Field Experiments: Design, Analysis and Interpretation
Solution Sets for Odd Number Exercises
(Stata Version)

Alan S. Gerber and Donald P. Green[*]

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*

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

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

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

> 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

[1] Devoto et al. 2011.

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

Alan S. Gerber and Donald P. Green*

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:
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

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

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

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

			$Y_i(1)$		
		15	20	30	
$Y_i(0)$	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

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

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

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

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

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

Alan S. Gerber and Donald P. Green*

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.

d) How does complete random assignment differ from block random assignment and clustered random assignment? Answer:

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

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

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

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 1: 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:

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

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.

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

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"

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

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

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]

```
In [1]: clear
        set seed 1234567

        qui input D Y
                0 1
                0 0
                0 0
                0 4
                0 3
                1 2
                1 11
```

Table 2: 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	?	?

```
                1 14
                1 0
                1 3
        end

In [2]: qui gen Y_star= Y+D*(-7)

        cap program drop ate
        program define ate, rclass
                args Y D
            sum `Y' if `D'==1, meanonly
            local Y_treat=r(mean)
            sum `Y' if `D'==0, meanonly
            local Y_con=r(mean)
            return scalar ate_avg = `Y_treat'-`Y_con'
        end

In [3]: tsrtest D r(ate_avg): ate Y_star D

Two-sample randomization test for theta=r(ate_avg) of ate Y_star D by D

Combinations:   252 = (10 choose 5)
Assuming null=0
Observed theta: -2.6

Minimum time needed for exact test (h:m:s):  0:00:00
Mode: exact

progress: |...|

 p=0.83730 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.20635 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.41270 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]
```

```
In [4]: // ate
        di r(obsvStat)

-2.6


In [5]: // p.value.onesided
        di r(lowertail)

.20634921
```

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.

Question 8

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.[2] 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)?

[2] 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.

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 3: 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?
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.

```
In [1]: rename treatment D
        rename pair block
        rename preexperimentbets covs

In [2]: // calculate probs under block assignment
        qui bysort block: egen probs=mean(D)

        // permuation to calculate F stat and one-side P value
        ritest D e(F), strata(block) reps(10000) right nodots: ///
        regress D covs


      Source |       SS           df       MS      Number of obs   =        34
-------------+----------------------------------   F(1, 32)        =      0.02
       Model |  .005024372         1  .005024372   Prob > F        =    0.8914
    Residual |  8.49497563        32  .265467988   R-squared       =    0.0006
-------------+----------------------------------   Adj R-squared   =   -0.0306
       Total |         8.5        33  .257575758   Root MSE        =    .51524

------------------------------------------------------------------------------
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        covs |  -.0000386   .0002809    -0.14   0.891    -.0006109    .0005336
       _cons |   .5137818   .1335793     3.85   0.001     .2416896     .785874
------------------------------------------------------------------------------


      command:  regress D covs
        _pm_1:  e(F)
  res. var(s):  D
   Resampling:  Permuting D
Clust. var(s):  __000005
     Clusters:  34
Strata var(s):  block
       Strata:  17

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _pm_1 |   .0189265    3736   10000  0.3736  0.0048  .3641064   .3831672
------------------------------------------------------------------------------
Note: Confidence interval is with respect to p=c/n.
Note: c = #{T >= T(obs)}



In [3]:// p.value
        di el(r(p),1,1)
```

```
.3736
```

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 3736 of the simulated experiments, implying a p-value of .3736.

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.

```
In [4]: rename experimentbets change

In [5]: tabstat change, by(D) stat(mean) save
        di "ATE ="%180.4f el(r(Stat2),1,1)-el(r(Stat1),1,1)




Summary for variables: change
     by categories of: D

       D |      mean
---------+----------
       0 |  571.4118
       1 |  461.2353
---------+----------
   Total |  516.3235
--------------------

ATE =                 -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:

```
In [6]: qui bysort block (D): gen pair_diff = change - change[_n+1]
        mean(pair_diff)


Mean estimation                     Number of obs   =         17

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------
   pair_diff |   110.1765   104.8377     -112.0695    332.4225
------------------------------------------------------------------
```

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.

```
In [7]: cap program drop ate_block
         program define ate_block, rclass
         args Y D probs
         tempvar ipw
         gen `ipw' = .
// calculate inverse probability weight under block assignment
         replace `ipw' = `D'/`probs' + (1-`D')/(1-`probs')
         qui reg `Y' `D' [iw=`ipw']
         return scalar ate=_b[`D']
         end

In [8]: ritest D r(ate), strata(block) reps(10000) nodots: ///
         ate_block change D probs

(34 missing values generated)
(34 real changes made)

      command:  ate_block change D probs
        _pm_1:  r(ate)
  res. var(s):  D
   Resampling:  Permuting D
Clust. var(s):  __00000A
     Clusters:  34
Strata var(s):  block
       Strata:  17

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _pm_1 |  -110.1765    3170   10000  0.3170  0.0047  .3078845   .3262222
------------------------------------------------------------------------------
Note: Confidence interval is with respect to p=c/n.
Note: c = #{|T| >= |T(obs)|}


In [9]: // ate
         di el(r(b),1,1)
```

```
-110.17647


In [10]: // p.value.twosided
         di el(r(p),1,1)

.317
```

A two-tailed test generates a p-value of 0.317, 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

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.

```
In [1]: clear
        set seed 1234567
        qui set obs 14

In [2]: qui input Y0 Y1
                0 0
                1 0
                2 1
                4 2
                4 0
                6 0
                6 2
                9 3
                14 12
                15 9
                16 8
                16 15
                17 5
                18 17 end
```

```
In [3]: qui gen int cluster = (_n+1)/2
```

```
In [4]: //ssc install tabstatmat  (install the package)
        // save tabstat summary result to matrix
        qui tabstat Y0, by(cluster) stat(mean) save
        qui tabstatmat Ybar0, nototal
        mat colnames Ybar0=Ybar0

        qui tabstat Y1, by(cluster) stat(mean) save
        qui tabstatmat Ybar1, nototal
        mat colnames Ybar1=Ybar1
```

```
In [5]: // function to calculate population variance
        cap program drop var_pop
        program define var_pop, rclass
                args varname
                tempvar x_dev
                qui sum `varname'
                local avg = r(mean)
                local length = r(N)
                gen `x_dev' = (`varname'-`avg')^2/`length'
                qui tabstat `x_dev', stat(sum) save
                return scalar variance_pop = el(r(StatTotal),1,1)
        end
```

```
In [6]: // function to calculate population covariance
        cap program drop cor_pop
        program define cor_pop, rclass
                args x y
                tempvar xy_dev
                qui sum `x'
                local avg_x = r(mean)
                local length = r(N)

                qui sum `y'
                local avg_y = r(mean)

                gen `xy_dev' = (`x'-`avg_x')*(`y'-`avg_y')
                qui tabstat `xy_dev', stat(sum) save
                return scalar cor_pop = el(r(StatTotal),1,1)/`length'
        end
```

```
In [7]: preserve
        clear
        qui set obs 7
        svmat Ybar0, names(col)
        svmat Ybar1, names(col)

In [8]: // var_Ybar0
        var_pop Ybar0
        scalar var_Ybar0=r(variance_pop)

        // var_Ybar1
        var_pop Ybar1
        scalar var_Ybar1=r(variance_pop)

        // cov_Ybar0
        cor_pop Ybar0 Ybar1

        scalar cov_Ybar0=r(cor_pop)

        scalar se_ate = sqrt((1/6)*((4/3)*var_Ybar0+(3/4)*var_Ybar1+2*cov_Ybar0))

        di "se_ate ="%8.6f se_ate

        restore

se_ate =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 $\{1, 14\}, \{2, 13\}, \{3, 12\} \dots \{7, 8\}$. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to the treatment.

```
In [9]: qui replace cluster = _n
        qui replace cluster = 15-cluster if (cluster>7)

        clear matrix
        // Ybar0
        qui tabstat Y0, by(cluster) stat(mean) save
        qui tabstatmat Ybar0, nototal
        mat colnames Ybar0=Ybar0

        // Ybar1
        qui tabstat Y1, by(cluster) stat(mean) save
        qui tabstatmat Ybar1, nototal
        mat colnames Ybar1=Ybar1
```

```
In [10]: preserve
         clear
         qui set obs 7
         svmat Ybar0, names(col)
         svmat Ybar1, names(col)
```

```
In [11]: // var_Ybar0 <- var.pop(Ybar0)
         var_pop Ybar0
         scalar var_Ybar0=r(variance_pop)

         // var_Ybar1 <- var.pop(Ybar1)
         var_pop Ybar1
         scalar var_Ybar1=r(variance_pop)

         // cov_Ybar0 <- cov.pop(Ybar0,Ybar1)
         cor_pop Ybar0 Ybar1
         scalar cov_Ybar0=r(cor_pop)

         // se_ate
         scalar se_ate = sqrt((1/6)*((4/3)*var_Ybar0+(3/4)*var_Ybar1+2*cov_Ybar0))
         di "se_ate ="%8.7f se_ate

         restore
```

```
se_ate =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

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

Alan S. Gerber and Donald P. Green*

Question 1

Important concepts: [10pts]

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:

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

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

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

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

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? [6 pts]
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 1: 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

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. [10pts]

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?

```
In [1]: clear
        clear matrix
        clear mata
        set matsize 11000
        set maxvar 32767
        set seed 1234567
        set more off
```

```
In [2]:
// loop to simulte random assignment and save to a matrix
        cap matrix drop z
        matrix z=J(40, 10000, .)

        qui forvalues i = 1/10000 {
        //create and save 50 permutations of treatment
                import delim
       "./data/chapter04/GerberGreenBook_Chapter4_Exercises_4-5.csv"
                tempvar teststat Z
                gen `Z' = .
                gen `teststat' = -1
                while `teststat' < 0.05{
                        tempvar rannum Zri t
                    gen `rannum'=uniform()
                        egen `Zri' = cut(`rannum'), group(2)
                        qui reg `Zri' x
                        gen `t' = _b[x]/_se[x]
                        replace `teststat' = 2*ttail(e(df_r),abs(`t'))
                }
                replace `Z' = `Zri'
                forvalues j = 1/40 {
                matrix z[`j', `i'] = `Z'[`j']
                }
                drop _all
        }
```

```
In [3]: qui import delim
        "./data/chapter04/GerberGreenBook_Chapter4_Exercises_4-5.csv", clear
        rename d D
        rename y Y
        matrix rowm = z * J(colsof(z), 1, 1/colsof(z))
        matrix colnames rowm=probs
        svmat double rowm, names(col)
        qui gen weights = (1/probs)*D +(1/(1-probs))*(1-D)
        tabstat weights, by(D) stat(v)
```

```
Summary for variables: weights
     by categories of: D (D)

       D |  variance
---------+----------
       0 |  .0004185
       1 |   .000625
---------+----------
```

```
   Total |   .0005084
--------------------
```

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.

