Behavioral Types: A New Perspective on Estimating Treatment Effects in Social Science Experiments with Binary Responses

by

David GrahamSquire

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Deborah Nolan, Chair
Professor Jasjeet S. Sekhon
Associate Professor Laura Stoker

Summer 2018

Behavioral Types: A New Perspective on Estimating Treatment Effects in Social Science Experiments with Binary Responses

Abstract

Behavioral Types: A New Perspective on Estimating Treatment Effects in Social Science Experiments with Binary Responses

by

David GrahamSquire

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Deborah Nolan, Chair

In the year 2000, Gerber and Green published the results of a field experiment which examined the impact of electoral campaigns on voter participation. Since this landmark work, more than one hundred similar studies have appeared in the political science literature. These randomized controlled trials are usually conducted within get-out-the-vote (GOTV) drives seeking to increase voter turnout. The surge in GOTV experiments was partly due to a statistical innovation that preceded Gerber and Green's publication, the average treatment effect for the treated (ATT), which allowed researchers to compare directly those treated to a similar group that was assigned to the control.

In this research we focus on settings common to many social science field experiments such as those of GOTV studies, where participants may comply, or not, with the treatment protocol assigned by the experimenter. For experiments with binary outcomes, we show that each individual in the study may be classified as one of a finite number of distinct *types*. We call these *behavioral types* because they characterize the individual's complete reaction, their measured response and how they receive treatment, to the assignment of each possible experimental group. In this context, the data is generated by randomly allocating these various behavioral types to the different levels of treatment. Thus, the model is parameterized by the unknown proportions of the different behavioral types so that many statistical aspects of the experiment, such as commonly studied average treatment effects, may be written as a function of these proportions.

Viewing the data as generated by these behavioral types changes the analysis of the experiment in two ways. First, it changes the perspective on what is being estimated. Instead of finding a particular treatment effect, the ultimate goal can be seen as estimating proportions of behavioral types. With this frame of reference, the effect of a certain treatment will be most accurately represented as the fraction of the experimental sample for which the treatment has an effect. Second, by clarifying the underlying data generating process, a behavioral-types approach directs the resulting statistical analysis.

We use a well cited example to introduce behavioral types before providing formal definitions. We present the ATT as a case study for how to apply a behavioral-types approach for a design known to many social science researchers. The understanding of the data generating process allows us to evaluate the bias and variance of the ATT estimator, and we show the variance depends on the choice of the sampling assumptions. We then provide rigorous definitions of a behavioral type and of *restrictions* which reduce the number of behavioral types in a population to a number where the proportions of each type may be estimated. We present three experimental designs and present a strategy to identify the proportions of each type and elucidate how treatment effects may be found from the proportions.

A behavioral-types approach is well suited to multi-treatment experiments because it distills often complex designs into an estimation problem of a manageable number of types. We apply the behavioral types approach to four published social science field experiments involving multiple levels of ordered treatment. For each, we show how the interpretations and the statistical analyses differ with a behavioral types approach, and can lead to different conclusions. Through the applications we illustrate how behavioral types provides insight into a range of experimental designs, such as those with spillover effects or partial ordering of treatment levels.

For two of the four applications we further examine the issue of joint significance by constructing multi-dimensional confidence regions for the proportion of behavioral types. We find that normal approximation methods perform poorly, but the shortcomings can be corrected by bootstrap methods. However, even the bootstrap regions may not attain the desired coverage levels, so we adjust our regions using a double bootstrap. We discuss other methods that merit further exploration.

To Regan

For the wild and the beautiful and the laugh out loud times. And the mundane and hard times too. It's all so much better with you.

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply indebted to my thesis advisor Deborah Nolan, for her insight, mathematical tenacity, and patience. Deb spent over 200 hours (I added them up) meeting with me, guiding, and making sure this thesis was done. I hope I can reciprocate her generosity somewhere down the line. Committee members Jasjeet Sekhon and Laura Stoker provided very needed feedback and helped frame this work within the fields of Causal Inference and Political Science. I had useful conversations with Philip Stark, Chad Schafer and Peng Ding about their research and I appreciate their eagerness to help. My graduate studies were split over two time periods and I thank Bin Yu for all her advising during the first epoch. I learned a lot from, and laughed a lot with, Bin and her mentorship was an important part of this process. I also learned much from Roger Purves and Hank Ibser and I am grateful for the efforts of all the other Statistics faculty who taught my course work (in chronological order), Steve Evans, Andrew Gelman, Jim Pitman, Peter Bickel, David Freedman and Terry Speed; you were all inspirational.

Timmy Lu from the Asian Pacific Environmental Network made the space for me, during very a very hectic campaign, to try out some of my ideas. Timmy was a pleasure to work with and my involvement with his get-out-the-vote program sparked my return to graduate school.

Tanya Smith, Laurel Lucia, Regan Brashear, Carl Nadler, Ryan Firestone, Elinor Graham (Thanks Mom!), Sharon Lerman, and Adam and Mike Graham-Squire all read key parts of my thesis and I am grateful for their helpful comments and eagle eyes. And Linda Lan was a true friend who kept me on task during the last year.

Though I wouldn't say studying for prelims was enjoyable, it was a bonding experience, and I'm thankful I got to go through it with fellow classmates Jon McAuliffe, Tao Shi, Xiaoyue Zhao, Ingileif Halgrimmsdottir, Chao Chen, Yong Cho, Apratim Guha, Noam Berger, Jonathan Grib, Aiyou Chen, Jason Reed and Michael Last. Henrik Bengtsson, Andrea Gordon, Jane Frydlyand, Gang Liang, Daphna Michaeli, Yu Chuan Tai, Katerina Kechris and Joel Hansen kept things fun. And I'm glad I got to share an office with Alexander Tsigler during the final year. His calming presence and dry humor helped keep up my spirits.

I enjoyed every interaction with the staff of the Statistics Department! Ryan Lovett, Luis Torres and Chris Paciorek were a wonderful resource and very responsive when I ran into computing problems. There are so many administrative hassles to get this degree but La Shana Porlaris either buffered me from them or showed me how to quickly negotiate my way

through. Mary Melinn was always there when I needed assistance, which was often. Laura Slakey kept the whole department afloat. And I appreciate the work of all the university staff, which keep this campus running, in ways I don't even know about.

Sylvia Allegretto, Ken Jacobs, Xiao Chen, Miranda Dietz, Laurel Lucia, Greg Watson, Jerry Kominski, Ian Perry, Dylan Roby, Barbara Campbell. Petra Rasmussen and Jack Needleman were incredibly supportive in helping me juggle my day job at The Labor Center. And they covered for me, especially during the final weeks, to allow the time to focus and wrap up this project. Ken was especially patient, and never scrunched up his face, or even blinked, when I told him I would need just a little more time to finish (even though he had numerous opportunities to do so).

Zephyr, Dominic, Jonah, Eva, Max and Ronan have brought so much joy during these recent years that I can't write this without mentioning them. And having Steve, Dad, Pam, Jerry, Corbin, Joan and Anne for family is one of the great blessings of my life.

Finally, I want to thank my sweetheart, Regan Brashear, who was with me at every step. Regan helped with so many parts of my thesis but most importantly helped me to *not* work on my thesis when I needed a break. Regs, I love you and I'm so lucky to have you in my life.

Chapter 1

Introduction

In the year 2000, Gerber and Green published the results of a field experiment which examined the impact of electoral campaigns on voter participation. Since this landmark work, more than one hundred similar studies can be found in the political science literature. These randomized controlled trials, where the subjects are registered voters, are usually conducted within get-out-the-vote (GOTV) drives seeking to increase voter turnout. For those assigned to treatment, campaigns attempt some form of contact via mail, phone, face-to-face conversation or social media while voters assigned to control are not contacted. The outcome, whether a subject votes or not, may be determined by examining publicly available voting records after the elections. The studies address a number of practical questions. Which medium of contact is the most cost effective for turning out additional votes? How close to Election Day should contact be made? What form of outreach works best for targeted demographic groups such as youth, women of color, or Latinos? The experiments are also outside of academia as campaigns conduct internal studies, such as testing different messaging strategies during a primary to see which will be most useful during a general election.

This surge in GOTV experiments was partly due to a statistical innovation that preceded Gerber and Green's publication: the average treatment effect for the treated (ATT). When some of the subjects assigned to treatment do not receive the treatment, statisticians have long warned about comparing the entire control group to the subset of the treatment group actually treated, as estimated effects will be impacted by selection bias. The remedy to this noncompliance was to compare the entire control group to the entire treatment group, whether the treatment was received or not. This intention-to-treat (ITT) effect measures the impact of being *assigned* to treatment. However, the average treatment effect for the treated (ATT) allowed researchers to compare directly those treated, on average, to a similar group that was assigned to the control, to obtain the average treatment effect on the treated. The innovation of ATT arose from applying the potential outcome framework of Neyman and Rubin (see Holland, 1986) to settings with noncompliance. With a few reasonable assumptions, the ATT may be found in all designs with a control and single treatment when subjects either are treated, or are not treated and effectively receive the control protocol.

This is a common setting for experiments in the social sciences, as well as medicine, since for ethical reasons, subjects must be allowed to not comply with the assigned treatment. For GOTV experiments, subjects who are assigned to treatment are reached or they are not. That is, the "treated" are subjects who the campaign was able to successfully contact. Thus the ATT provides a measure of the increased voter turnout of those actually reached, a more direct measure of the impact of the campaign than the ITT effect. Gerber and Green were the first political scientists to recognize the usefulness of the ATT and how researchers could carry out experiments within existing GOTV campaigns.

Our work focuses on experiments with a finite number of experimental groups, where subjects may or may not comply with the assigned treatment, and the special case where outcomes are binary (as they are in GOTV campaigns). Our main finding is that in this context we may group subjects by their compliance behavior to treatment assignment and their response behavior to treatment assignment, so that every subject may be classified as one of a finite number of distinct types. We call these "types" *behavioral types*. A behavioral type thus refers to an intrinsic behavioral trait of an individual, describing how they receive and respond to the different treatment assignments. Furthermore, the observed data is generated by a simple model which is completely parameterized by the proportion of each of the behavioral types in the experimental sample. This brings a new perspective on properties of interest, such as the ITT or ATT, as they may be viewed as a function of the proportion of these behavioral types. From this perspective, we show,causal effects are less about the impact of a certain intervention on a population and more about the number of individuals in the population whose compliance and response behaviors lead them to be impacted. The behavioral-types approach applies to a number of settings, as long as there is a finite number of treatment conditions and a finite number of ways for subjects to comply with their assigned treatment. We demonstrate the wide applicability of this approach in a series of examples.

In Chapter 2 we review the potential outcomes framework and delve into the derivation of the ATT in the special case when outcomes are binary. We use the well-cited work of Angrist, Imbens, and Rubin (1996) as a case study to introduce the notion of behavioral types. In this setting, with a control, one treatment and the presence of noncompliance, every subject may be associated with one of five distinct types. We uncover how the ATT and ITT may be expressed as the proportions of these behavioral types.

In Chapter 3 we focus on the properties of the ATT estimator. We derive the bias of the estimator and examine it's variance, which hinges on the assumption of whether the control and treatment are drawn from an infinite superpopulation or if treatment is assigned, without replacement, given a finite sample. We quantify how the asymptotic variance differs under the two assumptions, again in terms of the proportions of behavioral types. However, even when the assumptions lead to a difference in the calculated variance, both assumptions lead to similar conclusions about the significance of the ATT estimate. That is, a difference in the variance estimators is found when p-values are very small. We also examine an alternative estimate of the causal effect based on randomization inference, which has fewer assumptions

and does not require large samples.

In Chapter 4 we present a rigorous definition of a behavioral type. There is an upper limit to the number of unique behavioral types, which may exist, in any experimental setting. Many of the assumptions made by researchers, such as those used to derive the ATT, may be seen as restrictions on the kinds of behavioral types, that may exist. These restrictions are key to limiting the number of unique behavioral types in an experiment to a manageable number so their proportions may be estimated. We present a methodical procedure to determine the number of distinct behavioral types in an experiment and identify the restrictions that eliminate certain types. We apply this method to three experimental designs, which we present as case studies. First, we consider the ATT setting of the previous chapter and show how the step-by-step process leads to the same five behavioral types identified in Chapter 2. Second, we choose a design with multiple levels of treatment, where the levels have a clear ordering. Third, we examine a proposed GOTV experiment that allows for the measurement of spillover effects. In each case study we show how the parameters of most interest may be expressed as functions of the proportions of the behavioral types and how these proportions may be estimated directly from the data.

In Chapter 5 we apply our methods to experiments with multiple levels of treatment. A behavioral-types approach is well suited to multi-treatment experiments because it distills these often complex designs into an estimation problem of a manageable number of types. We choose four published social science field experiments and re-analyze them through the lens of behavioral types, comparing our conclusions to those of the authors. The first experiment contains treatments that are ordered in terms of the severity of the intervention, from weakest to strongest, and this order is seen in the observed response rates for each treatment group. The authors arrive at unambiguous conclusions, and our analysis concurs. The second experiment has a similar design, but the conclusions are not nearly as strong as those of the first experiment. While we agree with the authors about the significance of the estimated effects, we differ on our interpretation of these effects due to our focus on behavioral types. What they see as only one of the two treatments having an effect we view as at least of two behavioral types must exist. The third experiment uses a complex design to estimate indirect or spillover effects. Though we are not able to identify all of the behavioral types proportions, our analysis leads to stronger evidence for spillover effects than the analysis of the authors. The fourth experiment is the only one of the four to involve noncompliance and is further complicated by a partial ordering of the treatments. The data needed to carry out an analysis with behavioral types is unavailable, but we show how such an analysis could be conducted if the observed response rates were summarized by treatment assigned and treatment received. For all four experiments, the behavioral-types approach results in inference of multiple parameters, which raise questions of joint significance. We address these concerns in the next chapter.

In Chapter 6 we examine the issue of joint significance by constructing multi-dimensional confidence regions for the behavioral types proportions. We take two of the experiments from Chapter 5 to use as case studies. We find that normal approximation methods perform

poorly, but the shortcomings can be corrected by bootstrap methods. However, even the bootstrap regions may not attain the desired coverage levels, so we further adjust our regions using a double bootstrap. We briefly discuss three other promising methods that merit further exploration.

In Chapter 7 we draw conclusions, note areas which could be improved, and discuss additional research possibilities.

Though we initially focused on the ATT, as it is an application well known by many social science researchers, the behavioral-types approach is quite general. Our aim is to demonstrate its applicability in a wide range of experimental settings: with noncompliance, multiple levels of treatment, partial orderings of the treatments, partial compliance, and designs which allow the measurement of indirect or spillover effects. In GOTV field experiments researchers often use some sort of linear modeling to address non-standard designs. Our hope is that rather than rely on these rather strong modeling assumptions, social science researchers analyzing data with binary outcomes can begin to view their research with behavioral types in mind. This approach is much less model dependent, so it can result in more persuasive conclusions and generate a deeper understanding of the underlying data generating process.

Chapter 2

Revisiting Potential Outcomes with Binary Responses

In randomized controlled trials (RCT), the difference between the observed responses of the treatment and control groups is caused by some combination of a true causal effect of the treatment and by the random assignment, which may lead to an imbalance in the average responses of the two groups even when the treatment has no effect.

The potential outcomes framework provides a mathematically rigorous definition of the average causal effect of a treatment and describes how it can be estimated from an experiment. The roots of the field began with Neyman (1923) who described the basis of causal inference from agricultural experiments. An important innovation, most widely attributed to Angrist and Imbens (1994), allowed for the measurement of the causal effects without requiring all subjects to comply with the assigned treatment. This innovation, the *average treatment effect for the treated* (*att*), isolates the impact of the treatment to only those who receive the treatment and has been a breakthrough for social science field experiments, where noncompliance is necessary for ethical studies with human subjects.

In this chapter we show that for an RCT where subjects may not comply with treatment, and where responses are binary, we may classify every individual as belonging to one of a finite number of *types* which leads to a novel parameterization of the model. We begin by describing the potential outcomes framework underpinning the estimates of causal effects. When outcomes are binary, we show that each individual in the experiment may be characterized as a certain "type" (for example, individuals who comply with whichever treatment is assigned and always have a positive response). We call these different types of individuals *behavioral types* and show that the proportion of the behavioral types in the experimental sample may be used to parameterize the underlying model. In this context, treatment effects such as the *att* may be interpreted as proportions of these types and we demonstrate how these proportions may be identified from the observed data. Our description in this chapter is more informal to make concepts more accessible. We develop precise definitions in Chapter 4.

The paper by Angrist and Imbens (1994) was followed two years later by Angrist et al. (1996), a longer exposition of the 1994 publication, written more explicitly within the context of potential outcomes. The follow-up publication also identified the assumptions needed for a causal interpretation of the estimator, $\widehat{att}$, to be plausible. In many ways this chapter mirrors the second article as we adopt its notation and restate a number of its assumptions, restricted to the case when response values are binary.

Though we present our findings in general terms, a widely used application has been to measure the impact of get-out-the-vote (GOTV) campaigns. Gerber and Green (2000) were early to apply estimates of *att* to the GOTV setting, and their landmark study has spurred over a hundred field experiments on electoral strategy (see Bedolla and Michelson, 2012). We often cite these studies in our examples.

2.1 Potential Outcomes and the Neyman-Rubin Causal Model

The basis of the potential outcome framework was first described by Neyman (1923), further clarified by Rubin (1974) and is often referred to as the Neyman-Rubin Causal Model as dubbed by Holland (1986). The key feature of the framework is discerning the source of randomness in the observed data. To highlight this we follow a common notation convention where random quantities are represented with uppercase symbols and non-random quantities such as fixed sample sizes and parameters are written in lowercase.

Consider a sample of n individuals participating in an experiment where each is assigned to treatment or control. Let Z_i be an indicator of whether individual i was randomly assigned to treatment. At this point we assume perfect compliance with treatment assignment, that is, each subject assigned to treatment receives the protocol of the treatment group and each subject assigned to control receives the protocol for the control group (we discuss imperfect compliance, where there is noncompliance for some subjects in the next section). Each individual is associated with an *observed outcome*, Y_i, which is measured from the experiment. In Neyman's framework each individual has two intrinsic *potential outcomes*, y_{i1} and y_{i0}, or y_{iz}, where z denotes the receipt of treatment. The observed outcome is

$$Y_i = y_{i1}Z_i + y_{i0}(1 - Z_i). \tag{2.1}$$

Of the quantities in Equation 2.1, the only source of randomness comes from the Z_i. With this understanding of how the observed data is generated, we define the *treatment effect* for individual i to be $y_{i1} - y_{i0}$ and the sample's *average treatment effect (ate)* as

$$\begin{aligned} ate &\equiv \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - y_{i0}) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_{i1} - \frac{1}{n}\sum_{i=1}^{n} y_{i0}. \end{aligned} \tag{2.2}$$

This is the difference between the average potential outcomes for treatment and the average potential outcomes for control. Since we typically cannot observe both y_{i0} and y_{i1} for subject i, what Holland (1986) calls *The Fundamental Problem of Causal Inference*, it is not possible to measure the individual treatment effects. However, we can find an unbiased estimate of *ate* from the difference in the sample averages of the treatment and control groups,

$$\overline{Y_1} - \overline{Y_0}.$$

Implicit in this estimation is the assumption that each subject only has two potential outcomes, and these outcomes depend on the assignment of the subject and are not influenced by the assignment of other individuals. More formally we state this as our first assumption.

Assumption 1 (SUTVA)**.** *The causal effect of any individual does not depend on the assignment of other individuals. This is commonly known as the Stable Unit Treatment Value Assumption (SUTVA) as first defined by Rubin (1974).*

Without this assumption the potential outcomes of any subject could depend on the assignment of the other $n-1$ subjects which leads to each individual having 2^n potential outcomes.

