$.
Answer:
The ATE of talking with a persuasive person is (3/4) - (1/2) = (1/4). The ATE of talking with an unpersuasive person is (1/4) - (1/2) = -(1/4).

b) How many random pairings are possible with four subjects?
Answer:
There are $4!/(2!(4-2)!) = 6$ pairings.

c) Suppose that the experimenter ignores the distinction between Y_{1i} and Y_{2i} and considers only two treatment conditions: the control condition prevents communication between pairs of players, and the treatment condition allows communication. Call the observed outcomes in the communication condition Y_{1i}^*. Across all possible random pairings of subjects, what is the average difference-in-means estimate when the average Y_{1i}^* is compared to the average Y_{0i} ? Does this number correspond to either of the two estimands defined in part (a)? Does it correspond to the average of these two estimands?
Answer:
The average difference-in-means estimate is $\frac{0-0.5-1+0+0.5+0.5}{6} = -\frac{1}{12}$. This does not correspond to any of the estimands defined in part a), nor does it correspond to the average of this estimands.

d) What is the probability that a persuasive subject is treated by communicating with an unpersuasive subject? What is the probability that an unpersuasive subject is treated by communicating with an unpersuasive subject?
Answer:
Subject 1 has a 1/6 chance of being assigned to communicate with a persuasive subject (subject 2) and has a 1/3 chance of being assigned to communicate with an unpersuasive subject (subjects 3 or 4). The same probabilities apply to Subject 2. Subject 3 has a 1/6 chance of communicating with an unpersuasive subject (subject 4). The same probabilities apply to Subject 4.

e) Briefly summarize why a violation of the non-interference assumption leads to biased difference-in-means estimates in this example.
Answer:
One's potential outcomes change depending on how the randomization happened to come out.

Bias occurs because the probability of encountering a persuasive or unpersuasive partner is related to potential outcomes.

f) Would bias be eliminated if the experimenter replicated this study (with four subjects) each day and averaged the results over a series of 100 daily studies?
Answer:
It depends. Replicating small experiments with the same combination of persuasive and unpersuasive subjects simply reproduces the bias described above, because each experiment is subject to the same bias. On the other hand, if one imagines replicating this study with a random draw of the four subject types (see part G below), no bias results because selecting one subject for treatment does not prevent a subject of the same type from being assigned to control.

g) Would bias be eliminated if the experimenter assembled 400 subjects at the same time (imagine 100 subjects for each of the four potential outcomes profiles in the table) and assigned them to pairs? Hint: Answer the question based on the intuition suggested by part (d).
Answer:
Bias becomes negligible as the size of a given experiment increases, because in a large experiment the probability of encountering a persuasive partner is nearly the same for both persuasive and unpersuasive subjects.

Question 8

Concerns about interference between units sometimes arise in survey experiments. For example, surveys sometimes administer a series of "vignettes" involving people with different attributes. A respondent might be told about a low-income person who is randomly described as white or black; after hearing the description, the respondent is asked to rate whether this person deserves public assistance. The respondent is then presented with a vignette about a second person, again randomly described as white or black, and asked about whether this person deserves public assistance. This design creates four experimental groups: (a) two vignettes about blacks, (b) two vignettes about whites, (c) a black vignette followed by a white vignette, and (d) a white vignette followed by a black vignette. Each respondent provides two ratings.

a) Propose a model of potential outcomes that reflects the ways that subjects might respond to the treatments and the sequences in which they are presented.
Answer:
One could model this scenario as a two-period within-subjects experiment. Call the black vignette $D_i = 1$ and the white vignette $D_i = 0$. Assume that respondents are affected only by the treatments they have received in the past or present; future treatments are irrelevant. For each respondent, the relevant potential outcomes are denoted $Y_{t-1,t}$, where the $t-1$ subscript refers to which (if any) treatment is administered in the preceding time period, t refers to the current time period. This notation allows us to use Y_{01} in period i, for example, to refer to that period's potential outcome if a respondent were given the white vignette in the preceding period and the black vignette in the current period. If no question is asked in the preceding period (because the outcome is the response to the first question), denote the outcome as Y_{*0} or Y_{*1}.

b) Using your model of potential outcomes, define the ATE or ATEs that a researcher might seek to estimate.
Answer:
One estimand is the average difference between Y_{*1} and Y_{*0}, which is the effect of the vignette

on responses to the first question. Another is the average difference between Y_{01} and Y_{00}, which is the effect of race on responses to the second question for those who are asked about whites in the first vignette. Similarly, one might consider the average difference between Y_{11} and Y_{10}, which is the effect of race on responses to the second question for those who are asked about blacks in the first vignette.

c) Suggest an identification strategy for estimating this (these) causal estimand(s) using the data from this experiment.
Answer:
Each of the estimands mentioned in part (b) can be estimated by using the between-subjects part of the design. To estimate $E[Y_{*1} - Y_{*0}]$, compare average responses to the first question among those randomly assigned black or white vignettes. To estimate $E[Y_{01} - Y_{00}]$, compare average responses to the second question among those randomly assigned black or white vignettes who earlier received a white vignette. To estimate $E[Y_{11} - Y_{10}]$, compare average responses to the second question among those randomly assigned black or white vignettes who earlier received a black vignette.

d) Suppose a researcher analyzing this experiment estimates the average "race effect" by comparing the average evaluation of the white recipient to the average evaluation of the black recipient. Is this a sound approach?
Answer:
This approach amounts to pooling the respondents' answers to both questions; if there are N respondents, this analysis analyzes 2N observations. If the three race effects defined above differ, this approach will estimate a weighted average of the three estimands, which may be uninterpretable as a causal effect.

Question 9

Use data from the hotspots experiment in Table 8.4 (these data are also available at http://isps.research.yale.edu/FEDAI) and the probabilities that each unit is exposed to immediate or spillover treatments (Table 8.5) to answer the following questions:

a) For the subset of 11 hotspot locations that lie outside the range of possible spillovers, calculate $E[Y_{01} - Y_{00}]$, the ATE of immediate police surveillance.

```
true_ate <- with(hotspots, mean(y01[prox500==0]) - mean(y00[prox500==0]))
true_ate

## [1] -5

ate_hat <- with(hotspots, mean(y[prox500==0 & assignment==1]) -
                  mean(y00[prox500==0 & assignment==0]))
ate_hat

## [1] 3.333333
```

The true ATE for the observations that lie outside the range of possible spillovers is -5. The estimated ATE using the observed random assignment is 3.33.

b) For the remaining 19 hotspot locations that lie within the range of possible spillovers, calculate $E[Y_{01} - Y_{00}]$, $E[Y_{10} - Y_{00}]$, and $E[Y_{11} - Y_{00}]$.

```
true_ate_01 <- with(hotspots, mean(y01[prox500==1]) - mean(y00[prox500==1]))
true_ate_10 <- with(hotspots, mean(y10[prox500==1]) - mean(y00[prox500==1]))
true_ate_11 <- with(hotspots, mean(y11[prox500==1]) - mean(y00[prox500==1]))
true_ate_01
```

```
## [1] -5
```

```
true_ate_10
```

```
## [1] 5
```

```
true_ate_11
```

```
## [1] -7
```

```
hotspots <- within(hotspots,{
  exposure[exposure == 11] <- "11"  # Indirect and Direct Treatment
  exposure[exposure == 10] <- "10"  # Indirect Treatment
  exposure[exposure == 01] <- "01"  # Direct Treatment
  exposure[exposure == 00] <- "00"  # Control

  # Calculate probability of assignment to exposure condition
  Q <- NA
  Q[exposure == "11"] <- prob11[exposure == "11"]
  Q[exposure == "10"] <- prob10[exposure == "10"]
  Q[exposure == "01"] <- prob01[exposure == "01"]
  Q[exposure == "00"] <- prob00[exposure == "00"]

  # Generate weights
  weights <- 1/Q
})

# Estimate E[Y_{01} - Y_{00}]:
fit.01 <- lm(y ~ exposure, weights=weights,
             subset(hotspots, prox500 > 0 & exposure %in% c("00", "01")))

# Estimate E[Y_{10} - Y_{00}]:
fit.10 <- lm(y ~ exposure, weights=weights,
             subset(hotspots, prox500 > 0 & exposure %in% c("00", "10")))

# Estimate E[Y_{11} - Y_{00}]:
fit.11 <- lm(y ~ exposure, weights=weights,
             subset(hotspots, prox500 > 0 & exposure %in% c("00", "11")))
```

Among observations that lie outside the range of possible spillovers, the ATE of direct treatment is -5, the ATE of indirect treatment is 5, and the ATE of direct and indirect treatment together is -7.

Table 2: Question 9c: Treatment Effect Estimates

	Crime Rate		
	(1)	(2)	(3)
exposure01	−16.033		
	(8.065)		
exposure10		−0.037	
		(9.074)	
exposure11			−9.606
			(7.725)
Constant	62.606	62.606	62.606
	(5.222)	(4.976)	(4.918)
N	12	14	11
R^2	0.283	0.00000	0.147

By comparing weighted averages, with weights equal to the inverse of the probability that an observation is assigned to its observed treatment condition, we obtain estimates for the three ATEs: -16.0, -0.04, -9.6, respectively.

c) Use the data at http://isps.research.yale.edu/FEDAI to estimate the average effect of spillover on nonexperimental units. Note that your estimator must make use of the probability that each unit lies within 500 meters of a treated experimental unit; exclude from your analysis any units that have zero probability of experiencing spillovers.

```
hotspot_nonexp <- within(hotspot_nonexp,{
  exposure[exposure==10] <- "10"
  exposure[exposure==0] <- "00"

  Q <- NA
  Q[exposure=="10"] <- prob10[exposure=="10"]
  Q[exposure=="00"] <- prob00[exposure=="00"]

  weights <- 1/Q
  })

fit.nonexp <- lm(y ~ exposure, weights=weights,
                 data=subset(hotspot_nonexp, prob10 > 0 & prob10 < 1))

fit.nonexp

##
```

```
## Call:
## lm(formula = y ~ exposure, data = subset(hotspot_nonexp, prob10 >
##     0 & prob10 < 1), weights = weights)
##
## Coefficients:
## (Intercept)   exposure10
##       4.286        4.602
```

The estimate of the spillover effects of treatment on the non-experimental units is 4.6.

Question 10

A doctoral student conducted an experiment in which she randomly varied whether she ran or walked 40 minutes each morning.[2] In the middle of each afternoon over a period of 26 days, she measured the following outcome variables: (1) her weight (minus a constant, for privacy's sake), (2) her score in a game of Tetris, (3) her mood on a 0-5 scale, with 5 being the most pleasant, (4) her energy level on a 0-5 scale, with 5 being the most energetic, and (5) whether she answered correctly a randomly selected problem from the math section of the GRE. Outcomes are missing for days 13 and 17. The data are listed below.

a) Suppose you were seeking to estimate the average effect of running on her Tetris score. Explain the assumptions needed to identify this causal effect based on this within-subjects design. Are these assumptions plausible in this instance? What special concerns arise due to the fact that the subject was conducting the study, undergoing the treatments, and measuring her own outcomes?
Answer:
Suppose the effect were defined as $Y_i(1)$ and $Y_i(0)$, where the subscript refers to day. This formulation assumes potential outcomes respond only to the treatments administered that day, with no carryover from days past and no anticipation of treatments to come. The no-anticipation assumption seems reasonable; more questionable is the assumption that Tetris scores respond only to today's treatment, not yesterday's. The potential outcomes above presuppose that the cognitive or physiologic effects of running disappear after a night's sleep. In order to relax this assumption, one could expand the schedule of potential outcomes to include pairs (or longer sequences) of treatments on successive days. There is a risk of an excludability violation when measuring one's own outcomes; what if the subject tries harder when playing tetris in the wake of a running treatment?

b) Estimate the effect of running on Tetris score. Use randomization inference to test the sharp null hypothesis that running has no immediate or lagged effect on Tetris scores.

```
library(ri)
Y <- hough$tetris
Z <- hough$run
N <- length(Z)
```

[2] Hough 2010.

```r
# exclude day 1 from analysis
Zlag <- c(NA,Z[2:N-1])
Ylag <- c(NA,Y[2:N-1])

# simple random assignment based on coin flips
randfun <- function() rbinom(N,1,.5)
numiter <- 10000
set.seed(343)
# random assignment follows the custom function "randfun"
perms <- genperms.custom(numiter=numiter,randfun=randfun)

## note on missing data: the default for LM is NA.omit=TRUE
## This default eliminates the first lag, and the two days with missing outcomes

### This code performs the estimation for parts b, c, and d

fit1 <- lm(Y~Z)$coefficients["Z"]              # regress Y on current Z
fit2 <- summary(lm(Y~Z+Zlag))$fstatistic[1]    # regress Y on current and lagged Z
fit3 <- lm(Ylag~Z)$coefficients["Z"]           # placebo fit: regress lagged Y on Z
fit4 <- lm(hough$energy~Z)$coefficients["Z"]   # consider current Z's effects on energy
fit5 <- lm(hough$gre~Z)$coefficients["Z"]      # consider current Z's effects on GRE

# initialize the five vectors of results
dist1 <- dist2 <- dist3 <- dist4 <- dist5 <- rep(NA,numiter)

for (i in 1:numiter) {
  Zri <- perms[,i]
        Zlagri <- c(NA,Zri[2:N-1]) # exclude day 1 from analysis

  dist1[i] <- lm(Y~Zri)$coefficients["Zri"]
  dist2[i] <- summary(lm(Y~Zri+Zlagri))$fstatistic[1]
  dist3[i] <- lm(Ylag~Zri)$coefficients["Zri"]
  dist4[i] <- lm(hough$energy~Zri)$coefficients["Zri"]
  dist5[i] <- lm(hough$gre~Zri)$coefficients["Zri"]
        }

mean(dist1 >= fit1)      # one-tailed p-value: does running increase Tetris scores

## [1] 0.0068

mean(dist2 >= fit2)      # one-tailed p-value: does running increase Tetris scores

## [1] 0.019

mean(abs(dist3) >= abs(fit3))  # two-tailed p-value: placebo fit

## [1] 0.8994
```

```
mean(dist4 >= fit4)      # one-tailed p-value: does running improve energy

## [1] 0.4563

mean(dist5 >= fit5)      # one-tailed p-value: does running improve GRE

## [1] 0.8164
```

Focusing solely on the immediate effect of running that day's Tetris scores, we see that on non-running days the average Tetris score is 12,806, as compared to 26,419 on running days, for a difference of 13,613. Randomization inference indicates that this observed difference has a one-tailed p-value of 0.0068. Using the F-statistic to assess the joint significance of immediate and one-period lagged effects, we obtain a p-value of 0.019, again allowing us to reject the null hypothesis of no effect.

c) One way to lend credibility to within-subjects results is to verify the no-anticipation assumption. Use the variable Run to predict the Tetris score *on the preceding day*. Presumably, the true effect is zero. Does randomization inference confirm this prediction?
Answer:
See code block above for calculations. As expected, the means are similar (21,740 for control and 20,727 for treatment), and the two-tailed p-value is 0.8994.

d) If Tetris responds to exercise, one might suppose that energy levels and GRE scores would as well. Are these hypotheses borne out by the data?
Answer:
No, energy has a difference-in-means of just 0.07 and a p-value of 0.4563; and GRE's effect goes (insignificantly, p=0.8164) in the wrong direction, with the treatment diminishing the probability of a right answer by 0.157.

Question 11

Return to the stepped-wedge advertising example in section 8.6 and the schedule of assigned treatments in Table 8.8.

a) Estimate $E[Y_{01} - Y_{00}]$ by restricting your attention to weeks 2 and 3. How does this estimate compare to the estimate of $E[Y_{11} - Y_{00}]$ presented in the text, which is also identified using observations from weeks 2 and 3?

```
# Recreate dataset
week <- c(rep("2", 8), rep("3", 8))
prob00 <- c(rep(0.5, 8), rep(0.25, 8))
prob01 <- c(rep(0.25, 8), rep(0.25, 8))
prob11 <- c(rep(0.25, 8), rep(0.50, 8))
Y <- c(9,5,2,3,3,8,3,1,
       4,7,10,10,3,10,4,3)
Z <- c("11", "00", "01", "00", "00", "11", "00", "01",
       "11", "01", "11", "01", "00", "11", "00", "11")
```

```r
# Estimate E[Y_{01}-Y_{00}]
mean01 <- weighted.mean(Y[Z=="01"], w=1/prob01[Z=="01"])
mean00 <- weighted.mean(Y[Z=="00"], w=1/prob01[Z=="00"])
ate01_00 <- mean01 - mean00
ate01_00

## [1] 1.5
```

The effect of immediate treatment appears to be 1.5, which is weaker than the effect mentioned in the text (4.13).

b) Estimate $E[Y_{11} - Y_{00}]$ without imposing the assumption that treatment effects disappear after two weeks by restricting your attention to week 2.

```r
# Estimate E[Y_{11}-Y_{00}]
mean11 <- weighted.mean(Y[Z=="11" & week=="2"], w=1/prob01[Z=="01" & week=="2"])
mean00 <- weighted.mean(Y[Z=="00" & week=="2"], w=1/prob01[Z=="00" & week=="2"])
ate11_00 <- mean11 - mean00
ate11_00

## [1] 5
```

Without imposing this assumption and focusing only on week two, the estimated ATE of immediate and lagged treatment is 5.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 9 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts:

a) Define CATE. Is a Complier average causal effect (CACE) an example of a CATE?
Answer:
CATE stands for conditional average treatment effect, or the ATE among a subgroup. Typically, the subgroup in question is defined by some observable covariate(s), such as the CATE for women over 40 years of age. One could, however, define a CATE for a latent group such as Compliers (those who take the treatment if and only if assigned to the treatment group). Therefore, a CACE is a CATE.

b) What is an interaction effect?
Answer:
An interaction refers to systematic variation in treatment effects. A treatment-by-covariate interaction refers to variation in ATEs that is a function of covariates. A treatment-by-treatment interaction refers to variation in the average effect of one randomized intervention that occurs as a function of other assigned treatments.

c) Describe the multiple comparisons problem and the Bonferroni correction.
Answer:
The multiple comparisons problem refers to the distortion in p-values that occurs when researchers conduct a series of hypothesis tests. When several hypothesis tests are conducted, the chances that at least one of them appears significant may be substantially greater than 0.05, the nominal size of each test. The Bonferroni correction reestablishes the proper size of each test when several hypothesis tests are conducted. If k tests are conducted at the 0.05 level, the Bonferroni-corrected target significance level is $0.05/k$.

Question 2

The standard error formula given in equation (3.4) suggests that, all else being equal, reducing variance in $Y_i(0)$ helps reduce sampling uncertainty. Referring to the procedure outlined in section 9.2, explain why the same principle applies to estimating bounds on treatment effect heterogeneity.

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

Answer:
Nonparametric tests of heterogeneity put an estimated lower bound on the variance of the subject-level treatment effect by sorting the observed $Y_i(0)$ and $Y_i(1)$ in ascending order and calculating the difference between them. The variance of this difference is the estimated lower bound. When the variance of $Y_i(0)$ is small, the variance of the differences between $Y_i(0)$ and $Y_i(1)$ is scarcely affected by whether $Y_i(0)$ is sorted in ascending or descending order. In the limiting case where the variance of $Y_i(0)$ is zero, sorting makes no difference at all; if one were sure that $Y_i(0)$ were constant, one could estimate unit-level treatment effects by subtracting the mean of $Y_i(0)$ from $Y_i(1)$.

Question 3

One way to reduce variance in $Y_i(0)$ is to block on a prognostic covariate. When blocking is used, the joint distribution of $Y_i(0)$ and $Y_i(1)$ is simulated within blocks using the bounding procedure described in section 9.2. Using the schedule of potential outcomes below, show how the maximum and minimum values of the covariance of $Y_i(0)$ and $Y_i(1)$ compare to the maximum and minimum values of the covariance of $Y_i(0)$ and $Y_i(1)$ for the dataset as a whole (i.e., had blocking not been used).