```
In [4]: qui reg Y D [pw=weights]
        global ate_restricted_RA = _b[D]

        di "ATE (Restricted Assignment)= " $ate_restricted_RA

ATE (Restricted Assignment)= 10.712532

In [5]: svmat z

        cap matrix drop y_dis
        matrix y_dis=J(10000, 1, .)

        forvalues i = 1/10000 {
                tempvar weight`i'
                gen `weight`i'' = (1/probs)*z`i' +(1/(1-probs))*(1-z`i')
                qui reg Y z`i' [pw=`weight`i'']
                matrix y_dis[`i', 1] = _b[z`i']

        }

In [6]: preserve
         svmat y_dis
         // p value
         count if abs(y_dis1) > abs($ate_restricted_RA)
         di r(N)/_N
         restore

  56

.0056
```

The IPW estimate of the ATE is 10.71, 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 0.0056, 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.

```
In [7]: qui reg Y x D [pw=weights]
         global ate_cov_restricted_RA = _b[D]
         di "ATE Controlling Covariance (Restricted Assignment)= "
                         $ate_cov_restricted_RA

ATE Controlling Covariance (Restricted Assignment)= 5.3186105

In [8]: cap matrix drop cov_dis
         matrix cov_dis=J(10000, 1, .)
         forvalues i = 1/10000{
                         tempvar weight`i'
                         gen `weight`i'' = (1/probs)*z`i' +(1/(1-probs))*(1-z`i')
                         qui reg Y x z`i' [pw=`weight`i'']
                         matrix cov_dis[`i', 1] = _b[z`i']

         }

In [9]: preserve
         svmat cov_dis

         // p.value controlling covariance
         count if abs(cov_dis1) > abs($ate_cov_restricted_RA)
         di r(N)/_N
         restore

  28

.0028
```

The IPW estimate of the ATE is 5.32, 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 0.0028, 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.

```
In [10]: /*-----------------------se_complete_RA

         // calculate ate_complete_RA
         qui reg Y D
         global ate_complete_RA = _b[D]
         di "ATE under Complete Assignment = " $ate_complete_RA

ATE under Complete Assignment = 10.7

In [11]: // RI under the null ate=ate_complete_RA
         cap drop Y0_sim Y1_sim Y_sim
         qui gen Y0_sim = Y
         qui gen Y1_sim = Y
         qui gen Y_sim = .
         qui replace Y0_sim = Y - $ate_complete_RA if D==1
         qui replace Y1_sim = Y + $ate_complete_RA if D==0

In [12]: capture program drop ate_complete_RA_ri
         program define ate_complete_RA_ri, rclass
                 replace Y_sim = Y0_sim*(1-D) + Y1_sim*(D)
                 regress Y_sim D
             return scalar Ys_complete_RA=_b[D]
         end

         tsrtest D r(Ys_complete_RA) using distout_complete_RA.dta, ///
         overwrite: ate_complete_RA_ri

Two-sample randomization test for
theta=r(Ys_complete_RA) of ate_complete_RA_ri by D

Combinations:   137846528820 = (40 choose 20)
Assuming null=0
Observed theta: 10.7

Minimum time needed for exact test (h:m:s):  392479:42:00
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.48995 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.50995 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.48995 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to distout_complete_RA.dta...done.


In [13]: // calculate se_complete_RA
         preserve
         use "distout_complete_RA.dta", clear
         qui drop if _n==1
         tabstat theta, stat(sd)
         restore
```

```
    variable |        sd
-------------+----------
       theta |  4.673591
------------------------
```

```
In [14]: /*--------------------se_cov_complete_RA

         // calculate ate_cov_complete_RA
         qui reg Y D x
         global ate_cov_complete_RA = _b[D]
         di "ATE Controlling Covariance under Complete Assignment= " $ate_cov_complete_RA
```

```
ATE Controlling Covariance under Complete Assignment= 5.3155362
```

```
In [15]: // RI under the null ate= ate_cov_complete_RA
         cap drop Y0_sim Y1_sim Y_sim
         qui gen Y0_sim = Y
         qui gen Y1_sim = Y
         qui gen Y_sim = .
         qui replace Y0_sim = Y - $ate_cov_complete_RA if D==1
         qui replace Y1_sim = Y + $ate_cov_complete_RA if D==0
```

```
In [16]: capture program drop ate_cov_complete_RA_ri
         program define ate_cov_complete_RA_ri, rclass
                 replace Y_sim = Y0_sim*(1-D) + Y1_sim*(D)
                 regress Y_sim D x
             return scalar Ys_cov_complete_RA=_b[D]
         end

         tsrtest D r(Ys_cov_complete_RA) using distout_cov_complete_RA.dta, ///
         overwrite: ate_cov_complete_RA_ri
```

```
Two-sample randomization test for
theta=r(Ys_cov_complete_RA) of ate_cov_complete_RA_ri by D

Combinations:   137846528820 = (40 choose 20)
Assuming null=0
Observed theta: 5.316

Minimum time needed for exact test (h:m:s):  408561:47:43
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.50495 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.49495 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.50495 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to distout_cov_complete_RA.dta...done.
```

```
In [17]: // calculate se_cov_complete_RA
         preserve
         use "distout_cov_complete_RA.dta", clear
         qui drop if _n==1
         tabstat theta, stat(sd)
         restore
```

```
    variable |        sd
-------------+----------
       theta |  1.577595
------------------------
```

```
In [18]: /*---------------------se_restricted_RA

         // calculate ate_restricted_RA
         di "ATE under Restricted Assignment= " $ate_restricted_RA
```

```
ATE under Restricted Assignment= 10.712532
```

```
In [19]: // RI under the null ate= ate_restricted_RA
         cap drop Y0_sim Y1_sim Y_sim
         qui gen Y0_sim = Y
         qui gen Y1_sim = Y
         qui gen Y_sim = .
         qui replace Y0_sim = Y - $ate_restricted_RA if D==1
         qui replace Y1_sim = Y + $ate_restricted_RA if D==0
```

```
In [20]: // clean space
         drop __00*
```

```
In [21]: cap matrix drop distout_restricted_RA
         matrix distout_restricted_RA=J(10000, 1, .)

         qui forvalues i = 1/10000 {
                 tempvar weight`i'
                 gen `weight`i'' = (1/probs)*z`i' +(1/(1-probs))*(1-z`i')
                 replace Y_sim = Y0_sim*(1-z`i') + Y1_sim*(z`i')
                 qui reg Y_sim z`i' [pw=`weight`i'']
                 matrix distout_restricted_RA[`i', 1] = _b[z`i']

         }
```

```
In [22]: /*se_restricted_RA*/
         preserve
         svmat distout_restricted_RA
         tabstat distout_restricted_RA, stat(sd)
         restore

    variable |        sd
-------------+----------
distout_re~1 |  4.139767
------------------------
```

```
In [23]: /*---------------------se_cov_restricted_RA

         // calculate ate_restricted_RA
         di "ATE Controlling Covariance(Restricted Assignment)=
          " $ate_cov_restricted_RA

ATE Controlling Covariance(Restricted Assignment)= 5.3186105
```

```
In [24]: // RI under the null ate= ate_cov_restricted_RA

         cap drop Y0_sim Y1_sim Y_sim
         qui gen Y0_sim = Y
         qui gen Y1_sim = Y
         qui gen Y_sim = .
         qui replace Y0_sim = Y - $ate_cov_restricted_RA if D==1
         qui replace Y1_sim = Y + $ate_cov_restricted_RA if D==0
```

```
In [25]: // clean space
         drop __00*
```

```
In [26]: cap matrix drop distout_cov_restricted_RA
         matrix distout_cov_restricted_RA=J(10000, 1, .)

         qui forvalues i = 1/10000 {
                 replace Y_sim = Y0_sim*(1-z`i') + Y1_sim*(z`i')
                 tempvar weight`i'
                 gen `weight`i'' = (1/probs)*z`i' +(1/(1-probs))*(1-z`i')
                 qui reg Y_sim x z`i' [pw=`weight`i'']
                 matrix distout_cov_restricted_RA[`i', 1] = _b[z`i']

         }
```

```
In [27]: /*se_restricted_RA*/
         preserve
         svmat distout_cov_restricted_RA
         tabstat distout_cov_restricted_RA, stat(sd)
         restore
```

```
    variable |        sd
-------------+----------
distout_co~1 |  1.592966
------------------------
```

Table 2: Summary of Estimated Standard Errors

	Without Covariates	With Covariates
Complete Random Assignment	4.674	1.578
Restricted Random Assignment	4.140	1.593

Without covariates and assuming complete randomization, we obtain a standard error of 4.674. Under restricted randomization, the standard error declines to 4.140. Including a covariate and assuming complete randomization, we obtain a standard error of 1.578. Under restricted randomization, the standard error remains essentially unchanged at 1.593. 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

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. [10pts]

Table 3: Question 7 Table

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

Table 4: 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

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.

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 5: 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

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.[1] 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. [10pts]

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 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.

```
In [1]: clear
        input ests shareoftotalN
        .00964 .049487
        -.007829 .1520981
        -.01362 .626616
        .008271 .171799
```

[1] Gerber and Green 2005.

```
        end

        qui gen overall_ate = ests*shareoftotalN
        total(overall_ate)

        ests  shareof~N

Total estimation                  Number of obs    =          4

--------------------------------------------------------------
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
 overall_ate |  -.0078273   .0090322     -.0365719    .0209173
--------------------------------------------------------------
```

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.