2.1.1 Compliance to Treatment Assignment and Compliance Types

In experiments with human subjects, those assigned to receive treatment may or may not comply with the treatment protocol. In this context, the *ate* may not be useful. If some fraction of the subjects never accept treatment, the average response of the treated, $\frac{1}{n}\sum_{i=1}^{n} y_{i1}$ may not be a meaningful feature to estimate. Instead we may measure the impact of treatment assignment on the outcome. To formalize this, and allow us to mirror the arguments of Angrist et al. (1996), which present treatment effects in terms of probabilistic expectation, we focus on the response of a randomly chosen member of the sample. Suppose one of the n subjects is randomly selected such that each may be chosen with probability $1/n$. Let Y be their response and Z their treatment assignment. We are interested in the impact of treatment assignment on outcome, commonly known as the *intention-to-treat* (*itt*) effect or

$$itt \equiv \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0). \tag{2.3}$$

We can evaluate both terms by conditioning on which subject is chosen so that

$$\begin{aligned} itt = &\sum_{i=1}^{n} \mathbb{E}(Y \mid Z = 1, \text{subject } i \text{ chosen}) \Pr(\text{subject } i \text{ chosen}) \\ &- \sum_{i=1}^{n} \mathbb{E}(Y \mid Z = 0, \text{subject } i \text{ chosen}) \Pr(\text{subject } i \text{ chosen}) \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i \mid Z_i = 1, \text{subject } i \text{ chosen}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i \mid Z_i = 0, \text{subject } i \text{ chosen})$$

$$itt = \frac{1}{n}\sum_{i=1}^{n} \left[\, \mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0) \,\right]. \tag{2.4}$$

Under perfect compliance $\mathbb{E}(Y_i \mid Z_i = 1)$ would equal y_{i1}, and $\mathbb{E}(Y_i \mid Z_i = 0)$ would be y_{i0}. So *itt* would be identical to *ate*. With noncompliance, we focus on the individuals who comply with treatment, and estimate the treatment effect for just those who are treated. This is called the *average treatment effect for the treated* (*att*), also known as the *local average treatment effect* or sometimes the *complier average treatment effect*. This insight was reported by Sommer and Zeger (1991), before it was noted by Angrist and Imbens (1994). Before we can understand *itt* and *att* in terms of the potential outcomes values, we must incorporate the degree to which individuals comply with treatment assignment. For simplicity, we contain ourselves to simple compliance of the two possible assignments where an individual either receives the treatment or control protocol, with nothing in between (in Section 4.2.3 we discuss an example with partial compliance where a subject may receive only part of the intended treatment).

When considering compliance, the two possible compliance actions to two possible experimental group assignments yield $2 \times 2 = 4$ distinct *compliance types*. An individual who adheres to protocol and complies with the experimental regimen to which they were assigned is a *complier* while one who always receives the opposite from what is assigned is known as a *defier*. An *alwaystaker* always takes the treatment regardless of what is assigned while a *nevertaker* never takes the treatment and only experiences the control condition. Just as with the potential outcomes, behavioral traits are deterministic and specific to each individual for the experiment being conducted. Since each individual is observed in only one of the two possible treatments, we are not able to completely determine their compliance type. For example, an individual assigned to treatment who receives treatment may be a complier or an alwaystaker; we don't know which.

More formally, for a sample of n subjects let Z_i again be a random indicator of whether the i^{th} subject is assigned to treatment. The treatment received by subject i, $d_i(z)$, is considered a non-random trait of subject i which is an indicator function of whether they receive the treatment protocol if assigned to z. Each subject i still has two potential outcomes, y_{i1} and y_{i0}, and y_{id} is the response of subject i if they receive d. Thus, the response depends on the treatment received $d_i(Z)$, that is, the response is unrelated to assignment once the received treatment is taken into account. The observed outcome is

$$Y_i = y_{i1}d_i(Z_i) + y_{i0}(1 - d_i(Z_i)). \tag{2.5}$$

As in Equation 2.1, the only source of randomness in Equation 2.5 comes from the assignment to treatment; Y_i is a function of Z_i. The treatment received and response of each of the four compliance types is summarized in Table 2.1.

Compliance Type	Treatment Assigned		Treatment Received $d_i(Z_i)$		Response $Y_i = y_{i1}d_i(Z_i) + y_{i0}(1 - d_i(Z_i))$
complier	if $Z_i = 0$	then	$d_i(0) = 0$	and	$Y_i = y_{i1}(0) + y_{i0}(1 - 0) = y_{i0}$
complier	if $Z_i = 1$	then	$d_i(1) = 1$	and	$Y_i = y_{i1}(1) + y_{i0}(1 - 1) = y_{i1}$
defier	if $Z_i = 0$	then	$d_i(0) = 1$	and	$Y_i = y_{i1}(1) + y_{i0}(1 - 1) = y_{i1}$
defier	if $Z_i = 1$	then	$d_i(1) = 0$	and	$Y_i = y_{i1}(0) + y_{i0}(1 - 0) = y_{i0}$
alwaystaker	if $Z_i = 0$	then	$d_i(0) = 1$	and	$Y_i = y_{i1}(1) + y_{i0}(1 - 1) = y_{i1}$
alwaystaker	if $Z_i = 1$	then	$d_i(1) = 1$	and	$Y_i = y_{i1}(1) + y_{i0}(1 - 1) = y_{i1}$
nevertaker	if $Z_i = 0$	then	$d_i(0) = 0$	and	$Y_i = y_{i1}(0) + y_{i0}(1 - 0) = y_{i0}$
nevertaker	if $Z_i = 1$	then	$d_i(1) = 0$	and	$Y_i = y_{i1}(0) + y_{i0}(1 - 0) = y_{i0}$

Table 2.1: Treatment received and response values as a function of treatment assigned for each of the four compliance types.

We state our earlier comment, that once the treatment received is known the response does not depend on the treatment assigned, as an explicit assumption.

Assumption 2 (Exclusion Restriction)**.** *The outcome only depends on the treatment received, regardless of the treatment assigned. This is often referred to as the "exclusion restriction".*

In most experimental settings, the researchers who administer the treatment have charge over who receives it. Subjects in the control group, typically, cannot be treated. We assume no crossover from control to treatment so that two of the compliance types are not possible.

Assumption 3. *There are no defiers.*

Assumption 4. *There are no alwaystakers.*

Of the four original compliance types only two are left: compliers and nevertakers.

Returning to our aim to localize the measurement of the treatment effect to those treated, we might consider comparing the average response of just the treated to the untreated, or

$$\mathbb{E}(Y \mid d(Z) = 1) - \mathbb{E}(Y \mid d(Z) = 0).$$

This is problematic, however. As explained by Freedman, Pisani, and Purves (1998, p. 4), using this estimate to represent the impact of treatment may overstate the effect as the values

of the potential outcomes may be correlated with the treatment received. For example, in a medical setting where doctors and patients have some understanding of how an individual responds to a certain treatment, those who actually receive the treatment may be those for whom the benefit is greatest while those in the control are the ones for whom the treatment is riskier or less likely to have impact. Instead, we consider the average treatment effect for the treated (*att*), defined earlier as the average treatment effect over the compliers. Returning to Y and Z referring to a subject chosen at random, let the event "complier" occur when the chosen subject is a complier. We define *att* as

$$att \equiv \mathbb{E}(Y \mid Z = 1, \text{complier}) - \mathbb{E}(Y \mid Z = 0, \text{complier}). \tag{2.6}$$

Since there isn't perfect compliance of the subjects, we focus on the group for which compliance is not an issue. The *att* provides an estimate of a treatment effect, albeit for a subset of the population. We estimate *att* by connecting it to *itt*. Starting with Equation 2.3, we expand *itt* by conditioning on the two compliance types:

$$\begin{aligned} itt &= \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) \\ &= [\mathbb{E}(Y \mid Z = 1, \text{complier}) \Pr(\text{complier}) + \mathbb{E}(Y \mid Z = 1, \text{nevertaker}) \Pr(\text{nevertaker})\,] \\ &\quad - [\mathbb{E}(Y \mid Z = 0, \text{complier}) \Pr(\text{complier}) + \mathbb{E}(Y \mid Z = 0, \text{nevertaker}) \Pr(\text{nevertaker})\,] \\ &= [\mathbb{E}(Y \mid Z = 1, \text{complier}) - \mathbb{E}(Y \mid Z = 0, \text{complier})\,] \Pr(\text{complier}) \\ &\quad + [\mathbb{E}(Y \mid Z = 1, \text{nevertaker}) - \mathbb{E}(Y \mid Z = 0, \text{nevertaker})] \Pr(\text{nevertaker}) \end{aligned}$$

The second term on the right hand side is 0 because a nevertaker does not receive the treatment so $d_i(0) = d_i(1) = 0$ and $y_{i\,d_i(1)} = y_{i\,d_i(0)} = y_{i0}$. Thus $\mathbb{E}(Y \mid Z = 1, \text{nevertaker}) - \mathbb{E}(Y \mid Z = 0, \text{nevertaker}) = 0$. This gives

$$itt = [\mathbb{E}(Y \mid Z = 1, \text{complier}) - \mathbb{E}(Y \mid Z = 0, \text{complier})\,] \Pr(\text{complier})$$

Dividing both sides by Pr(complier) yields

$$\frac{itt}{\Pr(\text{complier})} = \mathbb{E}(Y \mid Z = 1, \text{complier}) - \mathbb{E}(Y \mid Z = 0, \text{complier})$$

or, as defined by 2.6,

$$\frac{itt}{\Pr(\text{complier})} = att. \tag{2.7}$$

This representation provides an elegant and intuitive understanding of the relationship between *itt* and *att*. It also brings out another assumption required to estimate *att*.

Assumption 5. *There are compliers among the n experimental subjects.*

Assumption 5 is needed for *att* from Equation 2.7 to be defined. It also makes intuitive sense; for there to be any hope of seeing an effect for the treated, some need to comply with treatment.

2.2 In Experiments with Binary Responses

If the response, Y_i, is binary, indicating whether the subject responds or not, the data observed from the experiment can be summarized in a table such as the one below.

	Assigned Control	Assigned Treatment	
	Received Control	Received Control	Received Treatment
No response	C_0	T_{00}	T_{01}
Responds	C_1	T_{10}	T_{11}
Total	c	t	

Table 2.2: Table of observations with binary responses.

The C_y in Table 2.2 is the total number of subjects assigned to the control with response value y, and T_{yd} is the total assigned to treatment with response y and treatment received d. The total assigned to each experimental condition are c and t with $n = c + t$. For now, we assume c and t are fixed in advance of the experiment.

We may estimate the *itt*, the expected difference in the response rate of those assigned to treatment minus the response rate of those assigned to control, with the observed difference

$$\widehat{itt} = \frac{T_{10} + T_{11}}{t} - \frac{C_1}{c}. \tag{2.8}$$

In estimating *att* from Equation 2.7, the estimated fraction of compliers is simply the percent of subjects in the treatment group who were treated:

$$\widehat{att} = \begin{cases} \dfrac{\dfrac{T_{10} + T_{11}}{t} - \dfrac{C_1}{c}}{\dfrac{T_{01} + T_{11}}{t}} & \text{if } T_{01} + T_{11} > 0 \\ 0 & \text{if } T_{01} + T_{11} = 0, \end{cases} \tag{2.9}$$

If $T_{01} + T_{11} = 0$, that is, no observed compliers, we set $\widehat{att} = 0$ so the estimator has finite expectation. This is different from Assumption 5, that there are some compliers among the the entire group of subjects. Even if the assumption holds, our definition of $\widehat{att}$ returns a finite estimate in the (unlikely) event of all compliers assigned to the control.

2.2.1 Response Types

Our initial presentation of *att* and *itt* in Section 2.1.1 was more general where the potential outcomes could take any real value. The restriction of only two possible outcomes, to respond or not, leads to an elegant representation of the treatment effects. By Assumption 2, the response is determined by the treatment received, $d_i(Z_i)$, so we may characterize the response behavior by whether one was treated or not. Since the received treatment and response are binary there are $2 \times 2 = 4$ *response types*.

always respond (AR) always responds whether treated or not

never respond (NR) never responds whether treated or not

treated respond (TR) if treated responds, if not treated doesn't respond

not treated respond (NTR) if not treated responds, if treated doesn't respond

Listing each separately highlights the plausibility of our final assumption.

Assumption 6 (Monotonicity)**.** *There are no "not treated respond".*

That is, there are no persons who respond if not treated but don't respond if treated. We think it reasonable to assume that if this portion of the population exists it is small enough to be negligible. More generally, the monotonicity assumption applies when the assigned experimental conditions form an ordered set, as do the possible outcomes. The assumption states that for any two treatments s and t such that $s \leq t$, then for every subject i in the sample we have $y_{is} \leq y_{it}$. With only two treatments and two responses such formal handling may seem unnecessary, but it's importance becomes clear in later chapters when we consider experiments with multiple treatments.

Example 2.2.1. We note that the six assumptions hold for field experiments in get-out-the-vote campaigns. Here, the treatment is outreach via face-to-face conversation, phone, mail, email or social media. The response is vote or not. Voters who answer the door when campaigners visit or pick up the phone are the ones who comply with treatment. Assumption 1, or the SUTVA, depends on no spillover effects between those in the treatment and control groups and this is often adhered to by using subjects in separate households and assuming that subjects do not come into contact with others in the experiment. If an individual is not slated to be treated, then there is no avenue for them to receive the treatment (again if no spillover) so Assumption 2, the exclusion restriction and Assumption 3, no defiers, seem reasonable. Assumption 4, no alwaystakers should hold, as voters who the campaign does not attempt to contact should have no way to receive the treatment. Assumption 5, that compliers exist, is seen directly from the table of results. As discussed above, Assumption 6, monotonicity, seems plausible, that there are no individuals who vote if not treated but who vote if treated (or at least if there are any "not treated respond" their numbers are negligible).

2.2.2 Combining Compliance Types and Response Types Leads to Distinct Behavioral Types

Combining the two compliance types with the three response types gives three types of compliers and three types of nevertakers:

complier-always-respond ($comAR$)

complier-never-respond ($comNR$)

complier-if-treated-respond ($comTR$)

nevertaker-always-respond ($nevAR$)

nevertaker-never-respond ($nevNR$)

~~**nevertaker-if-treated-respond** ($nevTR$)~~

The last is crossed out because nevertakers are not treated, and so a nevertaker-if-treated-respond behaves just as a nevertaker-never-respond. Thus every individual in the experiment can be classified as one of five different joint *behavioral types* corresponding with the five different combined outcomes of $(Z_i, d_i(Z_i), Y_i)$ below.

$comAR$**:** if Z_i=0 then $(d_i(Z_i), Y_i) = (0, 1)$, if Z_i=1 then $(d_i(Z_i), Y_i) = (1, 1)$.

$comNR$**:** if Z_i=0 then $(d_i(Z_i), Y_i) = (0, 0)$, if Z_i=1 then $(d_i(Z_i), Y_i) = (1, 0)$.

$comTR$**:** if Z_i=0 then $(d_i(Z_i), Y_i) = (0, 0)$, if Z_i=1 then $(d_i(Z_i), Y_i) = (1, 1)$.

$nevAR$**:** if Z_i=0 then $(d_i(Z_i), Y_i) = (0, 1)$, if Z_i=1 then $(d_i(Z_i), Y_i) = (0, 1)$.

$nevNR$**:** if Z_i=0 then $(d_i(Z_i), Y_i) = (0, 0)$, if Z_i=1 then $(d_i(Z_i), Y_i) = (0, 0)$.

Of the five behavioral types, complier-if-treated-respond is the only one with response affected by the treatment assignment. In the next section we show that estimating the *itt* is equivalent to estimating the proportion of complier-if-treated-respond in the population; in the Neyman-Rubin Causal framework, when outcomes are binary, we don't estimate different treatment effects for a population as much as we estimate the *proportion of the population for which the treatment has an effect*. This is a key contrast between the behavioral-types approach and the commonly applied linear modeling. Instead of a treatment effect as a coefficient of a linear model, it is the frequency of some behavioral type. Under this approach, estimating treatment effects such as *itt* and *att* come directly from estimating the distribution of the behavioral types.

2.2.3 Behavioral Types and the Interpretation of Treatment Effects

The link between the behavioral types and treatment effects of *itt* and *att* can be seen by examining the table of observations. With the compliance and response to each treatment assignment understood for each behavioral type, we can show where each appear in the results. Table 2.3 shows the table of observations again, and just below it, which behavioral types appear in each cell. Connecting the behavioral types to their location in the

	Assignment Control	Assignment Treatment	
	Received Control	Received Control	Treatment
No response	C_0	T_{00}	T_{01}
Responds	C_1	T_{10}	T_{11}
Total	c	t	

	Assignment Control	Assignment Treatment	
	Received Control	Received Control	Treatment
No response	$nevNR$ $comNR$ $comTR$	$nevNR$	$comNR$
Responds	$nevAR$ $comAR$	$nevAR$	$comAR$ $comTR$

Table 2.3: Location of the five behavioral types in the table of observations.

table of observations reveals that C_1 is simply the number of nevertaker-always-respond and complier-always-respond assigned to control. And T_{10} is the number of nevertaker-always-respond that appear in the treatment. Alternatively, from the perspective of parameter estimation, we see that we can estimate the proportion of each behavioral type in the experiment from the table of observations. Thus $C_1/c - T_{10}/t$ estimates the proportion of complier-always-respond. Since c and t are fixed, the table of observations has just four degrees of freedom. We can estimate the fraction of each of the five behavioral types with

$$\widehat{p}_{nevNR} = \frac{T_{00}}{t} \tag{2.10}$$

$$\widehat{p}_{comNR} = \frac{T_{01}}{t} \tag{2.11}$$

$$\widehat{p}_{nevAR} = \frac{T_{10}}{t} \tag{2.12}$$

$$\widehat{p}_{comAR} = \frac{C_1}{c} - \frac{T_{10}}{t} \tag{2.13}$$

$$\widehat{p}_{comTR} = \frac{T_{10} + T_{11}}{t} - \frac{C_1}{c} \tag{2.14}$$

To demonstrate the connection between the proportion of behavioral types and the treatment effects we begin with a bit of notation. Let n_{comAR} be the number of complier-always-respond within the experiment and define the totals for the other behavioral types in the same manner so that $n_{comAR} + ... + n_{nevNR} = n$. Define $p_{comAR} = n_{comAR}/n$, etc. Then $\widehat{p}_{comAR}$ is an unbiased estimator of p_{comAR} as are the other estimates of the proportion of behavioral types. We connect the causal estimates with behavioral types by noting that the estimands of Equations 2.8 and 2.14 have the same formula. That is,

$$\widehat{itt} = \widehat{p}_{comTR}.$$

The observed *itt* is just the estimated fraction of complier-if-treated-respond types. This makes intuitive sense. The intention-to-treat effect is the difference in the response rate of the treatment from that of the control. From Table 2.3 this difference is exactly the proportion of complier-if-treated-respond types as they are the only behavioral type to change response value based on treatment assignment. To confirm this, more formally, from Equation 2.3,

$$itt = \frac{1}{n}\sum_{i=1}^{n} \left[\, \mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0) \,\right].$$

For complier-if-treated-respond, $\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)$ equals 1 as they respond only if assigned to treatment. For the other four behavioral types this difference is 0 so that

$$\begin{aligned} itt &= \frac{1}{n}\#(comTR) \\ &= \frac{n_{comTR}}{n} \\ &= p_{comTR}. \end{aligned}$$

It immediately follows from Equation 2.7 that

$$\begin{aligned} att &= \frac{itt}{\Pr(\text{complier})} \\ &= \frac{p_{comTR}}{p_{comAR} + p_{comNR} + p_{comTR}}, \end{aligned}$$

while from Equation 2.9 we have

$$\widehat{att} = \begin{cases} \dfrac{\widehat{p}_{comTR}}{\widehat{p}_{comAR} + \widehat{p}_{comNR} + \widehat{p}_{comTR}} & \text{if } T_{01} + T_{11} > 0 \\ \\ 0 & \text{if } T_{01} + T_{11} = 0. \end{cases}$$

This insight, that the two causal effects of primary interest are simply proportions of the population, helps clarify the underlying model when calculating the variance of $\widehat{att}$ in the next chapter, and also provides direction when we extend results to experiments with multiple levels of treatment in the Chapters 4 and 5. The tactic of showing where each behavioral type lands in the table of observations is useful to both estimate the proportion of each type and to show their connection to the treatment effects. We reuse this approach a number of times in subsequent chapters to identify the proportions of the behavioral types, which parameterize the underlying data generating process.

