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

Alan S. Gerber and Donald P. Green*

April 8, 2018

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

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

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

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

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

which is what we get when we make the same substitution into equation (7.6).
A more complete version:
Equation [7.5]:

$$E[Y_i(1) - Y_i(0)] = E[R_i(1)] \cdot E[Y_i(1)|R_i(1) = 1]) + (1 - E[R_i(1)] \cdot E[Y_i(1)|R_i(1) = 0]$$
$$- E[R_i(0)] \cdot E[Y_i(0)|R_i(0) = 1] - (1 - E[R_i(0)] \cdot E[Y_i(0)|R_i(0) = 0]$$

Assuming that the subjects are either Always-Reporters or Never-Reporters and that $\forall i, \tau_i = ATE$, then:

$$ATE_{AR} = ATE_{NR}$$
$$E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1] = E[Y_i(1)|R_i(1) = 0] - E[Y_i(0)|R_i(0) = 0]$$

And so:

$$\begin{aligned}
E[Y_i(1) - Y_i(0)] &= E[R_i(1)] \cdot E[Y_i(1)|R_i(1) = 1] + (1 - E[R_i(1)]) \cdot E[Y_i(1)|R_i(1) = 0] \\
&\quad - E[R_i(0)] \cdot E[Y_i(0)|R_i(0) = 1] - (1 - E[R_i(0)]) \cdot E[Y_i(0)|R_i(0) = 0] \\
&= E[R_i(1)] \cdot \{E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1]\} \\
&\quad + (1 - E[R_i(0)]) \cdot \{E[Y_i(1)|R_i(1) = 0] - E[Y_i(0)|R_i(0) = 0]\} \\
&= (E[R_i(1)] + 1 - E[R_i(0)]) \cdot (E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1]) \\
&= (1) \cdot (E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1]) \\
&= E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1]
\end{aligned}$$

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.

Using the general formula for the ATE,

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

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

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

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

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

From the table we see that the ATE is 1. Now let's assume that subject 1 and 3 are assigned to treatment and subject 2 and 4 to control. Subject 2âĂŹs outcome is missing. Let's assume that

the outcome measure can range from 0 to 10, in which case the extreme value bounds substitute 0 or 10.

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

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

Question 5

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

Answer:

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

Question 6

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

[1] Angrist, Bettinger, and Kremer 2006.

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

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

##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## [1,] 0.2318797 0.2318797 0.2905849 0.2665066 0.2905849 0.3271479
## [2,] 0.2846098 0.3141727 0.3158772 0.3275862 0.3471120 0.3471120

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

## (Intercept)      vouch0
##  46.9208148   0.6827378

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

## (Intercept)      vouch0
##  46.9654430   0.6580337
```

The estimated ATE from the weighted regression is 0.658, which is very similar to the unweighted estimate. This estimate suggests that assignment to the voucher increased reading scores by an average of 0.658 scale points (which is fairly small given a standard deviation of 5.6). None of the probabilities used for weights is outside the 0-1 range; the weights vary from 0.23 to 0.34.

Question 7

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

Answer:

Each of these practices may produce biased estimates. Subjects who "understand what hypothesis the experiment is testing" may have distinctive potential outcomes; discarding these observation may lead to bias, especially if they are more likely to suspect the hypothesis when assigned to the treatment group. Subjects who seem to not be taking the experiment seriously or fail to follow

directions may also have distinctive potential outcomes, and behavior that might cause them to be expelled may differ depending on experimental assignment.

Question 8

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

Table 1: Question 8 Table

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

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

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

[2] Ditlmann and Lagunes 2010.

group's missing values:

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

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

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

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

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

The identification problem stems from the fact that our experiment only provides direct evidence about the second quantity, since the only ones who request identification from whites are those with potential outcomes $r_i(W) = 1$ and $r_i(L) = 1$. The first quantity presents more of an empirical challenge because we observe outcomes from two types of retail clerks, those with potential outcomes $r_i(W) = 1$ and $r_i(L) = 1$ and with potential outcomes $r_i(0) = 1$ and $r_i(L) = 1$. However, the first term in the expression above refers only to the potential outcomes among clerks whose potential outcomes are $r_i(W) = 1$ and $r_i(L) = 1$.

Trimming bounds make use of the fact that we can estimate the relative shares of Always-Ask clerks ($r_i(W) = 1$ and $r_i(L) = 1$) and If-Latino-Ask clerks ($r_i(0) = 1$ and $r_i(L) = 1$) in the subject pool. The share of Always-Ask in the entire subject pool can be estimated based on the non-missingness rate in the White treatment group. The difference in non-missingness rates in the two experimental groups estimates the share of subjects who are If-Latino-Ask clerks:

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

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

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

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

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

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

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

The estimate is:

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

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

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

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

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

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

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

## [1] 0.04190119 0.18824265
```

Question 9

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

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

b) Another identical experiment is performed in a different developing country. This time the attrition problem is as follows: households that were offered cash grants are more likely to live at the same address years later, when researchers return in order to measure outcomes. Of the 1,000 households assigned to the treatment group, 900 are found when researchers return to measure outcomes, as opposed to just 700 of the 1,000 households in the control group. What are the implications of this type of attrition for the analysis and interpretation of the experiment?
Answer:
This type of attrition may be a source of bias. Migration (missingness) may be related to potential education outcomes, and the treatment (or lack thereof) may cause some households to relocate. For example, if students with lower potential education outcomes tend to migrate when their incomes are low, the treatment has the effect of causing some lower-performing students to remain in the non-missing sample, thereby reducing the estimated effect of the treatment based on a comparison of non-missing subjects in treatment and control. In this case, a researcher might turn to trimming bounds on the assumption that those who would have been available for an interview if assigned to the control group would also have been available for an interview if assigned to the treatment group.

Question 10

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

a) The first six rows of the table consider scenarios in which missingness is unrelated to potential outcomes. Each scenario varies the rate of missingness and the number of observations gathered in the second round of data collection. Compare the two estimators (one based on initial data collection only, the other based on both rounds of data collection) in terms of bias, precision, and the width of the extreme value bounds.
Answer:
The difference-in-means estimates based on the first round of data collection are the same for each of the simulations; what varies are the results of the second round of data collection. Because missingness is unrelated to potential outcomes, the point estimates are nearly identical across all simulations (all are unbiased), regardless of which estimator we choose. The standard errors increase as we move to the second round of data collection, because the two-round estimator is putting extra weight on imprecisely estimated quantities from the second round.

Across all pairs of simulations, the SE for the second round is reduced substantially when more observations are gathered in the second round. Even though the two-round approach does not reduce the standard error, it does markedly reduce the spread of the extreme value bounds, even when relatively few observations are gathered in the second round. Second round sampling does better in this regard because a smaller proportion of the sample has its missing values filled in with extreme values.

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

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

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