```
In [2]: clear
        qui import delim ./data/chapter04/Gerber_Green_AAAPSS_2005, clear
        qui bysort strata: egen blockpr = mean(treat2)
        qui gen q = blockpr*treat2 + (1-blockpr)*(1-treat2)
        regress vote02 treat2 [aw=1/q]

(sum of wgt is   1.8814e+06)

      Source |       SS           df       MS      Number of obs   =   940,715
-------------+----------------------------------   F(1, 940713)    =     57.98
       Model |  14.4110635         1  14.4110635   Prob > F        =    0.0000
```

```
    Residual |  233826.144   940,713  .248562679   R-squared       =    0.0001
-------------+----------------------------------   Adj R-squared   =    0.0001
       Total |  233840.555   940,714  .248577734   Root MSE        =    .49856

------------------------------------------------------------------------------
      vote02 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      treat2 |   -.007828   .0010281    -7.61   0.000    -.0098429    -.005813
       _cons |   .4661975   .0007269   641.31   0.000     .4647727    .4676223
------------------------------------------------------------------------------
```

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

Question 10

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

Alan S. Gerber and Donald P. Green*

Question 1

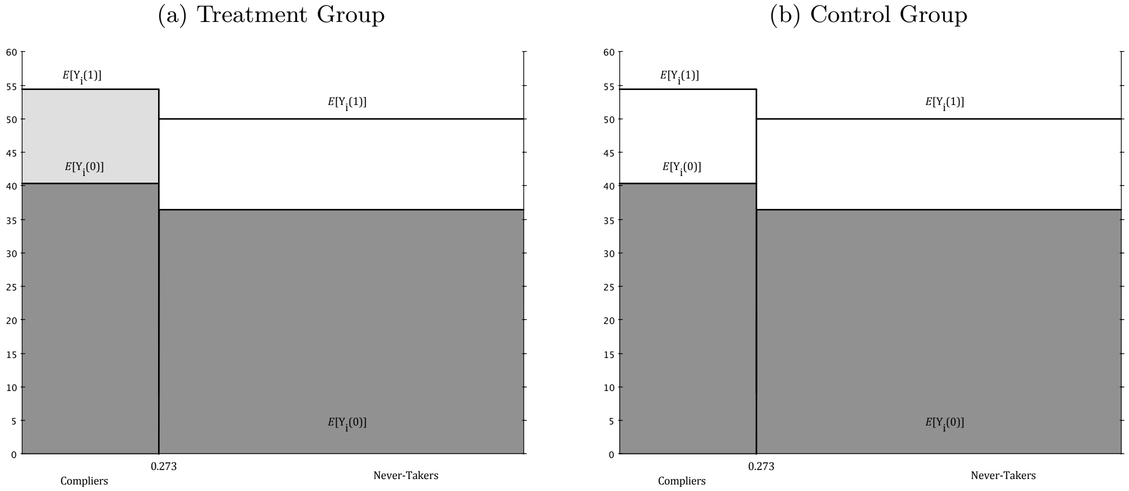
Using the data in Table 5.2: [5 pts]

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 $E[Y_i(1)|d_i(1) = 1]$.

$$
\begin{aligned}
E[d_i(1)] &= \frac{395}{1445} = 0.273 \\
E[(Y_i(0)|d_i(1) = 0)] &= 36.48 \\
E[(Y_i(0)|d_i(1) = 1)] &= \frac{37.54 - (.727 * 36.48)}{0.273} = 40.36 \\
E[(Y_i(1)|d_i(1) = 1)] &= 54.43
\end{aligned}
$$

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

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. [8 pts.]

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, the ITT is a function of the ITT_D: the ITT_D is in both the numerator (multiplied by the $CACE$) and the denominator.

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

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." [5 pts.]
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 among them. 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

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. [5 pts.]
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 exclustion 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 1: 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

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.[1] 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. [10 pts.]

Table 2: Question 9 Table

	Control	Treatment group #1	Treatment group #2
		(minimal effort)	(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]

[1] Gerber and Green 2005.

- $D_i(2) = 1, D_i(1) = 1$ [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) =

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

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

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

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

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

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

(EV, minimal - EV,control) =

$$\begin{aligned}&E(Y(1)|\text{ easy to reach}) * Pr(\text{easy to reach}) -\\&E(Y(0)|\text{ easy to reach}) * Pr(\text{easy to reach})\\&= E(Y(1) - (Y(0)|\text{ easy to reach}) * Pr(\text{easy to reach})\\&= (\text{ATE}|\text{ easy to reach}) * Pr(\text{easy to reach}).\end{aligned}$$

$$\text{ATE}|\text{ easy to reach} = \frac{\text{(EV, minimal - EV,control)}}{Pr(\text{easy to reach})}$$

Similarly,

(EV, maximal - EV,minimal) =

$$\begin{aligned}&E(Y(1)|\text{ hard to reach}) * Pr(\text{hard to reach}) -\\&E(Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})\\&= E(Y(1) - (Y(0)|\text{ hard to reach}) * Pr(\text{hard to reach})\\&= (\text{ATE}|\text{ hard to reach}) * Pr(\text{hard to reach}).\end{aligned}$$

$$\text{ATE}|\text{ hard to reach} = \frac{\text{(EV, maximal - EV,minimal)}}{Pr(\text{hard to reach})}$$

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

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).[2] Based on the results presented below, calculate the following: [10 pts.]

[2] Nickerson 2005, 2008.

Table 3: 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?

```
In [1]: clear
        qui set obs 7723
        // ssc install egenmore (install the package)
        qui egen z = repeat(), values("baseline")
        qui replace z = "treatment" in 2573/5144
        qui replace z = "placebo" in 5145/7723
        qui egen d = fill(0,0)
        qui replace d = 1 in 2573/3058
        qui replace d = 1 in 5145/5614
        qui egen y = fill(1,1)
        qui replace y=0 in 804/2572
        qui replace y=0 in 2763/3058
        qui replace y=0 in 3742/5144
        qui replace y=0 in 5285/5614
        qui replace y=0 in 6293/7723

In [2]: qui sum d if z=="treatment"
        scalar pr_c_treatment =  r(mean)
        di %8.3f pr_c_treatment

   0.189

In [3]: qui sum d if z=="placebo"
        scalar pr_c_placebo =  r(mean)
        di %8.4f pr_c_placebo

  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 fairly small and not statistically significant. These rates of compliance are consistent with the null hypothesis that both groups have the same proportion of Compliers.

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

```
In [4]: // rate.nt.treatment
        qui sum y if z=="treatment" & d==0
        di %8.4f r(mean)

  0.3274

In [5]: // rate.nt.placebo
        qui sum y if z=="placebo" & d==0
        di %8.4f r(mean)

  0.3215
```

Yes, the turnout rate among the encouragement never takers is 32.7% versus 32.2% for the placebo group. If $D_i(1)$ is the same for all subjects whether the $Z = 1$ means that they are assigned to the placebo or the encouragement, and the subjects are randomly assigned to each group, then the groups of untreated placebo and encouragement subjects are formed by random assignment from the same pool of subjects (the non-compliers). A prediction that follows from this claim is that 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. One implication of this is that perhaps the pattern of subject compliance is not the same for the two treatments.

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

```
In [6]: qui sum y if z=="placebo"
        scalar y_placebo = r(mean)
        qui sum y if z=="baseline"
        scalar y_baseline = r(mean)

In [7]: scalar itt_placebo = y_placebo - y_baseline
        scalar cace_placebo = itt_placebo/pr_c_placebo

In [8]: // Estimate the CACE of receiving the placebo
        disp %8.3f cace_placebo

   0.027
```

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). The fact that the placebo group has higher turnout than the control group makes the GOTV vs. placebo comparison more conservative.

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.

```
In [9]: qui sum y if z=="treatment"
        scalar y_treat = r(mean)
        qui sum y if z=="baseline"
        scalar y_base= r(mean)
        scalar itt_treatment = y_treat - y_base
        scalar cace_treatment1 = itt_treatment/pr_c_treatment
        disp %8.7f cace_treatment1

0.1440329
```

```
In [10]: qui sum y if z=="treatment" & d==1
         scalar yd_treat = r(mean)
         qui sum y if z=="placebo" & d==1
         scalar yd_placebo = r(mean)
         scalar cace_treatment2 = yd_treat - yd_placebo
         disp cace_treatment2

.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 conservative estimates of the CACE in this case.

Question 12

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

Alan S. Gerber and Donald P. Green*

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 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 receive the treatment if assigned to it ($D_i(1) = 1$). In the case of one-sided non-compliance, this subgroup is the Compliers. For two-sided non-compliance this is composed of Always-Takers and 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

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

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 Compliers, 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

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): $ITT_D = \pi_C + \pi_{AT} - (\pi_D + \pi_{AT}) = \pi_C - \pi_D$, which in this case implies that the $ITT_D = 0$.

Question 6

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

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.[1] 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.

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

[1] Galiani, Rossi, and Schargrodsky 2010.

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 volution 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 that 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

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.[2] 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 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: $\widehat{CACE} = \frac{365-360.5}{0.798-0.139} = 6.82$

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.

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

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

Alan S. Gerber and Donald P. Green[*]

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 $r_i(Z_i)$, which indicates whether outcomes will be observed given a treatment assignment, has no effect on the potential outcomes of $Y_i(Z_i)$.

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.

Question 2

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

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 $5 - 4.25 = 0.75$, which is biased.

Question 4

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 hyopthesis is true, our degree of belief in the hypothesis grows if neither test shows signs of asymmetry.

Question 6

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

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, 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

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

Alan S. Gerber and Donald P. Green*

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

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

Question 5

In their study of spillover effects, Sinclair, McConnell, and Green sent mailings to ran-domly 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.

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

[1] Sinclair, McConnell, and Green 2010.

effects?