2.3 An Alternative Parameterization Based on the Observations

Our model to generate $(C_0, C_1, T_{00}, T_{01}, T_{10}, T_{11})$ is indexed by six parameters: c, t, p_{comAR}, p_{comNR}, p_{comTR} and p_{nevAR}. The proportion p_{nevAR} is omitted because the five proportions must sum to 1. To evaluate the variance of $\widehat{att}$ we present an alternative parameterization based on the observations. Define the proportions from the table of observations as $\widehat{q}_i = C_i/c$ and $\widehat{p}_{ij} = T_{ij}/t$. We may rearrange Equations 2.10 to 2.14 to give:

$$\begin{aligned} \widehat{p}_{comAR} &= \widehat{q}_1 - \widehat{p}_{10} \\ \widehat{p}_{comNR} &= \widehat{p}_{01} \\ \widehat{p}_{comTR} &= \widehat{p}_{10} + \widehat{p}_{11} - \widehat{q}_1 \\ \widehat{p}_{nevAR} &= \widehat{p}_{10}. \end{aligned}$$

These equations describe our estimates for the proportion of behavioral types in terms of the proportions from the table of observations. Alternatively, we can represent the proportions of observations as linear combinations of the behavioral types.

$$\begin{aligned} \widehat{q}_1 &= \widehat{p}_{comAR} + \widehat{p}_{nevAR} \\ \widehat{p}_{01} &= \widehat{p}_{comNR} \\ \widehat{p}_{10} &= \widehat{p}_{nevAR} \end{aligned}$$

$$\widehat{p}_{11} = \widehat{p}_{comAR} + \widehat{p}_{comTR}$$

Here $\widehat{p}_{00} = \widehat{p}_{nevNR}$ is redundant as the four $\widehat{p}_{ij}$ sum to 1. These equations demonstrate the one-to-one relationship between $(\widehat{p}_{comAR}, \widehat{p}_{comNR}, \widehat{p}_{comTR}, \widehat{p}_{nevAR})$ and $(\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ and we may choose to parameterize our model with either

$$(p_{comAR}, p_{comNR}, p_{comTR}, p_{nevAR}) \text{ or } (q_1, p_{01}, p_{10}, p_{11}).$$

Furthermore, under the parameterization $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$, we can write att as a function $h()$ of $\mathbf{p}$ where

$$att = h(\mathbf{p}) = \frac{p_{10} + p_{11} - q_1}{p_{01} + p_{11}}. \tag{2.15}$$

which we estimate with

$$\widehat{att} = \begin{cases} h(\widehat{\mathbf{p}}) = \dfrac{\widehat{p}_{10} + \widehat{p}_{11} - \widehat{q}_1}{\widehat{p}_{01} + \widehat{p}_{11}} & \text{if } T_{01} + T_{11} > 0 \\ 0 & \text{if } T_{01} + T_{11} = 0. \end{cases}$$

In the next chapter, when we examine the properties of $\widehat{att}$, we use parameterization from this section.

2.4 Discussion

In this chapter we review the key concepts of the potential outcomes model focusing on the experimental design with a control group and one treatment, where the subjects may or may not receive the assigned treatment. In our setting, treatment assigned, treatment received and responses are binary variables. We show that in such settings, subjects can be grouped by their compliance and response behaviors to the assigned treatment. As a result, each subject may be classified as one of a finite number of distinct types. We further show that the assumptions typically made in the analysis of experimental data imply restrictions on the number of types. These restrictions reduce the total number to five. The five remaining *behavioral types* describe how an individual receives and responds to the different treatment assignments.

We show that the formulas for the two causal effects of primary interest, the intention-to-treat effect and the average treatment effect for the treated, can be written in terms of the proportions the behavioral types among the experimental subjects. This is a noteworthy shift in how causal effects are often interpreted. The causal effects we estimate do not measure average treatment effects for a population as much as they reflect the *proportion of the population for which the treatment has an effect.*

Furthermore, under the potential outcomes framework, the observed experimental data is generated by a simple model which is completely parameterized by the unknown proportion

of the five behavioral types in the experiment and the mechanism for assigning subjects to treatment (along with known parameters such as the sample sizes of the control and treatment groups). This understanding brings a new perspective, not just for the *itt* or *att*, but for any inferential feature of the experiment, as it may be viewed as a function of the proportion of these behavioral types. We return to this key point throughout the research. This understanding of the data process, where observations are generated by randomly allocating the five behavioral types to the treatment and control, directs how we evaluate the bias and variance of $\widehat{att}$ in the next chapter.

The work of Angrist et al. (1996) describing the average treatment effect for the treated is a useful case study to introduce the concepts of behavioral types. It is a well-cited work with which many readers may be familiar. More importantly, it states clear and explicit assumptions which restrict possible values for the received treatment and response (such as "no defiers" and the exclusion restriction). In Chapter 4 we show that such assumptions are essential to reducing the number of behavioral types to a number which may be identified in the data. In this study with a control and one level of treatment we reduced eight possible types to five. As we show in Chapters 4 and 5, a behavioral-types approach is well suited for analysis of multi-treatment experiments as it distills often complex designs into an estimation problem of a manageable number of types.

Chapter 3

Properties of the Average Treatment Effect for the Treated Estimator with Binary Responses

In most social science randomized controlled studies, the two experimental groups originate from a finite pool of subjects and, conditioning on the study subjects, any variability in the observed outcomes lies in the random assignment to treatment and control. That is, the assignment to treatment is done at random without replacement and, unless there are further assumptions, any inference about the treatment effects is only made within the experimental sample. On the other hand, if the sample is thought to be drawn at random from an infinite superpopulation, then the treatment and control groups are independent of one another, as are the random counts T_{ij} and C_i. This was the assumption made by Angrist and Imbens (1994) though the variance for $\widehat{att}$ for the finite case has been described elsewhere (for a recent work see Sekhon and Shem-Tov, 2017). This distinction, between measurement of a finite *sample average treatment effect for the treated* (*satt*) versus an infinite *population average treatment effect for the treated* (*patt*) was made, generally speaking, as early as Neyman (1923) and the impact of the sampling assumption on inference continues to be studied as in Imbens and Rubin (2015, Ch 6) and Hartman, Grieve, Ramsahai, and Sekhon (2015).

We examine the difference the two modeling assumptions have on the conclusions of the experiment. While the sampling assumptions have no impact on the parameter identification or the estimates of the treatment effects, they do impact the variance of the estimators. We begin with the infinite population assumption and show how the asymptotic variance of $\widehat{att}$ may be found from the delta method. We then address the finite sample case. Next, we conduct simulations to verify that for most sample sizes of interest, the actual variance of $\widehat{att}$ may be well approximated by it's asymptotic limit. we compare the two sampling schemes to show how the variance and standard error estimates for $\widehat{att}$ hinge on the sampling assumptions. However, for GOTV experiments, we show the sampling assumptions likely have

little impact on the conclusions of most studies. Finally, we highlight a simpler approach to estimating treatment effects that introduces a number of ideas, well aligned with behavioral types, which are useful when we extend our results to RCTs with multiple treatments in Chapter 4. While we include many of the calculations within the text, we place the proofs in the final section of the chapter.

3.1 The Bias and Variance of $\widehat{att}$, when Observations Originate from an Infinite Population

To evaluate the bias and variance of $\widehat{att}$ we must first understand the properties of $\widehat{p}$. In this section we begin by describing the underlying data generating process for the the observed results under the infinite population assumptions. We then examine the bias and variance of the parameter estimates, $\widehat{p}$. Finally we determine the bias and asymptotic variance of $\widehat{att}$.

Suppose an experiment has c control and t treatment subjects such that individuals assigned to the two experimental conditions are drawn from an infinite population of possible subjects. The table of observations, when tabulated by the assigned treatment, received treatment and response is

	Assigned Control	Assigned Treatment	
	Received Control	Received Control	Received Treatment
No response	C_0	T_{00}	T_{01}
Responds	C_1	T_{10}	T_{11}
Total	c	t	

where C_y is the total number of subjects assigned to the control with response value y and T_{yd} is the total assigned to control with response y and treatment received d. The total assigned to each treatment condition are c and t with $c + t = n$. Under these sampling assumptions, C_1 follows a binomial(c, q_1) distribution where q_1 equals the fraction who respond if assigned to the control group (that is, the proportion of complier-always-respond and nevertaker-always-respond in the population). Similarly the distribution of $(T_{01}, T_{01}, T_{10}, T_{11})$ is multinomial$(t, p_{00}, p_{01}, p_{10}, p_{11})$, independent of (C_0, C_1), where

p_{00} is the infinite population proportion of nevertaker-never-respond, or p_{nevNR},
p_{01} is the infinite population proportion of complier-never-respond, or p_{comNR},
p_{10} is the infinite population proportion of nevertaker-always-respond, or p_{nevAR},
p_{11} is the infinite population proportion of compliers who either always respond or

respond if treated, or p_{comAR} and p_{comTR}.

Since $p_{00} + p_{01} + p_{10} + p_{11} = 1$ we only concern ourselves with p_{01}, p_{10} and p_{11}. We use this parameterization, $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$, corresponding with the table of observed results as in described Section 2.3, for the remainder of this chapter.

Again, we calculate the expected value and variance of $\widehat{\mathbf{p}}$ under the infinite population assumptions before examining the properties $\widehat{att}$. However, from the expectation properties of the binomial and multinomial distributions, we see that $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ is unbiased. That is, under the infinite population assumption, which we denote with *pop*,

$$\mathbb{E}_{pop}(\widehat{\mathbf{p}}) = \mathbf{p} = (q_1, p_{01}, p_{10}, p_{11}).$$

And we evaluate the variance of $\widehat{\mathbf{p}}$ in the next section.

3.1.1 Variance of $\widehat{\mathbf{p}}$ under the Infinite Population Assumption

Under the infinite sampling model,

$$\begin{aligned}\mathrm{Var}_{pop}(\widehat{\mathbf{p}}) &= \mathrm{Var}_{pop}(\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11}) \\ &= \mathrm{Var}_{pop}\left(\frac{C_1}{c}, \frac{T_{01}}{t}, \frac{T_{10}}{t}, \frac{T_{11}}{t}\right).\end{aligned}$$

Since C_1 is binomial(c, q_1) and independent of $(T_{01}, T_{01}, T_{10}, T_{11})$, which is multinomial$(t, p_{00}, p_{01}, p_{10}, p_{11})$, the covariance matrix of $\widehat{\mathbf{p}}$ is readily available from the properties of the binomial and multinomial distributions.

$$\mathrm{Var}_{pop}(\widehat{\mathbf{p}}) = \begin{pmatrix} \frac{q_1(1-q_1)}{c} & 0 & 0 & 0 \\ 0 & \frac{p_{01}(1-p_{01})}{t} & \frac{-p_{01}p_{10}}{t} & \frac{-p_{01}p_{11}}{t} \\ 0 & \frac{-p_{01}p_{10}}{t} & \frac{p_{10}(1-p_{10})}{t} & \frac{-p_{10}p_{11}}{t} \\ 0 & \frac{-p_{01}p_{11}}{t} & \frac{-p_{10}p_{11}}{t} & \frac{p_{11}(1-p_{11})}{t} \end{pmatrix}$$

The zeros derive from the independence of C_i and T_{ij} and thus the independence of $\widehat{q}_1$ and $\widehat{p}_{ij}$. If we denote the fraction of the subjects assigned to treatment as $\beta \equiv t/n$, then

$$\mathrm{Var}_{pop}(\widehat{\mathbf{p}}) = \frac{1}{n}\left\{\frac{1}{\beta}\begin{pmatrix} \frac{\beta}{(1-\beta)}q_1(1-q_1) & 0 & 0 & 0 \\ 0 & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ 0 & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ 0 & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix}\right\} \quad (3.1)$$

$$\doteq \frac{1}{n}\mathbf{\Sigma_{pop}} \tag{3.2}$$

where $\mathbf{\Sigma_{pop}}$ is the matrix within the curly braces of Equation 3.1. With the mean and covariance of $\widehat{\mathbf{p}}$ spelled out, we proceed to evaluating the bias and variance of $\widehat{att}$.

3.1.2 Bias of $\widehat{att}$ under the Infinite Population Assumption

Unlike $\widehat{\mathbf{p}}$, the estimate $\widehat{att}$ is biased. To describe the limiting behavior of the bias we imagine an infinite sequence of experiments indexed by the total number of subjects, n, where the fraction assigned to treatment is approximately equal as n increases. We assume response values are binary to give the table of observations, $(C_0, C_1, T_{01}, T_{01}, T_{10}, T_{11})$, and suppose Assumptions 1 - 6 of Chapter 2, which lead to the five distinct behavioral types and allow us to identify *att*.

Proposition 3.1.1. *Suppose:*

i *The total assigned to treatment, $t(n)$ is such that $t(n)/n \to \beta$ and the total assigned to control is $c(n) = n - t(n)$.*

ii *Assume infinite population sampling, that is, C_1 is binomial($c(n)$, q_1) and $(T_{01}, T_{01}, T_{10}, T_{11})$ is multinomial($t(n), p_{00}, p_{01}, p_{10}, p_{11}$).*

iii *The average treatment effect for the treated is defined as*

$$att = \mathrm{h}(\boldsymbol{p}) = \frac{p_{10} + p_{11} - q_1}{p_{01} + p_{11}},$$

which is estimated by

$$\widehat{att} = \begin{cases} \mathrm{h}(\widehat{\boldsymbol{p}}) = \dfrac{\dfrac{T_{10} + T_{11}}{t(n)} - \dfrac{C_1}{c(n)}}{\dfrac{T_{01} + T_{11}}{t(n)}} & \text{if } T_{01} + T_{11} > 0 \\ 0 & \text{if } T_{01} + T_{11} = 0. \end{cases}$$

Then the order of the bias is $1/n$, or

$$\mathbb{E}_{pop}(\widehat{att}) = att + O(\tfrac{1}{n}).$$

Note: though not stated explicitly, each random quantity above, such as $\widehat{att}$, C_1, etc., is indexed by n. We remove the index for ease of notation.

We leave the proof to the end of the chapter.

3.1.3 Variance of $\widehat{att}$ under the Infinite Population Assumption

Evaluating the variance of $\widehat{att}$ directly is challenging, as it is a ratio of dependent random variables. Instead we may approximate the variance via the delta method (see Bishop, Fienberg, and Holland, 1975, p. 493). From equation 2.15 we may write att as a multivariate function, h() of $\mathbf{p}$, that is,

$$att = \mathrm{h}(\mathbf{p}) = \frac{p_{10} + p_{11} - q_1}{p_{01} + p_{11}}$$

To find the asymptotic variance we apply the delta method for convergence in distribution and obtain the following result.

Proposition 3.1.2. *Suppose the assumptions of Proposition 3.1.1 then*

$$\sqrt{n}(\widehat{att} - att) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\boldsymbol{p})' \, \mathbf{\Sigma_{pop}} \, \nabla \mathrm{h}(\boldsymbol{p})).$$

We postpone the proof to section 3.8. Under the infinite population assumption, Proposition 3.1.2 gives the asymptotic variance of $\widehat{att}$ but doesn't inform how applicable the variance approximation will be for fixed n. In Section 3.3 we show that for large samples, or at least for the sample sizes of many published social science experiments, they asymptotic variance is a useful approximation to the actual variance, that is,

$$\mathrm{Var}_{pop}(\widehat{att}) \approx \nabla \mathrm{h}(\mathbf{p})' \, \frac{1}{n} \mathbf{\Sigma_{pop}} \nabla \mathrm{h}(\mathbf{p}). \tag{3.3}$$

3.2 The Bias and Variance of $\widehat{att}$, when Observations Originate from a Finite Sample

In the same manner of Section 3.1, we evaluate the bias and variance of $\widehat{\mathbf{p}}$ and then turn to the properties of $\widehat{att}$. First, we describe the data generating process for the table of observations, $(C_0, C_1, T_{00}, T_{01}, T_{10}, T_{11})$.

Under the finite sample assumption, the C_i and T_{ij} result from how the five behavioral types are randomly split between the treatment and control. In this setting, the number of each behavioral type found in each experimental condition follows a multivariate hypergeometric distribution (see Johnson, Kotz, and Balakrishnan, p. 171). This is an extension of the more common hypergeometric distribution, often described as imagining an urn with n objects of which n_1 are red, n_2 are green, $n_1 + n_2 = n$, and objects are indistinguishable except for color. When drawing t times without replacement from the urn, the number of objects drawn of a certain color is described by a distribution depending on n_1, n_2 and t. In

the multivariate hypergeometric distribution, the description is similar but there are three or more different colors for the objects.

In our experiment, with binary outcomes and possible noncompliance of treatment, there are n total subjects and five different behavioral types. Each behavioral type is analogous to a different color. We set $n_{type\,i}$ as the number of $type\,i$ individuals in the finite sample. We have

$$n = n_{comAR} + n_{comNR} + n_{comTR} + n_{nevAR} + n_{nevNR}.$$

Of these, t individuals are chosen without replacement and assigned to the treatment group while $c = n - t$ remain in control. Let T_{comAR} be the random variable representing the number of complier-always-respond who are assigned to treatment so $C_{comAR} = n_{comAR} - T_{comAR}$. Then T_{comAR} and the four other $T_{type\,i}$ each follow a multivariate hypergeometric distribution where, if $samp$ indicates finite sampling,

$$\begin{aligned}
\mathbb{E}_{samp}(T_{type\,i}) &= t\ \frac{n_{type\,i}}{n}, \\
\mathrm{Var}_{samp}(T_{type\,i}) &= \frac{t\,c\,n_{type\,i}(n - n_{type\,i})}{n^2(n-1)}\ \text{, and} \\
\mathrm{Cov}_{samp}(T_{type\,i}, T_{type\,j}) &= -\frac{t\,c\,n_{type\,i}\,n_{type\,j}}{n^2(n-1)} \quad \text{for } i \neq j.
\end{aligned}$$

For the $C_{type\,i}$ representing the number in the control group, since $C_{type\,i} = n_{type\,i} - T_{type\,i}$ we can calculate the expectation and covariance in terms of $T_{type\,i}$ so that

$$\begin{aligned}
\mathbb{E}_{samp}(C_{type\,i}) &= c\ \frac{n_{type\,i}}{n}, \\
\mathrm{Var}_{samp}(C_{type\,i}) &= \mathrm{Var}_{samp}(T_{type\,i}) = \frac{t\,c\,n_{type\,i}(n - n_{type\,i})}{n^2(n-1)}\ \text{, and} \\
\mathrm{Cov}_{samp}(C_{type\,i}, C_{type\,j}) &= \mathrm{Cov}_{samp}(T_{type\,i}, T_{type\,j}) = -\frac{t\,c\,n_{type\,i}\,n_{type\,j}}{n^2(n-1)} \quad \text{for } i \neq j.
\end{aligned}$$

Furthermore, $T_{type\,i}$ and $C_{type\,j}$ have covariance

$$\mathrm{Cov}_{samp}(T_{type\,i}, C_{type\,i}) = -\mathrm{Var}_{samp}(T_{type\,i}) = -\frac{t\,c\,n_{type\,i}(n - n_{type\,i})}{n^2(n-1)}\ \text{, and} \tag{3.4}$$

$$\mathrm{Cov}_{samp}(T_{type\,i}, C_{type\,j}) = -\mathrm{Cov}_{samp}(T_{type\,i}, T_{type\,j}) = \frac{t\,c\,n_{type\,i}\,n_{type\,j}}{n^2(n-1)} \quad \text{for } i \neq j. \tag{3.5}$$

From the table of observations, as seen in Table 2.3, we are not able to identify the $T_{type\,i}$ and $C_{type\,j}$ for each behavioral type. However, along the lines of the reparameterization in Section 2.3, T_{ij} and C_i may be written as the combinations of $T_{type\,i}$ and $C_{type\,j}$ as follows.

$$C_0 = C_{nevNR} + C_{comNR} + C_{comTR} \tag{3.6}$$

$$C_1 = C_{nevAR} + C_{comAR} \tag{3.7}$$

$$T_{00} = T_{nevNR} \tag{3.8}$$
$$T_{01} = T_{comNR} \tag{3.9}$$
$$T_{10} = T_{nevAR} \tag{3.10}$$
$$T_{11} = T_{comAR} + T_{comTR} \tag{3.11}$$

With the connection between the table of observations and the behavioral types in each cell delineated, we can solve for the expectations and covariance structure among C_i and T_{ij}.