Table 1: Question 3 Table

Block	Subject	Yi(0)	Yi(1)
A	A-1	0	2
A	A-2	1	5
A	A-3	1	3
A	A-4	2	1
B	B-1	2	3
B	B-2	3	3
B	B-3	4	9
B	B-4	4	7


```
block <- c(rep("A", 4), rep("B", 4))
Y0 <- c(0,1,1,2,2,3,4,4)
Y1 <- c(2,5,3,1,3,3,9,7)

# function for calculating population covariances
cov.pop <- function(x,y){sum((x-mean(x))*(y-mean(y)))/(length(x))}

# Ignoring blocks
Y1.lowtohigh <- sort(Y1)
Y1.hightolow <- sort(Y1, decreasing=TRUE)

cov.min <- cov.pop(Y0, Y1.hightolow)
cov.max <- cov.pop(Y0, Y1.lowtohigh)
cov.min

## [1] -3.141
```

```
cov.max

## [1] 3.234

# Including blocks
Y1.lowtohigh.block <- c(sort(Y1[block=="A"]), sort(Y1[block=="B"]))
Y1.hightolow.block <- c(sort(Y1[block=="A"],decreasing=TRUE),
                        sort(Y1[block=="B"],decreasing=TRUE))
cov.min.block <- cov.pop(Y0, Y1.hightolow.block)
cov.max.block <- cov.pop(Y0, Y1.lowtohigh.block)
cov.min.block

## [1] -0.01562

cov.max.block

## [1] 2.984
```

The lowest and highest covariances under simple random assignment are -3.14 and 3.23. In order to find the lowest and highest covariances under blocked assignment, sort the potential outcomes within blocks before calculating the covariances for all observations. Under blocked random assignment, the lowest covariance is -0.02, and the highest covariance is 2.98. Taking advantage of the blocks reduces the range of possible covariances.

Question 4

Suppose that a researcher compares the CATE among two subgroups, men and women. Among men (N = 100), the ATE is estimated to be 8.0 with a standard error of 3.0, which is significant at $p < 0.05$. Among women (N = 25), the CATE is estimated to be 7.0 with an estimated standard error of 6.0, which is not significant, even at the 10% significance level. Critically evaluate the researcher's claim that "the treatment only works for men; for women, the effect is statistically indistinguishable from zero." In formulating your answer, address the distinction between testing whether a single CATE is different from zero and testing whether two CATEs are different from each other.

Answer:

The researcher's interpretation of the results ignores the fact that the estimated CATE for women is almost as large (7 versus 8) as the estimated CATE among men. The difference between the estimated CATEs (8-7=1) is much smaller than the apparent standard error of the difference (which is the square root of the sum of the estimated standard errors, or 6.7). An alternative interpretation is that both of the CATEs are the same, but the CATE among men is estimated with greater precision because the male sample is much larger than the female sample.

Question 5

The table below shows hypothetical potential outcomes for an experiment in which low-income subjects in a developing country are randomly assigned to receive (i) loans to aid their small businesses; (ii) business training to improve their accounting, hiring, and inventory-management skills; (iii) both; or (iv) neither. The outcome measure in business income during the subsequent

year. The table also includes a pre-treatment covariate, an indicator scored 1 if the subject was judged to be proficient in these basic business skills.

Table 2: Question 5 Table

Subject	$Y_i(loan)$	$Y_i(training)$	$Y_i(both)$	$Y_i(Neither)$	Prior business skills
1	2	2	3	2	0
2	2	3	2	1	0
3	5	6	6	4	1
4	3	1	5	1	1
5	4	4	5	0	0
6	10	8	11	10	1
7	1	3	3	1	0
8	5	5	5	5	1
Average	4	4	5	3	0.5

a) What is the ATE of the loan if all subjects were also to receive training?
Answer:
The relevant comparison is the average potential outcomes under "both" to the average potential outcome under only "training." The ATE is 5-4=1.

b) What is the ATE of the loan if no subjects receive training?
Answer:
The relevant comparison is the average potential outcomes under "loan" to the average potential outcome under only "neither." The ATE is 4-3=1.

c) What is the ATE of the training if all subjects also receive a loan?
Answer:
The relevant comparison is the average potential outcomes under "both" to the average potential outcome under only "loan." The ATE is 5-4=1.

d) What is the ATE of the training if no subjects receive a loan?
Answer:
The relevant comparison is the average potential outcomes under "training" to the average potential outcome under only "neither." The ATE is 4-3=1.

e) Suppose subjects were randomly assigned to one of the four experimental treatments in equal proportions. Use the table above to fill in the expected values of the four regression coefficients for the model and interpret the results:

$$Y_i = \alpha_0 + \alpha_1 Loan_i + \alpha_2 Training_i + \alpha_3 (Loan_i * Training_i) + e_i \quad (1)$$

The four coefficients are $\alpha_0 = 3$, the average outcome under "neither"; $\alpha_1 = 1$, the ATE of loan when there is no training; $\alpha_2 = 1$, the ATE of training when there is no loan; and $\alpha_3 = 0$ the change in the effect of training that occurs when our focus switches from those who receive no loan to those who receive a loan. Note that this interaction term can also be interpreted as the change in the ATE of loans that we observe when we move from the untrained subgroup to the trained subgroup.

$$Y_i = 3 + 1 * Loan_i + 1 * Training_i + 0 * (Loan_i * Training_i) + e_i \quad (2)$$

f) Suppose a researcher were to implement a block randomized experiment, such that two subjects with business skills are assigned to receive loans, and two subjects without business skills are assigned to receive loans, and the rest are assigned to control. No subjects are assigned to receive training. The researcher estimates the model

$$Y_i = \gamma_0 + \gamma_1 Loan_i + \gamma_2 Skills_i + \gamma_3 (Loan_i * Skills_i) + e_i \quad (3)$$

Over all 36 possible random assignments, the average estimated regression is as follows:

$$Y_i = 1.00 + 1.25 Loan_i + 4.00 Skills_i - 0.50 (Loan_i * Skills_i) \quad (4)$$

Interpret the results and contrast them with the results from part (e). (Hint: the block randomized design does not affect the interpretation. Focus on the distinction between treatment-by-treatment and treatment-by-covariate interactions.)
Answer:
The key thing to bear in mind when interpreting these results is that the interaction between loans and skills is a treatment-by-covariate interaction because skills are not randomly assigned. The results seem to suggest that loans are more effective amongst those without skills (CATE=4) than among those with skills (CATE = 4 - 0.5 = 3.5). These CATEs may describe the ATEs in these two skill groups, but the change in CATEs does not necessarily imply that a random increase in skill would diminish the effects of loans.

Question 6

Rind and Bordia studied the tipping behavior of lunchtime patrons of an "upscale Philadelphia restaurant" who were randomly assigned to four experimental groups.[1] One factor was server sex (male or female), and a second factor was whether the server draws a "happy face" on the back of the bill presented to customers.[2] Download the data located at http://isps.research.yale.edu/FEDAI.

a) Suppose you ignored the sex of the server and simply analyzed whether the happy face treatment has heterogeneous effects. Use randomization inference to test whether $Var(\tau_i) = 0$ by testing whether $Var(Y_i(1)) = Var(Y_i(0))$. Construct the full schedule of potential outcomes by assuming that the treatment effect is equal to the observed difference-in-means between $Y_i(1)$ and $Y_i(0)$. Interpret your results.

```
library(ri)
# generate a treatment indicator
Z <- as.integer(rindb$happyface) - 1
Y <- rindb$tip
```

[1] Rind and Bordia 1996.

[2] The authors took steps to ensure the blindness of the servers to the happy face condition, which was determined only moments before the bill was delivered. The authors also instructed waitstaff to deliver bills and walk away, so that there would be no additional interaction with customers. It is not clear whether the sex of the server was randomly assigned.

```r
probs <- genprobexact(Z)
ate <- estate(Y,Z,prob=probs)

numiter <- 10000
set.seed(343)
perms <- genperms(Z,maxiter=numiter)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 5.19137106437769e+25 to perform exact estimation.

# generate a schedule of potential outcomes under the assumption
# that the treatment effect equals the estimated ATE for all subjects
Ys <- genouts(Y,Z,ate=ate)

# compare variances in treatment and control groups
testvar <- var(Y[Z==1]) - var(Y[Z==0])
vardist <- rep(NA,numiter)

# generate the sampling distribution of the difference in variances
for (i in 1:numiter) vardist[i] <- var(Y[perms[,i]==1]) - var(Y[perms[,i]==0])

# p-value for var(Y1)>Var(Y0)
mean(vardist >= testvar)

## [1] 0.2398

# p-value for var(Y1)<>Var(Y0)
mean(abs(vardist) >= abs(testvar))

## [1] 0.472
```

We constructed a simulation of 10,000 random assignments and for each assessed the difference in variances between treatment and control group. The observed difference is 53.31. However, this absolute difference has a p-value of 0.472. We cannot reject the null hypothesis that the observed difference in variances is the produce of random sampling variability. The failure to reject the null is not surprising given the low power of this test, which does not focus on any specific model of heterogeneous treatment effects.

b) Write down a regression model that depicts the effect of the sex of the waitstaff, whether they write a happy face on the bill, and the interaction of these factors.
Answer:
Using tip percentage as the outcome and a binary variable for sex (female=1) and for the use of a happy face (face=1), a regression model is as follows:

$$Y_i = \gamma_0 + \gamma_1 Sex_i + \gamma_2 Face_i + \gamma_3 (Sex_i * Face_i) + e_i \tag{5}$$

c) Estimate the regression model in (b) and test the interaction between waitstaff sex and the happy face treatment. Is the interaction significant?
Answer:

```
# generate indicator of waitstaff sex
female <- as.integer(rindb$female) - 1

# regression with interaction between happyface and waitstaff sex
lmmodelint <- lm(Y~Z+female+Z*female)
# regression model without interaction
lmmodel <- lm(Y~Z+female)

summary(lmmodelint)

##
## Call:
## lm(formula = Y ~ Z + female + Z * female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.406  -5.726  -0.684   5.286  39.419
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.406      2.287   9.360 1.01e-14 ***
## Z             -3.630      3.163  -1.147   0.2544
## female         6.378      3.163   2.016   0.0469 *
## Z:female       8.887      4.447   1.999   0.0488 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.48 on 85 degrees of freedom
## Multiple R-squared:  0.2476,Adjusted R-squared:  0.2211
## F-statistic: 9.325 on 3 and 85 DF,  p-value: 2.14e-05

#Confirm p-value with RI

# use estimated coefficients from base model to impute potential outcomes
Y00 <- Y - lmmodel$coefficients["Z"]*Z - lmmodel$coefficients["female"]*female
Y10 <- Y00 + lmmodel$coefficients["Z"]
Y01 <- Y00 + lmmodel$coefficients["female"]
Y11 <- Y00 + lmmodel$coefficients["Z"] + lmmodel$coefficients["female"]

f.obs <- waldtest(lmmodelint,lmmodel)$F[2]
f.sims <- rep(NA,numiter)
```

```r
for (i in 1:numiter) {
  Z.sim <- perms[,i]

        # realized values of Y reflect single or compound treatments
  # and the potential outcomes that they reveal
        Y.sim <- Y00
        Y.sim[Z.sim == 0 & female == 1] <- Y01[Z.sim == 0 & female == 1]
  Y.sim[Z.sim == 1 & female == 0] <- Y10[Z.sim == 1 & female == 0]
  Y.sim[Z.sim == 1 & female == 1] <- Y11[Z.sim == 1 & female == 1]

        # regressions based on two nested models: with and without interaction
        lmmodelint.sim <- lm(Y.sim~Z.sim + female + female*Z.sim)
        lmmodel.sim <- lm(Y.sim~Z.sim + female)
        # calculate the F-statistic by comparing two nested models
        f.sims[i] <- waldtest(lmmodelint.sim,lmmodel.sim)$F[2]
}

# calculate the p-value by comparing the observed F-statistic
# to the F-statistic under the null of constant & additive effects
mean(f.sims >= f.obs)

## [1] 0.0483
```

The regression reported above suggests a positive interaction between the happyface treatment and female, implying that female waitstaff receive much more return from happyfaces than male waitstaff. The two-sided p-value from the regression is 0.049, which is similar to the result from randomization inference ($p = 0.0483$). A two-sided test is appropriate here because the direction of the effect was not predicted ex ante. Thinking back to section (a), the specific interaction posited by this regression sets the stage for a more powerful test of treatment effect heterogeneity.

Question 7

In their 2004 study of racial discrimination in employment markets, Bertrand and Mullainathan sent resumes with varying characteristics to firms advertising job openings. Some firms were sent resumes with putative African American names, while other firms received resumes with putatively Caucasian names. The researchers also varied other attributes of the resume, such as whether the resume was judged to be of high or low quality (based on labor market experience, career profile, gaps in employment, and skills listed).[3] The table below shows the rate at which applicants were called back by employers, by the city in which the experiment took place and by the randomly assigned attributes of their applications.

[3] Bertrand and Mullainathan 2004, p. 994.

Table 3: Question 7 Table

	Boston				Chicago			
	Low-quality resume		High-quality resume		Low-quality resume		High-quality resume	
	Black	White	Black	White	Black	White	Black	White
% Received Call	7.01	10.15	8.5	13.12	5.52	7.16	5.28	8.94
(N)	(542)	(542)	(541)	(541)	(670)	(670)	(682)	(682)

a) For each city, interpret the apparent treatment effects of race and resume quality on the probability of receiving a follow-up call.
Answer:
For Boston, the effect of (white) race is 10.15 - 7.01 = 3.14 when resume quality is low and 13.12 - 8.50 = 4.62 when resume quality is high. For Chicago, the effect of (white) race is 7.16 - 5.52 = 1.64 when resume quality is low and 8.94 - 5.28 = 3.66 when resume quality is high. Note that another, equally valid way to interpret the table is to assess the effect of resume quality for each race, but the substantive focus of this study is on race effects.

b) Propose a regression model that assesses the effects of the treatments, interaction between them, and interactions between the treatments and the covariate, city.
Answer:
This model is similar to the interactive regression specifications described above, but it contains treatment-by-treatment interactions (race x resume) and treatment-by-covariate interactions (race x city, resume x city) and a higher order interaction (race x resume x city) that allows for the possibility that the race x resume interaction differs by city. Here, City is scored 1 if Chicago. Race = 1 if white. Resume =1 if high quality. Notice that the "saturated" regression model contains eight parameters, one for each cell of the table.

$$Y_i = \gamma_0 + \gamma_1 Race_i + \gamma_2 Resume_i + \gamma_3 City_i + \gamma_4(Race_i * Resume_i) + \gamma_5(Race_i * City_i) + \gamma_6(Resume_i * City_i) + \gamma_7(Race_i * Resume_i * City_i) + e_i$$

c) Estimate the parameters in your regression model. Interpret the results (This can be done by hand based on the percentages given in the table.)
Answer:
Because there as many parameters as experimental groups, the estimated coefficients reproduce the percentages given in the table:

$$Y_i = 7.01 + 3.14 Race_i + 1.49 Resume_i - 1.49 City_i + 1.48(Race_i * Resume_i) - 1.50(Race_i * City_i) - 1.73(Resume_i * City_i) + 0.54(Race_i * Resume_i * City_i) + e_i$$

```
Y <- c(rep(1, 38), rep(0, 542-38), rep(1, 55), rep(0, 542-55),
       rep(1, 46), rep(0, 541-46), rep(1, 71), rep(0, 541-71),
       rep(1, 37), rep(0, 670-37), rep(1, 48), rep(0, 670-48),
       rep(1, 36), rep(0, 682-36), rep(1, 61), rep(0, 682-61))
```

```
boston <- c(rep(1, 542+542+541+541), rep(0, 670+670+682+682))
chicago <- 1-boston
lowquality <- c(rep(1, 542+542), rep(0, 541+541), rep(1, 670+670), rep(0, 682+682))
highquality <- 1-lowquality
black<- c(rep(1, 542), rep(0,542), rep(1, 541), rep(0,541),
          rep(1, 670), rep(0,670), rep(1, 682), rep(0,682))
white <- 1-black

# All the models are
# Y ~ race + quality + city + race*quality + race*city + quality*city + race*quality*city
# In principle, there are 8 possibilities... here are 4.

fit_1 <- lm(Y ~ white + highquality + chicago + white*highquality +
              white*chicago + highquality*chicago + white*highquality*chicago)
fit_2 <- lm(Y ~ black + highquality + chicago + black*highquality +
              black*chicago + highquality*chicago + black*highquality*chicago)
fit_3 <- lm(Y ~ white + highquality + boston + white*highquality +
              white*boston + highquality*boston + white*highquality*boston)
fit_4 <- lm(Y ~ black + lowquality + chicago + black*lowquality +
              black*chicago + lowquality*chicago + black*lowquality*chicago)
stargazer(fit_1, fit_2, fit_3, fit_4, style = "apsr", notes = c("Coeffcients depend on stude
```


Question 8

In Chapter 3, we analyzed data from Clingingsmith, Khwaja, and Kremer's study of Pakistani Muslims who participated in a lottery to obtain a visa for the pilgrimage to Mecca.[4] By comparing lottery winners to lottery losers, the authors are able to estimate the effects of the pilgrimage on various attitudes, including views about people from other countries. Winners and losers were asked to rate the Saudi, Indonesian, Turkish, African, European, and Chinese people on a five-point scale ranging from very negative (-2) to very positive ($+2$). Adding the responses to all six items creates an index ranging from -12 to $+12$. The key results are presented in the table below.

a) Explain the meaning of "absolute difference in variances."
Answer:
The term "difference in variances" is the observed difference between the variance of outcomes in the treatment group and the variance of outcomes in the control group. The term "absolute" refers to the absolute value of this difference.

b) Describe how one could use randomization inference to test the null hypothesis of constant treatment effects.
Answer:
One method is to create a full schedule of potential outcomes under the null hypothesis of constant treatment effects. For example, we could assume that all subjects have a treatment effect equal to the observed ATE. In order to obtain untreated potential outcomes for the treatment group, we subtract off the ATE from the observed treated potential outcomes. In

[4] Clingingsmith, Khwaja, and Kremer 2009.

Table 4: Question 7C Table

	Y			
	(1)	(2)	(3)	(4)
white	0.031*		0.016	
	(0.016)		(0.015)	
black		−0.031*		−0.046***
		(0.016)		(0.016)
highquality	0.015	0.030*	−0.002	
	(0.016)	(0.016)	(0.015)	
lowquality				−0.030*
				(0.016)
chicago	−0.015	−0.030*		−0.042***
	(0.016)	(0.016)		(0.016)
boston			0.015	
			(0.016)	
white:highquality	0.015		0.020	
	(0.023)		(0.021)	
white:chicago	−0.015			
	(0.022)			
black:highquality		−0.015		
		(0.023)		
black:lowquality				0.015
				(0.023)
black:chicago		0.015		0.010
		(0.022)		(0.022)
highquality:chicago	−0.017	−0.012		
	(0.022)	(0.022)		
white:highquality:chicago	0.005			
	(0.031)			
black:highquality:chicago		−0.005		
		(0.031)		
white:boston			0.015	
			(0.022)	
highquality:boston			0.017	
			(0.022)	
white:highquality:boston			−0.005	
			(0.031)	
lowquality:chicago				0.012
				(0.022)
black:lowquality:chicago				0.005
				(0.031)
Constant	0.070***	0.101***	0.055***	0.131***
	(0.012)	(0.012)	(0.010)	(0.012)
N	4,870	4,870	4,870	4,870
R^2	0.008	0.008	0.008	0.008
Adjusted R^2	0.006	0.006	0.006	0.006
Residual Std. Error (df = 4862)	0.271	0.271	0.271	0.271
F Statistic (df = 7; 4862)	5.359***	5.359***	5.359***	5.359***

*p < .1; **p < .05; ***p < .01

Coeffcients depend on student's parameterization

Table 5: Question 8 table

	Control group	Treatment group
N	448	510
Mean	1.868	2.343
Variance	5.793	6.902
Absolute difference in variances	1.109	

order to obtain the treated potential outcomes for the control group, we add the apparent ATE. We then simulate a large number of possible random assignments; for each random assignment, we calculate the absolute difference between the variance of the treatment group and the variance of the control group. We obtain p-values by determining where the observed absolute difference falls in the sampling distribution under the null hypothesis.

c) Assume that researchers applied the method you proposed in part (b) and simulated 100,000 random assignments, each time calculating the absolute difference in variances; they find that 25,220 of these differences are as large or larger than 1.109, the absolute difference in variances observed in the original sample. Calculate the p-value implied by these results. What do you conclude about treatment effect heterogeneity in this example?
Answer:
The p-value is $25220/100000 = 0.2522$. The difference in observed variances is not inconsistent with the null hypothesis of homogeneous effects.

d) Suppose that this experiment were partitioned into subgroups defined according to whether the subjects had traveled abroad in the past. Suppose that the CATE among those who had previously traveled abroad were 0 and that the CATE among those who had not traveled abroad were 1.0. Suppose this difference in CATEs were significant at $p < .05$. Does this result imply that randomly encouraging people to travel abroad eliminates the Hajj's effect?
Answer:
Not necessarily. This regression reports a treatment-by-covariate interaction, which describes the CATEs for two subgroups that may or may not have similar potential outcomes. Randomly encouraging travel is designed to create groups with the same expected potential outcomes; this design tests whether travel causes the effect of the Hajj to change.