```
In [1]: clear
        qui set obs 18859
        qui egen z_ind = fill(0,0)
        qui replace z_ind = 1 in 9534/18859

        // ssc install egenmore (uncomment to install the package)
        qui egen z_zip = repeat(), values("none")
        qui replace z_zip = "half" in 6218/12482
        qui replace z_zip = "all" in 12483/18859

        qui egen Y = fill(1,1)
        qui replace Y=0 in 1022/6217
        qui replace Y = 0 in 6744/9533
        qui replace Y=0 in 10154/12482
        qui replace Y=0 in 13799/18859

In [2]: qui mean Y if z_ind==1 & z_zip=="half"
        scalar ate_treat_half = _b[Y]
        qui mean Y if z_ind==0 & z_zip=="half"
        scalar ate_untreat_half = _b[Y]
        qui mean Y if z_ind==0 & z_zip=="none"
        scalar ate_untreat_none = _b[Y]
        qui mean Y if z_ind==1 & z_zip=="all"
        scalar ate_treat_all = _b[Y]

In [3]: // ate.fristhand.half
        disp ate_treat_half - ate_untreat_half

.05161591


In [4]: // ate.secondhanf.untreated
        disp ate_untreat_half - ate_untreat_none

-.00560227


In [5]: // ate.secondhand.treated
        disp ate_treat_all - ate_treat_half

-.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

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 persusaive person is (3/4) - (1/2) = (1/4). The ATE of talking with an unpersusaive 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

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.

```
In [1]: qui import delimited "./data/chapter08/GerberGreenBook_Chapter8_Table_8_4_8_5.csv",

In [2]: qui mean y01 if prox500==0
        scalar mean_y01 = _b[y01]
        qui mean y00 if prox500==0
        scalar mean_y00 = _b[y00]
        scalar true_ate = mean_y01-mean_y00

In [3]: disp true_ate

-5

In [4]: qui mean y if prox500==0 & assignment ==1
        scalar mean_y = _b[y]
        qui mean y00 if prox500==0 & assignment==0
        scalar mean_y00_0 = _b[y00]
        scalar ate_hat = mean_y - mean_y00_0

In [5]: disp ate_hat

3.3333333
```

The true ATE for the observations that lie outside the range of possibile 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 $E[Y_{11} - Y_{00}]$.

```
In [6]: qui mean y01 if prox500==1
        scalar mean_y01 = _b[y01]
        qui mean y00 if prox500==1
        scalar mean_y00 = _b[y00]
        qui mean y10 if prox500==1
        scalar mean_y10 = _b[y10]
        qui mean y11 if prox500==1
        scalar mean_y11 = _b[y11]

In [7]: scalar true_ate_01 = mean_y01 - mean_y00
        scalar true_ate_10 = mean_y10 - mean_y00
        scalar true_ate_11 = mean_y11 - mean_y00

In [8]: disp  true_ate_01

-5

In [9]: disp  true_ate_10

5

In [10]: disp  true_ate_11

-7

In [11]: qui gen q =.
         qui replace q=prob10 if exposure==10
         qui replace q=prob11 if exposure==11
         qui replace q=prob01 if exposure==01
         qui replace q=prob00 if exposure==00
         qui gen weights = 1/q

In [12]: // fit.01
         qui regress y i.exposure if ///
         prox500>0 & (exposure==0 | exposure==1) [aweight=weights]
         estimates store m1, title(Model 1)

In [13]: // fit.10
         qui regress y i.exposure if ///
         prox500>0 & (exposure==0 | exposure==10) [pweight=weights]
         estimates store m2, title(Model 2)

In [14]: // fit.11
         qui regress y i.exposure if ///
         prox500>0 & (exposure==0 | exposure==11) [pweight=weights]
         estimates store m3, title(Model 3)

In [15]: estout m1 m2 m3, ///
         cells(b(star fmt(3)) se(par fmt(3))) ///
         legend label varlabels(_cons Constant) ///
         stats(N r2)
```

Among observations that lie outside the range of possibile 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.

```
In [16]: clear
         qui import delimited "./data/chapter08/GerberGreenBook_Chapter8_Exercise_9c.csv", c

In [17]: qui gen q=.
         qui replace q=prob10 if exposure==10
         qui replace q=prob00 if exposure==0
         qui gen weights = 1/q

In [18]: regress y i.exposure if (prob10>0) & (prob10<1) [pweight=weights]

(sum of wgt is   1.5089e+02)

Linear regression                               Number of obs     =         71
                                                F(1, 69)          =      56.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5496
                                                Root MSE          =     2.0976

------------------------------------------------------------------------------
             |               Robust
```

```
            y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 10.exposure |   4.602226   .6098582     7.55   0.000     3.385592     5.81886
       _cons |   4.285784   .5233643     8.19   0.000     3.241701    5.329867
------------------------------------------------------------------------------
```

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

Question 10

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?

```
In [1]: clear
        qui set obs 16
        //(uncomment to install the package)
        //ssc install egenmore
        qui egen week = repeat(), values("2")
        qui replace week = "3" in 9/l
        qui egen prob00 = fill(0.5,0.5)
        qui replace prob00=0.25 in 9/l
        qui egen prob01 = fill(0.25,0.25)
        qui egen prob11 = fill(0.25,0.25)
        qui replace prob11 = 0.5 in 9/l

In [5]: input int y str2 z
        9 "11"
        5 "00"
        2 "01"
        3 "00"
        3 "00"
        8 "11"
        3 "00"
        1 "01"
        4 "11"
        7 "01"
        10 "11"
        10 "01"
        3 "00"
        10 "11"
        4 "00"
        3 "11" end
```

```
In [6]: gen prob01z = prob01 if z=="01"
        replace prob01z = prob01 if prob01z==.
        //difference in the way that R and STATA weighted mean functions
        gen y01z = y if z=="01"
        egen mean01 = wtmean(y01z), weight(1/prob01z)
        gen prob00z = prob00 if z=="00"
        replace prob00z = prob00 if prob01z==.
        gen y00z= y if z=="00"
        egen mean00 = wtmean(y00z), weight(1/prob00z)
        gen ate01_00 = mean01 - mean00
        disp ate01_00

1.5
```

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

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.

```
In [7]: gen prob01z2 = prob01 if z=="01" & week=="2"
        gen y11z2 = y if z=="11" & week=="2"
        replace prob01z2 = .25 if prob01z2==.
        egen mean11 = wtmean(y11z2), weight(1/prob01z2)
        gen prob00z2 = prob01 if z=="00" & week=="2"
        replace prob00z2 = prob01 if prob00z2==.
        gen y00z2 = y if z=="00" & week=="2"
        drop mean00
        egen mean00 = wtmean(y00z2), weight(1/prob00z2)

        gen ate11_00 = mean11 - mean00
        disp ate11_00

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*

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 disortion 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

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

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

```
In [1]: clear
        set obs 8
        egen block = repeat(), values("A")
        replace block ="B" in 5/8

In [3]: input int y0 int y1
                0 2
                1 5
                1 3
                2 1
                2 3
                3 3
                4 9
                4 7 end

In [6]: // function to calculate population covariance
        cap program drop cov_pop
        program define cov_pop, rclass
        args x y
        tempvar xy_dev
        qui sum `x'
        local avg_x = r(mean)
        local length = r(N)

        qui sum `y'
        local avg_y = r(mean)
        gen `xy_dev' = (`x'-`avg_x')*(`y'-`avg_y')
        qui tabstat `xy_dev', stat(sum) save
        return scalar cor_pop = el(r(StatTotal),1,1)/`length'
        end

        qui egen rank_y1=rank(y1), unique
        qui gen id=_n
```

```
In [7]: vlookup id, generate(y1_lowtohigh) key(rank_y1) value(y1)
        replace id = 9-id
        vlookup id, generate(y1_hightolow) key(rank_y1) value(y1)
```

```
In [8]: cov_pop y0 y1_hightolow
        di "cov.min ="%8.3f r(cor_pop)
```

```
cov.min =  -3.141
```

```
In [9]: cov_pop y0 y1_lowtohigh
        di "cov.min ="%8.3f r(cor_pop)
```

```
cov.min =   3.234
```

```
In [10]: qui replace id=_n
         qui replace id=. if block=="B"
         qui egen rank_y1_A = rank(y1), unique by(block)
         qui replace rank_y1_A =. if block=="B"
         qui vlookup id, generate(y1_hightolow_block_A) key(rank_y1_A) value(y1)
         qui replace id = _n-4
         qui replace id =. if block=="A"
         qui egen rank_y1_B = rank(y1), unique by(block)
         qui replace rank_y1_B =. if block=="A"
         qui vlookup id, generate(y1_hightolow_block_B) key(rank_y1_B) value(y1)
         qui gen y1_hightolow_block = y1_hightolow_block_A
         qui replace y1_hightolow_block = y1_hightolow_block_B if block=="B"
         qui replace id=5-_n
         qui replace id=. if block=="B"
         qui vlookup id, generate(y1_lowtohigh_block_A) key(rank_y1_A) value(y1)
         qui replace id = 9-_n
         qui replace id =. if block=="A"
         qui vlookup id, generate(y1_lowtohigh_block_B) key(rank_y1_B) value(y1)
         qui gen y1_lowtohigh_block = y1_lowtohigh_block_A
         qui replace y1_lowtohigh_block = y1_lowtohigh_block_B if block=="B"
```

```
In [11]: cov_pop y0 y1_lowtohigh_block
         di "cov.min ="%8.4f r(cor_pop)



cov.min = -0.0156


In [12]: cov_pop y0 y1_hightolow_block
         di "cov.min ="%8.3f r(cor_pop)



cov.min =   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 advantange of the blocks reduces the range of possible covariances.