Thus, under the finite sample assumptions, the model is parameterized by n_{nevNR}, n_{nevAR}, n_{comNR}, n_{nevAR}, n_{comTR} and t. We write

$$p_{nevNR}(n) = \frac{n_{nevNR}}{n}, \quad p_{nevNR}(n) = \frac{n_{nevNR}}{n},$$

and so forth, as the proportion for the behavioral types. Here, in contrast to the infinite sampling model, each proportion is a multiple of $1/n$ and is not fixed as n increases. This distinction becomes important when we explore the limiting behavior of the finite sample model as $n \to \infty$. However, for ease of notation, we drop the index by n, and write p_{nevNR} for $p_{nevNR}(n)$, etc. for the remainder of the section, and then address the issue more formally when we discuss the asymptotic distribution of $\widehat{att}$ in Propositions 3.2.1 and 3.2.2. With these proportions we may parameterize the model with p_{nevAR}, p_{comNR}, p_{nevAR}, p_{comTR}, c and t. For convenience, we choose the proportion parameterization of $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$ from Section 2.3, which is estimated by $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$.

Using Equations 3.6 - 3.11 we may evaluate the bias and variance of $\widehat{\mathbf{p}}$. Since each $C_{type\,i}/c$ and $T_{type\,i}/t$ are unbiased estimators for $p_{type\,i}$, $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ is an unbiased estimator for $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$. We determine the variance of $\widehat{\mathbf{p}}$ in the next section.

3.2.1 Variance of $\widehat{\mathbf{p}}$ under the Finite Population Assumption

The relationships in Equations 3.6 to 3.11 are useful for calculating the covariances among $\widehat{p}_{ij}$ and $\widehat{q}_i$. Here, the parameters p_{ij} and q_i come from the proportions of behavioral types in the sample and are multiples of $1/n$. The expectations of $\widehat{p}_{10}$ and $\widehat{p}_{11}$ are $p_{10} = n_{nevAR}/n$ and $p_{11} = (n_{comAR} + n_{comTR})/n$, and the two have covariance

$$\begin{aligned}
\mathrm{Cov}_{samp}(\widehat{p}_{10}\ ,\ \widehat{p}_{11}) &= \frac{1}{t^2}\mathrm{Cov}_{samp}(T_{10}\ ,\ T_{11}) \\
&= \frac{1}{t^2}\mathrm{Cov}_{samp}(T_{nevAR},\ T_{comAR} + T_{comTR}) \\
&= \frac{1}{t^2}\left[\mathrm{Cov}_{samp}(T_{nevAR},\ T_{comAR}) + \mathrm{Cov}_{samp}(T_{nevAR},\ T_{comTR})\right] \\
&= \frac{1}{t^2}\left[-\frac{t\,c\,n_{nevAR}\,n_{comAR}}{n^2(n-1)}) + -\frac{t\,c\,n_{nevAR}\,n_{comTR}}{n^2(n-1)}\right]
\end{aligned}$$

$$= -\frac{c\, n_{nevAR}}{tn^2(n-1)} \left[n_{comAR} + n_{comTR}\right]$$

$$= -\frac{c}{t(n-1)} \left[\frac{n_{nevAR}}{n}\right] \left[\frac{n_{comAR} + n_{comTR}}{n}\right]$$

$$= -\frac{c}{t(n-1)} p_{10}p_{11}\,.$$

This amount equals the covariance if sampling from an infinite population multiplied by a factor of $\frac{c}{n-1}$. This also holds for the covariance among the other $\widehat{p}_{ij}$ while the covariance between $\widehat{q}_i$ and $\widehat{p}_{ij}$ may be solved using Equations 3.4 and 3.5. The resulting covariance matrix of $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$, under finite sampling assumptions is

$$\mathrm{Var}_{samp}(\widehat{\mathbf{p}}) = \begin{pmatrix} \frac{tq_1(1-q_1)}{c(n-1)} & \frac{q_1p_{01}}{n-1} & \frac{-p_{10}(1-q_1)}{n-1} & \frac{p_{10}-q_1(1-p_{11})}{n-1} \\ \frac{q_1p_{01}}{n-1} & \frac{cp_{01}(1-p_{01})}{t(n-1)} & \frac{-cp_{01}p_{10}}{t(n-1)} & \frac{-cp_{01}p_{11}}{t(n-1)} \\ \frac{-p_{10}(1-q_1)}{n-1} & \frac{-cp_{01}p_{10}}{t(n-1)} & \frac{cp_{10}(1-p_{10})}{t(n-1)} & \frac{-cp_{10}p_{11}}{t(n-1)} \\ \frac{p_{10}-q_1(1-p_{11})}{n-1} & \frac{-cp_{01}p_{11}}{t(n-1)} & \frac{-cp_{10}p_{11}}{t(n-1)} & \frac{cp_{11}(1-p_{11})}{t(n-1)} \end{pmatrix}$$

$$= \frac{1}{n-1} \begin{pmatrix} \frac{tq_1(1-q_1)}{c} & q_1p_{01} & -p_{10}(1-q_1) & p_{10}-q_1(1-p_{11}) \\ q_1p_{01} & \frac{cp_{01}(1-p_{01})}{t} & \frac{-cp_{01}p_{10}}{t} & \frac{-cp_{01}p_{11}}{t} \\ -p_{10}(1-q_1) & \frac{-cp_{01}p_{10}}{t} & \frac{cp_{10}(1-p_{10})}{t} & \frac{-cp_{10}p_{11}}{t} \\ p_{10}-q_1(1-p_{11}) & \frac{-cp_{01}p_{11}}{t} & \frac{-cp_{10}p_{11}}{t} & \frac{cp_{11}(1-p_{11})}{t} \end{pmatrix}$$

. (3.12)

As before with $\beta = t/n$ and $1-\beta = c/n$ this becomes

$$\mathrm{Var}_{samp}(\widehat{\mathbf{p}}) = \frac{1}{n-1} \begin{pmatrix} \frac{\beta q_1(1-q_1)}{(1-\beta)} & q_1p_{01} & -p_{10}(1-q_1) & p_{10}-q_1(1-p_{11}) \\ q_1p_{01} & \frac{(1-\beta)p_{01}(1-p_{01})}{\beta} & \frac{-(1-\beta)p_{01}p_{10}}{\beta} & \frac{-(1-\beta)p_{01}p_{11}}{\beta} \\ -p_{10}(1-q_1) & \frac{-(1-\beta)p_{01}p_{10}}{\beta} & \frac{(1-\beta)p_{10}(1-p_{10})}{\beta} & \frac{-(1-\beta)p_{10}p_{11}}{\beta} \\ p_{10}-q_1(1-p_{11}) & \frac{-(1-\beta)p_{01}p_{11}}{\beta} & \frac{-(1-\beta)p_{10}p_{11}}{\beta} & \frac{(1-\beta)p_{11}(1-p_{11})}{\beta} \end{pmatrix}$$

$$= \frac{1}{n-1}\frac{(1-\beta)}{\beta}\begin{pmatrix} \frac{\beta^2 q_1(1-q_1)}{(1-\beta)^2} & \frac{\beta q_1 p_{01}}{1-\beta} & \frac{-\beta p_{10}-q_1(1-p_{11})}{1-\beta} & \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} \\ \frac{\beta q_1 p_{01}}{1-\beta} & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ \frac{-\beta p_{10}(1-q_1)}{1-\beta} & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix}$$

$$= \frac{1}{n}\left\{ \frac{n}{n-1}\frac{(1-\beta)}{\beta}\begin{pmatrix} \frac{\beta^2 q_1(1-q_1)}{(1-\beta)^2} & \frac{\beta q_1 p_{01}}{1-\beta} & \frac{-\beta p_{10}(1-q_1)}{1-\beta} & \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} \\ \frac{\beta q_1 p_{01}}{1-\beta} & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ \frac{-\beta p_{10}(1-q_1)}{1-\beta} & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix} \right\} \tag{3.13}$$

$$\doteq \frac{1}{n}\mathbf{\Sigma_{samp}}, \tag{3.14}$$

where $\mathbf{\Sigma_{samp}}$ denotes the matrix within the curly braces of Equation 3.13. Notice that except for the first row and column, the other nine entries of the matrix equal of $\mathbf{\Sigma_{pop}}$ multiplied by a factor of $\frac{n}{n-1}(1-\beta)$. We further explore this connection in Section 3.4.

3.2.2 Bias of $\widehat{att}$ under the Finite Population Assumption

Similar to the findings of Proposition 3.1.1, the estimate $\widehat{att}$ is biased. Again, we assume response values are binary to give the table of observations $(C_0, C_1, T_{00}, T_{01}, T_{10}, T_{11})$, and suppose Assumptions 1 - 6 of Chapter 2, which lead to the five distinct behavioral types and allow us to identify *att*. As discussed at the beginning of Section 3.2, in the finite setting, the parameters are not fixed as they must be multiples of $1/n$. This requires additional specifications on the parameters, $\mathbf{p}(n)$, which must be indexed by n.

Proposition 3.2.1. *Suppose:*

i The total assigned to treatment, $t(n)$ is such that $t(n)/n \to \beta$ and the total assigned to control is $c(n) = n - t(n)$.

ii Assume finite population sampling as described in Section 3.2. That is, the five different behavioral types are randomly allocated without replacement. This results in the counts $(C_0, C_1, T_{00}, T_{01}, T_{10}, T_{11})$ found via the relationships in Equations 3.6 - 3.11.

iii *The proportion parameters for each finite n, $\boldsymbol{p}(n) = (q_1(n), p_{01}(n), p_{10}(n), p_{11}(n))$, converge to the limiting vector $\boldsymbol{p}^\infty = (q_1^\infty, p_{01}^\infty, p_{10}^\infty, p_{11}^\infty)$ as follows:*

$$\begin{aligned} q_1(n) &= q_1^\infty + O(1/n) \\ p_{01}(n) &= p_{01}^\infty + O(1/n) \\ p_{10}(n) &= p_{10}^\infty + O(1/n) \\ p_{11}(n)) &= p_{11}^\infty + O(1/n) \end{aligned}$$

iv *The average treatment effect for the treated is defined as*

$$att(n) = \mathrm{h}(\boldsymbol{p}(n)) = \frac{p_{10}(n) + p_{11}(n) - q_1(n)}{p_{01}(n) + p_{11}(n)},$$

which is estimated by

$$\widehat{att} = \begin{cases} \mathrm{h}(\widehat{\boldsymbol{p}}) = \dfrac{\dfrac{T_{10} + T_{11}}{t(n)} - \dfrac{C_1}{c(n)}}{\dfrac{T_{01} + T_{11}}{t(n)}} & \text{if } T_{01} + T_{11} > 0 \\ 0 & \text{if } T_{01} + T_{11} = 0. \end{cases}$$

Then the order of the bias is $1/n$, or

$$\mathbb{E}_{samp}(\widehat{att}) = att(n) + O(\tfrac{1}{n}).$$

Again we postpone the proof to the end of the chapter. The bias from Propositions 3.1.1 and 3.2.1 are of the same order though the underlying distributions, and the methods needed to prove the results differ.

3.2.3 Variance of $\widehat{att}$ under the Finite Population Assumption

The challenges described in Section 3.1.3 to directly evaluate the variance of $\widehat{att}$ still apply in the finite sample case. Furthermore, the proof of Proposition 3.1.2, which establishes the asymptotic normality of $\widehat{att}$ under infinite sampling, required that $\widehat{\mathbf{p}}$ converge to a normal distribution. That condition does not hold, in general, for finite sampling. However, in a recent work, Li and Ding (2017) show that many treatment effect estimates, under finite sampling, are asymptotically normal. We use their findings, which we describe in the final section of this chapter, to show this holds for $\widehat{att}$ as well.

Proposition 3.2.2. *Suppose the assumptions of Proposition 3.2.1 then*

$$\sqrt{n}(\widehat{att} - att) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\boldsymbol{p})' \, \boldsymbol{\Sigma}_{\mathbf{samp}} \, \nabla \mathrm{h}(\boldsymbol{p})).$$

Proposition 3.2.2 gives the limiting variance of $\widehat{att}$ under the finite population assumption. For large n, as we show in the next section, we have

$$\mathrm{Var}_{samp}(\widehat{att}) \approx \nabla \mathrm{h}(\mathbf{p})' \frac{1}{n} \mathbf{\Sigma_{samp}} \nabla \mathrm{h}(\mathbf{p}). \tag{3.15}$$

3.3 Accuracy of Variance Approximations to $\widehat{att}$

The asymptotic variances of Propositions 3.1.2 and 3.2.2 may not be adequate for many social science field experiments with sample sizes in the thousands. To find how large of a sample is needed to be be able to use the approximation we carry out a simulation study. We choose a large number of representative points of the parameter space and simulate $\widehat{att}$, using both sampling assumptions, to obtain their variance and compare to the approximations.

First we create an evenly spaced grid to span the possible parameter space by returning to our original parameterization of p_{comAR}, p_{comNR}, p_{comTR}, p_{nevAR} and β ($\beta = t/n$, the fraction assigned to treatment). With five dimensions the total points of the grid may grow prohibitively large. To reduce this, we take advantage of constraints to limit the parameter space. Each of the five values are between 0 and 1 and satisfy

$$0 \leq p_{comTR} + p_{comAR} + p_{comNR} + p_{nevAR} \leq 1$$

We also restrict the parameters to ranges usually found in social science experiments so that the grid is computed as follows:

- $0 \leq p_{comTR} \leq .20$ by increments of .001,
- $.01 \leq p_{comAR} \leq .70$ by increments of .01,
- $.01 \leq p_{comNR} \leq .70$ by increments of .01,
- $.01 \leq p_{nevAR} \leq .80$ by increments of .01,
- $.05 \leq \beta \leq .90$ by increments of .05,

where p_{comTR} is examined more closely as it is the main behavioral type of interest. Additionally we restrict the total fraction of compliers by

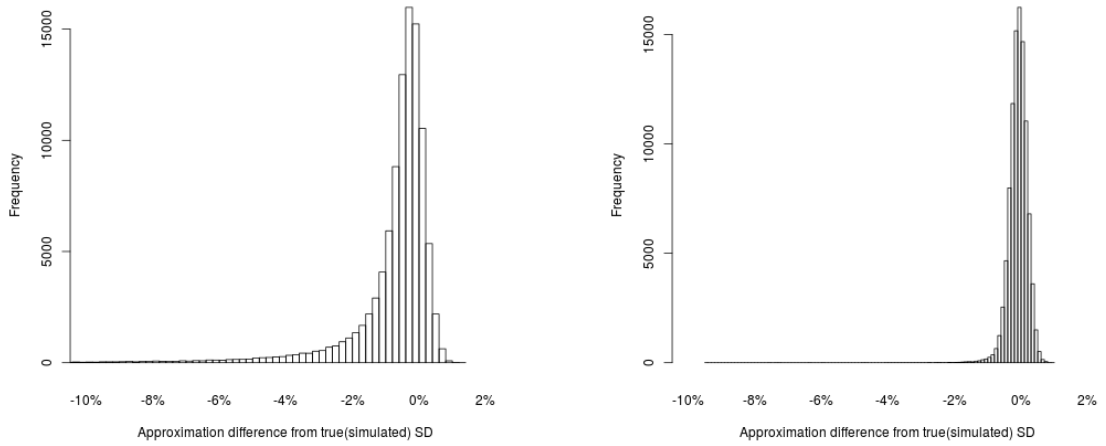
$$p_{comTR} + p_{comAR} + p_{comNR} \leq .95,$$

as our concern is with experiments with noncompliance. In total, this yields a grid of over 400 million points. For each of these points we vary n by two possible sample sizes: 1,000 and 10,000. We shall refer to this grid again in Section 3.4

To make our analysis more manageable we randomly select 100,000 of the 400 million points of the parameter grid. Under the infinite population assumptions, for each parameter

point we obtain the "true" value of the variance by simulating the experiment one hundred thousand times, and obtaining one hundred thousand observed values for $\widehat{att}$. We use these to compute the standard deviation at the point. This is compared to the approximate standard deviation using the asymptotic amount from Proposition 3.1.2.

Figure 3.1 shows a histogram of the difference between the approximation and the true value found from simulation. In (a), where n is 1,000, the histogram has a mode at zero with a left tail. While the approximation of $\text{SD}_{pop}(\widehat{att})$ tends to be within a few percentage points of the true value, for some parameter points it may understate the SD by 5–10%. In (b), where the parameter values are the same but n is increased 10,000, the histogram is centered around zero the approximations are largely within 1% of the true value. Though difficult to see in (b) there are a handful of points where the the approximation is smaller by more than 5%. This occurs when β is 0.05, that is this is smallest treatment group on the grid and the fraction of compliers is very small, less than 6%. These extreme parameter points are rarely encountered in social science field experiments and the approximation seems adequate for sample sizes of more than 10,000.

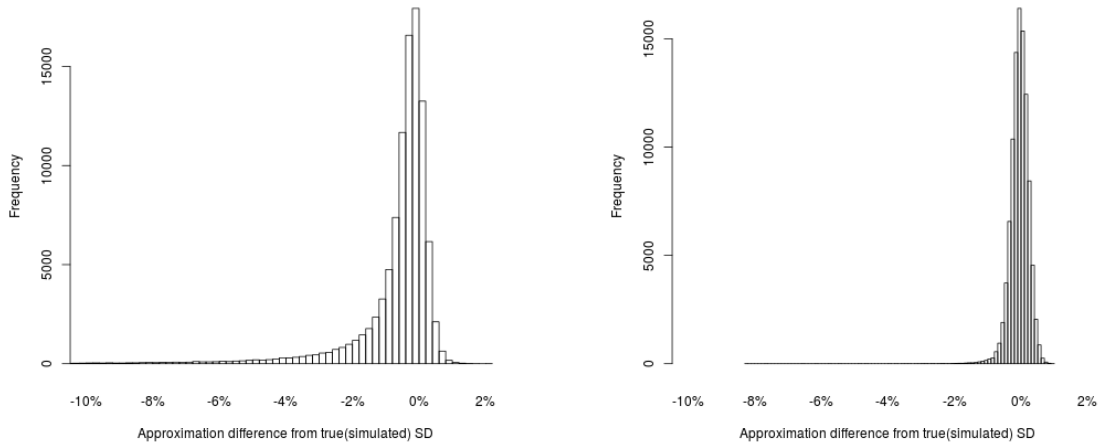


(a) Sample size of 1,000

(b) Sample size of 10,000

Figure 3.1: Difference between the approximation of $\text{SD}_{pop}(\widehat{att})$ and the value found via simulation when the sample size is 1,000 or 10,000. Each value of the histogram represents a different point of the parameter space. 100,000 of the the over 400 million points of the parameter grid have been randomly selected for the plot.

In Figure 3.2 we see nearly identical histograms for the finite sample case and we draw the same conclusions: using the asymptotic variance to approximate the standard deviation is well-suited to sample sizes of more than 10,000 and may be a good approximation at samples sizes of 1,000.