Question 9

An example of a two-factor design that encounters one-sided noncompliance may be found in Fieldhouse et al.'s study of voter mobilization in the United Kingdom.[5] In this study, the first factor is whether each voter was mailed a letter encouraging him or her to vote in the upcoming election. The second factor is whether each voter was called with an encouragement to vote. Noncompliance occurs in the case of phone calls, as some targeted voters cannot be reached when called. The experimental design consists of four groups: a control group, a mail-only group, a phone-only group, and a group targeted for both mail and phone. The following table shows the results by assigned experimental group.

[5] Fieldhouse et al. 2010.

Table 6: Question 9 Table

	Control	Mail Only	Phone Only	Mail and Phone
N	5179	4367	3466	2287
Number Contacted by Phone	0	0	2003	1363
Among those Assigned to this Experimental Group, Percent who Voted	0.397	0.403	0.397	0.418
Among those Contacted by Phone, Percent who Voted	NA	NA	0.465	0.468

a) Show that, under certain assumptions, this experimental design allows one to identify the following parameters: (i) the ATE of mail, (ii) the Complier average causal effect (CACE) of phone calls, (iii) the CATE of mail among those who comply with the phone call treatment, (iv) the CATE of mail among those who do not comply with the phone call treatment, and (v) the CACE of phone calls among those who receive mail.
Answer:

The ATE of mail is identified using the core assumptions of chapter 2 (random assignment, non-interference, and excludability). Excludability in this holds that the only way that random assignment of mail affects outcomes is through the mail treatment itself. In order identify the CACE of phone calls, we must invoke the assumptions of Chapter 5, since this is a case of one-sided non-compliance. Again, the exclusion restriction holds that the only way that the assignment of phone calls affects outcomes is through actual phone contacts. The CACE here is ATE among those who receive phone calls if assigned to the treatment group. The CATE of mail is identified in the same way as an ATE, except that it is restricted to those who actually receive phone calls; the same goes for the ATE among those who are not treated when called. The CACE of phone calls among those who receive mail is identified among the same group of compliers as the CACE above, since mail is assigned and received randomly (because we assume full compliance with the mail treatment). However, the ATE of the calls among Compliers may differ from the ATE among Compliers who also receive mail, due to a treatment-by-treatment interaction.

b) Using the identification strategies you laid out in part (a), estimate each of the five parameters using the results in the table.

The estimated ATE of mail is 40.3 - 39.7 = 0.6 percentage points. The estimated CACE of phone calls is the estimated ITT divided by the share of compliers: (39.7 - 39.7)/ (2003/3466) = 0. The estimated CATE of mail among those who receive a call is 46.8 - 46.5 = 0.3 percentage points. In order to figure out the CATE of mail among those who did not comply when called, we must first back out the voting rates given the numbers presented above. For example, the overall voting rate in the treatment group of 41.8 is a weighted average of the voting rates among the contacted and uncontacted. Thus, 41.8 = 46.8(1363/2287) + X(923/2287). Solving for X gives 34.4, and repeating the same calculation for the control group gives 30.4. Therefore, the estimated effect of mail for this subgroup is 4.0 percentage points. The CACE of phone calls among those who receive mail is the estimated ITT divided by the contact rate: (41.8 - 40.3) / (1363/2287) = 2.5 percentage points

c) In Chapters 5 and 6, we discussed the use of instrumental variables regression to estimate CACEs when experiments involve noncompliance. Here, we can apply instrumental variables regression to a factorial experiment in which one factor encounters noncompliance. With the replication dataset at http://isps.research.yale.edu/FEDAI, use instrumental variables regression to estimate the parameters of the Vote equation in the following three-equation regression model:

$$PhoneContact_i = \alpha_0 + \alpha_1 Mail_i + \alpha_2 PhoneAssign_i + \alpha_3 (PhoneAssign_i * Mail_i) + e_i$$
$$PhoneContact_i * Mail_i = \gamma_0 + \gamma_1 Mail_i + \gamma_2 PhoneAssign_i + \gamma_3 (PhoneAssign_i * Mail_i) + \epsilon_i$$
$$Vote_i = \beta_0 + \beta_1 Mail_i + \beta_2 PhoneContact_i + \beta_3 (PhoneContact_i * Mail_i) + u_i$$

Interpret the regression estimates in light of the five parameters you estimated in part (b). Which causal parameters does instrumental variables regression estimate or fail to estimate?

```
fieldhouse <- within(fieldhouse,{
  mail <- m
  phone <- p
  phone_contact <- c
  vote <- y
})

fit.iv <- ivreg(vote ~ mail + phone_contact + phone_contact*mail
                | mail + phone + mail*phone, data=fieldhouse)
summary(fit.iv)
```

```
##
## Call:
## ivreg(formula = vote ~ mail + phone_contact + phone_contact *
##     mail | mail + phone + mail * phone, data = fieldhouse)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.428 -0.403 -0.397  0.597  0.603
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.396988   0.006810   58.30   <2e-16 ***
## mail                0.006035   0.010068    0.60     0.55
## phone_contact      -0.000178   0.018614   -0.01     0.99
## mail:phone_contact  0.025334   0.028231    0.90     0.37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.49 on 15296 degrees of freedom
## Multiple R-Squared: 0.00105,Adjusted R-squared: 0.00085
## Wald test: 1.13 on 3 and 15296 DF,  p-value: 0.337
```

The intercept is the voting rate in the control group. The coefficient for "phone contact" is the estimated CACE for phones when no mail is assigned. The effect for "mail" is the ATE for mail when no phone calls are assigned. The coefficient for "mail:phone contact" is the extent to which the apparent CACE of phone calls increases when we move from the no-mail to the mail group. These estimates reproduce the estimates generated by hand above. Notice that IV regression does not report the effect of mail for non-compliers.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 10 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts:

a) Suppose that equations (10.1), (10.2), and (10.3) depict the true causal process that generates outcomes. Referring to these equations, define the direct effect of Z_i on Y_i and the indirect effect that Z_i transmits through M_i to Y_i.
Answer:
The direct effect is the causal influence that is transmitted from Z_i to Y_i without passing through M_i, and the indirect effect is the causal influence that passes from Z_i to Y_i through M_i. The direct effect of Z_i on Y_i is the parameter d in equation (10.3). The indirect or "mediated" effect is the product ab.

b) Explain why the equation Total effect = Direct effect + Indirect effect breaks down when the parameters of equations (10.1), (10.2), and (10.3) vary across subjects.
Answer:
The indirect or "mediated" effect is the product ab, but when these two parameters vary, their expected product is not in general equal to the product of their expectations. Thus, one cannot estimate the average a_i using equation (10.1) and multiply it by the estimate of the average b_i from equation (10.3) in order to obtain an estimated whose expected value is $E[a_i b_i]$.

c) Suppose that the effect of M_i on Y_i varies from one subject to the next. Show that the indirect effect of Z_i on Y_i is zero when the treatment effect of Z_i on M_i is zero for all subjects.
Answer:
When a_i is zero for all subjects, the expected product of a_i and b_i is zero: $E[a_i b_i] = aE[b_i] = 0E[b_i] = 0$.

d) Explain why the complex potential outcome $Y_i(M_i(0), 1)$ defies empirical investigation.
Answer:
The expression $Y_i(M_i(0), 1)$ denotes the potential outcome that would occur given two inputs: $Z_i = 1$ (i.e., the subject is assigned to the treatment group) and M_i were the value it would take on if $Z_i = 0$. These are two incompatible conditions, since Z_i is either 1 or 0. When $Z_i = 1$, for instance, the outcome we observe is $Y_i(M_i(1), 1)$; when $Z_i = 0$, the outcome we observe is $Y_i(M_i(0), 0)$.

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

e) Explain the distinction between the indirect effect that Z_i transmits to Y_i through M_i given in equations (10.15) and (10.16) and the causal effect of M_i, defined using $Y_i(m, z)$ notation as $Y_i(1, 0) - Y_i(0, 0)$ or $Y_i(1, 1) - Y_i(0, 1)$. (Hint: Look closely at how the mediator takes on its value).
Answer:
Equations 10.15 and 10.16 involve complex potential outcomes, which are inherently unobservable. The causal effect of M holding Z constant involves two potentially observable potential outcomes. The difference is that in the latter comparison, we are not trying to set the value of the mediator to its potential outcome in the wake of a manipulation of Z. Instead, we are just setting M to a value and holding Z constant.

Question 2

When researchers use an encouragement design to study mediation, what assumptions must they make in order to satisfy the CACE Theorem from Chapter 6?
Answer:
The CACE theorem assumes non-interference, excludability, and monotonicity. The latter two assumptions may be especially problematic in the context of mediation analysis. Excludability implies that potential outcomes for Y_i respond solely to M_i, whereas the regression framework of equation (10.3) allows Y_i to respond to both M_i and Z_i. When the mediating variable M_i is binary, the monotonicity assumption implies that there are no Defiers (subjects for whom $M_i = 1$ if and only if they are assigned to the control group).

Question 3

Consider the following schedule of potential outcomes for 12 observations. This table illustrates a special situation in which the disturbance e_{1i} is unrelated to the disturbance e_{3i}.

Table 1: Question 3 Table

Observation	Yi(m = 0, z = 0)	Yi(m = 0, z = 1)	Yi(m = 1, z = 0)	Yi(m = 1, z = 1)	Mi(z = 0)	Mi(z = 1)
1	0#	0*	0	0	0	0
2	0	0*	0#	0	0	1
3	0	0	0#	0*	1	1
4	0#	1*	0	1	0	0
5	0	1*	0#	1	0	1
6	0	1	0#	1*	1	1
7	1#	0*	1	1	0	0
8	1	0*	1#	1	0	1
9	1	0	1#	1*	1	1
10	0#	1*	1	1	0	0
11	0	1*	1#	1	0	1
12	0	1	1#	1*	1	1

a) What is the average effect of Z_i on M_i?

Answer:
The average effect of Z on M is the average difference between the last two columns on p.339: $\frac{1}{3}$

b) Use yellow to highlight the cells in the table of potential outcomes to indicate which potential outcomes for Y_i correspond to $Y_i(M_i(0), 0)$. Use green to highlight the cells in the table of potential outcomes to indicate which potential outcomes for Y_i correspond to $Y_i(M_i(1), 1)$. Put an asterisk by the potential outcomes for Y_i in each row that correspond to the complex potential outcome $Y_i(M_i(0), 1)$. Put a pound sign by the potential outcomes for Y_i in each row that correspond to the complex potential outcome $Y_i(M_i(1), 0)$.

c) What is the average total effect of Z_i on Y_i?
Answer:
This difference is green minus yellow = 8/12 - 4/12 = 1/3

d) What is the average direct effect of Z_i on Y_i holding M_i constant at $M_i(0)$? Hint: see equation (10.13).
Answer:
This difference is asterisk minus yellow = 7/12 - 4/12 = 1/4

e) What is the average direct effect of Z_i on Y_i holding M_i constant at $M_i(1)$? Hint: see equation (10.14).
Answer:
This difference is green minus pound sign = 8/12 - 5/12 = 1/4

f) What is the average indirect effect that Z_i transmits through M_i to Y_i when $Z_i = 1$? Hint: see equation (10.15).
Answer:
This difference is green minus asterisk = 8/12 - 7/12 = 1/12

g) What is the average indirect effect that Z_i transmits through M_i to Y_i when $Z_i = 0$? Hint: see equation (10.16).
Answer:
This difference is pound sign minus yellow = 5/12 - 4/12 = 1/12

h) In this example, does the total effect of Z_i equal the sum of its average direct and indirect effect?
Answer:
Yes because the average of the direct effects is 1/4 and the average of the indirect effects is 1/12, which sums to the total effect, 1/3

i) What is the average effect of M_i on Y_i when $Z_i = 0$?
Answer:
This is the 3rd column minus the 1st column: 6/12 - 3/12 = 3/12

j) Suppose you were to randomly assign half of these observations to treatment ($Z_i = 1$) and the other half to control ($Z_i = 0$). If you were to regress Y_i on M_i and Z_i, you would obtain unbiased estimates of the average direct effect of Z_i on Y_i and the average effect of M_i on Y_i. (This fact may be verified using the R simulation at http://isps.research.yale.edu/FEDAI.) What special features of this schedule of potential outcomes allows for unbiased estimation?
Answer:
See simulation below and following question for answer.

```
rm(list=ls())       # clear objects in memory
library(ri)

# schedule of potential outcomes for problem 10.3
Z <- c(0,0,0,0,0,0,1,1,1,1,1,1)

Y0M0 = c(0,0,0,0,0,0,1,1,1,0,0,0)
Y1M0 = c(0,0,0,1,1,1,0,0,0,1,1,1)
Y0M1 = c(0,0,0,0,0,0,1,1,1,1,1,1)
Y1M1 = c(0,0,0,1,1,1,1,1,1,1,1,1)
M0 = c(0,0,1,0,0,1,0,0,1,0,0,1)
M1 = c(0,1,1,0,1,1,0,1,1,0,1,1)

# verify column averages
mean(Y0M0)

## [1] 0.25

mean(Y1M0)

## [1] 0.5

mean(Y0M1)

## [1] 0.5

mean(Y1M1)

## [1] 0.75

# simulate all possible random assignments
perms <- genperms(Z)

# stores estimates from equation 10.3
coefmat <- matrix(NA,ncol(perms),3)
# stores estimates from equation 10.2
tcoefmat <- matrix(NA,ncol(perms),2)
# stores estimates from equation 10.1
mcoefmat <- matrix(NA,ncol(perms),2)

for (i in 1:ncol(perms)) {
        Zri <- perms[,i]
        M <- M0*(1-Zri) + M1*Zri
        Y <- Y0M0*(1-Zri)*(1-M) + Y1M0*(Zri)*(1-M) +
          Y0M1*(1-Zri)*(M) + Y1M1*(Zri)*(M)
        coefmat[i,] <- lm(Y~M+Zri)$coefficients
```

```
        tcoefmat[i,] <- lm(Y~Zri)$coefficients
        mcoefmat[i,] <- lm(M~Zri)$coefficients
        }

# results omit instances of perfect colinearity between M and Z
# report the avg coefficients from a regression of Y on M and Z
colMeans(na.omit(coefmat))

## [1] 0.25 0.25 0.25

# report the avg coefficients from a regression of Y on Z
colMeans(na.omit(tcoefmat))

## [1] 0.3333333 0.3333333

# report the avg coefficients from a regression of M on Z
colMeans(na.omit(mcoefmat))

## [1] 0.3333333 0.3333333
```

k) In order to estimate average indirect effect that Z_i transmits through M_i to Y_i, estimate the regressions in equations (10.1) and (10.3) and multiply the estimates of a and b together.[1] Use the simulation to show that this estimator is unbiased when applied to this schedule of potential outcomes. Why does this estimator, which usually produces biased results, produce unbiased results in this example?
Answer:


```
# Estimates of a, from simulation above
as <- mcoefmat[,2]
# Estimates of b, from simulation above
bs <- coefmat[,3]
# This average is very nearly 1/12.
mean(as*bs, na.rm = TRUE)

## [1] 0.08224401
```

The simulation confirms that the results are unbiased (excluding random assignments that result in perfect collinearity between Z and M) for the direct and total effects. The reason is that the special conditions (1) constant direct and indirect effects on Y and (2) no relationship between unobserved causes of Y and unobserved causes of M. In effect, M is as good as randomly assigned in this special case.

[1] Text mistakenly has "multiply estimates of a and c together.

Question 4

Earlier we indicated that in Bhavnani's experiment, the pathway between random reservations for women and voter turnout appears to be zero, suggesting that we may be able to rule out this mediator as a possible pathway.

a) With the replication dataset at http://isps.research.yale.edu/FEDAI, use randomization inference to test the sharp null hypothesis of no treatment effect on turnout in 2002 for any subject.

```
# treatment: reservations for women candidates
Z <- as.integer(bhav$controltreat) - 1
# an intermediate outcome: turnout
Y <- bhav$turnout

# generate probability of treatment
probs <- genprobexact(Z)
# estimate the ITT (ATE of assignment to reservation)
ate <- estate(Y,Z,prob=probs)

numiter <- 10000
perms <- genperms(Z,maxiter=numiter)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 5.87105332845865e+30 to perform exact estimation.

# create potential outcomes under the sharp null of no effect for any unit
Ys <- genouts(Y,Z,ate=0)

# generate the sampling distribution based on the
# schedule of potential outcomes implied by the null hypothesis
distout <- gendist(Ys,perms,prob=probs)
# obtain p value
p.value.two.sided <- mean(abs(distout) >= abs(ate))

ate

## [1] -0.6234801

p.value.two.sided

## [1] 0.6028
```

b) Following the steps described in Chapter 9, use randomization inference to test the null hypothesis that $Var(\tau_i) = 0$.

```r
testvar <- var(Y[Z==1]) - var(Y[Z==0])

varlist <- rep(NA,numiter)

for (i in 1:dim(perms)[2]) {
        Zri <- perms[,i]
        varlist[i] <- var(Y[Zri==1]) - var(Y[Zri==0])
}
# p-value for one-tailed comparison
mean(varlist >= testvar)

## [1] 0.3306

# p-value for testing unequal variances
mean(abs(varlist) >= abs(testvar))

## [1] 0.7477
```

c) It is tempting to include voter turnout in 1997 as a covariate when assessing the relationship between reservations and turnout in 2002, but is turnout in 1997 a pre-treatment covariate? Explain why or why not.
Answer:
No. Turnout in 1997 occurs after random assignment and may be affected by randomly assigned reservations for women candidates in the 1997 election.

Question 5

In most places in the United States, you can only vote if you are a registered voter. You become a registered voter by filling out a form and, in some cases, presenting identification and proof of residence. Consider a jurisdiction that requires and enforces voter registration. Imagine a voter registration experiment that takes the following form: unregistered citizens are approached at their homes with one of two randomly chosen messages. The treatment group is presented with voter registration forms along with an explanation of how to fill them out and return them to the local registrar of voters. The control group is presented with an encouragement to donate books to a local library and receives instructions about how to do so. Voter registration and voter turnout rates are compiled for each person who is contacted using either script. In the table below, Treatment = 1 if encouraged to register, 0 otherwise; Registered = 1 if registered, 0 otherwise; Voted = 1 if voted, 0 otherwise; and N is the number of observations).