Question 4

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:

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

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.

$$Y_i = 3 + 1 * Loan_i + 1 * Training_i + 0 * (Loan_i * Training_i) + e_i \tag{2}$$

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

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

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

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

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 = 1.25) than among those with skills (CATE = 1.25 - 0.5 = 0.75). 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

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.

```
In [1]: qui import delim ./data/chapter09/Rind_Bordia_JASP_1996, clear

In [2]: gen Z =.
        qui replace Z = 1 if happyface==1
        qui replace Z = 0 if happyface==0

        rename tip Y

        capture program drop var_difference
        program define var_difference, rclass
                sum Y if Z==1, detail
                local var_treat = r(Var)
                sum Y if Z==0, detail
                local var_control = r(Var)
                return scalar vardiff= `var_treat'-`var_control'
        end

        tsrtest Z r(vardiff): var_difference

Two-sample randomization test for theta=r(vardiff) of var_difference by Z

Combinations:   5.19137106438e+25 = (89 choose 44)
Assuming null=0
Observed theta: 53.31

Minimum time needed for exact test (h:m:s):  1.66e+18:00:00
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.23448 [one-tailed test of Ho:  theta(Z==0)<=theta(Z==1)]
 p=0.76542 [one-tailed test of Ho:  theta(Z==0)>=theta(Z==1)]
 p=0.47205 [two-tailed test of Ho:  theta(Z==0)==theta(Z==1)]


In [3]: //  p-value for var(Y1)>Var(Y0)
        di %8.4f r(uppertail)

  0.2345


In [4]: //  p-value for var(Y1)<>Var(Y0)
        di %8.3f r(twotail)

   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 samping 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:

```
In [5]: rename female female_factor
        gen female = .
        replace female = 1 if female_factor==1
        replace female = 0 if female_factor==0

        gen zfemale = Z*female
```

```
In [6]: //lmmodelint: regression with interaction between
        //happyface and waitstaff sex
        regress Y Z female zfemale
```

```
      Source |       SS           df       MS      Number of obs   =        89
-------------+----------------------------------   F(3, 85)        =      9.32
       Model |  3072.39611         3  1024.13204   Prob > F        =    0.0000
    Residual |  9335.52582        85  109.829716   R-squared       =    0.2476
-------------+----------------------------------   Adj R-squared   =    0.2211
       Total |  12407.9219        88  140.999113   Root MSE        =     10.48

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |  -3.629627   3.163098    -1.15   0.254    -9.918714     2.65946
      female |   6.378199   3.163098     2.02   0.047      .089112    12.66729
     zfemale |   8.887078   4.446646     2.00   0.049     .0459551     17.7282
       _cons |   21.40571   2.286916     9.36   0.000     16.85871    25.95272
------------------------------------------------------------------------------
```

```
In [7]: //lmmodel: regression model without interaction
        regress Y Z female
```

```
      Source |       SS           df       MS      Number of obs   =        89
-------------+----------------------------------   F(2, 86)        =     11.59
       Model |  2633.69091         2  1316.84545   Prob > F        =    0.0000
    Residual |  9774.23103        86  113.653849   R-squared       =    0.2123
```

```
-------------+----------------------------------   Adj R-squared   =    0.1939
       Total |  12407.9219        88  140.999113   Root MSE        =    10.661

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |   .8673363   2.261535     0.38   0.702    -3.628446    5.363119
      female |   10.87516   2.261535     4.81   0.000      6.37938    15.37094
       _cons |   19.05503   1.995134     9.55   0.000     15.08883    23.02122
------------------------------------------------------------------------------
```

```
In [8]: scalar coeff_z = _b[Z]
        cap drop Y0 Y1
        gen Y0 = Y - coeff_z * Z
        gen Y1 = Y + coeff_z*(1- Z)


        qui regress Y Z female zfemale
        qui test zfemale
        global f_obs = r(F)

In [9]: capture program drop wald_f
        program define wald_f, rclass
                tempvar Y_sim zsimfemale
                gen `Y_sim' = Y1 * Z + Y0 * (1 - Z)
                gen  `zsimfemale' = female*Z
                qui reg `Y_sim' Z female `zsimfemale'
                test `zsimfemale'
                return scalar f_sims = r(F)
        end

In [10]: tsrtest Z r(f_sims) using 9_6_fsims.dta, overwrite: wald_f
```

```
Two-sample randomization test for theta=r(f_sims) of wald_f by Z

Combinations:   5.19137106438e+25 = (89 choose 44)
Assuming null=0
Observed theta: 3.994

Minimum time needed for exact test (h:m:s):  6.84e+19:00:00
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.04740 [one-tailed test of Ho:  theta(Z==0)<=theta(Z==1)]
 p=0.95250 [one-tailed test of Ho:  theta(Z==0)>=theta(Z==1)]
 p=0.04740 [two-tailed test of Ho:  theta(Z==0)==theta(Z==1)]

Saving log file to 9_6_fsims.dta...done.


In [11]: di %8.4f r(uppertail)

  0.0474
```

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 ($p = 0.0474$). 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).[1] 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.

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.

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

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.

$$\begin{aligned} Y_i = {} & \gamma_0 + \gamma_1 Race_i + \gamma_2 Resume_i + \gamma_3 City_i + \\ & \gamma_4 (Race_i * Resume_i) + \gamma_5 (Race_i * City_i) + \gamma_6 (Resume_i * City_i) + \gamma_7 (Race_i * Resume_i * City_i) + e_i \end{aligned}$$

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:

$$\begin{aligned} Y_i = {} & 7.01 + 3.14 Race_i + 1.49 Resume_i - 1.49 City_i + \\ & 1.48 (Race_i * Resume_i) - 1.50 (Race_i * City_i) - 1.73 (Resume_i * City_i) + 0.54 (Race_i * Resume_i * City_i) + e_i \end{aligned}$$

Additional response (Boston = 1; Black = 1; Low quality = 1)

$$\begin{aligned} Y_i = {} & 8.94 - 3.66 Race_i - 1.78 Resume_i + 4.18 City_i + \\ & 2.02 (Race_i * Resume_i) - 0.96 (Race_i * City_i) - 1.19 (Resume_i * City_i) - 0.54 (Race_i * Resume_i * City_i) + e_i \end{aligned}$$

Additional response (Boston = 1; Black = 1; High quality = 1)

$$\begin{aligned} Y_i = {} & 7.16 - 1.64 Race_i + 1.78 Resume_i + 2.99 City_i - \\ & 2.02 (Race_i * Resume_i) - 1.50 (Race_i * City_i) + 1.19 (Resume_i * City_i) + 0.54 (Race_i * Resume_i * City_i) + e_i \end{aligned}$$

```
In [1]: clear
        qui set obs 4870
        qui egen y = fill(1,1)
        qui replace y = 0 in 39/542
        qui replace y = 0 in 598/1084
        qui replace y = 0 in 1131/1625
        qui replace y = 0 in 1697/2166
        qui replace y=0 in 2204/2836
        qui replace y=0 in 2885/3506
        qui replace y=0 in 3543/4188
        qui replace y=0 in 4250/4870
        qui egen boston = fill(1,1)
        qui replace boston = 0 in 2167/4870
        qui gen chicago = 1-boston
        qui egen lowquality = fill(1,1)
        qui replace lowquality = 0 in 1085/2166
        qui replace lowquality = 0 in 3507/4870
        qui gen highquality = 1-lowquality
        qui egen black = fill(1,1)
        qui replace black = 0 in 543/1084
        qui replace black = 0 in 1626/2166
        qui replace black = 0 in 2837/3506
        qui replace black = 0 in 4188/4870
        qui replace black = 0 in 4189/4870
        qui gen white = 1-black

        qui gen whitehighquality = white*highquality
        qui gen whitechicago = white*chicago
        qui gen highqualitychicago = highquality*chicago
        qui gen whitehighqualitychicago = white * highquality * chicago

In [2]: // fit_1
        qui regress y white highquality chicago whitehighquality ///
        whitechicago highqualitychicago whitehighqualitychicago
        estimates store m1, title(Model 1)

        gen blackhighquality = black*highquality
        gen blackchicago = black*chicago
        gen blackhighqualitychicago = black * highquality * chicago

In [3]: // fit_2
        qui regress y black highquality chicago blackhighquality ///
        blackchicago highqualitychicago blackhighqualitychicago
        estimates store m2, title(Model 2)

In [4]: // fit_3
        gen whiteboston = white*boston
        gen highqualityboston = highquality*boston
        gen whitehighqualityboston = white * highquality * boston
        qui regress y white highquality boston whitehighquality ///
        whiteboston highqualityboston whitehighqualityboston

        estimates store m3, title(Model 3)

In [5]: // fit_4
        gen blacklowquality = black*lowquality
        gen lowqualitychicago = lowquality*chicago
        gen blacklowqualitychicago = black * lowquality * chicago
        qui regress y black lowquality chicago blacklowquality ///
        blackchicago lowqualitychicago blacklowqualitychicago

        estimates store m4, title(Model 4)
```

```
In [6]: estout m1 m2 m3 m4, cells(b(star fmt(3)) se(par fmt(3))) ///
        starlevels( * 0.10 ** 0.05 *** 0.010) ///
        legend label varlabels(_cons constant) ///
        stats(N r2 r2_a rmse F, fmt(0 3 3 3 3) ///
        label(N R-squared Ajusted-R2 Residual_Std_Error F-Statistic))
```

```
------------------------------------------------------------------------------------
                          Model 1         Model 2         Model 3         Model 4
                             b/se            b/se            b/se            b/se
------------------------------------------------------------------------------------
white                       0.031*                          0.016
                          (0.016)                         (0.015)
highquality                 0.015           0.030*         -0.002
                          (0.016)         (0.016)         (0.015)
chicago                    -0.015          -0.030*                         -0.042***
                          (0.016)         (0.016)                         (0.016)
whitehighquality            0.015                           0.020
                          (0.023)                         (0.021)
whitechicago               -0.015
                          (0.022)
highqualitychicago         -0.017          -0.012
                          (0.022)         (0.022)
whitehighqualitych~o        0.005
                          (0.031)
black                                      -0.031*                         -0.046***
                                          (0.016)                         (0.016)
blackhighquality                           -0.015
                                          (0.023)
blackchicago                                0.015                           0.010
                                          (0.022)                         (0.022)
blackhighqualitych~o                       -0.005
                                          (0.031)
boston                                                      0.015
                                                          (0.016)
whiteboston                                                 0.015
                                                          (0.022)
highqualityboston                                           0.017
                                                          (0.022)
whitehighqualitybo~n                                       -0.005
                                                          (0.031)
lowquality                                                                 -0.030*
                                                                          (0.016)
blacklowquality                                                             0.015
                                                                          (0.023)
lowqualitychicago                                                           0.012
                                                                          (0.022)
blacklowqualitychi~o                                                        0.005
                                                                          (0.031)
constant                    0.070***        0.101***        0.055***        0.131***
                          (0.012)         (0.012)         (0.010)         (0.012)
------------------------------------------------------------------------------------
N                            4870            4870            4870            4870
```