(a) Sample size of 1,000 (b) Sample size of 10,000

Figure 3.2: Difference between the approximation of $\text{SD}_{samp}(\widehat{att})$ and the value found via simulation when the sample size is 1,000 or 10,000. Each value of the histogram represents a different point of the parameter space.

3.4 Comparing the Asymptotic Variances of the Infinite Population and Finite Sample Assumptions

To understand the connection between the asymptotic variances of $\widehat{att}$ under the infinite and finite sampling, we must first understand the connection between the variance matrices for $\widehat{\mathbf{p}}$ under the different assumptions. For ease of notation, we shall refer to the asymptotic variances as $\text{Var}_{pop}(\widehat{att})$ and $\text{Var}_{samp}(\widehat{att})$ as found in Propositions 3.1.2 and 3.2.2, though we recognize for any finite sample this will not represent the true variance of $\widehat{att}$. To characterize the relationship between the two, starting with Equation 3.13,

$$\mathbf{\Sigma_{samp}} = \frac{n}{n-1}\frac{(1-\beta)}{\beta}\begin{pmatrix} \frac{\beta^2 q_1(1-q_1)}{(1-\beta)^2} & \frac{\beta q_1 p_{01}}{1-\beta} & \frac{-\beta p_{10}(1-q_1)}{1-\beta} & \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} \\ \frac{\beta q_1 p_{01}}{1-\beta} & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ \frac{-\beta p_{10}(1-q_1)}{1-\beta} & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{\mathbf{samp}} = \tfrac{n}{n-1}\tfrac{(1-\beta)}{\beta}\left\{\beta\boldsymbol{\Sigma}_{\mathbf{pop}} + \begin{pmatrix} \frac{\beta^2 q_1(1-q_1)}{(1-\beta)^2} - \frac{\beta q_1(1-q_1)}{(1-\beta)} & \frac{\beta q_1 p_{01}}{1-\beta} & \frac{-\beta p_{10}(1-q_1)}{1-\beta} & \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} \\ \frac{\beta q_1 p_{01}}{1-\beta} & 0 & 0 & 0 \\ \frac{-\beta p_{10}(1-q_1)}{1-\beta} & 0 & 0 & 0 \\ \frac{\beta(p_{10}-q_1(1-p_{11}))}{1-\beta} & 0 & 0 & 0 \end{pmatrix}\right\}$$

bringing $\frac{(1-\beta)}{\beta}$ into the brackets gives

$$= \tfrac{n}{n-1}\left\{(1-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \begin{pmatrix} \frac{\beta q_1(1-q_1)}{(1-\beta)} - \frac{q_1(1-q_1)}{1} & q_1 p_{01} & -p_{10}(1-q_1) & p_{10}-q_1(1-p_{11}) \\ q_1 p_{01} & 0 & 0 & 0 \\ -p_{10}(1-q_1) & 0 & 0 & 0 \\ p_{10}-q_1(1-p_{11}) & 0 & 0 & 0 \end{pmatrix}\right\}.$$

The $(1,1)$ entry of the right matrix is simplified to

$$\boldsymbol{\Sigma}_{\mathbf{samp}} = \tfrac{n}{n-1}\left\{(1-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \begin{pmatrix} \frac{(2\beta-1)}{(1-\beta)} q_1(1-q_1) & q_1 p_{01} & -p_{10}(1-q_1) & p_{10}-q_1(1-p_{11}) \\ q_1 p_{01} & 0 & 0 & 0 \\ -p_{10}(1-q_1) & 0 & 0 & 0 \\ p_{10}-q_1(1-p_{11}) & 0 & 0 & 0 \end{pmatrix}\right\}.$$

And if we denote the right matrix to $\boldsymbol{\Sigma}_{\mathbf{gap}}$, as it represents the "gap" between $\boldsymbol{\Sigma}_{\mathbf{samp}}$ and $(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}}$, we have

$$\boldsymbol{\Sigma}_{\mathbf{samp}} = \frac{n}{n-1}\left\{(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\}$$

$$\boldsymbol{\Sigma}_{\mathbf{samp}} = (\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}} + \frac{1}{n-1}\left\{(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\}. \tag{3.16}$$

This is the key relation which links the $\widehat{att}$ variances under the two assumptions. We compare the asymptotic variances by substituting 3.16 into the formula for $\mathrm{Var}_{samp}(\widehat{att})$ from Proposition 3.2.2.

$$\mathrm{Var}_{samp}(\widehat{att}) = \nabla \mathrm{h}(\mathbf{p})' \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{samp}} \nabla \mathrm{h}(\mathbf{p})$$

$$
\begin{aligned}
&= \nabla \mathrm{h}(\mathbf{p})' \frac{1}{n}\left[(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}} + \frac{1}{n-1}\left\{(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\}\right] \nabla \mathrm{h}(\mathbf{p}) \\
&= \frac{1}{n}\nabla \mathrm{h}(\mathbf{p})' (1-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} \nabla \mathrm{h}(\mathbf{p}) + \frac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \boldsymbol{\Sigma}_{\mathbf{gap}} \nabla \mathrm{h}(\mathbf{p}) \\
&\quad + \frac{1}{n(n-1)}\nabla \mathrm{h}(\mathbf{p})' \left\{(1-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\} \nabla \mathrm{h}(\mathbf{p}) \\
&= (1-\beta)\mathrm{Var}_{pop}(\widehat{att}) + \frac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \boldsymbol{\Sigma}_{\mathbf{gap}} \nabla \mathrm{h}(\mathbf{p}) + O(\tfrac{1}{n(n-1)})
\end{aligned}
$$

$$
\mathrm{Var}_{samp}(\widehat{att}) = (1-\beta)\mathrm{Var}_{pop}(\widehat{att}) + \frac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \boldsymbol{\Sigma}_{\mathbf{gap}} \nabla \mathrm{h}(\mathbf{p}) + O(n^{-2}) \tag{3.17}
$$

Equation 3.17 shows that $\mathrm{Var}_{samp}(\widehat{att})$ can be split into three parts. The first term is a fraction of $\mathrm{Var}_{pop}(\widehat{att})$ which decreases as β, which is similar to a finite population correction factor, increases. The second term is based on $\boldsymbol{\Sigma}_{\mathbf{gap}}$ which is determined by the covariance between the observations in the control with the observations in the treatment. An inspection of $\nabla \mathrm{h}(\mathbf{p})$ and $\boldsymbol{\Sigma}_{\mathbf{gap}}$ shows that, for values of β greater than 0.5, as β increases the second term is positive and tend to larger numbers, perhaps offsetting the decrease in size of the first term. The third term becomes less important as n increases in size but may have some degree of impact if n is small.

This is a familiar form which we see in the much simpler comparison of treatment and control means seen in introductory statistics courses. For example, Freedman et al. (1998, p. 512) recommend that for randomized trials one may compute the variance of the difference in means test statistic as if the treatment and control groups were selected independently. They argue that if the variance is computed properly, accounting for the dependency between the groups, the finite population correction factor reduces the variance but the negative correlation between the means increases it. They conclude that practitioners should use the infinite population assumptions as the adjustments for the finite population cancel each other. We would like to know if the same argument holds for $\mathrm{Var}(\widehat{att})$ for the two sampling schemes.

From Equation 3.17, even if we ignore the third term, it's not clear which of the assumptions lead to a smaller variance. This is addressed in the following proposition and corollary.

Proposition 3.4.1. *The difference in the asymptotic variances of $\widehat{att}$ under infinite and finite sampling has the following form.*

$$
\begin{aligned}
\nabla \mathrm{h}(\boldsymbol{p})' \tfrac{1}{n}\boldsymbol{\Sigma}_{pop} \nabla \mathrm{h}(\boldsymbol{p}) - \nabla \mathrm{h}(\boldsymbol{p})' \tfrac{1}{n}\boldsymbol{\Sigma}_{samp} \nabla \mathrm{h}(\boldsymbol{p}) &= \frac{p_{comTR}(p_{comAR} + p_{comNR})}{n(p_{comAR} + p_{comNR} + p_{comTR})^3} \\
&\quad - \tfrac{1}{n(n-1)}\nabla \mathrm{h}(\boldsymbol{p})' \left\{(\mathbf{1}-\beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\} \nabla \mathrm{h}(\boldsymbol{p})
\end{aligned}
$$

Since the second term of the right hand side is $O(n^{-2})$, we have the following corollary which more directly relates to the variances.

Corollary 3.4.2.

$$\text{Var}_{pop}(\widehat{att}) - \text{Var}_{samp}(\widehat{att}) \approx \frac{p_{comTR}(p_{comAR} + p_{comNR})}{n(p_{comAR} + p_{comNR} + p_{comTR})^3}$$

Corollary 3.4.2 is notable for a number of reasons. First, the parameters β and p_{nevAR} do not appear; only the proportions of the three complier types matter. Second, as the proportion of complier-always-respond and complier-never-respond are interchangeable in the equation we may write the ratio as

$$\frac{\Pr(comTR \mid \text{complier})[\,1 - \Pr(comTR \mid \text{complier})\,]}{n \Pr(\text{complier})}.$$

The form of this equation, for the difference in the variances under the finite and infinite population models, is quite similar to one found by Imbens and Rubin (2015, p. 440) for general estimators of average causal effects, though their example does not address noncompliance. The denominator indicates the difference between the variance is largest when the fraction of compliers is small and while the denominator shows the difference is maximized when $comTR$ make up half of all compliers. On the contrary, the difference between the variances is minimized with more compliers and when the $comTR$ either make up a very large, or very small portion of the compliers. Finally, Corollary 3.4.2 shows that for large enough n, we have

$$\text{Var}_{pop}(\widehat{att}) \geq \text{Var}_{samp}(\widehat{att}),$$

as the difference between them is positive. This still does not answer the primary question of whether the choice of sampling assumption impacts the conclusions of the experiment. For instance, the formulas approximating the variances indicate $\text{Var}_{pop}(\widehat{att})$ and $\text{Var}_{samp}(\widehat{att})$ also increases as the fraction of compliers decrease. Whether this increase is large enough that the difference between the variances is meaningful is addressed in the next section.

3.5 The Impact of the Sampling Assumptions on the Conclusions about $\widehat{att}$

In Section 3.3 we show the asymptotic variance for $\widehat{att}$ well-approximates the actual variance and in Section 3.4 we calculate how the variance differs between finite and infinite sampling. While the finite sample assumptions lead to a smaller variance, in practice this only matters if the difference in the calculated standard error leads to a noticeable difference in the

significance of $\widehat{att}$. In this section we explore the degree to which the sampling assumption choice leads to different conclusions for hypothesis tests about the existence of att.

To address this question we return to the "parameter grid" described in Section 3.3. The grid allows us to evaluate the standard deviation of $\widehat{att}$ over an adequately dense representation, consisting of over 400 million points, of the parameter space of interest. To investigate the significance for $\widehat{att}$ we take two further steps. First, we transform the grid so the parameters correspond with the percentages observed in the table of observed results, that is, the parameters are $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$ and β) (the fraction assigned to treatment). Second, instead of viewing the points on the grid as the model parameters, we imagine them representing observed data, $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ and β). Every point of the grid represents a possible observed contingency table and, taken over the entire grid, we have an adequate representation of the space of all observed results. For each point of the grid we may then calculate a standard error and p-value for $\widehat{att}$ under each of the two sampling assumptions, and compare the p-values to reveal at which points they differ.

We begin by comparing the standard errors (SE) for $\widehat{att}$ under the two assumptions, calculating the ratio

$$\mathrm{SE}_{samp}(\widehat{att}) \,/\, \mathrm{SE}_{pop}(\widehat{att}),$$

for each point of the grid. Figure 3.3 shows a histogram of these ratios, where we have selected five percent of the total points (to make the plotting feasible). We see the ratio is primarily between 0.9 and 1, though there is a portion where the ratio is between 0.85 and 0.9. This is a difference which could, for example, lead to a noticeably different p-value for a null hypothesis of att equaling zero.

However, despite the difference in the standard error for the two assumptions, the p-values calculated for each point of the grid are quite similar. Figure 3.4(a) shows a scatter plot of the p-values under the two sampling assumption, indicating that most points lie along the identity line. Figure 3.4(b) shows the same plot, magnified to focus on p-values of less than 0.1. Even when the standard errors differ, it makes little difference for the p-values. Further analysis reveals that the points for which the standard errors are most different, for the two sampling assumptions, correspond with observed values which are highly significant. For example, the choice of the sampling scheme may lead to a p-values of 0.01 versus a p-values of 0.005.

In summary, the sampling assumptions *do* make a difference on the standard error but the difference seems to mostly impact observed values with highly significant p-values. These p-values are so significant that the choice of the modeling assumption matters little for the overall conclusions. While we kept n fixed at 100,000, for the figures shown, similar findings held with samples of size 1,000 and 10,000.

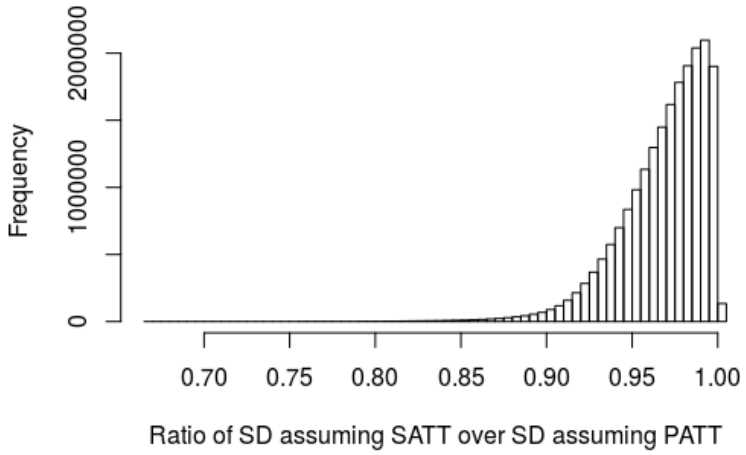
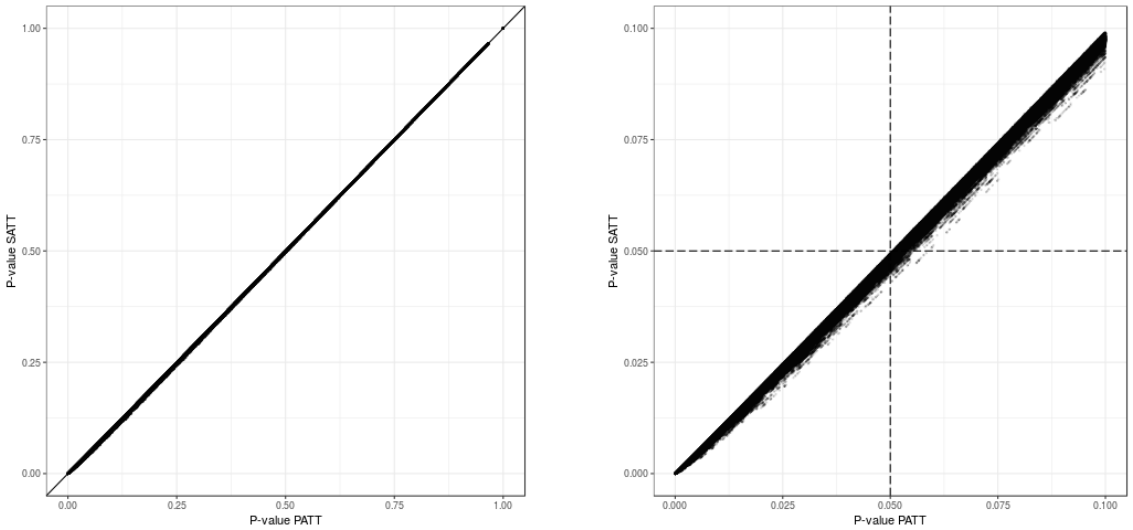


Figure 3.3: Histogram of the ratio of $\text{SE}_{samp}(\widehat{att}) \,/\, \text{SE}_{pop}(\widehat{att})$ for the differing parameter values of $(\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11}, \beta)$. Five percent of the over 400 million points of the grid have been randomly selected for this plot.

Impact of Sampling Assumptions on GOTV Field Experiments

While the previous section highlights the regions of the parameter space where $\text{SE}_{pop}(\widehat{att})$ and $\text{SE}_{samp}(\widehat{att})$ differ, for much of the space there is little practical difference between the two. To see if the different assumptions lead to different conclusions, in GOTV studies, we examined a recent survey of the literature from Green and Gerber (2015). In their Chapter 6, they summarize findings of door-to-door canvassing experiments, citing 22 reports which cover 24 separate field experiments with designs where we may measure $\widehat{att}$ (some articles contain multiple experiments that meet our criteria, some articles have none). From the journal articles and from GOTV micro-data stored at Yale University's Institution for Social and Policy Studies, we were able to obtain the values of $(C_0, C_1, T_{01}, T_{01}, T_{10}, T_{11})$ of 11 experiments, as shown in Table 3.1. We show the standard error under both infinite and finite cases and the p-value from testing the null hypothesis of att equaling zero.

Except for the Bridgeport experiment by Green et al., the underlying sampling assumptions have little impact on the overall significance of the estimate. For the range of results found in these experiments, the standard errors are nearly equal, with the SEs calculated under the finite sampling assumption slightly below the SEs calculated for the infinite case. The Bridgeport experiment has the largest gap between the SEs as the finite case SE is smaller by 4.3%. However, this only changes the p-value from 0.010 to 0.007. The Bridgeport experiment also yields the most significant p-value. This confirms our findings from the previous section; the SEs do differ under the two sampling assumptions but this occurs when the



(a) All points. (b) Focusing on p-values of less than 0.1

Figure 3.4: Scatter plot of p-values computed for *att* under infinite sampling versus p-values for *att* under finite sampling. Ten percent of the the over 400 million points of the grid have been randomly selected for this plot.

findings are already quite strong.

3.6 Estimating Attributable Effects

The landmark paper by Gerber and Green (2000) became a source of controversy when Imai (2005) brought to light troubles with the randomization of treatment assignment for some of the voters in the study. A series of published exchanges led to a spirited dialogue about the strength of the conclusions of the original work. Bowers and Hansen (2005) entered the discussion with suggestions on how the assignment, though imperfect, could still be incorporated into the modeling. Though their main focus was on recovering estimates of treatment effects when controlling for certain covariates, their paper provides a useful and thorough review of the Neyman-Rubin Causal Model when applied to settings with binary outcomes. In doing so, they find a confidence interval, for a treatment effect, that does not depend on assuming the large sample sizes required to apply the delta method. By recognizing the outcome (respond or not) as a direct function of treatment assignment a much simpler analysis can be derived.

Their insights were largely the inspiration for this work and they also make a reference to "types" in the same vein as our *behavioral types*. We take a moment to review their initial analysis. To distinguish this from the approach of Angrist et al. (1996), we express

Publication, Location and/or Treatment	n	$\widehat{att}$	SE *pop*	SE *samp*	p-value *pop*	p-value *samp*
Michelson (2003)						
Dos Palos, Civic Duty	1,903	4.38%	2.57%	2.52%	.088	.081
Dos Palos, Ethnic Solidarity	1,909	3.81%	2.54%	2.49%	.134	.126
Gerber and Green (2000)						
New Haven, Civic Duty	25,571	9.05%	4.30%	4.29%	.035	.035
New Haven, Neighbor Solidarity	25,467	5.11%	4.09%	4.08%	.212	.210
New Haven, Election is Close	25,514	12.12%	4.22%	4.20%	.041	.039
Matland and Murray (2012)						
Brownsville	11,424	7.27%	4.26%	4.23%	.088	.086
Green, Gerber, and Nickerson (2003)						
Bridgeport	1,806	13.94%	5.41%	5.18%	.010	.007
Columbus	2,478	9.97%	7.97%	7.81%	.211	.201
Detroit	4,954	8.33%	4.60%	4.54%	.070	.067
Minneapolis	2,827	10.24%	8.72%	8.62%	.240	.235
Raleigh	2,208	6.35%	6.45%	6.31%	.026	.023

Table 3.1: Results of Get-out-the-vote field experiments and observed p-values for $\widehat{att}$ when assuming observations derive from an infinite population or from a finite sample.

the outcome observed by individual i as W_i resulting from Z_i such that

$$W_i = y_{i1}(d_i(Z_i))) + y_{i0}(1 - d_i(Z_i)))$$

If we set w_{i1} to the outcome for $Z_i = 1$ and w_{i0} to the outcome when $Z_i = 0$–so that $w_{i1} = w_{i0}$ for nevertakers–then

$$W_i = w_{i1}Z_i + w_{i0}(1 - Z_i)$$

This eliminates the need to concern ourselves with the received treatment $d_i()$, so that instead of five types of individuals there are only three:

always-respond (AR) are the complier-always-respond and nevertaker-always-respond,
if Z_i=0 then $W_i = 1$, if Z_i=1 then $W_i = 1$.

never-respond (NR) consist of the complier-never-respond and nevertaker-never-respond,
if Z_i=0 then $W_i = 0$, if Z_i=1 then $W_i = 0$.

complier-if-treated-respond$(comTR)$ are just as before,
if Z_i=0 then $W_i = 0$, if Z_i=1 then $W_i = 1$.

