Experiments and Causality

David Reiley, David Broockman, D. Alex Hughes, Micah Gell-Redman, Scott Guenther, & David

2021-11-23

Contents

Live Session Introduction

This is the live session work space for the course. Our goal with this repository, is that we're able to communicate *ahead of time* our aims for each week, and that you can prepare accordingly.

Bloom's Taxonomy

An effective rubric for student understanding is attributed to Bloom (1956). Referred to as *Bloom's Taxonomy*, this proposes that there is a hierarchy of student understanding; that a student may have one *level* of reasoning skill with a concept, but not another. The taxonomy proposes to be ordered: some levels of reasoning build upon other levels of reasoning.

In the learning objective that we present in for each live session, we will also identify the level of reasoning that we hope students will achieve at the conclusion of the live session.

1. **Remember** A student can remember that the concept exists. This might require the student to define, duplicate, or memorize a set of concepts or facts.
2. **Understand** A student can understand the concept, and can produce a working technical and non-technical statement of the concept. The student can explain why the concept *is*, or why the concept works in the way that it does.
3. **Apply** A student can use the concept as it is intended to be used against a novel problem.
4. **Analyze** A student can assess whether the concept has worked as it should have. This requires both an understanding of the intended goal, an application against a novel problem, and then the ability to introspect or reflect on whether the result is as it should be.
5. **Evaluate** A student can analyze multiple approaches, and from this analysis evaluate whether one or another approach has better succeeded at achieving its goals.

6. **Create** A student can create a new or novel method from axioms or experience, and can evaluate the performance of this new method against existing approaches or methods.

Chapter 1

Importance of Experimentation

Why do we conduct experiments?

What is the value of making a causal statement?

This is a data science program, with enough data and a savvy enough model, can't we just generate a causal statement that will be right? Can't I generate a statement that converges in probability to the *correct* value?

1.1 Learning Objectives

At the end of this live session, students will be able to

1. *Remember* (or find) the goals of the course, the assessment structure, and the learning model.
2. *Define*, in non-technical language, what it means for an action to cause an outcome.
3. *Understand* the difference between a causal statement, and an association statement.
4. *Apply* the framework of causal thinking against a series of studies to determine whether the study has achieved the goal that it intends.

1.2 Class Introductions

In no more than 2 minutes, could each student please:

- Introduce themselves, announcing their name as they would like it to be pronounced;
- Tell us where in the world they are studying;
- State what semester they are in the program;
- Any descriptive features that they would like the class to know about them (for example, gender pronouns); and,
- [Instructor's choice]

1.3 Course Plan

The course is built out into three distinct phases

- **Part 1** Develops causal theory, potential outcomes, and a permutation-based uncertainty measurement
- **Part 2** Further develops the idea of a treatment effect, and teaches how the careful design of experiments can improve the efficiency, and easy of analysis
- **Part 3** Presents practical considerations when conducting an experiment, including problems that may arise, and how to design an experiment in anticipation of those problems.

1.4 Course Logistics

- bCourses
  - Learning Modules attached to weeks
  - Modules contain async lectures, coding exercises, and quizzes
- GitHub
  - All the course materials are available in a GitHub repository
  - We have protected the `main` branch, so you can't do anything destructive
  - Use that as empowerment! This is your class, propose changes that you would like to see!
- Github Classroom
  - Assignments will all be applied programming assignments against simulated and real data
  - All assignment code will be distributed through GitHub Classroom
- Gradescope
  - All assignments will be submitted to Gradescope where we'll read your solutions and provide scores and feedback

1.4.1 Learning model for the class

The course assignments are designed to put what we have learned in reading, async, and live session into practice in code. In our ideal version of your studying, we would have you working hard together with your classmates in a study group on the assignments, coming to office hours to talk candidly about what is and isn't working, and then *every single student* arriving at a full solution.

1.4.2 Feedback model for the class

We want to get you feedback *very* quickly after you turn your assignments.

1. We will release a solution set the day that you turn your assignment in
2. We will hold a problem set debrief office hour the Friday (i.e. next day) after the problem set it submitted
3. We will have light-feedback on your assignments within 7 days of when you submitted them.
4. You should bring your assignment to office hours after you have turned it in so that we can talk about any differences between your approach, and the instructors approach.