```
R-squared                  0.008          0.008          0.008          0.008
Ajusted-R2                 0.006          0.006          0.006          0.006
Residual_Std_Error         0.271          0.271          0.271          0.271
F-Statistic                5.349          5.349          5.349          5.349
--------------------------------------------------------------------------------
* p<0.10, ** p<0.05, *** p<0.010
```

Question 8

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.[2] 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.

Table 4: 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.

[2] Fieldhouse et al. 2010.

Answer:

(i) **ATE of mail.** 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.

(ii) **Complier average causal effect (CACE) of phone calls.** 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.

(iii) **CATE of mail among those who comply with the phone call treatment.** The CATE of mail is identified in the same was as an ATE, except that it is restricted to those who actually receive phone calls

(iv) **CATE of mail among those who do not comply with the phone call treatment.** Same as above, but among those who are not treated when called.

(v) **CACE of phone calls among those who receive mail.** 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.

(i) **ATE of mail.** The estimated ATE of mail is 40.3 - 39.7 = 0.6 percentage points.

(ii) **Complier average causal effect (CACE) of phone calls.** The estimated CACE of phone calls is the ITT divided by the share of compliers: (39.7 - 39.7)/ (2003/3466) = 0.

(iii) **CATE of mail among those who comply with the phone call treatment.** The CATE of mail among those who receive a call is 46.8 - 46.5 = 0.3 percentage points.

(iv) **CATE of mail among those who do not comply with the phone call treatment.** 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.

(v) **CACE of phone calls among those who receive mail.** The CACE of phone calls among those who receive mail is the 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:

$$\begin{aligned} PhoneContact_i &= \alpha_0 + \alpha_1 Mail_i + \alpha_2 PhoneAssign_i + \alpha_3(PhoneAssign_i * Mail_i) + e_i \\ PhoneContact_i * Mail_i &= \gamma_0 + \gamma_1 Mail_i + \gamma_2 PhoneAssign_i + \gamma_3(PhoneAssign_i * Mail_i) + \epsilon_i \\ Vote_i &= \beta_0 + \beta_1 Mail_i + \beta_2 PhoneContact_i + \beta_3(PhoneContact_i * Mail_i) + u_i \end{aligned}$$

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?

```
In [1]: import delim ./data/chapter09/Fieldhouse_et_al_unpublished_2010_expanded,clear

In [2]: rename m mail
        rename p phone_assign
        rename c phone_contact
        rename y vote
        rename c_m phone_contact_mail
        rename p_m phone_assign_mail


        gen mail_phone_contact = mail*phone_contact
        gen mail_phone_assign = mail*phone_assign

        ivregress 2sls vote mail ///
        (mail_phone_contact phone_contact = mail phone_assign mail_phone_assign)
```

```
Instrumental variables (2SLS) regression          Number of obs   =     15,300
                                                  Wald chi2(3)    =       3.38
                                                  Prob > chi2     =     0.3367
                                                  R-squared       =     0.0010
                                                  Root MSE        =     .49001

----------------------------------------------------------------------------------
              vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+--------------------------------------------------------------
mail_phone_contact |   .0253338   .0282274     0.90   0.369    -.0299908    .0806584
     phone_contact |  -.0001781   .0186117    -0.01   0.992    -.0366565    .0363002
              mail |   .0060348   .0100671     0.60   0.549    -.0136962    .0257659
             _cons |   .3969878    .006809    58.30   0.000     .3836424    .4103332
----------------------------------------------------------------------------------
Instrumented:  mail_phone_contact phone_contact
Instruments:   mail phone_assign mail_phone_assign
```

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*

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 $E[a_ib_i]$.

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: $E[a_ib_i] = aE[b_i] = 0E[b_i] = 0$.

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 $Y_i(M_i(0), 0)$.

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

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

$Y_i(1,0) - Y_i(0,0)$ 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

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.

```
In [1]: set more off
        input z Y0M0 Y1M0 Y0M1 Y1M1 M0 M1
        0 0 0 0 0 0 0
        0 0 0 0 0 0 1
        0 0 0 0 0 1 1
        0 0 1 0 1 0 0
        0 0 1 0 1 0 1
        0 0 1 0 1 1 1
        1 1 0 1 1 0 0
        1 1 0 1 1 0 1
        1 1 0 1 1 1 1
        1 0 1 1 1 0 0
        1 0 1 1 1 0 1
        1 0 1 1 1 1 1
        end

In [2]: tabstat Y0M0 Y1M0 Y0M1 Y1M1, stat(mean)

        gen M = .
        gen Y = .



   stats |      Y0M0      Y1M0      Y0M1      Y1M1
---------+----------------------------------------
    mean |       .25        .5        .5       .75
--------------------------------------------------


In [3]: //coefmat
        capture program drop coef
        program define coef, rclass
                replace M = M0*(1-z) + M1*z
                replace Y = Y0M0*(1-z)*(1-M) +
                        Y1M0*(z)*(1-M) + Y0M1*(1-z)*(M) + Y1M1*(z)*(M)
                qui reg Y M z
                return scalar coy = _b[_cons]
                return scalar com = _b[M]
                return scalar coz = _b[z]
                return scalar nocoli = _se[z]
        end

        qui tsrtest z r(coy) using co_y.dta, overwrite: coef
        qui tsrtest z r(com) using co_m.dta, overwrite: coef
        qui tsrtest z r(coz) using co_z.dta, overwrite: coef
        qui tsrtest z r(nocoli) using nocoli.dta, overwrite: coef

In [4]: // tcoefmat
        capture program drop tcoef
        program define tcoef, rclass
                replace M = M0*(1-z) + M1*z
                replace Y = Y0M0*(1-z)*(1-M) + Y1M0*(z)*(1-M) + Y0M1*(1-z)*(M) + Y1M1*(z)*(M
                qui reg Y z
                return scalar tcoy = _b[_cons]
                return scalar tcoz = _b[z]
        end

        qui tsrtest z r(tcoy) using tco_y.dta, overwrite: tcoef
        qui tsrtest z r(tcoz) using tco_z.dta, overwrite: tcoef
```

```
In [5]: // mcoefmat
        capture program drop mcoef
        program define mcoef, rclass
                replace M = M0*(1-z) + M1*z
                qui reg M z
                return scalar mcom = _b[_cons]
                return scalar mcoz = _b[z]
        end

        qui tsrtest z r(mcom) using mco_m.dta, overwrite: mcoef
        qui tsrtest z r(mcoz) using mco_z.dta, overwrite: mcoef
```

```
In [6]: // colMeans(na.omit(coefmat))
        preserve
        qui use "co_y.dta", clear
        qui rename theta co_y

        qui merge 1:1 _n using "co_m.dta"
        qui rename theta co_m

        qui drop _merge
        qui merge 1:1 _n using "co_z.dta"
        qui rename theta co_z
        qui drop _merge

        //check colinearity
        qui merge 1:1 _n using "nocoli.dta"
        qui rename theta nocoli
        qui drop _merge


        qui drop if _n == 1
        // omit instances of perfect colinearity between M and Z
        qui drop if nocoli==0

        tabstat co_y co_m co_z,stat(mean)

        restore
```

```
   stats |      co_y      co_m      co_z
```

```
---------+------------------------------
    mean |       .25       .25       .25
----------------------------------------
```

```
In [7]: // colMeans(na.omit(tcoefmat)))
        preserve
        qui use "tco_y.dta", clear
        qui rename theta tco_y

        qui merge 1:1 _n using "tco_z.dta"
        qui rename theta tco_z
        qui drop _merge

        qui /*check coli*/
        qui merge 1:1 _n using "nocoli.dta"
        qui rename theta nocoli
        qui drop _merge

        // drop the observation statistics
        qui drop if _n == 1

        tabstat tco_y tco_z,stat(mean)
        restore
```

```
   stats |     tco_y     tco_z
---------+--------------------
    mean |  .3333333  .3333333
------------------------------
```

```
In [8]: //colMeans(na.omit(mcoefmat)))
        preserve
        qui use "mco_m.dta", clear
        qui rename theta mco_m

        qui merge 1:1 _n using "mco_z.dta"
        qui rename theta mco_z
        qui drop _merge

        // drop the observation statistics
        qui drop if _n == 1
        tabstat mco_m mco_z,stat(mean)
        restore
```

```
   stats |     mco_m     mco_z
---------+--------------------
    mean |  .3333333  .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:

```
In [9]: preserve

         qui use "mco_z.dta", clear
         qui rename theta mco_z

         qui merge 1:1 _n using "co_z.dta"
         qui rename theta co_z
         qui drop _merge

         // check colinearity
         qui merge 1:1 _n using "nocoli.dta"
         qui rename theta nocoli
         qui drop _merge

         qui gen asbs = mco_z*co_z
         qui drop if _n == 1
         // omit instances of perfect colinearity between M and Z
         qui drop if nocoli==0


         tabstat asbs,stat(mean)
         restore



    variable |      mean
-------------+----------
        asbs |   .082244
------------------------
```

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

[1] Text mistakenly has “multiply estimates of a and c together.

unobserved causes of Y and unobserved causes of M. In effect, M is as good as randomly assigned in this special case.