Bowers and Hansen also assume there are no (what we call) *not-treated-respond* subjects. The primary exercise is to estimate the number of complier-if-treated-respond in the treatment group. Since we may omit the treatment received variable, the table of observations

collapses to

	Assigned Control	Assigned Treatment
No response	C_0	T_0
Responds	C_1	T_1
	c	t

As in the previous section we view the experimental group as consisting of three types which are randomly assigned to either the treatment or control so that

$$T_1 = T_{AR} + T_{comTR}$$

The T_{comTR} is what Rosenbaum (2001) refers to as the "attributable effect", the increase in observed responses which may be attributed to the causal effect of the treatment. To estimate the effect, note that if the number of $comTR$ in the treatment group is subtracted from T_1 then

$$T_1 - T_{comTR} = T_{AR},$$

and we are left with number of always-respond assigned to treatment. Since C_1 is the number of always-respond in the control group if T_{comTR} is known, say $T_{comTR} = k$, then $T_1 - k = T_{AR}$ is a hypergeometric random variable equaling the number of always-respond who end up assigned to treatment.

Formally, under the null hypothesis that $T_{comTR} = k$, the random quantity $T_1 - k$ follows a hypergeometric distribution with parameters

$$\begin{aligned} \text{Number of objects in the urn} &= t + c = n \\ \text{Number of always-responder objects in the urn} &= T_{AR} + C_{AR} \\ &= T_1 - k + C_1 \\ \text{Number of draws from the urn} &= t \end{aligned}$$

This allows us to compute p-values for different values of k. By doing so for all values of k from 0 to T_1 we take as our 95% confidence interval only those values of k with p-values below 5%. The estimate of the attributable effect is the value of k with the highest p-value, also known as the Hodges-Lehmann Point Estimate (Hodges and Lehmann, 1964). Furthermore, by side stepping issues of compliance there is no need for Assumptions 2 (exclusion restriction) and 5 (some compliers) of Chapter 2. Assumptions 5 (some compliers) and 6 (monotonicity) are combined and renamed as "non-negativity", that is, there are no individuals who respond when assigned to control, but not to treatment. Thus, we have an approach which gives estimates and confidence bounds for the impact of the treatment without assuming large sample theory. Such randomization inference methods are have many advantages, as discussed by Keele, Small, and Grieve (2017)

3.7 Discussion

In this chapter we evaluate the bias and variance of $\widehat{att}$ under two different sampling assumptions. The first assumes experimental groups are drawn from an infinite superpopulation and the second assumes a finite sample of subjects are randomly allocated, without replacement, between the treatment and control groups. The sampling assumptions *do* matter. Though the parameter estimates are not dependent on the different assumptions, the two sampling schemes lead to a different covariance structure among the observations, resulting in different variances for $\widehat{att}$. We find an approximate value for this difference. As noted in the previous chapter, because the model is parameterized by the proportion of behavioral types, the variance formulas under the two sampling assumptions, and their difference, are also functions of these proportions. However, from a practical point of view, for most social science field experiments, the choice of the sampling assumption will not make a difference in the conditions of the experiment. The difference in sampling assumptions impacts the variance of $\widehat{att}$ when the value of the estimate is highly significant, that is, when there is already strong evidence that $\widehat{att} > 0$. In such a situation, the two sampling assumptions are the difference between "highly significant" findings and "extremely significant" findings.

For the remainder of this research we assume that observations are from a finite sample, and we contain ourselves to inference within the experimental sample.

Under the finite sample assumptions, the underlying data generating process for the table of observations is very similar to a multivariate hypergeometric distribution, as the five behavioral types are randomly allocated between the treatment and control, without replacement. In the final section we show how randomization inference may be useful for small sample sizes as it does not depend on the large sample sizes needed for the approximations in the earlier parts of the chapter.

3.8 Proofs

3.8.1 Proof of Proposition 3.1.1

For this proof we assume the infinite population setting which we restate for clarity. Suppose an experiment has c control and t treatment subjects such that individuals assigned to the two experimental conditions are drawn from an infinite population of possible subjects. The table of observations, when tabulated by the assigned treatment, received treatment and response is

	Assigned Control	Assigned Treatment	
	Received Control	Received Control	Received Treatment
No response	C_0	T_{00}	T_{01}
Responds	C_1	T_{10}	T_{11}
Total	c	t	

where C_y is the total number of subjects assigned to the control with response value y and T_{yd} is the total assigned to control with response y and compliance d. The total assigned to each treatment condition are c and t with $c + t = n$. Under these sampling assumptions, C_1 follows a binomial(c, q_1) distribution where q_1 equals the fraction who respond if assigned to the control group (that is, the proportion of complier-always-respond and nevertaker-always-respond in the population). Similarly the distribution of $(T_{01}, T_{01}, T_{10}, T_{11})$ is multinomial$(t, p_{00}, p_{01}, p_{10}, p_{11})$ and independent of (C_0, C_1).

We begin our proof with four lemmas which, as shall be shown, are the only parts of the proof which hinge on the sampling assumptions. We prove the same lemmas for the finite sample setting in the proof of Proposition 3.2.1 in Section 3.8.3.

In this section, for ease of notation, we drop the *pop* subscript in $\mathbb{E}_{pop}()$ as all expectations are taken with respect to the infinite population assumption.

Lemma 3.8.1.

$$\mathbb{E}\left(T_{10} \mid T_{01} + T_{11}\right) = \frac{p_{10}}{p_{00} + p_{10}} \left(t - (T_{01} + T_{11})\right)$$

Proof. If $T_{01} + T_{11}$, the total assigned in the treatment group who receive the treatment, is known then the number who don't receive, $t - (T_{01} + T_{11})$, is also known. Among those not receiving treatment, each responds, independently with probability $p_{10}/(p_{00} + p_{10})$. Thus T_{10} has binomial distribution with the number of trials equaling $t - (T_{01} + T_{11})$ and success probability of $p_{10}/(p_{00} + p_{10})$. Multiplying the two yields the expectation. □

Lemma 3.8.2.

$$\mathbb{E}\left(T_{11} \mid T_{01} + T_{11}\right) = \frac{p_{11}}{p_{01} + p_{11}} \left(T_{01} + T_{11}\right)$$

Proof. Since t is fixed, conditioning on those who receive treatment $T_{01} + T_{11}$ is the same as conditioning on those who do not, $T_{00} + T_{10}$, so that

$$\mathbb{E}\left(T_{11} \mid T_{01} + T_{11}\right) = \mathbb{E}\left(T_{11} \mid T_{00} + T_{10}\right),$$

and we may employ the same argument from the proof of Lemma 3.8.2. If given $T_{00} + T_{10}$, T_{11} has binomial distribution with $T_{00} + T_{10}$ trials and chance of success $p_{11}/(p_{01} + p_{11})$ which gives the desired results. □

The definition of $\widehat{att}$ from equation 2.9 set the estimator to zero if $T_{01} + T_{11} = 0$. More generally, we define the operator *, for any ratio, such that the ratio equals zero when the denominator is zero. That is for any any x and y,

$$\left(\frac{x}{y}\right)^* = \begin{cases} \frac{x}{y}, & \text{if } y \neq 0 \\ 0, & \text{if } y = 0. \end{cases}$$

. We can add and multiply quantities with the * operator, that is

$$\left(\frac{w + x}{y}\right)^* = \left(\frac{w}{y}\right)^* + \left(\frac{x}{y}\right)^*$$

$$\left(\frac{wx}{y}\right)^* = w \left(\frac{x}{y}\right)^*.$$

We use both properties in the proofs of the following to lemmas involving the quantity $\left(\frac{t}{T_{01}+T_{11}}\right)^*$, that is,

$$\left(\frac{t}{T_{01} + T_{11}}\right)^* = \begin{cases} \frac{t}{T_{01}+T_{11}}, & \text{if } T_{01} + T_{11} \geq 1 \\ 0, & \text{if } T_{01} + T_{11} = 0. \end{cases}$$

Lemma 3.8.3.

$$\mathbb{E}\left(\frac{C_1}{T_{01} + T_{11}}\right)^* = c\, q_1\, \mathbb{E}\left(\frac{1}{T_{01} + T_{11}}\right)^*$$

Proof. Using the multiplicative property of the * operator and the independence of C_y and the T_{yd}, we may separate the expectation of the numerator and denominator. C_1 is binomial with expectation $c\, q_1$. □

Lemma 3.8.4.

$$\mathbb{E}\left(\frac{t}{T_{01}+T_{11}}\right)^{*} = \frac{1}{p_{01}+p_{11}} + O(\tfrac{1}{t})$$

Proof. Since $T_{01}+T_{11}$ is distributed as binomial(t, $p_{01}+p_{11}$), it suffices to show that if X is a binomial(n,p) random variable then then

$$\mathbb{E}\left(\frac{n}{X}\right)^{*} = \frac{1}{p} + O(\tfrac{1}{n}). \tag{3.18}$$

We prove Equation 3.18 by first noting that

$$\begin{aligned}\mathbb{E}\left(\frac{n}{X}\right)^{*} &= 0\cdot\Pr(X=0) + \frac{n}{1}\Pr(X=1) + \frac{n}{2}\Pr(X=2) + ... \\ &= n\sum_{i=1}^{n}\frac{1}{i}\Pr(X=i). \end{aligned}\tag{3.19}$$

Now consider a variable Y with the positive binomial distribution, binomial^{+}(n,p), that is for $y = 1, 2, ..., n$,

$$\begin{aligned}\Pr(Y=y) &= \frac{\Pr(X=y)}{1-\Pr(X=0)} \\ &= \frac{1}{1-(1-p)^n}\Pr(X=y).\end{aligned}$$

Then

$$\mathbb{E}\left(\frac{n}{Y}\right) = \frac{n}{1-(1-p)^n}\sum_{i=1}^{n}\frac{1}{i}\Pr(X=i)$$

and from Equation 3.19

$$\mathbb{E}\left(\frac{n}{X}\right)^{*} = (1-(1-p)^n)\,\mathbb{E}\left(\frac{n}{Y}\right) \tag{3.20}$$

The first moment of the inverse of positive binomial random variable has been examined a number of times in the statistics literature. Perhaps most immediate to our needs is the finding from Znidaric (2005) that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{Y}\right) &= \frac{np}{(np+1-p)^2}\left(1 - \frac{3(n-1)p(1-p)}{(np+1-p)^2} + \frac{4(n-1)p(1-p)(1-2p)}{(np+1-p)^3}\right) + O(\tfrac{1}{n^2}) \\ &= \frac{np}{(np+1-p)^2}\left(1 + O(\tfrac{1}{n}) + O(\tfrac{1}{n^2})\right) + O(\tfrac{1}{n^2}).\end{aligned}$$

We multiply both sides of the equation by n and simplify in terms of the order of n. This gives

$$\mathbb{E}\left(\frac{n}{Y}\right) = \frac{n^2p}{(np+1-p)^2}\left(1 + O(\tfrac{1}{n}) + O(\tfrac{1}{n})\right) + O(\tfrac{1}{n})$$

$$
\begin{aligned}
&= \frac{n^2 p}{(np + 1 - p)^2} \left(1 + O(\tfrac{1}{n})\right) + O(\tfrac{1}{n}) \\
&= \frac{p}{(p + \frac{1-p}{n})^2} \left(1 + O(\tfrac{1}{n})\right) + O(\tfrac{1}{n}) \\
&= \frac{1}{p} \left(1 + O(\tfrac{1}{n})\right) \left(1 + O(\tfrac{1}{n})\right) + O(\tfrac{1}{n}) \\
\mathbb{E}\left(\frac{n}{Y}\right) &= \frac{1}{p} + O(\tfrac{1}{n}) \qquad (3.21)
\end{aligned}
$$

Finally, by substituting this result into Equation 3.20 to we have

$$
\begin{aligned}
\mathbb{E}\left(\frac{n}{X}\right)^* &= (1 - (1-p)^n)\, \mathbb{E}\left(\frac{n}{Y}\right) \\
&= (1 - (1-p)^n) \left(\frac{1}{p} + O(\tfrac{1}{n})\right) \\
&= \frac{1}{p} + O(\tfrac{1}{n})
\end{aligned}
$$

which is the sufficient condition we need to show from Equation 3.18. □

With these four lemmas we return to our our main objective, to prove the following.

Proposition. *Suppose:*

i *The total assigned to treatment, $t(n)$ is such that $t(n)/n \to \beta$ and the total assigned to control is $c(n) = n - t(n)$.*

ii *Assume infinite population sampling, that is, C_1 is binomial($c(n)$, q_1) and $(T_{01}, T_{01}, T_{10}, T_{11})$ is multinomial($t(n), p_{00}, p_{01}, p_{10}, p_{11}$).*

iii *The average treatment effect for the treated is defined as*

$$att = \mathrm{h}(\boldsymbol{p}) = \frac{p_{10} + p_{11} - q_1}{p_{01} + p_{11}},$$

which is estimated by

$$
\widehat{att} = \begin{cases} \mathrm{h}(\widehat{\boldsymbol{p}}) = \dfrac{\dfrac{T_{10} + T_{11}}{t(n)} - \dfrac{C_1}{c(n)}}{\dfrac{T_{01} + T_{11}}{t(n)}} & \text{if } T_{01} + T_{11} > 0 \\ 0 & \text{if } T_{01} + T_{11} = 0. \end{cases}
$$

Then the order of the bias is $1/n$*, or*

$$\mathbb{E}_{pop}(\widehat{att}) = att + O(\tfrac{1}{n}).$$

Note: though not stated explicitly, each random quantity above, such as $\widehat{att}$*,* C_1*, etc., is indexed by* n*. We remove the index for ease of notation.*

Proof. For ease of notation we write c for $c(n)$ and t for $t(n)$ and drop the *pop* subscript in $\mathbb{E}_{pop}()$ as all expectations are taken with respect to the infinite population assumption.

$$\begin{aligned}\mathbb{E}(\widehat{att}) &= \mathbb{E}\left(\frac{\frac{T_{10}+T_{11}}{t} - \frac{C_1}{c}}{\frac{T_{01}+T_{11}}{t}}\right)^* \\ &= \mathbb{E}\underbrace{\left(\frac{T_{10}}{T_{01}+T_{11}}\right)^*}_{(a)} + \mathbb{E}\underbrace{\left(\frac{T_{11}}{T_{01}+T_{11}}\right)^*}_{(b)} - \mathbb{E}\underbrace{\left(\frac{\frac{t}{c}C_1}{T_{01}+T_{11}}\right)^*}_{(c)} \end{aligned} \tag{3.22}$$

We aim to understand how the three components from Equation 3.22 correspond with their analouges of Equation 2.15, that is, how (a), (b) and (c) correspond to

$$\frac{p_{10}}{p_{01}+p_{11}}, \frac{p_{11}}{p_{01}+p_{11}} \text{ and } \frac{q_1}{p_{01}+p_{11}}$$

.

We address each component in turn, by taking a double expectation, conditioning on $T_{01} + T_{11}$. Starting with (a),

$$\begin{aligned}(a) &= \mathbb{E}\left(\tfrac{T_{10}}{T_{01}+T_{11}}\right)^* \\ &= \mathbb{E}\left[\mathbb{E}\left(\left(\tfrac{T_{10}}{T_{01}+T_{11}}\right)^* \mid T_{01}+T_{11}\right)\right] \\ &= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(T_{10} \mid T_{01}+T_{11}\right)\right] \\ &= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \frac{p_{10}}{p_{00}+p_{10}}\left(t-(T_{01}+T_{11})\right)\right] \text{ , by Lemma 3.8.1} \\ &= \frac{p_{10}}{p_{00}+p_{10}}\mathbb{E}\left(\frac{t-(T_{01}+T_{11})}{T_{01}+T_{11}}\right)^* \\ &= \frac{p_{10}}{p_{00}+p_{10}}\mathbb{E}\left[\left(\frac{t}{T_{01}+T_{11}}\right)^* - \left(\frac{T_{01}+T_{11}}{T_{01}+T_{11}}\right)^*\right].\end{aligned}$$

The first term of the expectation is given by Lemma 3.8.4 and the second term is 1 unless $T_{01} + T_{11}$=0 so that

$$(a) = \frac{p_{10}}{p_{00}+p_{10}}\left(\frac{1}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right) - \left(1 - \Pr(T_{01}+T_{11}=0)\right)\right)$$

$$= \frac{p_{10}}{p_{00}+p_{10}} \left(\frac{1}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right) - 1 + (1-p_{01}-p_{11})^t \right)$$
$$= \frac{p_{10}}{p_{00}+p_{10}} \left(\frac{1}{p_{01}+p_{11}} - 1 + O\left(\tfrac{1}{t}\right) \right)$$

Recall that $p_{00}+p_{01}+p_{10}+p_{11}=1$. Then $\frac{1}{p_{01}+p_{11}} - 1 = \frac{1-(p_{01}+p_{11})}{p_{01}+p_{11}} = \frac{p_{00}+p_{01}}{p_{01}+p_{11}}$ so that,

$$(a) = \frac{p_{10}}{p_{00}+p_{10}} \left(\frac{p_{00}+p_{01}}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right) \right)$$
$$= \frac{p_{10}}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right)$$

For (b) we have

$$(b) = \mathbb{E}\left(\frac{T_{11}}{T_{01}+T_{11}} \right)^*$$
$$= \mathbb{E}\left[\mathbb{E}\left(\left(\frac{T_{11}}{T_{01}+T_{11}} \right)^* \mid T_{01}+T_{11} \right) \right]$$
$$= \mathbb{E}\left[\left(\frac{1}{T_{01}+T_{11}} \right)^* \mathbb{E}\left(T_{11} \mid T_{01}+T_{11} \right) \right]$$
$$= \mathbb{E}\left[\left(\frac{1}{T_{01}+T_{11}} \right)^* \frac{p_{11}}{p_{01}+p_{11}} \left(T_{01}+T_{11} \right) \right] \text{ , by Lemma 3.8.2}$$
$$= \frac{p_{11}}{p_{01}+p_{11}} \mathbb{E}\left(\frac{T_{01}+T_{11}}{T_{01}+T_{11}} \right)^*$$
$$= \frac{p_{11}}{p_{01}+p_{11}} (\, 1 - \Pr(T_{01}+T_{11}=0) \,).$$

For (c),

$$(c) = \frac{t}{c} \mathbb{E}\left(\frac{C_1}{T_{01}+T_{11}} \right)^*$$
$$= t\, q_1\, \mathbb{E}\left(\frac{1}{T_{01}+T_{11}} \right)^* , \text{ by Lemma 3.8.3}$$
$$= q_1\, \mathbb{E}\left(\frac{t}{T_{01}+T_{11}} \right)^*$$
$$= \frac{q_1}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right)$$

We may now insert each of the components into Equation 3.22.

$$\begin{aligned}
\mathbb{E}(\widehat{att}) &= (a) + (b) - (c) \\
&= \frac{p_{10}}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right) + \frac{p_{11}}{p_{01}+p_{11}}(\,1 - \Pr(T_{01}+T_{11}=0)\,) - \frac{q_1}{p_{01}+p_{11}} + O\left(\tfrac{1}{t}\right) \\
&= \frac{p_{10}+p_{11}-q_1}{p_{01}+p_{11}} - \frac{p_{11}}{p_{01}+p_{11}}\Pr(T_{01}+T_{11}=0) + O\left(\tfrac{1}{t}\right) \\
&= \frac{p_{10}+p_{11}-q_1}{p_{01}+p_{11}} - \frac{p_{11}}{p_{01}+p_{11}}(1-p_{01}-p_{11})^t + O\left(\tfrac{1}{t}\right) \\
&= att + O\left(\tfrac{1}{t}\right)
\end{aligned}$$

As $O(1/t)$ is $O(1/n)$ we have the desired result.