Table 2: Question 5 Table

Treatment	Registered	Voted	N
0	0	0	400
0	0	1	0
0	1	0	10
0	1	1	90
1	0	0	300
1	0	1	0
1	1	0	100
1	1	1	100

a) Estimate the average effect of Treatment (Z_i) on Registered (M_i). Interpret the results.
Answer:
The registration rate is 40% in the treatment group and 20% in the control group, for an ATE of 0.20, or 20 percentage points.

b) Estimate the average total effect of treatment on voter turnout (Y_i).
Answer:
The turnout rate is 20% in the treatment group and 18% in the control group, for an ATE of 0.02, or 2 percentage points.

c) Regress Y_i on X_i and M_i. What does this regression seem to indicate? List the assumptions necessary to ascribe a causal interpretation to the regression coefficient associated with M_i. Are these assumptions plausible in this case?
Answer:


```
Y <- c(rep(0, 400), rep(1, 0), rep(0, 10), rep(1, 90),
       rep(0, 300), rep(1, 0), rep(0, 100), rep(1, 100))
Z <- c(rep(0, 500), rep(1, 500))
M <- c(rep(0, 400), rep(0, 0), rep(1, 10), rep(1, 90),
       rep(0, 300), rep(0, 0), rep(1, 100), rep(1, 100))
summary(lm(Y~ Z + M))
```

```
##
## Call:
## lm(formula = Y ~ Z + M)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.708 -0.048 -0.048  0.064  0.404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04800    0.01213   3.957 8.12e-05 ***
## Z           -0.11200    0.01676  -6.683 3.90e-11 ***
## M            0.66000    0.01829  36.092  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2586 on 997 degrees of freedom
## Multiple R-squared:  0.5667,Adjusted R-squared:  0.5659
## F-statistic: 652.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

The results seem to suggest that registration has a strong effect on voter turnout, which makes intuitive sense; however, registration per se is not randomly assigned, and so this regression estimator may be biased. The regression also seems to indicate that the treatment exerts a negative effect on turnout holding registration constant. This finding makes no sense substantively; intuitively, one would think that the treatment should, if anything, have a positive effect net of its indirect via registration because the act of encouraging someone to register may also make them more interested in voting. Because Z and M are correlated, the inclusion of M (a post-treatment covariate) may lead to biased estimation of BOTH causal effects.

d) Suppose you were to assume that the treatment has no direct effect on turnout; its total effect is entirely mediated through registration. Under this assumption and monotonicity, what is the Complier average causal effect of registration on turnout?
Answer:
As noted above, the estimated ITT is 0.02, and the estimated ITT_d is 0.20, so the ratio of the two quantities is $0.02/0.20 = 0.10$. Among Compliers (those who register if and only if encouraged), the ATE of registration is a 10 percentage point increase in turnout.

Question 6

Fellner, Sausgruber, and Traxler (2009) collaborated with an Austrian tax collection agency to examine the conditions under which people who own televisions pay the mandatory annual fee when requested to do so via an official letter from the agency.[2] The researchers randomly varied the content of the mailings so that it emphasized either (1) a threat of prosecution for tax evasion, (2) a fairness appeal to pay one's fair share rather than forcing others to bear one's tax burden, or (3) information stating the descriptive norm that 94% of households comply with this tax. These interventions seem to accentuate three mediators: fear of punishment, concern for fairness, and conformity with perceived norms. There are two outcome measures. One is whether the recipient responded to the request for an explanation for non-payment by mailing in a prepaid envelope. The other outcome, which is a subset of the first, is payment of the registration fee. The table above presents an excerpt of the results.

[2]Fellner, Sausgruber, and Traxler 2009.

Table 3: Question 6 Table

	No mail	Standard letter	Letter with threat	Letter with norms	Letter with threat & norms	Letter with appeal to fairness	Letter with threat & fairness
Payment of registration fee	1.58%[3]	0.0862	0.0967	0.0823	0.097	0.0819	0.0932
Any response from recipient	N/A	0.4309	0.4501	0.407	0.4277	0.3882	0.4281
N	2586	6858	6694	6825	6960	6920	6750

a) This experiment included two control groups, one that received no letter and another that received a standard letter. Explain how the use of two control groups aids the interpretation of the results.
Answer:

The use of the standard letter helps the researchers assess the effect of the specific content of the various letters, holding constant the receipt of an official letter. For example, by comparing the STANDARD LETTER to the LETTER WITH THREAT, the researcher is able to assess the effects of threat among those who receive a letter of some sort. The NO MAIL group enables the researcher to assess the effect of receiving some sort of letter. If the aim is to assess the policy implications of sending out a given type of letter as opposed to nothing at all, the appropriate control group is the NO MAIL condition.

b) Analyze the data using the statistical model of your choice, and assess the effectiveness of threats, assertion of norms, and appeals to fairness.


```
rm(list=ls())
condition <- c(rep("No Mail", 2586), rep("Standard", 6858),
               rep("Threat", 6694), rep("Norms", 6825),
               rep("Threat+Norms", 6960), rep("Fairness", 6920),
               rep("Threat+Fairness", 6750))
no_mail <- as.numeric(grepl(pattern="Mail", condition))
standard <- as.numeric(grepl(pattern="Standard", condition))
threat <- as.numeric(grepl(pattern="Threat", condition))
norms <- as.numeric(grepl(pattern="Norms", condition))
fairness <- as.numeric(grepl(pattern="Fairness", condition))

Y <- c(rep(1, round(0.0158*2586)), rep(0, 2586 - round(0.0158*2586)),
       rep(1, round(0.0862*6858)), rep(0, 6858 - round(0.0862*6858)),
       rep(1, round(0.0967*6694)), rep(0, 6694 - round(0.0967*6694)),
       rep(1, round(0.0823*6825)), rep(0, 6825 - round(0.0823*6825)),
       rep(1, round(0.0970*6960)), rep(0, 6960 - round(0.0970*6960)),
       rep(1, round(0.0819*6920)), rep(0, 6920 - round(0.0819*6920)),
       rep(1, round(0.0932*6750)), rep(0, 6750 - round(0.0932*6750)))
```

```
fit.1 <- lm(Y~no_mail + standard + threat + norms +
              fairness + threat:norms + threat:fairness - 1)

summary(fit.1)

##
## Call:
## lm(formula = Y ~ no_mail + standard + threat + norms + fairness +
##     threat:norms + threat:fairness - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09698 -0.09665 -0.08618 -0.08194  0.98415
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## no_mail           0.015855   0.005477   2.895   0.0038 **
## standard          0.086177   0.003363  25.623   <2e-16 ***
## threat            0.096654   0.003404  28.393   <2e-16 ***
## norms             0.082344   0.003371  24.425   <2e-16 ***
## fairness          0.081936   0.003348  24.472   <2e-16 ***
## threat:norms     -0.082015   0.005839 -14.045   <2e-16 ***
## threat:fairness -0.085405   0.005856 -14.585   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2785 on 43586 degrees of freedom
## Multiple R-squared:  0.08915,Adjusted R-squared:  0.089
## F-statistic: 609.4 on 7 and 43586 DF,  p-value: < 2.2e-16
```

This regression models the proportion of people paying fees as an additive function of each letter's content. No intercept is included, so we include dummy variables for each of the core treatments (NOMAIL, STANDARD, THREAT, NORMS, FAIRNESS) and interactions between THREAT and NORMS and between THREAT and FAIRNESS. These interaction terms are coded 1 if the letter contains both of these ingredients and 0 otherwise. Regression suggests that both of these interactions are strongly negative, which implies that the addition of a second ingredient undercuts the effects of the first ingredient. For example, the effect of THREAT is an increase of 9.7 percentage points; the effect of FAIRNESS is an increase of 8.2 percentage points. However, when threat and fairness appear in same letter, the effect is 9.7 + 8.2 - 8.5 = 9.4 percentage points, which is a slightly smaller effect than THREAT alone.

c) What light do these results shed on the question of why people respond (or fail to respond) to requests to pay taxes?
Answer:
If one isolates the core treatments (THREAT, NORMS, and FAIRNESS), it appears that THREAT is most effective, and THREAT is the only core treatment that is more effective than the STANDARD letter. Neither FAIRNESS nor NORMS seems particularly effective by

themselves, nor do they appear to enhance the effectiveness of THREAT appeals.

Question 7

Several experimental studies conducted in North America and Europe have demonstrated that employers are less likely to reply to job applications from ethnic minorities than from non-minorities.

a) Propose at least two hypotheses about why this type of discrimination occurs.
Answer:
Hypothesis 1: Employers believe that ethnic minorities are less productive; according to this hypothesis, discrimination occurs because of rational economic calculations, not hostility toward ethnic minorities. Hypothesis 2: Employers tend to be hostile to ethnic minorities and discriminate against them in order to maintain "social distance" from them. Hypothesis 3: Employers themselves believe ethnic minorities to be as productive as non-minorities and do not discriminate out of animus toward them, but employers believe that their current employees look down on ethnic minorities and defer to their employees' tastes.

b) Propose an experimental research design to test each of your hypotheses, and explain how your experiment helps identify the causal parameters of interest.
Answer:
There is no ideal way to test these hypotheses, because each of them involves individual beliefs or tastes, which are unobserved. Some suggestive evidence, however, may be generated by experimentally inducing changes to beliefs or accommodating tastes. In order to test hypothesis 1, the application letter could provide evidence of qualifications and work experience attesting to the applicant's productivity; the point of this test is to see whether stereotypes about productivity can be overcome by applicant-specific information. The hostility hypothesis is more difficult to test, since it involves an interaction between the employer's attitudes and the minority treatment. In principle, one could conduct an unrelated survey of employers in order to gauge their attitudes toward various groups and assess whether their pattern of discrimination toward the fictitious applicants coincides with their general attitudes as expressed in response to the survey. Regarding the last hypothesis, one might devise a treatment that signals that the applicant is an especially likable and friendly person who fits in well in any situation.

c) Create a hypothetical schedule of potential outcomes, and simulate the results of the experiment you proposed in part (b). Analyze and interpret the results.
Answer:

Table 4: Hypothetical schedule of potential outcomes for Question 7

Employer Type	Outcome	Y(Non-minority)	Y(Minority)	Y(Productive Minority)	Y(Likeable Minority)
Hostile	Grants Interview	50	25	25	30
Hostile	No Interview	950	975	975	970
Accepting	Grants Interview	100	75	100	80
Accepting	No Interview	900	925	900	920

The above table simulates potential outcomes for 1000 people who, in response to a survey, express hostility toward minorities and 1000 people who are accepting of them. Each of these blocks could be randomly divided into four experimental groups, each of which receives one of the treatments. Suppose the results of the experiment were close to the expected proportions given above. The numbers above imply that employers in each block discriminate against minorities. Both groups are 2.5 percentage points more likely to interview a non-minority applicant than a minority applicant; since hostile employers are (for unknown reasons) less likely to interview any applicant, the ethnicity cue has a much larger effect on the odds they will grant an interview than it does on the odds that an accepting employer will grant an interview. Cues that the candidate is productive have no effect on hostile employers but eliminate the difference between minority and non-minority candidates among accepting employers. This treatment-by-covariate interaction (not necessarily causal, but suggestive) suggests that animus causes hostile employers to disregard applicants' qualifications; among the accepting, a showing of qualifications overcomes the presupposition that ethnic candidates are less productive. The likability treatment has little effect, suggesting that the consideration of who will "fit in" to the employment environment plays a small role in the decision to interview.

Question 8

Sometimes is it difficult and costly to conduct a long-term evaluation of policies or programs. For example, many states have instituted civics education requirements in high schools on the grounds that this type of curriculum makes for a more knowledgeable and involved citizenry. However, it is often impossible to track students after they leave high school. Suppose you were asked to evaluate the impact of a recommended civics curriculum that is being considered by a state that currently does not have a civics requirement. You may randomly assign a large number of schools and students to different curricula, but you can only measure outcomes up to the point at which students leave school.

a) Propose one or more mediating variables that you think explain why civics classes affect the attitudes and behaviors of students after they leave school.
Answer:
One hypothesis is that civics teaches students about the importance of public affairs. Another hypothesis is that civics provides information about how to get involved in community activities and politics.

b) Propose a research design that would shed light on whether your hypothesized mediating variables are affected by civics classes.
Answer:
Randomly assign 10^{th} grade students to three groups: a no-civics group, a group that is exposed to a yearlong curriculum that emphasizes the importance of public affairs, and a group that is exposed to a yearlong curriculum that exposes students to a variety of local community service and political opportunities. Interview students at the end of their 10^{th}, 11^{th}, and 12^{th} grade years about their interest in public affairs and willingness to volunteer for local community service or political activities.

c) One problem with measuring short term outcomes is that effects may dissipate over time. Although your study cannot address this issue directly because long-term outcomes cannot be measured, suggest ways in which your design could at least shed some light on the rate at which

effects decay over time.
Answer:
Decay could be studied by assessing whether the treatment effect observed immediately after the yearlong class diminishes when students are reinterviewed after 11th grade (one year later) and after 12th grade (two years later).

Question 9

Researchers who attempt to study mediation by adding or subtracting elements of the treatment confront the practical and conceptual challenge of altering treatments in ways that isolate the operation of a single causal ingredient. Carefully compare the four mailings from the Gerber et al. (2008) study, which are reproduced in the appendix to this chapter.

a) Discuss the ways in which the treatments differ from one another.
Answer:
The four treatments are: Civic Duty, Hawthorne, Self, and Neighbors. Civic Duty emphasizes citizens' responsibilities to participate in the Democratic process. Hawthorne simply informs subjects that they are under study. Self and Neighbors reveal voter history: the self treatment informs subjects of their past voter history and the neighbors treatment informs subjects of their own past voter history and that of their neighbors. Also, Self and Neighbors promise to send an updated vote history.

b) How might these differences affect the interpretation of Table 10.2?
Answer:
The largest difference is between the control group and the neighbors treatment. The reasons why the neighbors treatment are so effective may be many. It could be that the treatment reminds subjects of their civic duty. It could be that the treatment reminds subjects that they are being studied. It could be that the treatment reminds subjects of their own voter behavior. The other treatments in the experiment explicitly vary these factors. This allows us to conclude that social pressure is indeed the causative ingredient in the neighbors treatment.

c) Suppose you were in charge of conducting one or more "manipulation checks" as part of this study. What sorts of manipulation checks would you propose, and why?
Answer:
The following manipulation checks would be helpful. For all treatment groups, a question such as "Have you received any mail encouraging you to vote in the past three months?" would verify that treatment subjects did receive more encouragements than control subjects. For the "Self" and "Neighbors" treatments, a question such as "Did you vote in the November 2004 election" might reveal if the treatments increased subjects' recall. Another idea: ask a random subset (so as not to disrupt voting habits among a large segment of the subject pool) whether voting is a matter of public record.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 11 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Important concepts:

a) Explain the distinction between a sample average treatment effect and a population average treatment effect. Why might a researcher be primarily interested in one rather than the other?
Answer:
Define a population as a set of subjects from which an experimental sample is drawn.Depending on how a sample is drawn, the ATE for the sample may be similar or different from the ATE for the broader population; large, random samples tend to have similar ATEs to their parent populations. Researchers may be interested in the ATE for the sample because their primary goal is to figure out how the subjects in the experiment respond to the treatment. Or researchers may be interested in the ATE for the population because they seek to draw generalizations about how the intervention would work were it applied to others in the population.

b) What is a meta-analysis? Why is meta-analysis a better way to summarize research findings than comparing the number of studies that show significant estimated treatment effects to the number of studies that show insignificant estimated treatment effects?
Answer:
Meta-analysis refers to statistical procedures designed to summarize the results of research literatures. Meta-analysis is sometimes described as a "systematic" method for constructing a literature review because it summarizes research findings based on a replicable formula. Specifically, when meta-analysis is used to pool several studies, each study's experimental result is weighted according to a formula that follows from an underlying statistical model. In this chapter, the model involves random sampling from a large population, and the formula (fixed effects meta analysis) weights each study to the inverse of its precision, or squared standard error. This procedure is superior to a count of studies that show significant or insignificant results because the latter potentially accords too much weight to small studies that produce statistically insignificant results and too little weight to large studies that convincingly demonstrate an effect when other, smaller studies fail to do so. Another advantage of meta-analysis over this head-count method is that meta-analysis generates a point estimate and confidence interval, which is more informative than a summary statement about statistical significance.

c) Using equations (11.2), (11.3), and (11.4), provide a hypothetical example to illustrate how uncertainty about the possibility of bias affects the way in which prior beliefs are updated in

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

light of new evidence.
Answer:
Suppose that a researcher were to conduct a study on the effects of SAT prep classes on SAT scores using an observational design that compares a national random sample of high school seniors who take the class to those who do not. The researcher's normal prior about the ATE is centered at 30 points with a standard deviation of 15 points. The researcher's normal prior about the bias of the design is 15 points with a standard deviation of 10 points. The study's results suggest that the course increases performance by 65 points with a standard deviation of 5 points. In other words, $g = 30$, $\sigma_g^2 = 225$, $\beta = 15$, $\sigma_\beta^2 = 100$, $x_e = 65$, and $\sigma_{x_e}^2 = 25$. Plugging these numbers into equation (11.3) gives:

$$\sigma_{\bar{\tau}|x_e}^2 = \frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_\beta^2 + \sigma_{x_e}^2}} = \frac{1}{\frac{1}{225} + \frac{1}{100+25}} = 80.36$$

Plugging these numbers into equation (11.4) gives:

$$p_1 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_g^2} = \frac{\sigma_\beta^2 + \sigma_{x_e}^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{100+25}{225+100+25} = 0.357 \quad p_2 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_\beta^2 + \sigma_{x_e}^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = 1 - p_1 = 0.643$$

Finally, plugging these numbers into equation (11.2) gives the posterior estimate:

$$\begin{aligned} E[\bar{\tau}|X_e = x_e] &= p_1 * g + p_2(x_e - \beta) \\ &= 0.357 * 30 + 0.643 * (65 - 15) = 42.87 \end{aligned}$$

In the absence of uncertainty about bias (i.e., if $\sigma_\beta^2 = 0$), the weight given to the new evidence (p_2) would have been much greater: $\frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{225}{225+0+25} = 0.9$. The posterior would have more strongly shaped by the observational results:

$$\begin{aligned} E[\bar{\tau}|X_e = x_e] &= p_1 * g + p_2(x_e - \beta) \\ &= 0.1 * 30 + 0.9 * (65 - 15) = 48 \end{aligned}$$

d) What does it mean to conduct a hypothesis test that compares two "nested" models?
Answer:
Models are said to be "nested" when one model can be written as a special case of another model. For example, if one conducts an experiment with three groups, a control group and two treatments, one could estimate the ATE of each treatment, or one could estimate a nested model in which both treatments are assumed to have the same ATE. When expressed in regression form (with indicator variables for each treatment), the first model is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \epsilon_i$$

and the second model is:

$$Y_i = \beta_0 + \beta_1 (D_{1i} + D_{2i}) + \epsilon_i$$

Question 2

Identify a research literature where an experiment was replicated at least once. Carefully consider the manner in which subjects and contexts were selected, as well as the methods used to administer treatments and measure outcomes. If publication bias is a concern, note that, too. Based on your close reading of these studies, assess whether the estimated ATE from each study can be said to constitute an independent random sample from some larger population.
Answer:
Answers will vary widely.

Question 3

Suppose one were to sample N subjects at random from a population of $N*$ people. An experiment is performed whereby m of the N subjects are assigned to receive a treatment, and the remaining $N-m$ are assigned to the control group. Suppose that sometime after the treatment is administered, outcomes are measured for all $N*$ people.

a) Suppose one estimates the population ATE by comparing the mean outcome among the m subjects in the treatment group to the mean outcome among the $N^* - m$ subjects who were not assigned to the treatment. Is this estimator unbiased?
Answer:
Yes. The subjects assigned to the treatment and control groups are each random samples from the pool of N^* subjects in the population. Therefore, they have the same expected potential outcomes.

b) Would the appropriate standard error of this difference-in-means estimator be equation (11.1), equation (3.4), or neither?
Answer:
The correct formula is a modified version of equation (3.4) in which N^* replaces N.

Question 4

Suppose that $N = 2m$. Using equations (3.4) and (11.1), show that $SE(SATE) \approx SE(PATE)$ when the treatment effect is constant across subjects. Hint: In this case, $Var(Y_i(0)) = Var(Y_i(1)) = Cov(Y_i(0), Y_i(1))$.
Answer:
Substituting for $Var(Y_i(1))$ and for $Cov(Y_i(0), Y_i(1))$ and imposing the constraint that $N = 2m$ allows us to re-write equation (3.4) as:

$$\begin{aligned} SE(\widehat{SATE}) &= \sqrt{\frac{1}{N-1}(Var(Y_i(0)) + Var(Y_1(0)) + 2 * Var(Y_i(0)))} \\ &= 2\sqrt{\frac{Var(Y_i(0))}{N-1}} \end{aligned}$$

Similarly, equation (11.1) may be re-written:

$$\begin{aligned} SE(\widehat{PATE}) &= \sqrt{\frac{Var(Y_i(0))}{m} + \frac{Var(Y_i(0))}{m}} \\ &= \sqrt{\frac{2 * Var(Y_i(0))}{N} + \frac{2 * Var(Y_i(0))}{N}} \\ &= 2\sqrt{\frac{Var(Y_i(0))}{N}} \end{aligned}$$

The two formulas converge as N increases.

Question 5

Using the Bayesian updating equations, show algebraically how the priors represented in Figure 11.1 combine with the experimental results depicted in order to form a posterior distribution with a mean of 8 and a standard deviation of 0.89.

Answer:

In this example, $g = 0$, $\sigma_g^2 = 4$, $\beta = 0$, $\sigma_\beta^2 = 0$, $x_e = 10$, and $\sigma_{x_e}^2 = 1$

Plugging these numbers into equation (11.3) gives:

$$\sigma_{\bar{\tau}|x_e}^2 = \frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_\beta^2 + \sigma_{x_e}^2}} = \frac{1}{\frac{1}{4} + \frac{1}{0+1}} = 0.8$$

Plugging these numbers into equation (11.4) gives:

$$p_1 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_g^2} = \frac{\sigma_\beta^2 + \sigma_{x_e}^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{0+1}{4+0+1} = 0.2$$

$$p_2 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_\beta^2 + \sigma_{x_e}^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = 1 - p_1 = 0.8$$

Finally, plugging these numbers into equation (11.2) gives the posterior estimate:

$$\begin{aligned} E[\bar{\tau}|X_e = x_e] &= p_1 * g + p_2(x_e - \beta) \\ &= 0.2 * 0 + 0.8 * (10) = 8 \end{aligned}$$

Thus, Figure 11.1 depicts the posterior as centered at 8 with a standard error of $\sqrt{0.8} = 0.89$.

Question 6

Prior to a 2006 primary election, Gerber, Green, and Larimer sent a sample of registered voters in Michigan an encouragement to vote that disclosed whether each person registered to vote at that address had voted in previous elections.[1] (The self mailing is described in Chapter 10.) A similar mailing was sent to Michigan voters in 2007, prior to municipal elections in small cities and towns,[2] and to Illinois voters in 2009, prior to a special congressional election.[3] For comparability, we restrict each of the samples to the set of households containing just one registered voter.

[1] Gerber, Green, and Larimer 2008.
[2] Gerber, Green, and Larimer 2010.
[3] Sinclair, McConnell, and Green 2012.

Table 1: Question 6 Table

Study	Number in the control group	Number voting in the control group	Number in the treatment group	Number voting in the treatment group
2006 Michigan	26481	8755	5310	2123
2007 Michigan	348277	88960	12391	3791
2009 Illinois	15676	2600	9326	1936

a) Estimate the ATE for each study.
Answer:

```
p_c_MI06 <- 8755/26481
p_t_MI06 <- 2123/5310
ATE_MI06 = p_t_MI06 - p_c_MI06
ATE_MI06

## [1] 0.06919727

p_c_MI07 <- 88960/348227
p_t_MI07 <- 3791/12391
ATE_MI07 = p_t_MI07 - p_c_MI07
ATE_MI07

## [1] 0.05048232

p_c_IL09 <- 2600/15676
p_t_IL09 <- 1936/9326
ATE_IL09 =  p_t_IL09 - p_c_IL09
ATE_IL09

## [1] 0.04173304
```

b) Estimate the standard error for each study. Use the standard errors (squared) to calculate the precision of each study.
Answer:

```
SE_MI06 <- sqrt((p_t_MI06 * (1-p_t_MI06))/5310 +
                (p_c_MI06 * (1-p_c_MI06))/26481)
SE_MI06

## [1] 0.007317643
```

```
SE_MI07 <- sqrt((p_t_MI07 * (1-p_t_MI07))/12391 +
                (p_c_MI07 * (1-p_c_MI07))/348227)
SE_MI07

## [1] 0.004205133

SE_IL09 <- sqrt((p_t_IL09 * (1-p_t_IL09))/9326 +
                (p_c_IL09 * (1-p_c_IL09))/15676)
SE_IL09

## [1] 0.005144331

prec_MI06 <- 1/SE_MI06^2
prec_MI07 <- 1/SE_MI07^2
prec_IL09 <- 1/SE_IL09^2

prec_MI06

## [1] 18674.87

prec_MI07

## [1] 56551.04

prec_IL09

## [1] 37786.98
```

c) Assuming that these three samples are random draws from the same population, calculate a precision-weighted average of the three studies. (Hint: weight each estimate by the inverse of its squared standard error.)
Answer:

```
weighted.mean(c(ATE_MI06, ATE_MI07, ATE_IL09), c(prec_MI06, prec_MI07, prec_IL09))

## [1] 0.05064947
```

d) Show that this estimate is identical to what one obtains by using the Bayesian updating formula recursively: use the results from the 2006 study as your priors, and update them based on the 2007 study to form a posterior mean and posterior variance based on equations (11.2) and (11.3); then update this posterior using the results from the 2009 study.
Answer:

Using the MI2006 study as the prior and updating based on MI2007 implies the following parameters:

$g = 0.069$, $\sigma_g^2 = 1/18675$, $\beta = 0$, $\sigma_\beta^2 = 0$, $x_e = 0.042$, and $\sigma_{x_e}^2 = 1/56550$

Plugging these numbers into equation (11.3) gives:

$$\sigma_{\bar{\tau}|x_e}^2 = \frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_\beta^2 + \sigma_{x_e}^2}} = \frac{1}{18675 + 56550} + \frac{1}{75225} = 0.000032934$$

Plugging these numbers into equation (11.4) gives:

$$p_1 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_g^2} = \frac{\sigma_\beta^2 + \sigma_{x_e}^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{0 + 1/56550}{1/18675 + 0 + 1/56550} = 0.248$$

$$p_2 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_\beta^2 + \sigma_{x_e}^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = 1 - p_1 = 0.752$$

Finally, plugging these numbers into equation (11.2) gives the posterior estimate based on the first two studies:

$$\begin{aligned} E[\bar{\tau}|X_e = x_e] &= p_1 * g + p_2(x_e - \beta) \\ &= 0.248 * 0.069 + 0.752 * (0.051 - 0) = 0.055 \end{aligned}$$

The next step uses this estimate as the prior and updates based on IL2009. That exercise uses the following parameters:

$g = 0.055$, $\sigma_g^2 = 1/75225$, $\beta = 0$, $\sigma_\beta^2 = 0$, $x_e = 0.042$, and $\sigma_{x_e}^2 = 1/37787$

Plugging these numbers into equation (11.3) gives:

$$\sigma_{\bar{\tau}|x_e}^2 = \frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_\beta^2 + \sigma_{x_e}^2}} = \frac{1}{75225 + 37787} = 0.00000884859$$

Plugging these numbers into equation (11.4) gives:

$$p_1 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_g^2} = \frac{\sigma_\beta^2 + \sigma_{x_e}^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{0 + 1/37787}{1/75225 + 0 + 1/37787} = 0.666$$

$$p_2 = \frac{\sigma_{\bar{\tau}|x_e}^2}{\sigma_\beta^2 + \sigma_{x_e}^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = 1 - p_1 = 0.334$$

Finally, plugging these numbers into equation (11.2) gives the posterior estimate based on all three studies:

$$\begin{aligned} E[\bar{\tau}|X_e = x_e] &= p_1 * g + p_2(x_e - \beta) \\ &= 0.666 * 0.055 + 0.334 * (0.042 - 0) = 0.0507 \end{aligned}$$

This estimate matches what we obtained earlier based on a precision-weighted average.

e) Use equation (11.3) to estimate the variance of the precision-weighted average. Take the square root of the variance in order to obtain the standard error. In order to estimate the 95% confidence interval, use the following procedure, which is based on a large-sample approximation. Obtain the lower bound of the interval by subtracting 1.96 times the standard error from the precision-weighted average; obtain the upper bound of the interval by adding 1.96 times the standard error to the precision-weighted average.
Answer:

$$\begin{aligned}\sqrt{\sigma^2_{\bar{\tau}}|x_e} &= \sqrt{\frac{1}{\sigma^2_g} + \frac{1}{\sigma^2_\beta + \sigma^2_{x_e}}} \\ &= \sqrt{\frac{1}{75225 + 37787}} \\ &= 0.00297\end{aligned}$$

The Lower bound is $0.051(1.96)(.00297) = 0.0448$. The upper bound is $0.051 + (1.96)(.00297) = 0.0565$.

These results are confirmed by the `rmeta` package:

```
suppressMessages(library(rmeta))
meta_analysis <- meta.summaries(d = c(ATE_MI06, ATE_MI07, ATE_IL09),
                                se = c(SE_MI06, SE_MI07, SE_IL09))
meta_analysis$summary

## [1] 0.05064947

meta_analysis$se.summary

## [1] 0.002974651
```

f) Explain why the confidence interval formed in part (e) is likely to understate the true amount of uncertainty associated with the estimate of the population ATE.
Answer:
The formulas used to calculate the estimates and standard errors assume that the studies represent independent random samples from the same population. If the studies are biased samples from the population of citizens and campaign contexts (perhaps because they all involve Northerners and low-salience political contests) or if the studies are non-independent (because they were all conducted or analyzed by some of the same authors), the true degree of uncertainty may be understated.

Question 7

According to the logistic regression coefficients reported in Table 11.2, the intercept in Region 1 is 8.531 and the slope is -1.978. Based on these numbers, what proportion of those offered a price of 100 shillings is expected to buy a bed net? How does this compare to the actual rate of purchases at this price?

Answer:
The logistic model for Region 1 is

$$\begin{aligned} Pr[Y_i = 1] &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 ln[D-i])}} \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 ln[100])}} \\ &= 0.359 \end{aligned}$$

The corresponding empirical value from this region is 0.340.

Question 8

Table 11.3 presents observed and predicted values for each of the experimental sites in the Kenya bed nets study. Data for this experiment may be found at http://isps.research.yale.edu/FEDAI.

a) Verify that mispredictions mostly occur in Region 4 by calculating this region's contribution to the total chi-square statistics for each model.
Answer:

```
##    1    2    3    4    5    6
## 0.67 2.17 0.45 7.10 2.14 1.91
##     1     2     3     4     5     6
##  1.14  2.53  0.50 13.36  2.53  2.05
```

For Model I, Region 4's mispredictions contribute 7.10 to the overall chi-square of 14.44; For Model II, Region 4's mispredictions contribute 13.36 to the overall chi-square of 22.10. In other words, the mispredictions in Region 4 are greater (using the chi-square metric) than the other regions' combined.

b) Re-estimate the two logistic regression models presented in Table 11.3, this time excluding Region 4. Reproduce Table 11.3.
Answer:

```
suppressMessages({
  library(foreign)
  library(arm)
  library(dplyr)
  library(knitr)
})
dupas <- read.dta("/Users/alex/Documents/Dropbox/Columbia/Spring 2014/Experiments2014/FEDAI
dupas <- within(dupas,{
  purchased <- as.numeric(purchasednet =="yes")
  log_price <- log(price)
  region <- cfw_id
})
```

```r
dupas_subset <- subset(dupas, price !=0 & region!=4)

model_1 <- glm(purchased ~ log_price*region, data = dupas_subset, family = "binomial")
model_2 <- glm(purchased ~ log_price + region, data = dupas_subset, family = "binomial")

dupas_subset$preds_1 <- invlogit(predict(model_1))
dupas_subset$preds_2 <- invlogit(predict(model_2))

table_11_3_mod <-
dupas_subset %>%
  group_by(region, price) %>%
  summarize(purchases = sum(purchased),
            non_purchases = sum(purchased==0),
            pred_purchases_1 = sum(preds_1),
            pred_nonpurchases_1 = sum(1-preds_1),
            chi_square_1 = (purchases- pred_purchases_1)^2/ pred_purchases_1 +
              ( non_purchases- pred_nonpurchases_1)^2/ pred_nonpurchases_1,
            pred_purchases_2 = sum(preds_2),
            pred_nonpurchases_2 = sum(1-preds_2),
            chi_square_2 = (purchases- pred_purchases_2)^2/ pred_purchases_2 +
              ( non_purchases- pred_nonpurchases_2)^2/ pred_nonpurchases_2)

kable(table_11_3_mod[,c(1:4, 5:7)],caption = "Modified Table 11-3 (Model 1)")
```

```r
kable(table_11_3_mod[,c(1:4, 8:10)],caption = "Modified Table 11-3 (Model 2)")
```

```r
# degrees of freedom: 19 rows in all, minus 5 intercepts and 5 slopes
model_1_chi_sq <- with(table_11_3_mod, sum(chi_square_1))
pvalue_1 = pchisq(model_1_chi_sq, 9, lower.tail = FALSE)

# degrees of freedom: 19 rows in all, minus 5 intercepts and 1 slope
model_2_chi_sq <- with(table_11_3_mod, sum(chi_square_2))
pvalue_2 = pchisq(model_2_chi_sq, 13, lower.tail = FALSE)

model_1_chi_sq

## [1] 7.322821

pvalue_1

## [1] 0.603548

model_2_chi_sq

## [1] 7.686641
```

Table 2: Modified Table 11-3 (Model 1)

region	price	purchases	non_purchases	pred_purchases_1	pred_nonpurchases_1	chi_square_1
1	70	16	13	15.441596	13.558404	0.0431911
1	100	16	31	16.920757	30.079243	0.0782891
1	130	12	37	12.291135	36.708865	0.0092049
1	190	5	23	3.822194	24.177806	0.4203157
1	250	2	28	2.524317	27.475683	0.1189097
2	40	46	15	47.582604	13.417396	0.2393083
2	80	40	30	35.429156	34.570844	1.1940434
2	120	18	46	21.212681	42.787319	0.7277873
2	200	10	49	9.775560	49.224440	0.0061763
3	50	42	16	43.139894	14.860106	0.1175590
3	90	33	27	31.052284	28.947716	0.2532181
3	150	18	40	18.042950	39.957050	0.0001484
3	210	9	39	9.764872	38.235128	0.0752124
5	60	27	10	26.083853	10.916147	0.1090665
5	110	12	25	15.218789	21.781211	1.1564439
5	140	11	18	8.697359	20.302641	0.8707845
6	50	14	5	14.837853	4.162147	0.2159737
6	100	11	7	8.729828	9.270172	1.1462956
6	150	4	14	5.432319	12.567681	0.5408934

Table 3: Modified Table 11-3 (Model 2)

region	price	purchases	non_purchases	pred_purchases_2	pred_nonpurchases_2	chi_square_2
1	70	16	13	14.852988	14.147012	0.1815747
1	100	16	31	16.687555	30.312445	0.0439238
1	130	12	37	12.498147	36.501852	0.0266533
1	190	5	23	4.114794	23.885206	0.2232390
1	250	2	28	2.846515	27.153485	0.2781326
2	40	46	15	47.744038	13.255962	0.2931645
2	80	40	30	35.468776	34.531224	1.1734669
2	120	18	46	21.135136	42.864864	0.6943623
2	200	10	49	9.652050	49.347950	0.0149967
3	50	42	16	44.054580	13.945420	0.3985211
3	90	33	27	31.294685	28.705315	0.1942350
3	150	18	40	17.510472	40.489528	0.0196039
3	210	9	39	9.140263	38.859737	0.0026587
5	60	27	10	25.287203	11.712797	0.3664814
5	110	12	25	15.497341	21.502659	1.3580892
5	140	11	18	9.215456	19.784544	0.5065353
6	50	14	5	14.607437	4.392563	0.1092608
6	100	11	7	8.761556	9.238444	1.1142560
6	150	4	14	5.631007	12.368993	0.6874859

```
pvalue_2

## [1] 0.8633989
```

c) Although excluding Region 4 from the analysis improves the apparent fit of both models, why might it be considered a questionable practice to exclude Region 4?
Answer:
The mispredictions in Region 4 were not expected before the results became known; the adjustments to the model are therefore post hoc. Since the overall degree of fit is not worse than would be expected by chance when Region 4 is included, it may be that the anomalies in that region are due to random sampling variability.

Question 9

Because the log transformation of price is undefined when price is zero, we excluded the zero price condition from the analysis of bed net purchases in Tables 11.2 and 11.3.

a) If we exclude zero prices from our experimental analysis, will our estimate of the causal effect of price be biased?
Answer:
No. The exclusion is based on the treatment not on the results. Because those receiving the zero price are a random subset of all subjects, excluding these observations does not lead to biased estimates of the ATE. Thinking back to Chapter 7, missingness here is unrelated to potential outcomes.

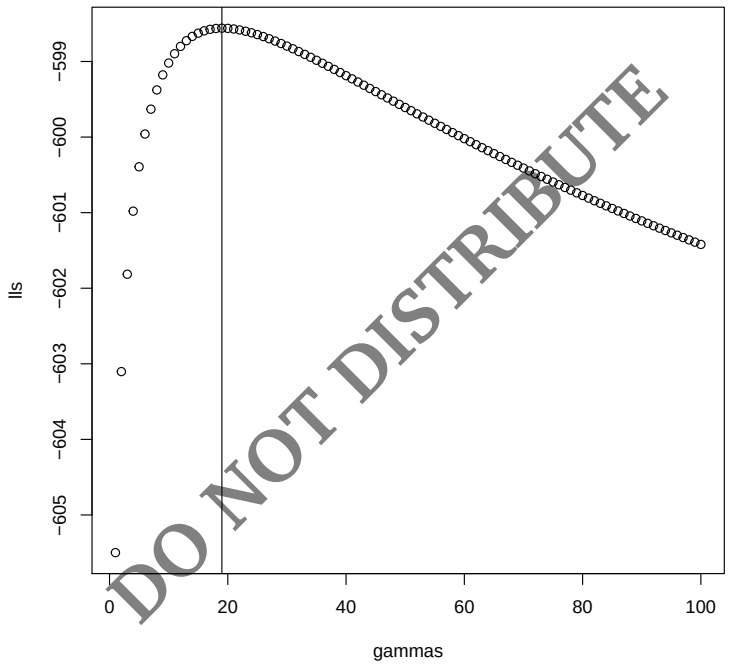
b) Suppose we reasoned that a nominal price of zero nevertheless involves some transaction cost, as villagers have to make the effort to redeem their vouchers. For a given subject, we may model the probability of making a purchase as:

$$Pr[Y = 1] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 ln[V_i + \gamma])}}$$

where γ represents the transaction cost of redeeming the voucher. In order to estimate γ, insert a positive value of γ, and use logistic regression to estimate the revised model; note the value of the log-likelihood for this model. Repeat this exercise for different values of γ. Obtain the "maximum likelihood estimate" of γ by finding the value of γ that maximizes the log-likelihood.
Answer:
We tried different values of γ until we came upon the value 19, which maximized the log-likelihood:

```
gammas <- seq(1:100)
lls <- rep(NA, length(gammas))


# Loop through possible values of gamma
for(i in 1:length(gammas)){
  dupas <- within(dupas,{
```

```
    log_price_star <- log(price+i)
  })
  fit <- glm(purchased ~ log_price_star + region, data = dupas, family = "binomial")
  lls[i] <- logLik(fit)
}
plot(gammas, lls)
abline(v = 19)

gammas[which(lls == max(lls))]