1.4.3 Office hour model for the class

- We will hold office hours Sunday through Thursday at 5:30.
- We will hold more than 10 hours of office hours every week; they will all be recorded, and any student is welcome in any office hour

1.5 Article Discussion

1.5.1 Predict or Cause

- What are a few examples that Athey raises of causal questions masquerading as prediction questions?
    1.
    2.
    3.
- Which of these examples is the most surprising to you?
- Is there something that is common to each of these examples? Is this a general phenomenon, or is Athey very clever in picking examples? Said differently, is Athey making a clever argument or is a lot of what we do as data scientists actually causal work in disguise?

1.5.2 Do the suburbs make you fat?

1. What is the causal claim being made in this article?
2. If you had to draw out this causal claim, using arrows, what would it look like?
3. Do you acknowledge the association that the authors present? Is there *actually* a difference between the BMI of people who live in cities and the suburbs?
4. If you acknowledge the association, does that compel you to believe the causal claim? Why or why not?
5. Name, and draw, five alternative *confounding* variables that might make you skeptical that the claimed relationship exists.
6. (Optional) Name, and draw two *mechanisms* that might exist between suburbs and BMI. Why does the existence (or not) of these mechanisms *not* pose a fundamental problem to the causal claim that the authors make?
7. At the conclusion of reading this paper, do you believe that there is a causal relationship between location and BMI? If so, what compels you to to believe this; if not, why are you not compelled to believe this?

1.5.3 Nike Shoes

1. What is the causal claim being made in this article?
2. If you had to draw out this causal claim, using arrows, what would it look like?
3. Do you acknowledge the association that the authors present? Is there actually a difference in the finish time between people who are running with the Nike shoes vs. other shoes?
4. If you acknowledge the association, does that compel you to believe the causal claim? Why or why not?
5. What are some of the confounding relationships that the authors identify? (Can you name four?) How do they adjust their analyses once they acknowledge the confounding problem?
6. At the conclusion of reading this paper, do you believe that there is a causal relationship between shoes and finish time? If so, what compels you to to believe this; if not, why are you not compelled to believe this?

1.5.4 What is Science: Feynman's View

In *Cargo Cult Science*, Richard Feynman poses a view of science that is about a seeking of the truth.

1. What is Feynman's view of science? What does he think makes something *scientific*?

2. What are ways that individuals fool themselves when they are working as scientists? What are ways that individuals fool themselves when they are working as data scientists?
3. How can we as (data) scientists, train ourselves not to be fooled?[1]

[1]This is a footnote, rendered into an html document.

Chapter 2

Apples to Apples

2.1 Learning Objectives

At the conclusion of this week's live session, student will be able to:

1. *Describe*, using the technical language of potential outcomes, what it means for an input to *cause* an output.
2. *Describe* the fundamental problem of causal inference.
3. *Apply* iid sampling as a method of producing an unbiased, consistent estimator of a population.
4. *Proove* that the average treatment effect estimator produces an unbaised, consistent estimator for the average treatment effect.

2.2 This Causes That

What does it mean for an action to cause an outcome? Don't worry about conducting the experiment, or any measurement concerns at this point, just engage with the concepts.

2.2.1 Damn fine coffee

Suppose that you're getting ready for class, and you want to make sure that you're at your best. So, you drink a cup of water, eat a small snack, and brew a small pot of coffee for while you're in class.

Why do you do this?

Presumably, you're doing this because you like each of these things, but also because you're interested in these things causing you to have a better class. If you framed this as as causal question, you might ask:

> If I drink a cup of coffee before class, will it cause me to be more alert?

What does it mean for coffee to cause alertness?

- Does coffee cause everyone to become more alert?
- Does coffee have to affect everyone equally in order for you to say it causes alertness?
- Could coffee have no effect for some people, and you would still say it causes alertness?

2.2.2 Meditation for focus

Suppose that you're getting ready for class, and you want ensure that you're at your best. So, you find a quiet place, and set your mind at ease with whatever form of meditation you think might be helpful.

> If I meditate before class, will it cause me to be more focused?

What does it mean for meditation to cause focus?

- Does meditation cause everyone to become more focused?
- Does meditation have to affect everyone equally?
- Some people are frustrated by not being able to quiet their thoughts, and actually find meditation frustrating. Can this be true, and still believe that meditation causes focus?