□

Remark. Aside from the lemmas, the sampling assumptions are only used to show that $\Pr(T_{01}+T_{11}=0) = O(1/t)$, which also folds for finite sampling. Thus to obtain this result for the finite sampling case we need only show that Lemmas 3.8.1 - 3.8.4 hold for the finite sample assumptions as well. This how we proceed to prove Proposition 3.2.1 later in this section.

3.8.2 Proof of Proposition 3.1.2

Proposition. *Suppose the assumptions of Proposition 3.1.1 then*

$$\sqrt{n}(\widehat{att} - att) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\boldsymbol{p})' \, \boldsymbol{\Sigma}_{\mathbf{pop}} \, \nabla \mathrm{h}(\boldsymbol{p})).$$

Proof. We first note that in this proof all random quantities, $\widehat{att}$, $\widehat{\mathbf{p}}$, C_y, T_{yd}, etc., are indexed by n. We leave the index implicit for ease of notation. Also, rather than addressing the discontinuity of $\widehat{att}$ when $T_{01}+T_{11}=0$ we observe that $\widehat{att}$ converges in probability to $\mathrm{h}(\widehat{\mathbf{p}})$ so it will suffice to show that

$$\sqrt{n}(\mathrm{h}(\widehat{\mathbf{p}}) - \mathrm{h}(\mathbf{p})) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\mathbf{p})' \, \boldsymbol{\Sigma}_{\mathbf{pop}} \, \nabla \mathrm{h}(\mathbf{p})).$$

To show this we employ the delta method for convergence in distribution (see Bishop et al., 1975, p. 493) and the first step is to show the asymptotic normality of $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$. Since C_1 is binomial$(c(n), q_1)$, it is the sum of $c(n)$ independent Bernoulli random variables, each with finite variance of $q_1(1-q_1)$. Using $\widehat{q}_1$ for $C_1/c(n)$, by the Central Limit Theorem we have

$$\sqrt{c^n}(\widehat{q}_1 - q_1) \xrightarrow{d} \mathrm{N}(0, q_1(1-q_1)).$$

We multiply the right hand side by $\sqrt{n/c^n}$ which goes to $\sqrt{1/\beta}$ by assumption. Then, by Slutsky's Theorem we have

$$\sqrt{n/c^n}\sqrt{c^n}(\widehat{q}_1 - q_1) \xrightarrow{d} \sqrt{1/\beta}Z,$$

where Z is $\mathrm{N}(0, q_1(1-q_1))$ so that

$$\sqrt{n}(\widehat{q}_1 - q_1) \xrightarrow{d} \mathrm{N}(0, \frac{q_1(1-q_1)}{\beta}).$$

Similarly, $(T_{00}, T_{01}, T_{10}, T_{11})$ is multinomial and the average counts converge to a normal distribution as $t(n) \to \infty$ (see Bishop et al., 1975, p 469). Using the same arguments as above we have

$$\sqrt{n}\,[\,(\widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11}) - (p_{01}, p_{10}, p_{11})\,] \xrightarrow{d} \mathrm{N}(0, \mathbf{\Sigma_{pop}}[2:4][2:4]).$$

where [2:4][2:4] denotes the the 3×3 sub-matrix of $\mathbf{\Sigma_{pop}}$ without the first row and first column. Taken together, since $\widehat{q}_1$ and $(\widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ are independent, $\widehat{\mathbf{p}} = (\widehat{q}_1, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})$ will converge to a normal distribution,

$$\sqrt{n}(\widehat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} \mathrm{N}(0, \mathbf{\Sigma_{pop}}).$$

With the asymptotic normality of $\widehat{\mathbf{p}}$ established, the final condition to use the delta method is to confirm the partial derivatives of h() exist for a neighborhood around $\mathbf{p}$. We see that

$$\nabla \mathrm{h}(\mathbf{p}) = \begin{pmatrix} \frac{\partial h}{\partial q_1} \\ \\ \frac{\partial h}{\partial p_{01}} \\ \\ \frac{\partial h}{\partial p_{10}} \\ \\ \frac{\partial h}{\partial p_{11}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{p_{01}+p_{11}} \\ \\ \frac{q_1 - p_{10} - p_{11}}{(p_{01}+p_{11})^2} \\ \\ \frac{1}{p_{01}+p_{11}} \\ \\ \frac{q_1 + p_{01} - p_{10}}{(p_{01}+p_{11})^2} \end{pmatrix}.$$

It is apparent that $\nabla \mathrm{h}(\mathbf{p})$ is continuous and exists $\mathbf{p}$ as long as $p_{01} + p_{11} \neq 0$. This condition holds by the "there are some compliers" assumption of Chapter 2. Thus, we may approximate $\mathrm{h}(\widehat{\mathbf{p}})$ with a first order Taylor's expansion of h() around $\mathbf{p}$, that is,

$$\mathrm{h}(\widehat{\mathbf{p}}) = \mathrm{h}(\mathbf{p}) + (\widehat{\mathbf{p}} - \mathbf{p})\,\nabla \mathrm{h}(\mathbf{p}) + o\left(\|(\widehat{\mathbf{p}} - \mathbf{p})\|\right),$$

so that the asymptotic distribution of $\mathrm{h}(\widehat{\mathbf{p}})$ is given by

$$\sqrt{n}(\mathrm{h}(\widehat{\mathbf{p}}) - \mathrm{h}(\mathbf{p})) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\mathbf{p})'\,\mathbf{\Sigma_{pop}}\,\nabla \mathrm{h}(\mathbf{p})).$$

This concludes the proof. □

3.8.3 Proof of Proposition 3.2.1

In the finite sample case, the underlying distribution generating the table of observations is multivariate hypergeometric. Here, the parameters are the totals assigned to control and

treatment, c and t such that $c + t = n$, and also the totals of each of the finite behavioral types: n_{comAR}, n_{comNR}, n_{comTR}, n_{nevAR} and n_{nevNR}. The totals of the behavioral types sum to n. We re-parameterize the model with $(q_1, p_{01}, p_{10}, p_{11})$ where

$$\begin{aligned} q_1 &= \frac{n_{comAR} + n_{nevAR}}{n} \\ p_{01} &= \frac{n_{comNR}}{n} \\ p_{10} &= \frac{n_{nevAR}}{n} \\ p_{11} &= \frac{n_{comAR} + n_{comTR}}{n} \end{aligned}$$

and $p_{00} = n_{nevNR}/n$ is redundant as $p_{00} = 1 - p_{01} - p_{10} - p_{11}$.

As with the proof of Proposition 3.1.1, the argument depends on the following four lemmas which are the finite sample analouges to Lemmas 3.8.1 – 3.8.4. In this section, for ease of notation, we drop the *samp* subscript in $\mathbb{E}_{samp}()$ as all expectations are taken with respect to the finite sample assumption.

Lemma 3.8.5.

$$\mathbb{E}\left(T_{10} \mid T_{01} + T_{11}\right) = \frac{p_{10}}{p_{00} + p_{10}} \left(t - (T_{01} + T_{11})\right)$$

Proof. T_{10} is the number of nevertaker-always-respond assigned to treatment. $T_{01} + T_{11}$ are the number of compliers assigned to treatment and if they are known, the number of compliers assigned to control is also given. Under these conditions, T_{10} is distributed as a hypergeometric distribution, counting number of successful events, from a pool of n_{nevAR} "successes", n_{nevNR} "failures" and $t - (T_{01} + T_{11})$ draws without replacement. Thus the expected number of successful draws is

$$\frac{n_{nevAR}}{n_{nevNR} + n_{nevAR}} \left(t - (T_{01} + T_{11})\right) = \frac{p_{10}}{p_{00} + p_{10}} \left(t - (T_{01} + T_{11})\right).$$

□

Lemma 3.8.6.

$$\mathbb{E}\left(T_{11} \mid T_{01} + T_{11}\right) = \frac{p_{11}}{p_{01} + p_{11}} \left(T_{01} + T_{11}\right)$$

Proof. First we note that since $T_{00} + T_{10} = t - (T_{01} + T_{11})$, if $T_{01} + T_{11}$ is known then $T_{00} + T_{10}$ is known, so that

$$\mathbb{E}\left(T_{11} \mid T_{01} + T_{11}\right) = \mathbb{E}\left(T_{11} \mid T_{00} + T_{10}\right).$$

We focus on the right hand side of the equation. T_{11} is the number of complier-always-respond and $comTR$ assigned to treatment. Given $T_{00} + T_{10}$ the number of nevertakers

assigned to treatment and the number of nevertakers assigned to control is known. As was argued in the proof of Lemma 3.8.5, under these conditions T_{11} is hypergeometric with $n_{comAR} + n_{comTR}$ "successes" and n_{comNR} "failures" and we draw $T_{01} + T_{11}$ times without replacement. Thus the expected number of successful draws is

$$\frac{n_{comAR} + n_{comTR}}{n_{comAR} + n_{comTR} + n_{comNR}}(T_{01} + T_{11}) = \frac{p_{11}}{p_{01} + p_{11}}(T_{01} + T_{11}).$$

□

Lemma 3.8.7.

$$\mathbb{E}\left(\frac{C_1}{T_{01} + T_{11}}\right)^* = c\, q_1\, \mathbb{E}\left(\frac{1}{T_{01} + T_{11}}\right)^*$$

Proof.

$$\begin{aligned}
\mathbb{E}\left(\frac{C_1}{T_{01} + T_{11}}\right)^* &= \mathbb{E}\left[\mathbb{E}\left(\left(\frac{C_1}{T_{01} + T_{11}}\right)^* \mid T_{01} + T_{11}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\left(\frac{C_{comAR} + C_{nevAR}}{T_{01} + T_{11}}\right)^* \mid T_{01} + T_{11}\right)\right] \\
&= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(C_{comAR} + C_{nevAR} \mid T_{01} + T_{11}\right)\right] \\
&= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(C_{comAR} \mid T_{01} + T_{11}\right)\right] \\
&\quad + \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(C_{nevAR} \mid T_{01} + T_{11}\right)\right]
\end{aligned}$$

We apply the Law of Iterated Expectations to each of the inner expectations, conditioning on T_{comAR} in the first and on T_{nevAR} in the second so that

$$\begin{aligned}
&= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(\mathbb{E}\left\{C_{comAR} \mid T_{01} + T_{11}, T_{comAR}\right\}\right)\right] \\
&\quad + \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(\mathbb{E}\left\{C_{nevAR} \mid T_{01} + T_{11}, T_{nevAR}\right\}\right)\right]
\end{aligned}$$

If T_{comAR} is known then C_{comAR} is known so that $\mathbb{E}\left\{C_{comAR} \mid T_{01} + T_{11}, T_{comAR}\right\}$ is simply C_{comAR}. Similarly, if T_{nevAR} is known then C_{nevAR} is known. Thus

$$\begin{aligned}
&= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(C_{comAR}\right)\right] + \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* \mathbb{E}\left(C_{nevAR}\right)\right] \\
&= \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* c\frac{n_{comAR}}{n}\right] + \mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^* c\frac{n_{nevAR}}{n}\right] \\
&= c\,\frac{n_{comAR} + n_{nevAR}}{n}\,\mathbb{E}\left[\left(\tfrac{1}{T_{01}+T_{11}}\right)^*\right]
\end{aligned}$$

$$= c\, q_1\, \mathbb{E}\left(\frac{1}{T_{01}+T_{11}}\right)^*$$

□

Lemma 3.8.8.

$$\mathbb{E}\left(\frac{t}{T_{01}+T_{11}}\right)^* = \frac{1}{p_{01}+p_{11}} + O(\tfrac{1}{t})$$

Proof. We may use a nearly identical proof to that of Lemma 3.8.4, which obtained the desired result for the infinite sampling case where $T_{01}+T_{11}$ was binomial$(t, p_{01}+p_{11})$. Here, $T_{01}+T_{11}$ is hypergeometric$(n, n(p_{01}+p_{11}), t)$ where n represents the total number objects, $n(p_{01}+p_{11})$ is the total "successes" and t is the number of draws. It suffices to show that if X is hypergeometric(N, Np, n) then

$$\mathbb{E}\left(\frac{n}{X}\right)^* = \frac{1}{p} + O(\tfrac{1}{n}).$$

We follow the exact argument in the proof of Lemma 3.8.4 so the only relation we need to show is the finite sample equivalent of Equation 3.21. That is

$$\mathbb{E}\left(\frac{n}{Y}\right) = \frac{1}{p} + O(\tfrac{1}{n}),$$

where Y is a positive hypergeometric(N, Np, n) random variable, that is, for $j = 1, 2, ...,$

$$\Pr(Y = j) = Pr(X = j)/\Pr(X > 0).$$

To prove this we use a result from Stephan (1945, p. 60) that

$$\mathbb{E}\left(\frac{1}{Y}\right) = \sum_{i=1}^{k} u_i + \mathbb{E}(R_k(Y)) \tag{3.23}$$

where

$$u_1 = \frac{(N+1)s_1}{(Np+1)(n+1)s_2} \tag{3.24}$$

and

$$s_0 = \Pr(X > 0)$$

$$s_1 = 1 - \sum_{j=0}^{i} \Pr(X' = j), \text{ where } X' \text{ is hypergeometric}(N+i, Np+i, n+i)$$

So for small i, $s_i \to 1$, rapidly, as $n \to \infty$ and from Equation 3.24 we have

$$u_1 = \frac{1}{pn} + o(1)$$

There is a recursive relationship between u_{i+1} and u_i which gives

$$\begin{aligned} u_2 &= \frac{(N+2)s_2}{(Np+2)(n+2)s_1} u_1 \\ &= \left(\frac{N}{Npn} + o(1)\right)\left(\frac{1}{pn} + o(1)\right) \\ &= \frac{1}{p^2n^2} + o(1). \end{aligned}$$

Furthermore, $\mathbb{E}(R_2(Y)) \leq 2u_2$. From Equation 3.23, expanding to two terms ($k = 2$), we have

$$\begin{aligned} \mathbb{E}\left(\frac{1}{Y}\right) &= u_1 + u_2 + \mathbb{E}(R_2(Y)) \\ &= \frac{1}{pn} + o(1) + \frac{1}{p^2n^2} + o(1) + 2\left(\frac{1}{p^2n^2} + o(1)\right) \\ &= \frac{1}{pn} + O\left(\frac{1}{n^2}\right) \end{aligned}$$

Multiplying both sides n gives the desired result. □

We return to our our main objective, to prove the following.

Proposition. *Suppose:*

i *The total assigned to treatment, $t(n)$ is such that $t(n)/n \to \beta$ and the total assigned to control is $c(n) = n - t(n)$.*

ii *Assume finite population sampling as described in Section 3.2. That is, the five different behavioral types are randomly allocated without replacement. This results in the counts $(C_0, C_1, T_{00}, T_{01}, T_{10}, T_{11})$ found via the relationships in Equations 3.6 - 3.11.*

iii *The proportion parameters for each finite n, $\boldsymbol{p}(n) = (q_1(n), p_{01}(n), p_{10}(n), p_{11}(n))$, converge to the limiting vector $\boldsymbol{p}^\infty = (q_1^\infty, p_{01}^\infty, p_{10}^\infty, p_{11}^\infty)$ as follows:*

$$\begin{aligned} q_1(n) &= q_1^\infty + O(1/n) \\ p_{01}(n) &= p_{01}^\infty + O(1/n) \\ p_{10}(n) &= p_{10}^\infty + O(1/n) \\ p_{11}(n)) &= p_{11}^\infty + O(1/n) \end{aligned}$$

iv The average treatment effect for the treated is defined as

$$att(n) = \mathrm{h}(\boldsymbol{p}(n)) = \frac{p_{10}(n) + p_{11}(n) - q_1(n)}{p_{01}(n) + p_{11}(n)},$$

which is estimated by

$$\widehat{att}(n) = \begin{cases} \mathrm{h}(\widehat{\boldsymbol{p}}) = \dfrac{\dfrac{T_{10}+T_{11}}{t(n)} - \dfrac{C_1}{c(n)}}{\dfrac{T_{01}+T_{11}}{t(n)}} & \text{if } T_{01}+T_{11} > 0 \\ 0 & \text{if } T_{01}+T_{11} = 0. \end{cases}$$

Then the order of the bias is $1/n$*, or*

$$\mathbb{E}_{samp}(\widehat{att}) = att(n) + O(\tfrac{1}{n}).$$

Proof. Our argument is nearly identical to the one provided in Section 3.8.1 of Proposition 3.1.1 if we substitute Lemmas 3.8.5 - 3.8.8 for Lemmas 3.8.1– 3.8.4 and note that $\Pr(T_{01} + T_{11} = 0) = O(1/t)$.

One final consideration is that the proportion parameters in the infinite sampling case, $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$, are now indexed by n, so the parameters in the prior proof must be replaced with $\mathbf{p}(n) = (q_1(n), p_{01}(n), p_{10}(n), p_{11}(n))$.

Noting these two changes, this proof follows along the lines of the proof for the infinite sampling case. □

3.8.4 Proof of Proposition 3.2.2

Proposition. *Suppose the assumptions of Proposition 3.2.1 then*

$$\sqrt{n}(\widehat{att} - att) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\boldsymbol{p})' \, \boldsymbol{\Sigma}_{\mathbf{samp}} \, \nabla \mathrm{h}(\boldsymbol{p})).$$

Proof. Preliminary note: The proportion parameters $\mathbf{p}(n) = (q_1(n), p_{01}(n), p_{10}(n), p_{11}(n))$ are indexed by n though for ease of exposition we drop the index and write them as $\mathbf{p} = (q_1, p_{01}, p_{10}, p_{11})$. This also holds for att, which is a limiting function of the components of $\mathbf{p}(n)$, but we ignore this important distinction in our argument below.

As in the proof of Proposition 3.1.2, we observe that the difference between $\widehat{att}$ and $\mathrm{h}(\widehat{\mathbf{p}})$ converges in probability to zero, or

$$\widehat{att} - \mathrm{h}(\widehat{\mathbf{p}}) \xrightarrow{P} 0,$$

so it suffices to show

$$\sqrt{n}(\mathrm{h}(\widehat{\mathbf{p}}) - \mathrm{h}(\mathbf{p})) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\mathbf{p})' \, \boldsymbol{\Sigma}_{\mathbf{samp}} \, \nabla \mathrm{h}(\mathbf{p})).$$

Earlier in the chapter we noted that, under finite sampling, the asymptotic normality of treatment effect estimators does not hold, in general. Recently though, Li and Ding (2017, pp. 1763-1764) establish conditions under which such estimators are asymptotically normal. We focus our attention on Theorems 3 and 5 of their article. The theorems apply to a range of experimental designs but, for clarity of our own arguement, we present their findings, and describe the required conditions, in terms of our setting, using the notation we have introduced in the last two chapters. We use results from Li and Ding to show the bivariate vector,

$$\left(\widehat{itt},\, \widehat{p}_{com}\right),$$

converges to a normal distribution. Here $\widehat{p}_{com}$ represents the observed fraction of compliers so we may write the bivariate vector as

$$\left(\widehat{p}_{10} + \widehat{p}_{11} - \widehat{q}_1,\, \widehat{p}_{01} + \widehat{p}_{11}\right),$$

which are the numerator and denominator for $\mathrm{h}(\widehat{\mathbf{p}})$. We then apply the delta method to the ratio to arrive at the desired asymptotic distribution.