## [1] 19
```



c) What is the substantive interpretation of the maximum likelihood estimates of γ and β_1? (Note that the standard errors using this method understate the true sampling variability because they are conditional on a particular choice of γ. Ignore the reported standard errors, and just interpret the estimates.)
Answer:
The maximum likelihood estimate of γ is 19 shillings, which suggests that redeeming the voucher involves some transaction cost even when the nominal price is very small. The coefficient on the treatment variable, $ln(price + 19)$, is -1.98, which suggests that for every one unit change in this rescaled version of the treatment, the log-odds of purchase declines by -1.98. For example, suppose that the offer price rises from 0 to 100 shillings. The $ln(price + 19)$ would change from 2.94 to 4.78, and this change would reduce the log-odds of purchase by 3.63.To illustrate what this means in terms of percentage points, suppose that a person has a 50% chance of making a purchase at a price of zero (50% implies a log-odds of 0). If that person were to be offered a

price of 100, the predicted probability of a purchase is $1/(1 + e^{(} - 3.63)) = 0.026$, or 2.6%.

Question 10

This chapter discussed the modeling issues that arise when randomly assigned treatments vary in intensity or duration. The table below considers a different case, where subjects are randomly assigned treatments but then choose to take different dosages. In this experiment, subjects were randomly assigned to receive one of two pre-recorded phone messages.[4] The treatment script encouraged people to vote and revealed whether members of the household voted in the past two elections. The control script encouraged people to recycle. Both calls were made the day before the election. Both calls were answered at similar rates. The table below presents voting rates for households that answered the phone call. Voting rates are broken down by how long the person answering the phone took to hang up after the recorded message began.

Table 4: Question 10 Table

	Duration of call before respondent disconnected				
Treatment group	1-10 seconds	11-20 seconds	21-30 seconds	31-40 seconds	Total
Call encouraged voting	16.6 (187)	17.4 (784)	19.7 (983)	24.3 (2,032)	21.4 (3,986)
Call encouraged recycling	17.5 (143)	18.3 (619)	18.9 (1,132)	19.8 (2,012)	19.2 (3,906)

Entries are percent voting, with Ns in parentheses. The sample is restricted to one-voter households. Both scripts were approximately 35 seconds long.

a) Focusing only on households than answered the phone, estimate the apparent average effect of assignment to the script that encouraged voting?
Answer:
The estimated ATE is 21.4 - 19.2 = 2.2 percentage points.

b) Does this table provide convincing evidence that "the longer a person listens to a recorded message that encourages voting, the more effective that message will be in terms of boosting voter turnout?" Why or why not?
Answer:
No. The length of time one listens before hanging up is not randomly assigned, and the people who listen for a given length of time to one script are not necessarily comparable in terms of potential outcomes to those who listen the same length of time to the other script. For example, it may be that people who are very interested in politics (and very likely to vote) listen to the entire voting script but do not listen to the entire recycling script. We cannot infer anything about the effect of listening duration from these results unless we impose some strong assumptions that do not follow from the experimental design.

[4] Gerber et al. 2010.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 12 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

Question 1

Stewart Page performed an audit study to measure the extent to which gay people encounter discrimination in the rental housing market.[1] Answer the following questions, which direct your attention to specific page numbers in the original article.

a) Who are the subjects in this experiment (p. 33)?
Answer:
The subjects are landlords who advertised rental housing in two Canadian cities (Windsor and London, Ontario; N=60 for each city) and Detroit Michigan (N=60). The landlords were selected based on advertisements they placed for rental property in the classified ads section of the most recently available newspaper in each city. It is not clear how the sample was drawn from the available ads, except that some restrictions are described: advertisements were excluded if there was no phone number listed in the ad or if the ad listed preferences or specific conditions for prospective tenants. In addition, once the experiment was underway subjects were discarded if they did not understand the meaning of the call (page 34). It is unclear, but it appears that subjects were dropped if the caller could not reach the landlord, either because the line was repeatedly busy or not answered. Finally, subjects were dropped if the person who was reached was either not in charge of renting the room or was authorized to give a definite answer regarding its availability (see p. 33). After these exclusions, there remained 180 landlords (assuming that no landlord in the sample was associated with more than one rental property).

b) What is the treatment (pp. 33-34)?
Answer:
Subjects were assigned to receive the control call (an inquiry about the current availability of the advertised apartment) or the treatment call (an inquiry about the availability of the apartment prefaced by the statement: "I guess it's only fair to tell you that I'm a gay person" (or "a lesbian"). For each city 30 calls were made by a male caller and 30 by a female caller (p. 33). Thus there was assignment into 4 groups (gender x sexual orientation) with 15 subjects in each group in each city. Both caller gender and the sexual orientation prompt may both be considered as treatments. According to the write up, there was no fixed script but calls were kept "as brief as possible, generally of only seconds in duration, and were limited to direct inquiries." (p. 33).

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock
[1]Page 1998.

c) One criticism of audit studies is that in addition to differing with respect to the intended treatment (in this case, sexual orientation of the renter), the treatment and control group also differ in other ways that might be related to the outcome variable. What is the technical name for the assumption that audit studies may violate?
Answer:
Audit studies may violate the exclusion restriction. Let Y be the stated availability of the apartment, D the presumed sexual orientation of the potential renter, and Z the assignment to the treatment or control script. If the script intended to convey a gay sexual orientation (D) strikes the landlord as disagreeably defensive or odd in the context of a call regarding the rental property, for example, then Z may affect Y through pathways other than the putative sexual orientation of the prospective tenant. Similarly, if those making the calls have a viewpoint regarding anti-gay bias, this may affect how the calls are delivered apart from conveying the sexual orientation of the potential renter.

d) Suppose that the experiment used one male caller to make calls that mentioned sexual orientation and another male caller to make calls that did not. How would this procedure affect your interpretation of the apparent degree of discrimination against gay men?
Answer:
There would be a potential violation of the exclusion restriction. If there are two callers, the difference in average outcomes for the straight and gay script groups will estimate the effect of the combination of the qualities of person 1 and the gay orientation script versus the qualities of person 2 and no gay orientation prompt. Unless there is no difference in how renters respond to person 1 versus person 2, this estimand is not the average effect of sexual orientation.

e) Take a careful look at the treatment and control scripts, and consider some ways that the treatment and control conditions might differ in addition to transmitting information about the potential renter's sexual orientation. Are the scripts the same length? Do both scripts seem similar in terms of tone and style? How might the incidental differences between scripts affect the generalizations that can be drawn from this study?
Answer:

Suppose the goal is to use the findings to draw conclusions about the gay person's real world experience when calling a landlord who is made aware of the caller's sexual orientation about the availability of an apartment. To evaluate generalizability, we consider whether the scripts parallel what a gay and straight person might actually say to a potential landlord and whether, in instances in which the scripts deviate from this, any deviations might affect potential outcomes.

The gay and straight scripts are different in several ways in addition to conveying different sexual orientations. The script with the sexual orientation prompt contains more information than the control script and is also longer. It is unlikely that these difference parallel real world differences in how a gay versus straight person will interact with a landlord, and so if these differences matter for landlord response this will compromise the generalizability of the experiment. However, it is unlikely that these differences are important in this particular context. Raising the issue of sexual orientation in a preliminary inquiry may convey a degree of assertiveness, political commitment, or a lifestyle that might have an independent effect on the desirability of the potential tenant apart from the particular question of sexual orientation. If sharing one's sexual orientation in a preliminary phone call is not a common feature in real rental experiences, and this script feature is considered odd or shocking by some landlords, the results may not generalize to the typical real world rental experience.

f) How might you design an experiment to eliminate some or all of these incidental differences between scripts?
Answer:
A key design challenge in this experiment is conveying sexual orientation in a brief interaction in a naturalistic way. A treatment script that used an indirect strategy would not have the same potential to carry the baggage (convey assertiveness, etc.) that is incidental to the intended treatment. For instance, a script that referred to the potential renter's boyfriend or girlfriend ("the location is great because my boyfriend/girlfriend goes to school or works in the neighborhood") might convey sexual orientation in a more subtle fashion. This script strategy also has the benefit of eliminating differences in script length and information content.

Given that any particular script for conveying sexual orientation might be less than ideal, the researcher might diversify and try a variety of indirect methods and see if the effects are different across scripts. If the scripts are equally effective, this suggests that the common element across the scripts (sexual orientation) rather than idiosyncratic features of the scripts is driving any results you observe.

g) Based on the description on pages 33-34, how are subjects assigned to the treatment groups? What is the implication if random assignment was not used?
Answer:
The allocation to groups is described as follows: "Calls to the same city were assigned to the two conditions by way of systematic alternation of telephone numbers." It is not entirely clear what this entails. Suppose it means that the list for a city was first sorted in ascending order by phone number and then the subjects were assigned to each of the 4 conditions in an alternating fashion such that the first number (and fifth and ninth etc.) was assigned to, say, the no gay prompt and male caller group.

If telephone numbers are randomly assigned to landlords and the order of assignment to the 4 treatment and control groups was independent of the telephone numbers, the alternation method is equivalent to random assignment.

However, it is (theoretically) possible that the allocation method is not equivalent to random assignment. First, subject potential outcomes may be correlated with phone numbers. If so, the sampling distribution produced by randomization inference under the sharp null will be incorrect. Second, if there is a relationship between the potential outcomes and telephone numbers, it is possible that conscious or unconscious bias might lead the researcher to assign some numbers to certain groups, which will lead to biased estimates.

Question 2

Over the past several decades, trust in government has declined. Among the possible culprits is the rise of confrontational TV news shows, which are thought by some to produce citizen disgust and disengagement. An influential study by Mutz and Reeves investigated the effects of uncivil political discourse by scripting and producing two versions of a candidate debate.[2] Subjects were randomly assigned to be shown either the uncivil (treatment) or civil (control) debate. After viewing the treatment or control video, subjects were asked about their level of trust in government.

a) Who are the experimental subjects in the first Mutz and Reeves experiment (p.4)?
Answer:

[2] Mutz and Reeves 2005.

Footnote 7 describes the subjects. The subjects were adults from the community and college students. The adult subjects were recruited through temporary employment agencies, and college students were recruited from political science courses in response to offers of extra credit. 75% of the subjects in the first experiment were college students.

b) Let the variable X_i categorize subjects according to whether they regularly watch political television shows ($X_i = 1$) or not ($X_i = 0$). Let the conditional average treatment effect be denoted $E[(Y_i(1) - Y_i(0))|X_i = 1]$ and $E[(Y_i(1) - Y_i(0))|X_i = 0]$. Does your intuition suggest that these CATEs will be similar or different? Why?
Answer:
Those who regularly watch political shows may differ from those who do not in several ways that are likely to be relevant for potential outcomes Y(1) and Y(0). Regular viewers of political shows are likely to be better informed and more interested in politics and government. As suggested by their viewing habits, regular viewers of political shows are, compared to those who avoid political shows, more likely to be engaged by, rather than shocked or offended by, the argumentative interactions typical of today's political shows. Further, it is plausible that such differences in knowledge and taste would affect the subgroup's ATE. For instance, if the theory linking incivility to trust is that those who are shocked by incivility will think less of politicians who display incivility and consequently have less trust in government, those who regularly watch political shows, who are presumably inured to or attracted by incivility might have a smaller or no negative treatment effect. Indeed, for those who like such things, the vigorous debate and spirited if sometimes nasty exchanges in the treatment condition may be viewed as how a strong democracy deliberates, leading to an increase in trust in government. On the other hand, those who watch political shows might watch the experimental treatments more intensely, leading to a larger negative treatment effect.

c) Write the expression for the average treatment effect as a weighted average of the CATEs of those who do and do not watch political TV shows.
Answer:

$$ATE = E[(Y_i(1) - Y_i(0))|X_i = 1] * E[X_i] + E[(Y_i(1) - Y_i(0))|X_i = 1] * (1 - E[X_i])$$

d) The researchers estimated the average treatment effect and found the uncivil video reduced trust in government. Suppose that only 5% of the general public watches shows that convey this treatment. To what extent does the experiment support a claim that exposure to uncivil political programs caused a decline in trust in government among the general public?
Answer:
Suppose that any effect of exposure to uncivil television shows is confined to those who view such shows. Any effect on trust would be due to a negative CATE for this subgroup (X=1). However, as C shows, a negative ATE does not imply a negative value for the $ATE|X = 1$ because the overall negative effect might be produced by a negative CATE for the X=0 subgroup. Given the small proportion of X=1 subjects, it is plausible that the ATE is essentially the CATE for the X=0 group. Further, there is no reason to suppose that the CATE for those who choose to regularly watch political shows will be the same as the CATE for those who do not, and for reasons provided in B. the CATE for those who watch political shows might be smaller, zero, or even positive. On the other hand, if the treatment of frequent viewers has important spillover effects on non-viewers, the treatment of a small fraction of the population could conceivably have a large aggregate effect on the whole population.

e) Critics of cable TV shows argue that the programs should be encouraged to be more civil. Can the estimated ATE be used to predict the effect of increasing the civility of cable shows on the overall public level of trust in government?
Answer:
There are a number of (familiar and fairly generic) reasons why estimates from the Mutz and Reeves laboratory experiment may not generalize. Among these is the concern that subjects viewing treatments in a lab are aware they are being monitored and therefore the viewing experience is artificial. More specific to the Mutz and Reeves application, although the estimated CATE for the X=1 subgroup might provide some guidance, the ATE is a mixture of CATEs and so might be a misleading guide to the effect of changing the content of cable shows.

f) Suppose that a company which tracks television viewers provides you a list of three million potential subjects, along with data on their TV viewing habits. How would you select the subjects for a follow-up experiment if you were interested in estimating how trust in government would change if political TV programs were to become more civil?
Answer:
Rather that assemble a random sample of the potential subjects, an alternative approach would be to oversample subjects with high, medium and low levels of consumption of political shows. Then perform the Mutz and Reeves experiment and measure the CATE for each of these groups. Focusing on those who watch political shows will provide an estimate of the treatment effect among those in the population actually exposed to the treatment. Selecting subjects with varying levels of prior exposure to political shows will permit the researcher to investigate the possibility that those who appear to be the most interested in political shows (who watch the most) have the smallest negative (or perhaps positive) treatment effects. Result of this sort should be interpreted with greatest caution, however, since those who watch different amounts of political television shows may differ in ways other than their television viewing: pre-treatment viewing habits are not randomly assigned. If the researcher is truly interested in measuring the causal effect of prior television viewing, a superior approach would be to produce variation in viewing through an encouragement to view political shows, and then see if the treatment effect if different across those randomly encouraged to watch political shows and those who are not so encouraged.

g) The researchers also measure whether aggressive shows are more engaging to audiences. They use multiple outcome measures: a survey item response and a physiological measure, galvanic skin response (see pp. 10-11 for a discussion). What is the rationale for using the physiological measure? What potential problem with survey response is it designed to address?
Answer:
Survey response may be inaccurate for a variety of reasons. Respondents may misreport their feelings and opinions to conform to researcher expectations (demand effects) or to provide socially normative responses. Respondents may not accurately perceive their own psychological states such as levels of arousal in response to a stimulus, and therefore they are unable to provide accurate reports. The most obvious potential problem with survey response in the Mutz and Reeves application is that it may be socially desirable to say that you did not find the uncivil show engaging. The rationale for using a physiological measure is to avoid these potential survey reporting pitfalls. The key assumptions are, first, that the physiological measure (galvanic skin response) is a more direct window into the subject's true response. It is an involuntary response to the stimulus and is uncontaminated by concerns about investigator expectations or social norms. Second, skin response is a proxy for level of engagement with what is being viewed.

Question 3

In an experiment designed to evaluate the effects of political institutions, Olken randomly assigned 49 villages in Indonesia to alternative political processes for selecting development projects.[3] Some villages were assigned to the status quo selection procedure (village meetings with low attendance), while others were assigned to use an innovative method of direct elections (a village-wide plebiscite). Consistent with expectations, participation in the plebiscite was 20 times greater than attendance at the village meetings. Olken examines the new procedure's effect on which projects are selected and how the villagers feel about the selection process. He finds that there are minimal changes in which projects are selected. However, a survey after the project selection found that the villagers who were assigned to the plebiscite reported much greater satisfaction with the project selection process, and were significantly more likely to view the selection as fair, and the project as useful and in accordance with their own and the people's wishes.

a) One part of this experiment focuses on whether the treatment influences which projects villages select. These results are reported in Figure 1, and the study is described on pp. 244-247. Describe the experimental subjects. What units are assigned to treatment versus control? What is the treatment?
Answer:

The subjects are 49 villages in Indonesia. Villages in three regions were randomly assigned to treatment or control: North Sumatra (5 plebiscite, 14 meeting), East Java (3 plebiscite, 7 meeting) and Southeast Sulawesi (9 plebiscite, 11 meeting). These villages are all eligible to propose development projects for possible funding through a government program. The treatment involves altering the process whereby villages select which projects they will propose for funding. The standard method (control group) involves assembling two lists of possible projects (a general project selected from the ideas produced by meetings attended by men or by both genders and a women's project selected from ideas produced by meetings of women). The final step in the project proposal process is to take the list of project ideas to sparsely attended general meetings (one to which the whole village is invited, one just for women) to select which two project ideas will be proposed. In the alternative decision process, which is the treatment, the final step in this process (the village-wide meeting) is replaced with a village-wide election (one election for the general project and one for the women's project) to determine which of the project ideas will be proposed. Further details of the election procedure are found on page 247.

b) Suppose that in Indonesia, the plebiscite method is rare, but the village meeting is very common. How would this affect your interpretation of the findings?
Answer:
If the treatment is novel, the treatment is the effect of a combination of two things, the introduction of a novel form of decision making and the introduction of the particular political structure. If the estimand of interest is, say, the consequences of varying the degree of participation holding the degree of novelty of the political process constant (which is arguably what the contrast between the plebiscite and status quo decision process is attempting to capture), the difference in average outcomes across groups will not estimate this. In addition, the novelty of the method may wear off, which suggests the effects will not generalize to long term effects.

c) The level of satisfaction is measured by survey responses. From the description on p. 250, can you tell who conducted the surveys and whether the interviewers were blinded as to the

[3] Olken 2010.

respondents' assignment to treatment or control? Why might survey measures of satisfaction be susceptible to bias?

Answer:

It is not entirely clear from the information contained in the data section of the article how the survey was designed and implemented. In particular, it is not clear who interviewers were, whether they were blinded as to subject treatment or control group status, and how subjects were assigned to interviewers. The use of survey response raises two important issues regarding the accuracy of variables measured by surveys. First, there is a danger that interviewers' biases may affect the measurement. When the interviewer is not blinded as to the respondent's assignment there is a danger the results may be shaped by intentional or unintentional favoritism. This can occur in several ways. There is often some discretion in how answers are coded. For example, respondents often do not use the categories supplied and the interviewer then asks follow-up questions to determine how to classify responses within the survey categories. Efforts to obtain responses may also vary with interviewer expectations about how the respondent is likely to answer the questions. Second, respondents may shape their responses to please the interviewer or to conform to a social expectation regarding proper response. Respondents may infer what answers would please the interviewer. In this context, if the respondents assume that the interviewer is connected to the development program, there might be a tendency to report a favorable response to the novel process introduced by the interviewer's presumed organization. Setting the interviewer aside, respondents may simply believe that any novel program represents a "gift" and to respond negatively would show ingratitude. This problem is heightened if the interviewer is assumed to provide the gift. Relatedly, the response might have a strategic component: the respondent might believe that a more favorable evaluation of the intervention will lead to additional benefits. Some of these difficulties can be remedied through survey design and implementation. The subjects should be randomly assigned to interviewers to prevent interviewers from sorting themselves to certain subjects. Interviewers should be blinded as to subject group to prevent biased coding or surveying. Ideally, respondents should be unaware the survey has any link to the program that is being evaluated.

d) There is no indication that the treatment and control villages had contact with each other. Imagine, however, that people regularly communicated across village lines. What assumption might be violated by this interaction? Discuss how cross-village communication might affect treatment effect estimates. What design or measurement strategy might address possible concerns?

Answer:

The interaction across villages violates the non-interference assumption. It is conjecture as to how the inference might affect the results. Assume that the researcher wishes to estimate the effects under the assumption of global non-interference. If projects may be viewed as substitutes (if one village does a water project, the neighboring village will not), communication may exaggerate the estimated effects of the intervention on project choice, since this is based on a comparison of the treatment and control group choices. On the other hand, if villages tend to copy one another's project choices, communication will attenuate the treatment effect. Communication across villages may affect the subjective assessment of the treatment intervention as well. For instance, learning of the introduction of a novel scheme of decision making in a neighboring village may lead to reduce satisfaction with the status quo institutions.

e) Olken concludes that, consistent with the views of many democratic theorists, participation in political decision making can substantially increase satisfaction with the political process

and political legitimacy. Does the experiment provide convincing evidence for this general proposition? What are some of the limitations noted by Olken (see pp. 265-266)? What additional limitations does the experiment have? How might you address these concerns in a future experiment?
Answer:
Olken discusses several limitations. Fist, the subjects are 49 villages in 3 Indonesian provinces and results may not generalize outside the subject pool. Second the study observed outcomes over a relatively short period of time. Satisfaction levels may change decay over time if actual project choices remain unchanged. There might be strategic adaptation to the new environment which might affect the results. Third, the study was small and so might have been insufficiently powered to detect some treatment effects. These concerns can be addressed by performing a larger study over a longer period of time with a broader subject population. Running the study for a longer period of time would also address the concern that the novel of the intervention is an important factor in the subject response to the introduction of the plebiscite.