2.2.3 Selling coffee and meditation

Suppose that you're an enterprising soul, and you want to sell a book about brewing coffee as a meditation. You reason that there must be a niche for this approach. To get the word out, you place a few flyers with tear off phone-numbers at the local yoga studios and tech incubators (good intuition to find those MIDS students).

> If shown a flyer for coffee-meditatin, will it cause someone to take my training?

What does it mean for for flyers to cause people to sign-up for the training?

- Does the flyer cause everyone to take the training?
- Does the flyer affect everyone equally?

2.2.4 Reflecting on Causes

Does anything unify questions of causes?

When you think about *{this}* causing *{that}*, do you think about it at a population level, a smaller group level, or at the individual level?

2.3 Potential Outcomes

Potential outcomes are a system of reasoning, and a corresponding notation, that allow us to talk about observable and un-observable characteristics of the world.

> What is your position on *ontology*? What does it mean for something to exist?

- Does *Field Experiments*, as a textbook, exist?
- Do Don Green and Alan Gerber, the authors of the textbook that we're reading, exist?
- Does David Reiley, the slower-talking Davids in the async, exist?
- Do I, your section, instructor, exist (or am I a deep fake in this room with you)?
- Can a concept exist, even if you can't hold it? Even if you haven't seen it?

2.3.1 Defining Potential Outcomes

For each of the following sets of notation: (1) Read the notation aloud, not as "Y sub i zero," but instead as "The potential outcome to control ...."

- $Y_i(0)$:
- $Y_i(1)$:

- $E[Y_i(0)]$:
- $E[Y_i(1)]$:
- $E[Y_i(0)|D_i = 0]$:
- $E[Y_i(1)|D_i = 1]$:
- $E[Y_i(0)|D_i = 1]$:
- $E[Y_i(1)|D_i = 0]$:

- Which of these concepts that you have just read aloud exist?
- Can a concept exist, even if you can't hold it? Even if you can't see it?

2.4 Using Independence

Suppose that you have a random variable that is defined as the function,

$$Y = \begin{cases} \frac{1}{10} & , 0 \leq y \leq 10 \\ 0 & , \text{otherwise} \end{cases}$$

- What is the expected value of this function?

$$\begin{aligned} E[Y] &= \int_0^{10} y \cdot f_y(y)\ dy \\ &= \int_0^{10} y \cdot \frac{1}{10}\ dy \\ &= \frac{1}{10} \int_0^{10} y\ dy \\ &= \frac{1}{10} \cdot \frac{1}{2} y^2 \Big|_0^{10} \\ &= \frac{1}{20} y^2 \Big|_0^{10} \\ &= \frac{1}{20} \cdot [(100) - (0)] \\ &= \frac{1}{20} \cdot 100 \\ &= \mathbf{5} \end{aligned}$$

- Why is the expected value a good characterization of a random variable?
- If you wanted to write down an estimator to produce a summary statistic for Y given a sample of data, what properties do the following estimators possess:
- $\hat{\theta}_1 = y_1$
- $\hat{\theta}_2 = \frac{1}{2} \sum_{i=1}^{2} y_i$

- $\hat{\theta}_3 = \frac{1}{n-1}\sum_{i=1}^{N} y_i$
- $\hat{\theta}_4 = \frac{1}{n}\sum_{i=1}^{N} y_i$

```
population_function <- function(sample_size) {
  runif(n=sample_size, min=0, max=10)
}
```

```
theta_1 <- function(data) {
  # take the first element

}

theta_2 <- function(data) {
  # sum the first two elements and divide by two

}

theta_3 <- function(data) {
  # sum the sample, and divide by 1 less than the sample size

}

theta_4 <- function(data) {
  # sum the sample, and divide by the sample size
  # honestly, just use the mean call.
  # clearly, this is a silly function to write, since you're just
  # providing an alias, without modification, to an existing function.

}

theta_4 <- function(data) {
  mean(data)
  }
```

```
theta_4(population_function(100))
```

```
## [1] 4.664467
```

Chapter 3

Quantifying Uncertainty

When we are working with a sample of data, estimates produced by an estimator might change – sometimes being higher than the *true* value, other times lower than the *true* value.

In Frequentist inference, we understand the variance in these estimates as *sampling based variance of the sample estimator*. In this week, we present a different inferential paradigm, **Randomization Inference**.

In randomization inference, there is no uncertainty about the parameter estimate that is generated in the experiment: The estimate that we observe is the estimate that we observe. Uncertainty, instead, comes from the acknowledgment that different *randomization* could have been realized, even from within the same sample.