Question 4

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:

```
In [1]: clear
        set obs 1000
        egen y = fill(0,0)
        replace y = 1 in 411/500
        replace y =1 in 901/1000
        egen z = fill(0,0)
        replace z = 1 in 501/1000
        egen m = fill(0,0)
        replace m = 1 in 401/500
        replace m = 1 in 801/1000
        regress y z m
```

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =    652.06
       Model |       87.22         2       43.61   Prob > F        =    0.0000
    Residual |       66.68       997  .066880642   R-squared       =    0.5667
-------------+----------------------------------   Adj R-squared   =    0.5659
       Total |       153.9       999  .154054054   Root MSE        =    .25861

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           z |      -.112     .01676    -6.68   0.000     -.144889    -.079111
           m |        .66   .0182867    36.09   0.000     .6241152    .6958848
       _cons |       .048     .01213     3.96   0.000     .0241967    .0718033
------------------------------------------------------------------------------
```

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

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 3: 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

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*

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 light of new evidence.
Answer:

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

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:

$$\sigma^2_{\bar{\tau}|x_e} = \frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_\beta^2+\sigma_{x_e}^2}} = \frac{1}{\frac{1}{225} + \frac{1}{100+25}} = 80.36$$

Plugging these numbers into equation (11.4) gives:

$$p_1 = \frac{\sigma^2_{\bar{\tau}|x_e}}{\sigma_g^2} = \frac{\sigma_\beta^2 + \sigma_{x_e}^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = \frac{100+25}{225+100+25} = 0.357 p_2 \quad = \frac{\sigma^2_{\bar{\tau}|x_e}}{\sigma_\beta^2 + \sigma_{x_e}^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\beta^2 + \sigma_{x_e}^2} = 1 - p_1 = 0.643$$

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

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

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:

$$\sigma^2_{\bar{\tau}|x_e} = \frac{1}{\frac{1}{\sigma^2_g} + \frac{1}{\sigma^2_\beta + \sigma^2_{x_e}}} = \frac{1}{\frac{1}{4} + \frac{1}{0+1}} = 0.8$$

Plugging these numbers into equation (11.4) gives:

$$\begin{aligned} p_1 &= \frac{\sigma^2_{\bar{\tau}|x_e}}{\sigma^2_g} = \frac{\sigma^2_\beta + \sigma^2_{x_e}}{\sigma^2_g + \sigma^2_\beta + \sigma^2_{x_e}} = \frac{0+1}{4+0+1} = 0.2 \\ p_2 &= \frac{\sigma^2_{\bar{\tau}|x_e}}{\sigma^2_\beta + \sigma^2_{x_e}} = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_\beta + \sigma^2_{x_e}} = 1 - p_1 = 0.8 \end{aligned}$$

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

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

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.

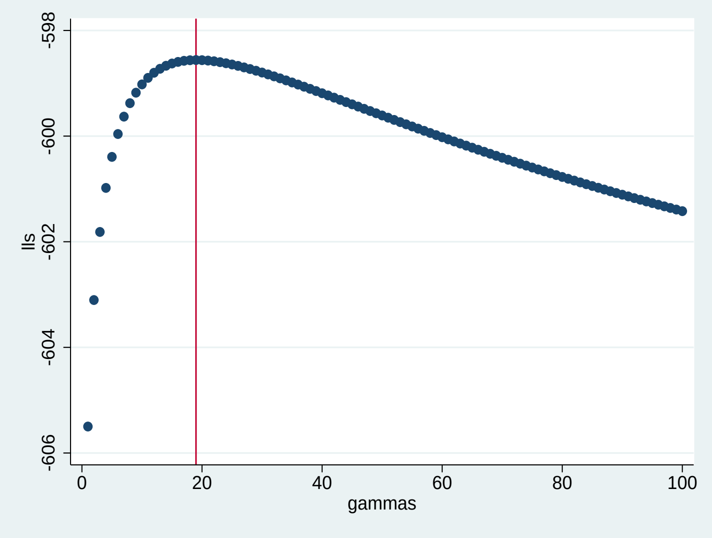
b) Suppose we reasoned that a nominal price of zero nevertheless involves some transaction cost, as villagers have to make the effort to redeem their vouchers. For a given subject, we may model the probability of making a purchase as:

$$Pr[Y = 1] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 ln[V_i + \gamma])}}$$

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:

```
In [1]: import delim ./data/chapter11/Chapter_11_Dupas_2010_Dataset,clear

In [2]: gen purchased =.
        replace purchased = 1 if purchasednet=="yes"
        replace purchased = 0 if purchasednet=="no"
        rename cfw_id region

        matrix t=J(100, 2, .)
        matrix colnames t=gammas lls

        forvalues i = 1/100 {
```

```
                        gen log_price_star_`i' = log(price + `i')
                        qui glm purchased log_price_star_`i' i.region,
                                family(binomial) link(logit)

                        matrix t[`i', 1] = `i'
                        matrix t[`i', 2] = `e(ll)'

        }

In [3]: svmat double t, names(col)
        qui sum lls
        list gammas if lls==r(max)

        scatter lls gammas, xline(19)
        graph export ../results/chapter11/exercise_11_9_graph.pdf

      +--------+
      | gammas |
      |--------|
  19. |     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

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 12 Exercises

Alan S. Gerber and Donald P. Green*

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).

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

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

[1] Page 1998.

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

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.[2] Some

[2] Olken 2010.

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 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

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

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

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

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:

```
In [1]: import delim ./data/chapter12/Howell_Peterson_BIP_2002, clear

In [2]: tabulate missing_y1math treat, column nof


missing_y1 |         treat
      math |         0          1 |     Total
-----------+----------------------+----------
         0 |     75.90      81.10 |     78.61
         1 |     24.10      18.90 |     21.39
-----------+----------------------+----------
     Total |    100.00     100.00 |    100.00
```

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:

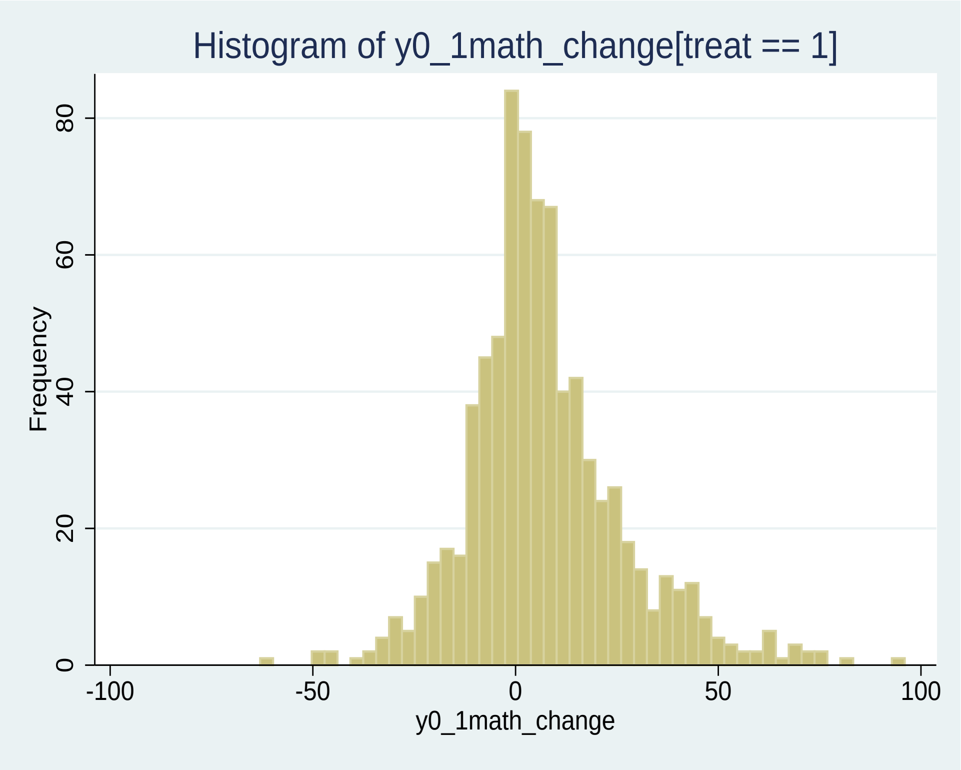
```
In [3]: tabstat y0_1math_change, by(treat) stat(mean) not


Summary for variables: y0_1math_change
     by categories of: treat

   treat |      mean
---------+----------
       0 |  6.486647
       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:

```
In [4]: histogram y0_1math_change if treat==1, bin(50) freq ///
        title("Histogram of y0_1math_change[treat == 1]")

In [5]: tabstat y0_1math_change if treat==1, stat(min max)
```

```
    variable |       min       max
-------------+--------------------
y0_1math_c~e |       -63        96
----------------------------------


In [6]: centile y0_1math_change if treat==1, centile(5 10 15 25 50 75 85 90 95)


                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
y0_1math_c~e |       781           5        -20             -23         -18
             |                    10        -13             -16         -11
             |                    15         -9             -11          -7
             |                    25         -4              -5          -2
             |                    50          4               3           5
             |                    75         16              14          18
             |                    85         25              22          27
             |                    90         32              28          36
             |                    95         43              40          48
```



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:

```
In [7]: centile y0_1math_change if treat==1, centile(93.6)



                                                          -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
y0_1math_c~e |       781       93.6          40              36          44
```

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:

```
In [8]: qui mean y0_1math_change if treat==1 & missing_y1math==0 & y0_1math_change < 40
        scalar l_b = _b[y0_1math_change]
        qui count if treat==1 & missing_y1math==0 & y0_1math_change < 40
        scalar l_b_count = r(N)
        qui count if treat==1

In [9]: disp %8.6f l_b

3.701513


In [10]: disp %8.7f 1-l_b_count/r(N)

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:

```
In [11]: qui mean y0_1math_change if treat==0
         disp %18.6f l_b - _b[y0_1math_change]

         -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:

```
In [12]: centile y0_1math_change if treat==1, centile(6.4)



                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
y0_1math_c~e |       781        6.4         -18             -21         -16
```

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:

```
In [13]: qui mean y0_1math_change if treat==1 & missing_y1math==0 & y0_1math_change > -18
         scalar u_b = _b[y0_1math_change]
         disp %8.6f u_b

9.707586


In [14]: qui count if treat==1 & missing_y1math==0 & y0_1math_change > -18
         scalar u_b_count = r(N)
         qui count if treat==1
         disp %8.7f 1-u_b_count/r(N)

