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:

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

$$= \sqrt{\frac{1}{6}\left\{\frac{Var(Y_i(0))}{6} + 6Var(Y_i(1)) + 2cov(Y_i(0),Y_i(1))\right\}}$$

$$cov(Y_i(0),Y_i(1)) = \frac{(10-15)(15-20)+(20-15)(30-20)+(20-15)(15-20)}{7} = \frac{50}{7}$$

$$= \sqrt{\frac{1}{6}\left\{\frac{\frac{100}{7}}{6} + 6\frac{300}{7} + 2\frac{50}{7}\right\}}$$

$$= 6.755$$

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

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]

```
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
```

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] -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

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

[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.

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

Table 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.

```
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

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.

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

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

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

## [1] 0.9766259
```

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

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

Question 12