3.1 Learning Objectives

At the end of this week's live session, students will be able to

1. *Understand* the sharp null, and how to apply it in an argument using randomization inference.
2. *Describe* how randomization creates uncertainty, and *assess* how this uncertainty differs from that in Frequentist paradigm
3. *Apply* the sharp null and randomization inference to data
4. *Assess* the assumptions necessary for Frequentist inference to produce nominal coverage on confidence intervals; *assess* the assumptions necessary for randomization inference to produce nominal coverage on confidence intervals; and, *evaluate* which of the two approaches is appropriate given a set of data.

5. *Describe* the concept of statistical power and what it means in the context of conducting an experiment.

3.2 Value of Theory

Suppose that you are evaluating the effect of coffee on students' alertness in class. You reason that drinking coffee will increase students' alertness in class.

Figure 3.1: "Damn Fine Coffee."

One might be a radical behaviorist (Skinner is perhaps the most famous in this line of thinking) that says, "In matters of human behavior, if I cannot see it, then I cannot reason or know about." If this is your view, then you would simply stop your investigation (and reasoning) at the conclusion of your experiment.

In many ways, experiments suited only to answer empirical, observable questions. These are the questions, and lens proposed by the radical behaviorist paradigm.

3.2.1 Limits of Behaviorist Reasoning

If you accept only that coffee has this effect, and that it is measurable, are you able to translate this knowledge to a new context?

- Suppose that your experiments finds an effect of coffee on alertness: those who drink a cup of coffee are more alert in class.
- Suppose, though, that you're out of coffee *tonight*. A radical behaviorist would simply say, "I know not what to drink then to increase my alertness."

3.2.2 Value of Theory

- Can you produce several theories (some of them might be silly) about why coffee might increase alertness in class?
  - Proposed theory #1:
  - Proposed theory #2:
  - Proposed theory #3:
- Does Feynman's approach to *Science* provide a method to adjudicate which of these theories is consonant with the evidence, and which are not consonant with the evidence? - Does Lakatos' approach to *Science* provide such a method?

3.2.3 Evaluating Theories

- What data might you be able to produce that would allow you to "drive a wedge" between the different theories?
- This ability to proactively deign an experiment to distinguish between theories is the goal you're striving to achieve, *and it is very hard to accomplish.*

3.3 Stating the sharp null

Continue with our idea of an experiment to evaluate if coffee produces alertness in class. Here, we are going to further develop this notional experiment into something that we might actually be able to conduct.

- What is the *sharp null* hypothesis that is at risk in this investigation?
- How, if at all, does this sharp null differ from the null hypothesis you might be more familiar with?
- Is the sharp null hypothesis a concept that ever makes sense? Is the sharp null hypothesis a concept that is ever, actually, true?

3.4 Randomization Inference

3.4.1 Stating the process of Randomization Inference

Randomization inference is a method of understanding the variability of results in an experiment that you have conducted. It specifically acknowledges several facts:

1. The sample of data that you collected or used in your experiment is, quite simply, the sample of data that you collected for your experiment. There might be a larger population; there might be an infinite population; or, there might not.
2. The observed outcomes that you observe are, quite simply, the outcomes that you observed. There is no uncertainty about having seen these.
3. When the experiment assigned some units to treatment and others to the control, it revealed some outcomes, for some people. Specifically, it revealed the potential outcoems to treatment, denoted $Y_i(1)$ for those who were assigned to the treatment group and the potential outcomes to control, denoted $Y_i(0)$ for those who were assigned to the control group.
4. The experimenter chose one *out of many possible* treatment assignments.
5. If the *sharp null hypothesis* were to be true (note the subjunctive verb tense there) then, the particular revelation of potential outcomes to treatment and control are inconsequential. Despite seeing only half the data (referred to as the **Fundimental Problem of Causal Inference**) we actually posess all the data. After all, if the sharp null were true, $Y_{Alex}(1) = Y_{Alex}(0)$, and $Y_{David}(1) = Y_{David}(0)$, $Y_i(1) = Y_i(0)$ for all of the $i = 1, \ldots, N$ people who are a part of the experiment.