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:

```
In [15]: qui mean y0_1math_change if treat==0
         disp %8.6f u_b - _b[y0_1math_change]

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

Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 13 Exercises

Alan S. Gerber and Donald P. Green*

Question 1

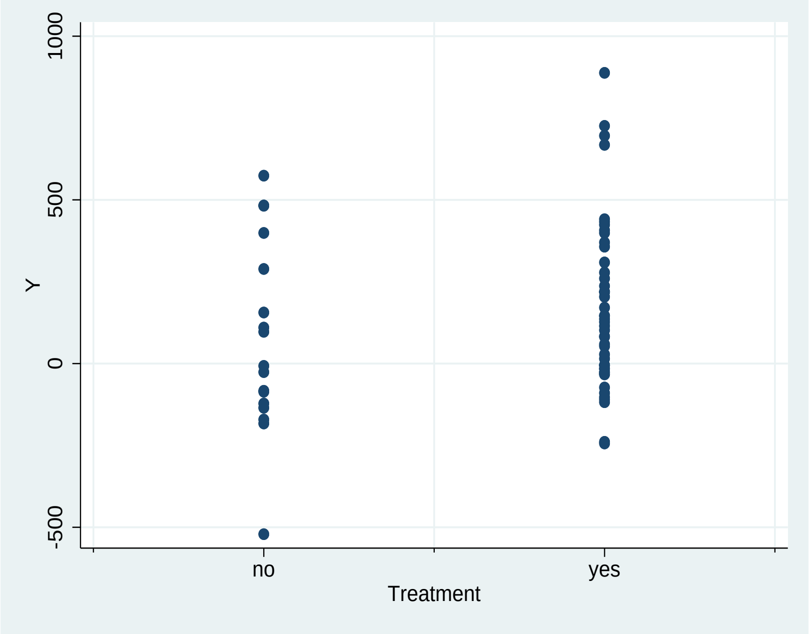
Middleton and Rogers report the results of an experiment in which ballot guides were mailed to randomly assigned precincts in Oregon prior to the 2008 November election. The guides were designed to encourage voters to support certain ballot measures and oppose others. Load the example dataset from `http://isps.research.yale.edu/FEDAI`. 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:

```
In [1]: qui import delim ./data/chapter13/Middleton_Rogers_AI_2010, clear

        rename relevant_measures_net Y
        gen int Z=.
        replace Z = 1 if treatment ==1
        replace Z = 0 if treatment ==0

        qui mean Y if Z==1
        scalar avg_treat = _b[Y]
        qui mean Y if Z==0
        scalar avg_control = _b[Y]
        global tau = avg_treat - avg_control

        disp "average treatment effect = " $tau

average treatment effect = 90.20098


In [2]: twoway (scatter Y Z), ///
        xlabel(0"no" 1 "yes") xmtick(-0.5(1)1.5,grid) xtitle("Treatment")
```

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

```
        graph export ../results/chapter13/exercise_13_1_a_graph.pdf
```



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.

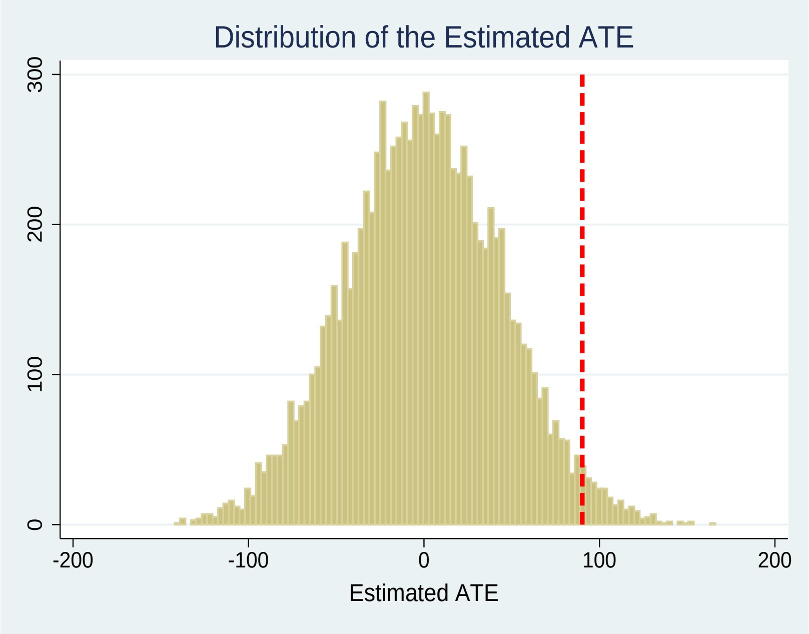
c) Use randomization inference to test whether the apparent difference-in-means could have occurred by chance under the sharp null hypothesis of no treatment effect for any precinct. Interpret the results. Answer:

```
In [3]: ritest Z ate_sim = _b[Z], ///
                reps(10000) sav(13_1_distout.dta, replace) right nodots: ///
                regress Y Z


      Source |       SS           df       MS      Number of obs   =        65
-------------+----------------------------------   F(1, 63)        =      1.50
       Model |  102140.815         1  102140.815   Prob > F        =    0.2248
    Residual |  4281899.43        63  67966.6576   R-squared       =    0.0233
-------------+----------------------------------   Adj R-squared   =    0.0078
       Total |  4384040.25        64  68500.6288   Root MSE        =     260.7

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |   90.20098   73.57996     1.23   0.225    -56.83684    237.2388
       _cons |   73.88235   63.23005     1.17   0.247    -52.47281    200.2375
------------------------------------------------------------------------------


      command:  regress Y Z
      ate_sim:  _b[Z]
  res. var(s):  Z
   Resampling:  Permuting Z
Clust. var(s):  __000003
     Clusters:  65
Strata var(s):  none
       Strata:  1

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ate_sim |   90.20098    1119   10000  0.1119  0.0032   .105785    .118242
------------------------------------------------------------------------------
Note: Confidence interval is with respect to p=c/n.
Note: c = #{T >= T(obs)}
```

```
In [4]: // one-tail p-value
        di %8.4f el(r(p), 1, 1)
```

```
  0.1119
```

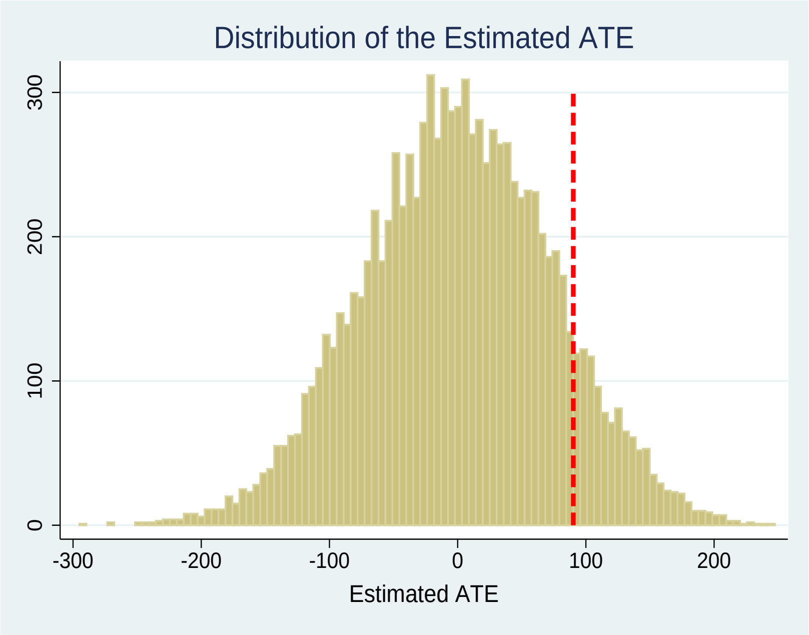
```
In [5]: set more off
        preserve
        use "13_1_distout", clear
        //historgam

        graph twoway (histogram ate_sim,frequency bin(100)) ///
        (scatteri 0 $tau 300 $tau, c(l) lc(red) lw(thick) lp(dash) m(i)), legend(off) ///
        b1title("Estimated ATE") title("Distribution of the Estimated ATE") ///
        xtitle("")

        graph export ../results/chapter13/exercise_13_1_c_graph.pdf

        restore
```



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 0.119. 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:

```
In [6]: clear
        clear matrix
        clear mata
        set matsize 11000
        set maxvar 32767
        set seed 67887975



        cap matrix drop z
        matrix z=J(65, 10000, .)

In [7]: // restircted RA loop
        qui forvalues i = 1/10000 {
                import delim ./data/chapter13/Middleton_Rogers_AI_2010, clear
                tempvar teststat Z
                gen `Z' = .
                gen `teststat' = 5
                while (abs(`teststat')>=0.5){
                        tempvar rannum ordering Zri
                    gen `rannum'=uniform()
                        egen `ordering' = rank(`rannum')
                        gen `Zri' = 1 if `ordering' <= 48
                        replace `Zri' = 0 if `ordering' > 48

                        qui reg dem_perf_06 `Zri'
                        replace `teststat' = _b[`Zri']
                }
                replace `Z' = `Zri'
                forvalues j = 1/65 {
                matrix z[`j', `i'] = `Z'[`j']
                }
                drop _all
        }

In [8]: import delim ./data/chapter13/Middleton_Rogers_AI_2010, clear
         rename relevant_measures_net Y
         gen int Z=.
         replace Z = 1 if treatment ==1
         replace Z = 0 if treatment ==0

In [9]: matrix rowm = z * J(colsof(z), 1, 1/colsof(z))
         matrix colnames rowm=probs
         svmat double rowm, names(col)

In [10]: // distribution of probabilities
         tabstat probs, stat(min p25 med mean p75 max)


    variable |       min       p25       p50      mean       p75       max
-------------+------------------------------------------------------------
       probs |     .7263     .7341      .737  .7384615     .7425     .7619
--------------------------------------------------------------------------



In [11]: svmat z

In [12]: cap matrix drop tau_dis
         matrix tau_dis=J(10000, 1, .)


         // calculate estimate distribution
         forvalues i = 1/10000{
                 tempvar weight`i'
                 gen `weight`i'' = z`i'/probs + (1 - z`i')/(1 - probs)
                 qui reg Y z`i' [pw=`weight`i'']
                 matrix tau_dis[`i', 1] = _b[z`i']

         }

In [13]: set more off

In [14]: preserve
         svmat tau_dis
         qui count if tau_dis1 > $tau
         // one tailed p-value
         di r(N)/_N

         graph twoway (histogram tau_dis1,frequency bin(100)) ///
         (scatteri 0 $tau 300 $tau, c(l) lc(red) lw(thick) lp(dash) m(i)), legend(off) ///
         b1title("Estimated ATE") title("Distribution of the Estimated ATE") ///
         xtitle("")

         graph export ../results/chapter13/exercise_13_1_d_graph.pdf

         restore

.0238
```



Randomization inference applied to 10,000 simulated restricted randomizations shows that one-tailed p-value of the estimated ATE is 0.0238. 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

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.