f) It is often claimed that short-term effects may diminish over time, but the short-run outcome measurements nevertheless reliably indicate the direction, if not the magnitude, of the long-term effects. However, if an institutional change is thought to be a durable feature of the political world, leaders and voters may change their behavior and the way they compete for power. Speculate on why the long-term effects of the plebiscite on satisfaction with the decision process might be negative despite the initial positive response.
Answer:
A more participatory process may lead over time to more political factions and more conflict and social tension, which may cause dissatisfaction. The short term positive response could be due to anticipated benefits of the new process and if the performance of the new system does not meet these expectations, this may lead to greater frustration and disappointment.

Question 4

In section 12.5, we considered a hypothetical experiment in which leaflets were distributed to publicize an audit that declared local government to be honest or corrupt. Suppose another experiment of this kind were conducted in 40 municipalities, half of which are honest and half corrupt. Half of the honest municipalities are randomly assigned to receive leaflets publicizing the auditor's finding of honesty, and half of the corrupt municipalities are randomly assigned to receive leaflets publicizing the auditor's report of corruption. Outcomes are the incumbent mayor's vote share in an upcoming election. The data from the experiment are used to estimate the following regression:

$$Voteshare_i = \beta_0 + \beta_1 Leaflet_i + \beta_2 Honest_i + \beta_3 Leaflet_i * Honest_i + u_i,$$

where $Voteshare_i$ is the incumbent's vote share (from 0 to 100 percent), $Leaflet_i$ is scored 1 if the municipality is randomly treated with a leaflet (0 otherwise), $Honest_i$ is scored 1 if the municipality receives an audit rating declaring it to be honest (0 if it was declared corrupt), and u_i is the disturbance term. Suppose the regression estimates (and estimated standard errors in parentheses) are as follows: $\hat{\beta}_0 = 30(4)$, $\hat{\beta}_1 = -15(5)$, $\hat{\beta}_2 = 25(5)$, $\hat{\beta}_3 = 35(7)$. Interpret the results, taking care not to assume that the average treatment effect of leaflets announcing the honest rating in honest municipalities is the same as the average treatment effect of leaflets announcing the honest rating in corrupt municipalities. (Hint: Use the regression coefficients to figure out what the regression results would be if honest and corrupt municipalities were analyzed separately. See section 9.4).

Answer:
From these four coefficients, we need to recover four group means: Treated versus untreated in honest municipalities and treated versus untreated in corrupt municipalities.
The intercept, β_0, is the average outcome in untreated corrupt municipalities: 30. $\beta_0 + \beta_1$ is the average outcome in the treated corrupt municipalities: 30 + -15 = 15. $\beta_0 + \beta_2$ is the average outcome in the untreated honest municipalities: 30 + 25 = 55. $\beta_0 + \beta_1 + \beta_2 + \beta_3$ is the average outcome in the treated honest municipalities: 30 + -15 + 25 + 35 = 75.
The average effect of treatment in the corrupt municipalities is a decrease of 15 points in the incumbent mayor's vote share. The average effect of treatment in the honest municipalities is an increase of 20 points in the incumbent mayor's vote share. These treatment effects both appear to be statistically significant, as does the difference between them.

Question 5

The Simester et al. study showed how incomplete outcome measurement can lead to erroneous conclusions. On that note, suppose researchers are concerned with the health consequences of what people eat and how much they weigh. Consider an experiment designed to measure the effect of a proposal to help people diet. Subjects are invited to a dinner and are randomly given regular-sized or slightly larger than regular-sized plates. Hidden cameras record how much people eat, and the researchers find that those given larger plates eat substantially more food than those assigned small plates. A statistical test shows that the apparent treatment effect is far greater than one would expect by chance. The authors conclude that a minor adjustment, reducing plate size, will help people lose weight.

a) How convincing is the evidence regarding the effect of plate size on what people eat and how much they weigh?
Answer:
The outcome measure is how much people eat at a single dinner. This may not be a good proxy for weight loss for a variety of reasons. Subject behavior may change along other dimensions (exercise behavior, snacking). The effects of plate size may wear off over time. Behavior at a dinner to which you are invited may differ from typical eating behavior.

b) What design and measurement improvements do you suggest?
Answer:
Several changes might improve the design. First, because other weight-related behavior may be altered in addition to the food consumption at the single dinner, the researchers would either need to obtain an accurate diary of food consumption and other activities, or else measure the variable of interest (that is, weight, after enough time has passed for digestion of the meal) directly. There are obvious limitations to these additional measures, as the diary may be inaccurate and weight is variable (lots of noise in Y) and the noise will dominate unlikely the treatment effect as weight is not likely to be affected appreciably by variation in consumption during a single meal. In any event, the study is likely far too short term to provide convincing evidence regarding weight loss. Finally, using a more naturalistic setting might also improve the study. Possibilities might be to use different size dishes to see how it affects portions in a cafeterias that people habitually eat in (this would assign groups of cafeteria regulars to treatment and control). Another possibility, this one at the household level, would be to give people a new set of (larger or slightly smaller) dishes to use in their home.

Question 6

As noted in section 12.1, experiments are sometimes motivated by a desire to test two rival explanations for an empirical regularity. Each of the three examples below features a clash between competing explanations. For each topic, propose an experiment that would, in principle, shed light on the causal influence of each explanation. Assume that you have a very large budget and a good working relationship with governments and other organizations that might implement your experiments.

a) Does imprisonment reduce crime because convicts have fewer opportunities to break the law, or does imprisonment deter crime by teaching prisoners about the penalties they face if they re-offend?
Answer:

The first explanation is at the aggregate level: the crime rates of whole towns may decrease if a larger proportion of criminals is incarcerated. The second explanation takes place at the level of the individual criminal: their experience in prison induces them not to re-offend.

A design that addresses both possible explanations will need to randomize increased incarceration at both levels. Suppose there are 500 municipalities and a list of (not currently incarcerated) criminals is available for all 500. First, we select a random 250 municipalities to receive treatment. In the treatment group municipalities, we randomly select half of the criminals to be locked up for 1 year.

We can then compare the crime rates (one year later) of untreated and treated municipalities. If there is a large decrease in the crime rates in treated municipalities, this would be good evidence for the first explanation (though it is possible that such harsh discipline deters *others* from engaging in crime, which is a third possible mechanism).

We can further compare the criminal activity of treated and untreated criminals in the year after the crackdown: If treated criminals are less likely to reoffend, then we have evidence for the second explanation.

Both, neither, or just one of the explanations might be true.

b) Do employers in the United States discriminate against black job applicants because they believe them to be less economically productive than whites, or do employers discriminate against black job applicants because they harbor negative attitudes toward black people in general?
Answer:

The approach that some audit studies take is to test the second mechanism – whether employers harbor negative attitudes towards black people in general – by providing employers with resumes that are identical except for the names, which are racially distinct. This approach tries to hold constant employers' beliefs about the economic productivity of the applicants while varying their race. The trouble is that employers may still believe that some unobservable quality of the black applicants (i.e., a quality not listed on the resume) will make them less productive.

One approach would be to gauge how much better the black applicant's resume must be in order to eliminate the racial gap. Another would be to attempt some sort of prejudice-reducing intervention aimed at employers, perhaps outside the employment setting.

c) Does face-to-face communication with voters before Election Day raise voter turnout because it reminds people about an upcoming election that they might otherwise forget, or because it conveys the importance of the choices that will be presented to voters?
Answer:

To tease apart these causal mechanisms, we can vary features of the treatment. Suppose that we vary both the timing and the content of treatment. There are two timings: one month before election day and one week before election day. There are 3 scripts: placebo, informational, and civic duty. In the informational script, the canvassers only reminds the voter that the election is coming up. In the civic duty script, the canvasser emphasizes the importance of voting.

The difference between the informational and civic duty scripts will shed light on whether the "importance of voters' choices" is a relevant causal mechanism. The difference between the placebo and the informational script will tell us how much causal work the reminder is doing. The timing will also help in this regard: the treatment effects due to "importance" should endure longer than the treatment effects due to "reminder". If the effects of the one-month importance script are similar to the one-week importance script, but the one-month reminder is much weaker than the one-week reminder, then we can conclude that the civic duty script works through the "importance" channel.

Question 7

In the Slemrod et al. experiment, measuring the outcome variables involved some effort and cost to match names and state tax return records. Outcome measurements were obtained for only a randomly selected portion of the households available to serve as control group observations.

a) Suppose that additional resources were made available to the researchers, and they gathered outcomes for randomly selected taxpayers who were not selected for treatment. (Assume that this was the only thing they could spend the money on.) How would including these additional observations in the control group affect the properties of the weighted difference-in-means estimator? Is it still unbiased? How does its standard error change?
Answer:
There are 6 types of households. For any of the 6 types, let N_t be the number of treatment households, and let N_c be the number of untreated households originally selected for the control group and let N_c^* be the number of additional households selected. The set of households originally selected for measurement from the full set of untreated households was a random sample, which implies that the $E[Y(0)]$ for the originally selected group of Nc households is the same as the $E[Y(0)]$ among the households left behind. The proposal is to take a random sample of these remaining households. Since the expected value of a random sample is the average of the group from which the random sample is drawn, the expected value of the additional control households is also equal to $E[Y(0)]$. Therefore the new difference of means estimator is an unbiased estimate of the CATE for each type of household.

Gathering additional households from the untreated for measurement and inclusion in the set of households used for the estimation of the treatment effect will increase the precision of the control group tax change estimates, and the average of the combined sample is an unbiased estimator for the change in tax payments for the subjects when they are untreated. Adding the new observations into the existing control group observations does not introduce bias. Using

formula 3.6, the estimated standard error for the estimates changes from $\sigma * (1/n_t + 1/n_c)$ to $\sigma * (1/n_t + 1/n_c^*)$.

b) Records are sometimes lost over time. Suppose that before the second round of outcome measurement were launched, some taxpayer records went missing. What additional assumption is necessary for the combined old and new control group outcome measurements to be an unbiased estimate of the same estimand as the old outcome measurements?
Answer:
The combined control group after the second round of sampling is a weighted average of a random sample of untreated households from the first round (which is an unbiased estimated of $E[Yi(0)]$, the average outcome when households are untreated) and the average of the households measured in the second round. The expected value of the households that can be measured in the second round is $E[Y_i(0)|R_i(0) = 1]$, where $R_i(0)$ denotes whether an household is missing or not when untreated, and $R_i(0) = 1$ if the household is not missing. Unbiasedness requires that the expected value of the second round random sample be $E[Y_i(0)]$, therefore the requirement for unbiasedness is $E[Y_i(0)|R_i(0) = 1] = E[Y_i(0)]$. This assumption is satisfied if the households are missing at random.

Question 8

According to social psychologists, performance on standardized tests may be affected by seemingly minor contextual features, such as the instructions read to those about to take a test and the similarity between the test-taker and other students taking the test at the same time. This literature implies that subtle asymmetries across treatment and control in how outcomes are measured may have a material effect on test scores. Suppose you were designing an experiment similar to the voucher experiment described in section 12.6. Instead of bringing students to a common testing center for testing, you have decided to use the standardized tests that students ordinarily take in their own schools.

a) What are some important potential sources of asymmetry in outcome measurement? Consider among other things how the test is administered, who proctors the test, who grades the test, the mixture of students in the room for a testing session, and whether the administration and grading is blinded to the subject's group status.
Answer:
If students take the tests in their respective schools, the list of issues raised in subsection (a) may lead to differences in test scores unrelated to academic achievement. First, conducting separate testing sessions opens the door to a number of measurement asymmetries. If different standardized tests are used for treatment and control groups, this clearly leads to measurement differences due to differences in the test. The testing sessions themselves may produce differences in student motivation, attention, and stress levels. These differences may be produced in a variety of ways. There may be differences in how the test is described to the students (whether it measures the results of effort or intelligence), including discussion of expected test performance. There may be differences in testing conditions, such as room temperature, noise levels, and crowding. Proctors in some testing centers may offer hints while others do not. Second, a testing regime may systematically favor one experimental group over the other. If those who administer the test are unblinded, they may treat students differently or explicitly favor one group over the other. If the tests are graded by unblinded graders, this could lead to fudging the results unconsciously or to cheating.

b) How would you design your study to reduce bias due to asymmetric outcome measurement of the treatment and control subjects?
Answer:
Conduct testing using the same tests, mixing control and treatment students together, assigned to randomly selected seating. This will ensure common test and no bias in the physical testing environments. If there is concern that an imbalance in T and C group students will create a less friendly environment for one group versus the other, effort to reduce the effect of the environment might be employed (no talking, space between desks). To avoid priming stereotypes, the instructions for the test should be limited to logistics and not include "welcoming remarks" that refer to private school kids or voucher winners, etc. Those who administrator the test (from first contact inviting the families to the session and on from there) should be blinded to the group assignment of the family. The test graders should be blinded.

c) Suppose you want to investigate the impact of the measurement asymmetries you discuss in part (a). Describe an experimental design to estimate the effect of the measurement asymmetries.
Answer:
The effect of the mix of students T and C group students taking the test during a session can be randomly varied (set up X classrooms and randomly vary the mix across each) to see if it affects test scores. The effect of crowding and other physical room conditions can be randomly varied as well. The effect of different descriptions of the test provided prior to the testing can be studied by randomly varying the descriptions.

The effect of blinding the proctors or graders can be studied by randomly providing some of the subjects (proctors and graders) the group assignments while leaving others unaware. In addition to test differences, classrooms could be videotaped to see if proctors are differentially helpful to one group versus the other. The key design challenge would be to conduct this exercise in an unobtrusive way and so as not tip off the subjects, so that the results provide insight into how unblinded agents behave in a natural setting. An interesting design, if it were possible, would be to provide the subjects (graders) with the group assignment of some but not all of the students under their supervision. In the case of proctoring, this would provide variation in blinding within a classroom, permitting the effect of blinding to be distinguished from other factors that might lead a teacher to favor one group over the other.

Question 9

As pointed out in section 12.4, sending resumes via email seems to have several advantages over typical face-to-face audit studies of racial discrimination. However, an email treatment is a more subtle method of communicating race than a face-to-face meeting. What if some employers do not notice the name on the job application or incorrectly guess the race of the applicant? For simplicity, assume that each human resource officer either concludes that the applicant is black or white. Suppose that when sent any white resume, a human resources officer has an 80% chance of surmising that it is from a white applicant. When sent any black resume, a human resources officer has a 90% chance of surmising that it is from a black applicant. Suppose that making a mistaken classification of a white resume is independent of making a misclassification of a black resume. Recall from Table 12.6 that 9.65% of the white resumes received callbacks, as opposed to 6.45% of the black resumes.

a) For definitional purposes, consider assignment to the white resume to be assignment to treatment, and consider assignment to the black resume to be assignment to control. To show how

misclassification is analogous to noncompliance, use the classification system in Chapter 6 to describe the four types of subjects: what proportion of subjects are Compliers, Never-Takers, Always-Takers, and Defiers?
Answer:
Compliers are the HR officers who think the applicant is white (D=1) when the "white" resume is sent (Z=1), and black (D=0) when the "black" resume is sent (Z=0) are 72% of the subject pool. Always Takers (HR thinks the candidate is white regardless of whether Z=1 or 0) are 8% Never takers: 18% Defiers: 2%

b) What is the ITT_D in this case?
Answer:

$$ITT_D = \pi_{compliers} - \pi_{defiers} = 0.72 - 0.02 = 0.70$$

c) What assumption(s) are needed to interpret the ratio of ITT/ITT_D as the Complier average causal effect? Suppose that when analyzing the data in Table 12.6, you assumed that these assumptions were satisfied; what would be your estimate of the CACE?
Answer:
The analyst could assume the absence of Defiers or, alternatively, that the treatment effect is the same for Defiers and Compliers. Under either assumption, the estimated CACE is: (9.65-6.45)/.7 = 4.57.

d) Does the rate of noncompliance have any bearing on the statistical significance of the relationship between race and interviews that the authors report in Table 12.6?
Answer:
No. The calculations in Table 12.6 are intent to treat effects, and the estimation of the ITT and calculation of its statistical significance does not involve the non-compliance rates. As suggested by part (c), the interpretation of the ITT may be affected by the compliance rate, however, since one reason for a small ITT is high rates of non-compliance. To convert an ITT into the CACE, requires either monotonicity or the assumption of homogenous treatment effects for defiers and and compliers. If either assumption holds, the rescaled ITT (ITT/c, where c is the estimated proportion of compliers minus the proportion of defiers, or the difference in the proportion treated in the treatment group minus the proportion treated in the control group)) is an estimate of the CACE. The standard error of the CACE is approximately equal to the ITT standard error divided by c, and the significance level of the CACE is approximately the same as that of the ITT.

e) What steps do Bertrand and Mullainathan take to reduce the rate of misclassification? Do they measure the rate of misclassification? What methods might you use to measure misclassification rates? What are some strengths and weaknesses of your proposal?
Answer:
Bertrand and Mullanathan compiled a list of the most racially distinctive names based on official records. To see if the names conveyed the information they intended, they performed a pilot study and found that individuals guessed the intended race with very high probability. They were however unable to directly measure how much misclassification by employers occurred in their experiment. Additional steps might be taken to investigate how much misclassification might have occurred. From the report provided in the paper, it appears that the pilot work to confirm the racial interpretation of the names did not involve HR workers and did not look at

the black and white names attached to the resumes. It is also unclear how resumes are evaluated by firms. For example, if resumes are sorted by putative race of applicant, this might be studied directly. Perhaps the best method to test the level of misclassification would be to work with a set of employers and have the HR office code each resume they process according to beliefs about the race of the applicant. Putting aside the feasibility of this proposal, introducing this coding might heighten attention to the racial "clues" in the resume. Having the HR worker fill out the form after processing the resume would avoid this issue, but only the first resume would avoid the potential distortion associated with the coding.

A simple modification of the pilot testing in Bertrand and Mullanathan would be to tests the names on HR workers and test names attached to resumes (with HR workers).

Question 10

One limitation of the restorative justice experiment described in section 12.6 is that one cannot identify the distinct effects of an apology or a no-show; instead, one can only estimate the effects of a treatment that is a combination of the two. Suppose that in order to identify the ATE of an apology as well as the ATE of a no-show, you assigned subjects randomly to one of three experimental groups: a control group, a standard encouragement group, and a strong encouragement group. The identification proof posits three different types of subjects: Compliers (those who show up when encouraged in any way), Reluctant-Compliers (those who show up only when strongly encouraged), and Never-Takers.

a) Write the expected outcome in the control group as a weighted average of the expected outcomes among Compliers, Reluctant-Compliers, and Never-Takers.
Answer:
Let the subjects offenders be of three types, T_i, where $T_i = 1$ for the Compliers, $T_i = 2$ for the Reluctant compliers, and $T_i = 3$ for the Never takers). Let there be three potential outcomes for each subject, Y(0), Y(-1), Y(1), for control group assignment, no show, and apology respectively.
Expected Outcome for the Control group:

$$E[Y_i(0)|T_i = 1] * P(T_i = 1) + E[Y_i(0)|T_i = 2] * P(T_i = 2) + E[Y_i(0)|T_i = 3] * P(T_i = 3),$$

where $P(X)$ is the proportion of the subjects for whom $T_i = x$

b) Write the expected outcome in the standard encouragement group as a weighted average of the expected outcomes among Compliers, Reluctant-Compliers, and Never-Takers. Your model should acknowledge that Compliers will offer an apology, but Reluctant-Compliers and Never-Takers will be no-shows.
Answer:
Expected Outcome for the Standard Encouragement group:

$$E[Y_i(1)|T_i = 1] * P(T_i = 1) + E(Y_i(-1)|T_i = 2] * P(T_i = 2) + E[Y_i(-1)|T_i = 3] * P(T_i = 3)$$

c) Write the expected outcome in the strong encouragement group as a weighted average of the expected outcomes among Compliers, Reluctant-Compliers, and Never-Takers. Your model should acknowledge that Compliers and Reluctant-Compliers will offer an apology, but Never-Takers will be no-shows.

Answer:
Expected Outcome, Strong Encouragement:

$$E[Y_i(1)|T_i = 1] * P(T_i = 1) + E[Y_i(1)|T_i = 2] * P(T_i = 2) + E[Y_i(-1)|T_i = 3] * P(T_i = 3)$$

d) Explain why the experimental design allows us to estimate the shares of the three types of subjects.
Answer:
From the strong encouragement group, we can estimate the percentage of $T_i = 3$ types, $P(T_i = 3)$ using the proportion that does not offer an apology. From the standard encouragement group, we can estimate the sum of $P(T_i = 2) + P(T_i = 3)$ using the proportion that does not offer an apology. From these two estimates, we can estimate $P(T_i = 2)$ and $P(T_i = 3)$. Since $P(T_i = 1) = 1P(T_i = 2)P(T_i = 3)$, we can estimate the share of the subject pool for each type.

e) Notice that in the three equations (a), (b), and (c) there are four parameters: the ATE of a no-show among Never-Takers, the ATE of a no-show among Reluctant-Compliers, the ATE of an apology among Compliers, and the ATE of an apology among Reluctant-Compliers. No matter how you manipulate the three equations, you cannot solve for each of the four parameters. In other words, with more unknown parameters than equations, you cannot identify either of the apology effects or either of the no-show effects. Suppose you assume instead that the ATE of a no-show is the same regardless of whether a Reluctant-Complier or Never-Taker is at fault and that the ATE of an apology is the same regardless of whether it comes from a Complier or Reluctant-Complier. Now you have reduced the number of unknowns to just two parameters. Revise your equations (a), (b), and (c) to reflect this assumption, and show that it allows you to identify the apology effect and the no-show effect.
Answer:

[ANSWER IS MISSING!]

Question 11

One reason for concern about attrition in the school voucher experiment described in section 12.7 was that, after the first year, the attrition rate was greater in the control group than the treatment group. Intuitively, the problem with comparing the treatment and control group outcomes is that the post-attrition control group is no longer the counter-factual for the post-attrition treatment group in its untreated state. The trimming bounds described in Chapter 7 attempt to extract from the post-attrition treatment group (which has a larger percentage of the randomly assigned group reporting) a subset of subjects who can be compared to the control group and used to bound the treatment effect. The dataset for this exercise at http://isps.research.yale.edu/FEDAI contains subjects of any race in the Howell and Peterson study who took a baseline math test. The outcome measure (Y_i) is the change in math scores that occurred between the baseline test and the test that was taken after the first year of the study.

a) What percentage of the control group is missing outcome data? What percentage of the treatment group is missing outcome data?
Answer:

```
with(howell,prop.table(table(missing_y1math, treat), 2))

##                treat
## missing_y1math         0         1
##              0 0.7590090 0.8110073
##              1 0.2409910 0.1889927
```

24% of the control group has missing outcome data, compared with 19% of the treatment group.

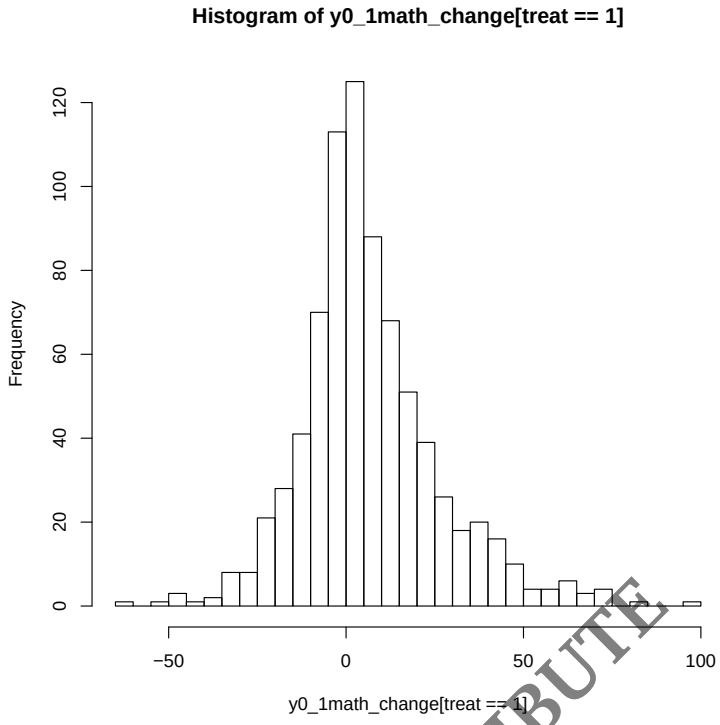
b) Among students with non-missing outcome data, what are the average outcomes for the control group and treatment group?
Answer:

```
with(howell, mean(y0_1math_change[treat==0],na.rm=TRUE))

## [1] 6.486647

with(howell, mean(y0_1math_change[treat==1],na.rm=TRUE))

## [1] 7.104994
```

6.487 for the control group, 7.105 for the treatment group.

c) What is the distribution of outcomes for the treatment group? What is the range of outcomes? What outcomes correspond to the 5%, 10%, 15%, 25%, 50%, 75%, 85%, 90%, and 95% percentiles?
Answer:


```
with(howell, hist(y0_1math_change[treat==1], breaks=50))
with(howell, c(min(y0_1math_change[treat==1], na.rm = TRUE),
               max(y0_1math_change[treat==1], na.rm = TRUE)))

## [1] -63  96

with(howell, quantile(y0_1math_change[treat==1],
                      probs = c(.05, .10, .15, .25, .50, .75, .85, .90, .95),
                      na.rm=TRUE))

##  5% 10% 15% 25% 50% 75% 85% 90% 95%
## -20 -13  -9  -4   4  16  25  32  43
```



d) To trim the top portion of the treatment group distribution, what value of Y_i is the 93.6 percentile of the treatment group? (The value 93.6 is the control group reporting rate divided by the treatment group reporting rate.)
Answer:

```
with(howell, quantile(y0_1math_change[treat==1],
                      probs = c(.936),na.rm=TRUE))

## 93.6%
##    40
```

The 93.6 percentile value of Y is 40.

e) What is the average value of the treatment group observations that are less than the 93.6 percentile value? Call this average treatment effect L_B. Confirm that the percentage of the original treatment group that remains is equal to the percentage of the control group with outcome data.
Answer:

```
howell <- within(howell,{
  treat_less_than_40 <- as.numeric(treat == 1 & y0_1math_change <40 & missing_y1math==0)
})

L_B <- with(howell, mean(y0_1math_change[treat==1 & treat_less_than_40==1]))
```

```
L_B

## [1] 3.701513

with(howell, mean(treat_less_than_40[treat==1]==0))

## [1] 0.2450675
```

The average value of the observations less than or equal to 40 is 3.70. There are 727 such values, and 1- (727/963) = 24.5%. The rate of missing for the control group is 24.1%.

f) Subtract the control group average from L_B.
Answer:

```
L_B - with(howell, mean(y0_1math_change[treat==0], na.rm=TRUE))

## [1] -2.785134
```

g) To trim the bottom portion of the treatment group distribution, what treatment group outcome corresponds to the 6.4 percentile? (The value 6.4 is calculated by subtracting 93.6 from 100.)
Answer:

```
with(howell, quantile(y0_1math_change[treat==1],
                      probs = c(0.064),na.rm=TRUE))

## 6.4%
##  -18
```

The 6.4 percentile value is -18.

h) What is the average value of the treatment group observations that are greater than the 6.4 percentile? Call this average treatment U_B. Confirm that the percentage of the original treatment group that remains after trimming is equal to the percentage of the control group with outcome data.
Answer:

```
howell <- within(howell,{
  treat_greater_than_18 <- as.numeric(treat == 1 & y0_1math_change >-18 & missing_y1math==0)
})

U_B <- with(howell, mean(y0_1math_change[treat==1 & treat_greater_than_18==1]))
U_B
```

```
## [1] 9.707586

with(howell, mean(treat_greater_than_18[treat==1]==0))

## [1] 0.2471443
```

The average of the values that remain after trimming off the lower 6.4% is 9.71. The percentage of those reporting with outcomes greater than -18 is 725/963=75.3% for a missing rate of 24.7%. This is approximately equal to the missing rate for the control group of 24.1%

i) Subtract the control group average from U_B.
Answer:

```
U_B - with(howell, mean(y0_1math_change[treat==0], na.rm=TRUE))

## [1] 3.220939
```

j) The lower and upper bounds that you calculated in parts (f) and (i) are designed to bound an ATE for a particular subgroup. Describe this subgroup.
Answer:
(3.22, -2.79) are the estimated bounds for the treatment effect for the always reporters.

Question 12

In private school voucher studies, treatment group observations are much more expensive than control observations. Assume the experiment is free from attrition and non-compliance. Suppose that the researchers have a fixed budget of $2M, each treatment group observation costs $2,000, and each control observation costs $200. The table below shows four possible ways to use the budget to form treatment and control groups. Let the standard deviation of outcomes in the treatment group and the control group be the same, and equal to s.

Table 1: Question 12 Table

	Option 1	Option 2	Option 3	Option 4
Treatment	950	750	600	900
Control	500	2500	4000	1000

Estimate the standard error for the difference-in-means estimator using the formula in equation (3.6), letting the number of observations assigned to treatment be n_t and the number of observations assigned to control be n_c. The standard error may be written:

$$s\sqrt{\frac{1}{n_c}+\frac{1}{n_t}}$$

a) In the table, which allocation of subjects to treatment and control produces the most precise estimate?
Answer:
The standard errors are S times: Option 1, .05525, Option 2: .041633, Option 3: .04378, Option 4: .045947. Therefore, Option 2 produces the most precise estimate.

There is a general method for minimizing the standard error subject to a budget constraint. Suppose the cost per observation in the control and treatment groups are p_c and p_t, respectively, and both groups have the same standard deviation. To minimize the standard error of the difference-in-means, assign subjects to groups in proportion to the square root of the cost ratio. The following questions illustrate the derivation behind this idea.

b) If n_t is the number of subjects you assign to the treatment group, how much money is spent on the treatment group?
Answer:
$p_t * n_t$

c) If n_c is the number of subjects you assign to the control group, how much money is spent on the control group?
Answer:
$p_c * n_c$

d) Express the budget B as the total spent on the treatment group and control group.

$$B = p_t * n_t + p_c * n_c$$

Set up a constrained maximization problem by defining the Lagrangian equation (Dixit 1990)

$$L(q, n_c, n_t) = s\sqrt{\frac{1}{n_c} + \frac{1}{n_t}} - q(B - n_t p_t - n_c p_c)$$

Take the partial derivative of L with respect to n_c, n_t, and q, and set each of the partial derivatives equal to zero. (If your calculus is rusty, use an online calculator to take derivatives.) The values of n_c and n_t that satisfy these conditions minimize the standard error subject to the budget constraint.

partial wrt n_c: (1) $-1/n_c^2 + q * p_c = 0$, which can be written $1/n_c^2 = q * p_c$ Partial wrt n_t: (2) $-1/n_t^2 + q * p_t = 0$, which can be written $1/n_t^2 = q * p_t$ Partial wrt to q: (3) $B = p_t * n_t + p_c * N_t$

e) Set the partial derivative with respect to n_t equal to the partial derivative with respect to n_c. Manipulate the resulting equation to show that

$$\frac{p_c}{p_t} = (\frac{n_t}{n_c})^2$$

Answer:
Three equations and three unknowns: n_c, n_t, q. The values which solve these three equations satisfy the necessary condition for minimizing the standard error subject to the budget constraint. From (1) and (2):

$$\frac{1}{p_c * n_c^2} = \frac{1}{p_t * n_t^2}$$
$$\frac{p_c}{p_t} = (\frac{n_t}{n_c})^2$$

From this result it follows that the ratio of the size of the treatment group to the size of the control group is equal to the inverse of the square root of the ratio of the costs of each type of observation. Thus, if a treatment group observation costs 10 times as much as a control group observation, the standard error minimizing division of resources places $\sqrt{10} \approx 3.2$ times as many observations in the control group.

f) When the cost of treatment and control group observations is the same, what is the appropriate way to allocate the budget to n_t and n_c?

Answer:
If $p_t = p_c$, then $\frac{p_c}{p_t} = 1$, which implies that the treatment and control groups should be the same size.

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 13 Exercises

Alan S. Gerber and Donald P. Green*

January 19, 2016

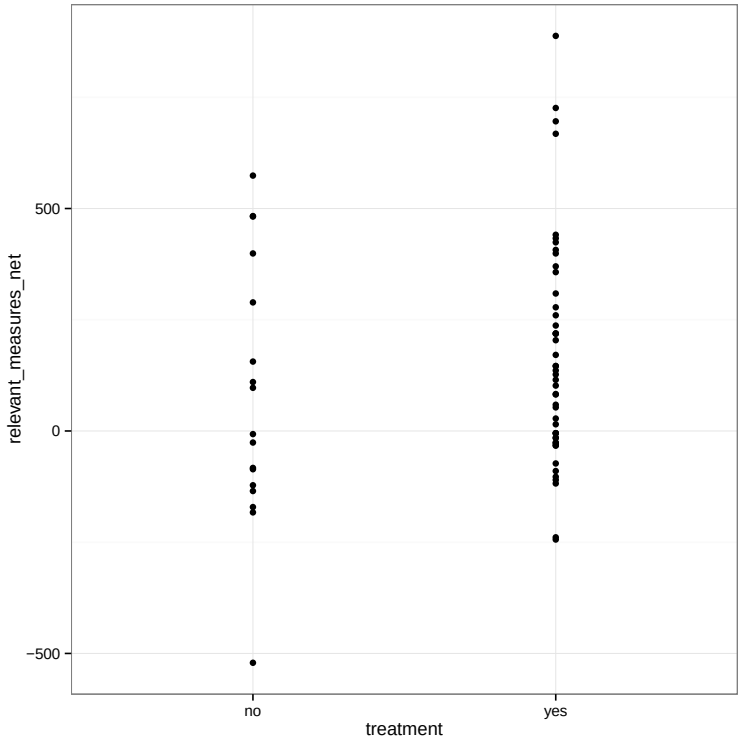
Question 1

Middleton and Rogers report the results of an experiment in which ballot guides were mailed to randomly assigned precincts in Oregon prior to the 2008 November election. The guides were designed to encourage voters to support certain ballot measures and oppose others. Load the example dataset from `http://isps.research.yale.edu/FEDAI`. The dataset contains election results for 65 precincts, each of which contains approximately 550 voters. The outcome measure is the number of net votes won by the sponsors of the guide across the four ballot measures that they endorsed or opposed. The treatment is scored 0 or 1, depending on whether the precinct was assigned to receive ballot guides. A prognostic covariate is the average share of the vote cast for Democratic candidates in 2006.

a) Estimate the average treatment effect, and illustrate the relationship between treatment and outcomes graphically using an individual values plot.
Answer:

```
with(middleton,
     mean(relevant_measures_net[treatment=="yes"])-
     mean(relevant_measures_net[treatment=="no"]))

## [1] 90.20098

ggplot(middleton, aes(x=treatment, y=relevant_measures_net)) +
  geom_point() + theme_bw()
```

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

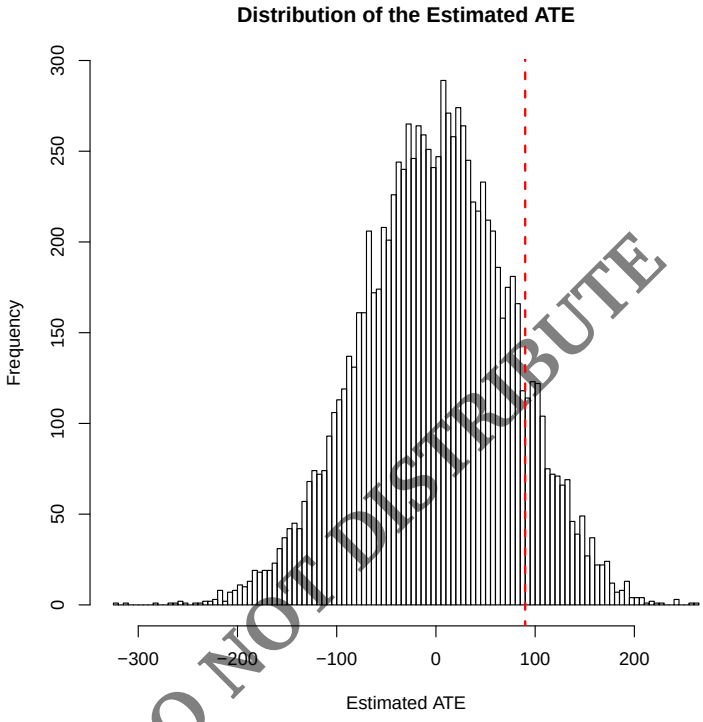


b) Interpret the graph in part (a).
Answer:
The mean of the treatment observations (164) is higher than the mean of the control observations (74), suggesting that the the treatment led to 90 more Democratic votes per precinct. The amount of dispersion around the mean is similar in both groups.

c) Use randomization inference to test whether the apparent difference-in-means could have occurred by chance under the sharp null hypothesis of no treatment effect for any precinct. Interpret the results. Answer:


```
library(ri)
middleton <-
  within(middleton,{
    Z <- as.numeric(treatment=="yes")
    Y <- relevant_measures_net
  })

# Conduct Randomization Inference
set.seed(1234567)
perms <- with(middleton, genperms(Z = Z))

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 1867897112363099 to perform exact estimation.
```

```
probs <- genprob(perms)
ate <- with(middleton, estate(Y=Y, Z=Z, prob = probs))
Ys <- with(middleton, genouts(Y = Y, Z = Z, ate = 0))
distout <- gendist(Ys=Ys, perms=perms, prob=probs)
result <- dispdist(distout, ate=ate)
result$greater.p.value

## [1] 0.1144
```

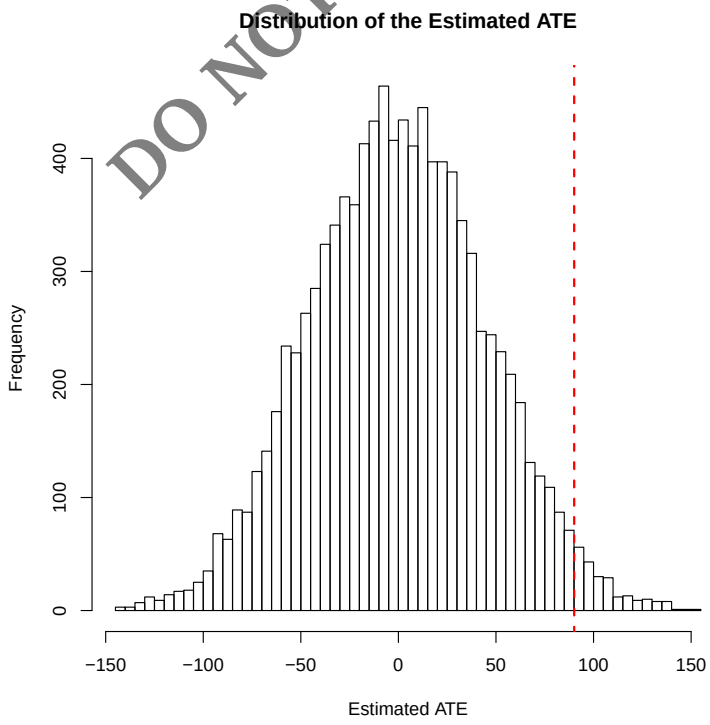


A one-tailed test is appropriate here given that the campaign sought to increase its votes. Randomization inference applied to 10,000 simulated randomizations shows that one-tailed p-value of the estimated ATE is 0.114. This figure is short of the conventional 0.05 threshold.

d) Suppose it were the case that when randomly assigning precincts, the authors used the following screening procedure: no random allocation was acceptable unless the average 2006 Democratic support score in the treatment group was within 0.5 percentage points of the average 2006 Democratic support score in the control group. Do all subjects have the same probability of being assigned to the treatment group? If not, re-estimate the ATE, weighting the data as described in Box 4.5. Redo your hypothesis test in part (c) subject to this restriction on the randomization. Interpret the results.
Answer:

```
# Write restricted RA function
restricted_ra <- function(){
  middleton$Z_sim <- ifelse(1:65 %in% sample(1:65, 48), 1, 0)
```

```
  # check condition
  fit <- lm(dem_perf_06 ~ Z_sim, data=middleton)
  if(abs(coef(fit)[2]) < 0.5){return(middleton$Z_sim)}
  # if condition is not met, call restricted_ra again
  return(restricted_ra())
}

set.seed(1234567)
perms <- genperms.custom(randfun = restricted_ra)
probs <- genprob(perms)
# Restricted randomization changes the probabilities that each unit enters into treatment.
# Here is the distribution of probabilities:
summary(probs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7281  0.7346  0.7370  0.7385  0.7430  0.7641

ate <- with(middleton, estate(Y=Y, Z=Z, prob = probs))
Ys <- with(middleton, genouts(Y = Y, Z = Z, ate = 0))
distout <- gendist(Ys=Ys, perms=perms, prob=probs)
result <- dispdist(distout, ate=ate)
result$greater.p.value

## [1] 0.022
```

Randomization inference applied to 10,000 simulated restricted randomizations shows that one-tailed p-value of the estimated ATE is 0.022. This figure allows us to reject the null hypothesis at the conventional 0.05 threshold. The p-value here is lower than when we assume unrestricted randomization because re-randomization functions as a form of blocking..

Question 2

Select a published article that presents the design and analysis of a field experiment. Based on the publication and any supplementary materials provided by the authors, try to fill in as much of the reporting checklist for research articles as you can. What pieces of information, if any, went unreported? Does the failure to address one or more items on the checklist affect the confidence that you place in the results they report?
Answer:
Answers to this question will vary.

Question 3

Conduct your own randomized experiment, based on one of the suggested topics in Appendix B.

a) Compose a planning document.

b) Take an online research ethics course, and obtain your certification to conduct human subjects research. Obtain approval for your study from the institutional review board at your college or university.

c) Conduct a small pilot study to work out any problems in administering the treatment or measuring outcomes.

d) Conduct the experiment. Construct a data file and supporting metadata.

e) Compose a research report.

Answer:
Answers to this question will vary.