3.4.2 Questions about Randomization Inference

- Where does uncertainty come from in an experiment that is evaluated using randomization inference?
- How is the ATE estimand defined?
- What is the feasible method that we use to write down an estimator (call it θ) for this quantity?
  - Which of the following properties does this feasible method possess?
    a. Unbiasedness: $E[\theta] = ATE$
    b. Convergence: $\theta \xrightarrow{p} ATE$, where $\xrightarrow{p}$ means converges in probabilty
    c. Efficiency: The mean squared error of θ is either (i) smaller than some other estimator, or (ii) as small is theoretically possible.

3.5 Applying Randomization Inference

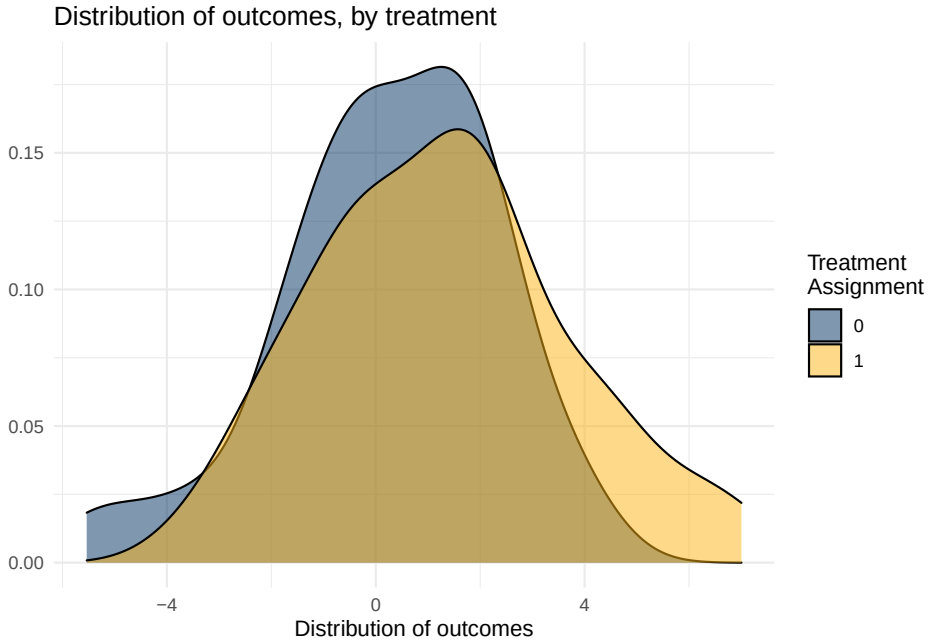
3.5.1 Make Data

```
set.seed(1)
d <- data.table(
  id = 1:100,
  D  = rep(0:1, each = 50),
  Y  = c(rnorm(n=50, mean=0, sd=2.5), rnorm(n=50, mean=1, sd=2.5))
)
```

3.5.2 Plot Data

In the following plot, are you able to assess whether there is a treatment effect simply by looking at the distributions?

```
ggplot(d) +
  aes(x=Y, fill=as.factor(D)) +
  geom_density(alpha=0.5) +
  labs(
    x        = 'Distribution of outcomes',
    y        = NULL,
    title    = 'Distribution of outcomes, by treatment',
    fill     = 'Treatment\nAssignment') +
  scale_fill_manual(
    values = c('#003262', '#FDB515')
  )
```



3.5.3 Classic Test

If you were to write a *classic* test against this data, given what you know about how it was generated, what would be the classic test? What do you learn from this test, and what is the interpretation?

```
d[ , t.test(Y ~ D)]
```

