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

Alan S. Gerber and Donald P. Green*

April 11, 2019

## Question 1

Important concepts: [10 points]

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

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

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



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

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

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



## Question 2

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

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

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

## Question 3

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

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

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

## Question 4

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

Table 1: Question 4 Table

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



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

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

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

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

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

## Question 5

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

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



Table 2: Question 5 Table

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

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

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

Beginning with Equation 3.4:

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

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

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



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

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

[1] Text mistakenly printed "less"

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

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

## Question 6

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



```
        import delim "Clingingsmith_et_al_QJE_2009dta.csv",clear
        set seed 1234567


(8 vars, 958 obs)


        rename success D
        rename views Y


        //findit tsrtest
        //package name:  st0158.pkg install

        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

        // ssc install tsrtest
        tsrtest D r(ate_avg) using 3_6_resam.dta, overwrite: ate Y D

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

Combinations:   8.4503047638e+285 = (958 choose 448)
Assuming null=0
Observed theta: .4748

Minimum time needed for exact test (h:m:s):  3.2e+278:00:0-1.3e+266
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.00190 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.99830 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.00360 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to 3_6_resam.dta...done.


        preserve
         use "3_6_resam.dta", clear
         global ate = theta[1]
         di "estimated ATE: "$ate
         drop if _n==1
         qui count if theta > $ate
         di "Count of Simulated ATE > estimated ATE: "  r(N)
         qui count if theta >= $ate
         di "One-side p-value: " %8.4f r(N)/_N
         qui count if abs(theta) > abs($ate)
         di "Count of absolute value Simulated ATE > estimated ATE: " r(N)
         qui count if abs(theta) >= abs($ate)
         di "Two-side p-value: " %8.4f r(N)/_N
         restore



estimated ATE: .4748337

(1 observation deleted)
```

```
Count of Simulated ATE > estimated ATE: 16


One-side p-value:   0.0019


Count of absolute value Simulated ATE > estimated ATE: 33


Two-side p-value:   0.0036
```

The estimated ATE is 0.4748337. The number of simulated ATEs under the sharp null hypothesis of no effect that were as large was 16, corresponding to a $p$-value of 0.0019. The number of simulated ATEs under the sharp null hypothesis of no effect that were as large in absolute value was was 33, corresponding to a $p$-value of 0.0036.

## Question 7

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



Table 3: Question 7 Table

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

```
set.seed(1234567)
D <- c(rep(0,5), rep(1, 5))
Y <- c(1,0,0,4,3,2,11,14,0,3)
```

```
Y_star <- Y + D*(-7)    # Subtracts 7 from "treatment" group

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

ate

## [1] -2.6

p.value.onesided

## [1] 0.2063492
```

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



## Question 8

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

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

```
D <-  titiunik$term2year
Y <- titiunik$bills_introduced
block <- titiunik$texas0_arkansas1

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

[2] Titiunik 2010.

Table 4: Question 8 Table

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

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

## [1] -16.74167

ate_arkansas

## [1] -10.09477
```

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

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

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

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

## [1] 9.345871

se_arkansas

## [1] 3.395979
```

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

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

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

## [1] -13.2168
```

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

d) Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimates of the overall ATE.
Answer:
The two states differ in terms of the probability that a given legislator will be assigned to the treatment. Therefore, we cannot pool the data without introducing a correlation between treatment assignment and the potential outcomes associated with the two states. In this study, the experiments take place within each state, and the analyst should pool the state-level results in order to obtain an overall result.

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

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

## [1] 4.74478
```

The overall standard error is (4.745).

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

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

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

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

## [1] -13.2168

p.value.twosided

## [1] 0.0071
```

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

## Question 9

Camerer reports the results of an experiment in which he tests whether large, early bets placed at horse tracks affect the betting behavior of other bettors.[3] Selecting pairs of long-shot horses running in the same race whose betting odds were approximately the same when betting opened, he placed two $500 bets on one of the two horses approximately 15 minutes before the start of the

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

race. Because odds are determined based on the proportion of total bets placed on each horse, this intervention causes the betting odds for the treatment horse to decline and the betting odds of the control horse to rise. Because Camerer's bets were placed early, when the total betting pool was small, his bets caused marked changes in the odds presented to other bettors. (A few minutes before each race started, Camerer canceled his bets.) While the experimental bets were still "live," were other bettors attracted to the treatment horse (because other bettors seemed to believe in the horse) or repelled by it (because the diminished odds meant a lower return for each wager)? Seventeen pairs of horses in this study are listed below. The outcome measure is the number of dollars that were placed on each horse (not counting Camerer's own wagers on the treatment horses) during the test period, which begins 16 minutes before each race (roughly 2 minutes before Camerer began placing his bets) and ends 5 minutes before each race (roughly 2 minutes before Camerer withdrew his bets). [10 points]

Table 5: Question 9 Table

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



a) One interesting feature of this study is that each pair of horses ran in the same race. Does this design feature violate the non-interference assumption, or can potential outcomes be defined so that the non-interference assumption is satisfied?
Answer:
This design feature violates non-interference if the estimand is defined as the difference between the following two potential outcomes: total bets on a given horse when experimental bets are placed on that horse versus no experimental bets on any horse in the race. One could avoid violating non-interference by redefining the estimand as the difference between the following two potential outcomes: total bets on a horse when experimental bets are placed on that horse versus experimental bets are placed on a competing horse in the same race.

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

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

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

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

numiter <- ncol(perms)

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

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

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

## [1] 0.3696
```

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

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

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

## [1] -110.1765
```

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

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

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

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

mean(pair_diffs)

## [1] 110.1765
```

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

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



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

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

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

ate

## [1] -110.1765

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

## [1] 0.3092
```

A two-tailed test generates a p-value of 0.3092, indicating that one cannot reject the sharp null of no effect for any unit. A two-tailed test is appropriate because some theories predict a positive

effect while others predict a negative effect: "were other bettors attracted to the treatment horse (because other bettors seemed to believe in the horse) or repelled by it (because the diminished odds meant a lower return for each wager)?" The appropriate null hypothesis in this case is no effect, which would be rejected if we observed either strongly positive or strongly negative differences between treatment and control horses.

## Question 10

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

The equation for the standard error under individual assignment:

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

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

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

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



## Question 11

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

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

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

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

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

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

## [1] 4.706192
```

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

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

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

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

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

## [1] 0.9766259
```

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

c) Why do the two methods of forming clusters lead to different standard errors? What are the implications for the design of cluster randomized experiments?
Answer:
The first method clusters the most similar villages together, and the second method clusters the most dissimilar villages together. As a result, the variances of the average within-cluster potential outcomes are much larger in the first method and smaller in the second. As a result, the second method produces a much narrower standard error of the ATE estimate. The implication for clustered design is that the more similar the observations with a cluster, the less precise the estimates we can produce. When possible, cluster heterogeneous observations together.

## Question 12

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

Table 6: Question 12 Table

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

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

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

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

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