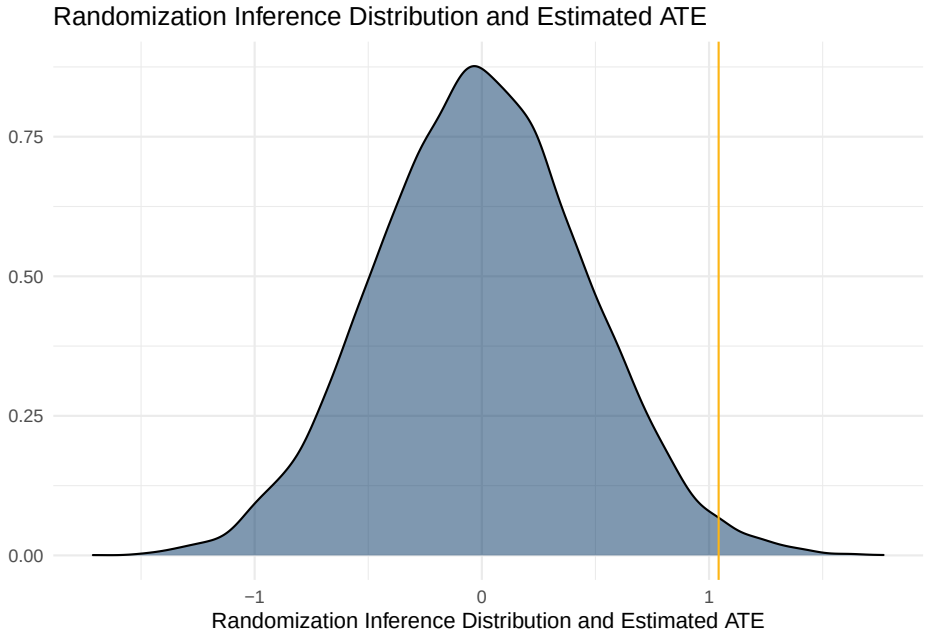
```
##
##  Welch Two Sample t-test
##
## data:  Y by D
## t = -2.309, df = 95.793, p-value = 0.02309
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9381728 -0.1462181
## sample estimates:
## mean in group 0 mean in group 1
##       0.2511207       1.2933161
```

3.5.4 Randomization Inference Test

Now, instead suppose that you were to conduct the randomization inference. What are the steps to the algorithm for producing a result using randomization?

1. State the null hypothesis
2. Compute the statistic of interest using the observed data
3. Fill in data, under the statement of the null hypothesis
4. Permute the treatment assignment labels to generate a new sample of the treatment assignment vector, and then estimate the statistic of interest
5. Repeat the permutation and estimation (step 4) process repeatedly to sample from the randomization inference distribution of the statistic
6. Examine randomization inference distribution

```
## 1. The sharp null is that tau = 0
## 2. Compute the statistic of interest
true_ate <- d[ , .(group_mean = mean(Y)), keyby = .(D)][ , group_mean[D==1] - group_mean[D==0]]
## 3, 4, 5. Permute the treatment assignment labels and repeatedly compute the statistic of inter
ri_distribution <- replicate(
  n=10000,
  expr = d[ , .(group_mean = mean(Y)), keyby = .(ri_treatment = sample(D))][ ,
            group_mean[ri_treatment==1] - group_mean[ri_treatment==0]]
  )
# 6. Examine distribution
ggplot() +
  geom_density(aes(x=ri_distribution), fill = '#003262', alpha = 0.5) +
  geom_vline(xintercept = true_ate, color = '#FDB515') +
  labs(
    x     = 'Randomization Inference Distribution and Estimated ATE',
    y     =  NULL,
    title = 'Randomization Inference Distribution and Estimated ATE')
```



How much of the randomization inference is more extreme than the treatment effect?

```
ri_p_value <- mean(abs(ri_distribution) > abs(true_ate))
ri_p_value
```

```
## [1] 0.0226
```

Notice that 0.023 of the randomization inference distribution is more extreme than the observed treatment effect. How does this compare to the t-test p-value that we calculated above?

3.6 Comparing Randomization Inference and Frequentist Inference

3.7 Statistical Power

Chapter 4

Blocking and Clustering

4.1 Learning Objectives

1.
2.
3.

Chapter 5

Covariates and Regression

5.1 Learning Objectives

1.
2.
3.

Chapter 6

Regression and Multifactor Experiments

6.1 Learning Objectives

1.
2.
3.

Chapter 7

Heterogeneous Treatment Effects

7.1 Learning Objectives

1.
2.
3.

Chapter 8

Treatment Noncompliance

8.1 Learning Objectives

1.
2.
3.

Chapter 9

Spillover and Interference

9.1 Learning Objectives

1.
2.
3.

Chapter 10

Causality from Observational Data

10.1 Learning Objectives

1.
2.
3.

Chapter 11

Problems and Diagnostics

11.1 Learning Objectives

1.
2.
3.

Chapter 12

Attrition, Mediation, and Generalizability

12.1 Learning Objectives

1.
2.
3.

Chapter 13

Applications of Experiments

13.1 Learning Objectives

1.
2.
3.

Chapter 14

Review of the Course

14.1 Learning Objectives

1.
2.
3.