We begin by specifying terms used for Theorems 3 and 5. Recall the notation from Chapter 2 with a control and just one treatment, so the "assignment", z, is 0 or 1. For assignment z, subject i will have binary response y_{iz} and a value for "treatment", d_{iz}, an indicator of whether the treatment was received. Therefore, for assignment z and $i = 1, 2, ..., n$ we have a two-dimensional potential outcome vector,

$$\mathbf{r}_{iz} = \left(y_{iz},\, d_{iz}\right)^{\mathsf{T}}.$$

(Note that Li and Ding refer to $\mathbf{r}_{iz}$ as $\mathbf{Y}_i(z)$). Furthermore, let

$$\mathbf{A}_0 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and define

$$\boldsymbol{\tau}_i(\mathbf{A}) = \sum_{z=0}^{1} \mathbf{A}_z \mathbf{r}_{iz}\,,$$

so that

$$\begin{aligned}
\boldsymbol{\tau}_i(\mathbf{A}) &= \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} y_{i0} \\ d_{i0} \end{pmatrix} \quad + \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_{i1} \\ d_{i1} \end{pmatrix} \\
&= \begin{pmatrix} y_{i1} - y_{i0} \\ d_{i1} - d_{i0} \end{pmatrix} \\
&= \begin{pmatrix} y_{i1} - y_{i0} \\ d_{i1} \end{pmatrix},
\end{aligned}$$

where the last equality holds because $d_{i0} = 0$, as subjects assigned to the control cannot receive the treatment. Thus, $\boldsymbol{\tau}_i(\mathbf{A})$ is the individual causal effect consisting of the treatment effect for subject i and an indicator of whether they are a complier. The average causal effect over the n subjects is

$$\begin{aligned}\boldsymbol{\tau}(\mathbf{A}) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\tau}_i(\mathbf{A}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_{i1}-y_{i0} \\ d_{i1}\end{pmatrix} \\ &\equiv \begin{pmatrix} itt \\ p_{com}\end{pmatrix},\end{aligned}$$

by the definition of itt from Chapter 2 and the description of p_{com} at the beginning of the proof. We estimate $\boldsymbol{\tau}(\mathbf{A})$ with

$$\widehat{\boldsymbol{\tau}}(\mathbf{A}) = \begin{pmatrix} \widehat{itt} \\ \widehat{p}_{com}\end{pmatrix}.$$

To show $\left(\widehat{itt},\, \widehat{p}_{com}\right)$ converges to a normal distribution we must show $\sqrt{n}\left(\widehat{\boldsymbol{\tau}}(\mathbf{A}) - \boldsymbol{\tau}(\mathbf{A})\right)$ is asymptotically normal. To meet the conditions of Theorems 3 and 5 we must evaluate the three following 2-by-2 matrices representing the covariances of the potential outcomes over the n subjects.

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i0}-\overline{y_0}, d_{i0}-\overline{d_0}\right)^{\intercal}\left(y_{i0}-\overline{y_0}, d_{i0}-\overline{d_0}\right)$$

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i1}-\overline{y_1}, d_{i1}-\overline{d_1}\right)^{\intercal}\left(y_{i1}-\overline{y_1}, d_{i1}-\overline{d_1}\right)$$

$$S_{01} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i0}-\overline{y_0}, d_{i0}-\overline{d_0}\right)^{\intercal}\left(y_{i1}-\overline{y_1}, d_{i1}-\overline{d_1}\right)$$

where $\overline{y_z}$ and $\overline{d_z}$ represent the average potential outcome values for the n subjects under assignment z. The conditions require each matrix to have a limiting value as $n \to \infty$. We demonstrate this for S_0^2, evaluating each entry. As $d_{i0} = 0$ we have

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i0}-\overline{y_0}, 0\right)^{\intercal}\left(y_{i0}-\overline{y_0}, 0\right),$$

so that the (1,1) entry of S_0^2 is,

$$S_0^2(1,1) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i0}-\overline{y_0}\right)^2.$$

We know $\overline{y_0} = q_1$ and that among the n subjects there are nq_1 always-respond types with $y_{i0} = 1$. The remaining $n(1-q_1)$ subjects have $y_{i0} = 0$. Thus,

$$\begin{aligned} S_0^2(1,1) &= \frac{1}{n-1}(nq_1(1-q_1)^2 + n(1-q_1)(0-q_1)^2) \\ &= \frac{1}{n-1}n(1-q_1)(q_1(1-q_1)+q_1^2) \\ &= \frac{1}{n-1}n(1-q_1)q_1, \end{aligned}$$

which has limiting value $q_1(1-q_1)$ as $n \to \infty$. For the other entries, because $d_{i0} = 0$ for all i,

$$S_0^2(1,2) = S_0^2(2,1) = S_0^2(2,2) = 0.$$

This means S_0^2 has limiting value. By evaluating S_1^2 and S_{01} in the same manner, it can be shown they too have similar limiting values of polynomials of q_1, p_{01}, p_{10} and p_{11}.

We now apply Theorem 5 of Li and Ding (which refers to conditions of Theorem 3) which we state below in terms of the quantities defined earlier in the proof.

Theorem 5. Under the setting of Theorem 3 (a completely randomized experiment with n units, two treatments and let $\boldsymbol{r}_{iz} \in \mathbb{R}^2$ be unit i's potential outcome (for assignment z). If S_0^2, S_1^2 and S_{01} have limiting values, $t(n)/n$ has positive limiting value and $\max_{0\le z\le 1}\max_{1\le i\le n}\|\boldsymbol{r}_{iz} - \overline{\boldsymbol{r}_z}\|/n \to 0$, then $n\mathrm{Var}_{samp}(\widehat{\boldsymbol{\tau}}(\boldsymbol{A}))$ has a limiting value, denoted by $\boldsymbol{V}$, and

$$\sqrt{n}\left(\widehat{\boldsymbol{\tau}}(\boldsymbol{A}) - \boldsymbol{\tau}(\boldsymbol{A})\right) \to \mathrm{N}(0,\, \boldsymbol{V})$$

Since $0 \le y_{iz}, d_{iz} \le 1$ the last condition on $\|\mathbf{r}_{iz} - \overline{\mathbf{r}_z}\|/n$ holds and we have shown the result of the theorem. Stated in terms of $\widehat{itt}$ and $\widehat{p}_{com}$, we have shown that

$$\left((\widehat{itt},\, \widehat{p}_{com})^\intercal - (itt,\, p_{com})^\intercal\right) \to \mathrm{N}((0,0)^\intercal, \mathbf{V}) \tag{3.25}$$

To evaluate $\mathbf{V}$, we have (where we drop the subscript *samp* of $\mathrm{Var}_{samp}()$ for ease of notation)

$$\begin{aligned} \mathbf{V} &= n\mathrm{Var}(\widehat{\boldsymbol{\tau}}(\mathbf{A})) \\ &= n\mathrm{Var}\begin{pmatrix} \widehat{itt} \\ \widehat{p}_{com} \end{pmatrix} \\ &= n\begin{pmatrix} \mathrm{Var}\left(\widehat{itt}\right) & \mathrm{Cov}\left(\widehat{itt}, \widehat{p}_{com}\right) \\ \mathrm{Cov}\left(\widehat{itt}, \widehat{p}_{com}\right) & \mathrm{Var}\left(\widehat{p}_{com}\right) \end{pmatrix}, \end{aligned} \tag{3.26}$$

where

$$\begin{aligned}
\mathrm{Var}\left(\widehat{itt}\right) &= \mathrm{Var}(\widehat{p}_{10} + \widehat{p}_{11} - \widehat{q}_1) \\
&= \mathrm{Var}(\widehat{p}_{10}) + \mathrm{Var}(\widehat{p}_{11}) + \mathrm{Var}(\widehat{q}_1) + 2\mathrm{Cov}(\widehat{p}_{10}, \widehat{p}_{11}) - 2\mathrm{Cov}(\widehat{p}_{10}, \widehat{q}_1) - 2\mathrm{Cov}(\widehat{p}_{11}, \widehat{q}_1) \\
\mathrm{Var}(\widehat{p}_{com}) &= \mathrm{Var}(\widehat{p}_{01} + \widehat{p}_{11}) \\
&= \mathrm{Var}(\widehat{p}_{01}) + \mathrm{Var}(\widehat{p}_{11}) + 2\mathrm{Cov}(\widehat{p}_{01}, \widehat{p}_{11}) \\
\mathrm{Cov}\left(\widehat{itt}, \widehat{p}_{com}\right) &= \mathrm{Cov}(\widehat{p}_{10} + \widehat{p}_{11} - \widehat{q}_1,\ \widehat{p}_{01} + \widehat{p}_{11}) \\
&= \mathrm{Var}(\widehat{p}_{11}) + \mathrm{Cov}(\widehat{p}_{10}, \widehat{p}_{01}) + \mathrm{Cov}(\widehat{p}_{10}, \widehat{p}_{11}) + \mathrm{Cov}(\widehat{p}_{01}, \widehat{p}_{11}) - \mathrm{Cov}(\widehat{q}_1, \widehat{p}_{01}) \\
&\quad - \mathrm{Cov}(\widehat{q}_1, \widehat{p}_{11})
\end{aligned}$$

and the variance and covariance formulas can be found from $\mathbf{\Sigma_{samp}}$ in Equation 3.13 and then used to solve for $\mathbf{V}$ from Equation 3.26.

With the normality of $(\widehat{itt}, \widehat{p}_{com})$ established, consider $\mathrm{g}(itt, p_{com}) = itt/p_{com}$ so that

$$\mathrm{g}(itt, p_{com}) = \mathrm{h}(\mathbf{p})$$

and

$$\mathrm{g}(\widehat{itt}, \widehat{p}_{com}) = \mathrm{h}(\widehat{\mathbf{p}}),$$

where the gradient, $\nabla \mathrm{g}$, is

$$\nabla \mathrm{g}(itt,\ p_{com}) = \begin{pmatrix} \frac{\partial \mathrm{g}}{\partial itt} \\ \\ \frac{\partial \mathrm{g}}{\partial p_{com}} \end{pmatrix} = \begin{pmatrix} \frac{1}{p_{com}} \\ \\ -\frac{itt}{p_{com}^2} \end{pmatrix}.$$

It can be shown that

$$\nabla \mathrm{h}(\mathbf{p})' \, \mathbf{\Sigma_{samp}} \, \nabla \mathrm{h}(\mathbf{p}) = \nabla \mathrm{g}(itt, p_{com})^\intercal \, \mathbf{V} \, \nabla \mathrm{g}(itt, p_{com})$$

Therefore, the desired result of

$$\sqrt{n}(\mathrm{h}(\widehat{\mathbf{p}}) - \mathrm{h}(\mathbf{p})) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{h}(\mathbf{p})' \, \mathbf{\Sigma_{samp}} \, \nabla \mathrm{h}(\mathbf{p})),$$

is equivalent to showing

$$\sqrt{n}\left(\mathrm{g}(\widehat{itt},\ \widehat{p}_{com}) - \mathrm{g}(itt,\ p_{com})\right) \xrightarrow{d} \mathrm{N}(0, \nabla \mathrm{g}(itt,\ p_{com})^\intercal \, \mathbf{V} \, \nabla \mathrm{g}(itt,\ p_{com})), \tag{3.27}$$

which we may be done via the delta method.

To apply the delta method we note the asymptotic normality of $(\widehat{itt},\ \widehat{p}_{com})$ established in Equation 3.25. From $\nabla \mathrm{g}$, the partial derivatives of g() are continuous and exist around

$(itt,\, p_{com})$ as long as $p_{com} \neq 0$. This holds by the "there are some compliers" assumption of Chapter 2. Thus, we approximate $\mathrm{g}(\widehat{itt},\, \widehat{p}_{com})$ with a first order Taylor's expansion of g() around $(itt,\, p_{com})$, that is,

$$\begin{aligned}\mathrm{g}(\widehat{itt},\, \widehat{p}_{com}) = \mathrm{g}(itt,\, p_{com}) + \Big((\widehat{itt},\, \widehat{p}_{com}) - (itt,\, p_{com})\Big)\, \nabla \mathrm{g}(itt,\, p_{com}) \\ + o\left(\|\Big((\widehat{itt},\, \widehat{p}_{com}) - (itt,\, p_{com})\Big)\|\right),\end{aligned}$$

so that the asymptotic distribution of $g(\widehat{itt},\, \widehat{p}_{com})$ is given by Equation 3.27. This concludes the proof. □

Note: In this argument we present the outcome of assignment as a two-dimensional vector of the response and the treatment received. We return to this notion in the next chapter when develop a formal definition for a behavioral type.

3.8.5 Proof of Proposition 3.4.1

Proposition. *The difference in the asymptotic variances of* $\widehat{att}$ *under infinite and finite sampling has the following form.*

$$\begin{aligned}\nabla \mathrm{h}(\boldsymbol{p})' \tfrac{1}{n} \boldsymbol{\Sigma}_{pop} \nabla \mathrm{h}(\boldsymbol{p}) - \nabla \mathrm{h}(\boldsymbol{p})' \tfrac{1}{n} \boldsymbol{\Sigma}_{samp} \nabla \mathrm{h}(\boldsymbol{p}) = \frac{p_{comTR}(p_{comAR} + p_{comNR})}{n(p_{comAR} + p_{comNR} + p_{comTR})^3} \\ - \tfrac{1}{n(n-1)} \nabla \mathrm{h}(\boldsymbol{p})' \left\{(\mathbf{1} - \beta)\boldsymbol{\Sigma}_{\mathbf{pop}} + \boldsymbol{\Sigma}_{\mathbf{gap}}\right\} \nabla \mathrm{h}(\boldsymbol{p})\end{aligned}$$

Proof. Before proceeding, we restate the definitions for $\nabla \mathrm{h}(\mathbf{p})$, $\boldsymbol{\Sigma}_{\mathbf{pop}}$ and $\boldsymbol{\Sigma}_{\mathbf{gap}}$, as these are employed early in the proof.

$$\nabla \mathrm{h}(\mathbf{p}) = \begin{pmatrix} \frac{\partial h}{\partial q_1} \\ \\ \frac{\partial h}{\partial p_{01}} \\ \\ \frac{\partial h}{\partial p_{10}} \\ \\ \frac{\partial h}{\partial p_{11}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{p_{01}+p_{11}} \\ \\ \frac{q_1 - p_{10} - p_{11}}{(p_{01}+p_{11})^2} \\ \\ \frac{1}{p_{01}+p_{11}} \\ \\ \frac{q_1 + p_{01} - p_{10}}{(p_{01}+p_{11})^2} \end{pmatrix},$$

$$\mathbf{\Sigma_{pop}} = \frac{1}{\beta}\begin{pmatrix} \frac{\beta}{(1-\beta)}q_1(1-q_1) & 0 & 0 & 0 \\ 0 & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ 0 & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ 0 & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix}$$

and

$$\mathbf{\Sigma_{gap}} = \begin{pmatrix} \frac{(2\beta-1)}{(1-\beta)}q_1(1-q_1) & q_1p_{01} & -p_{10}(1-q_1) & p_{10}-q_1(1-p_{11}) \\ q_1p_{01} & 0 & 0 & 0 \\ -p_{10}(1-q_1) & 0 & 0 & 0 \\ p_{10}-q_1(1-p_{11}) & 0 & 0 & 0 \end{pmatrix}.$$

We begin by using Equation 3.16 to substitute $\mathbf{\Sigma_{samp}}$ in terms of $\mathbf{\Sigma_{pop}}$, $\mathbf{\Sigma_{gap}}$ and β.

$$\begin{aligned}
&\nabla \mathrm{h}(\mathbf{p})' \tfrac{1}{n}\mathbf{\Sigma}_{pop} \nabla \mathrm{h}(\mathbf{p}) - \nabla \mathrm{h}(\mathbf{p})' \tfrac{1}{n}\mathbf{\Sigma}_{samp} \nabla \mathrm{h}(\mathbf{p}) \\
&= \tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\mathbf{\Sigma}_{pop} - \mathbf{\Sigma}_{samp}\right] \nabla \mathrm{h}(\mathbf{p}) \\
&= \tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\mathbf{\Sigma}_{pop} - (\mathbf{1}-\beta)\mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}} - \tfrac{1}{n-1}\left\{(\mathbf{1}-\beta)\mathbf{\Sigma_{pop}} + \mathbf{\Sigma_{gap}}\right\}\right] \nabla \mathrm{h}(\mathbf{p}) \\
&= \tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\beta\mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}} - \tfrac{1}{n-1}\left\{(\mathbf{1}-\beta)\mathbf{\Sigma_{pop}} + \mathbf{\Sigma_{gap}}\right\}\right] \nabla \mathrm{h}(\mathbf{p}) \\
&= \tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\beta\mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}}\right] \nabla \mathrm{h}(\mathbf{p}) \; - \; \tfrac{1}{n(n-1)}\nabla \mathrm{h}(\mathbf{p})' \left\{(\mathbf{1}-\beta)\mathbf{\Sigma_{pop}} + \mathbf{\Sigma_{gap}}\right\} \nabla \mathrm{h}(\mathbf{p})
\end{aligned}$$

so the second term is the same as the second term we are hoping to show. Thus, to complete the proof it is enough to show

$$\tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\beta\mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}}\right] \nabla \mathrm{h}(\mathbf{p}) = \frac{p_{comTR}(p_{comAR} + p_{comNR})}{n(p_{comAR} + p_{comNR} + p_{comTR})^3}. \tag{3.28}$$

As we see from their definitions, the only entries where both a and a are nonzero is the $(1,1)$ position. This term of $\beta\mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}}$ equal

$$\frac{\beta}{1-\beta}q_1(1-q_1) - \frac{2\beta-1}{1-\beta}q_1(1-q_1) \; = \; -\frac{\beta-1}{1-\beta}q_1(1-q_1) \; = \; q_1(1-q_1).$$

Thus we return to showing Equation 3.28 holds with

$$\tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\beta \mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}}\right] \nabla \mathrm{h}(\mathbf{p})$$

$$= \tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \begin{pmatrix} q_1(1-q_1) & -q_1 p_{01} & p_{10}(1-q_1) & -p_{10} + q_1(1-p_{11}) \\ -q_1 p_{01} & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{11} \\ p_{10}(1-q_1) & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{11} \\ -p_{10} + q_1(1-p_{11}) & -p_{01}p_{11} & -p_{10}p_{11} & p_{11}(1-p_{11}) \end{pmatrix} \nabla \mathrm{h}(\mathbf{p}),$$

a quadratic form yielding a sum of ten terms. We simplify $\nabla \mathrm{h}(\mathbf{p})$ by denoting the fraction of compliers as p_c which equals $p_{01} + p_{11} = p_{com_{AR}} + p_{com_{NR}} + p_{com_{TR}}$. Also note that $q_1 - p_{10} - p_{11} = -p_{com_{TR}}$ and that $q_1 + p_{01} - p_{10} = p_{com_{AR}} + p_{com_{NR}}$. Substituting these gives

$$\nabla \mathrm{h}(\mathbf{p})' = \left(-\frac{1}{p_c}\,,\, -\frac{p_{com_{TR}}}{p_c^2}\,,\, \frac{1}{p_c}\,,\, \frac{p_{com_{AR}} + p_{com_{NR}}}{p_c^2} \right).$$

So that the quadric form written as the sum of ten terms is

$$\begin{aligned} &= \frac{1}{n\,p_c^4} \left\{ q_1(1-q_1)p_c^2 \,+\, p_{01}(1-p_{01})p_{com_{TR}}^2 \,+\, p_{10}(1-p_{10})p_c^2 \,+\, p_{11}(1-p_{11})(p_{com_{AR}} + p_{com_{NR}})^2 \right. \\ &- \, 2q_1 p_{01} p_{com_{TR}} p_c \,-\, 2p_{10}(1-q_1)p_c^2 \,-\, 2(-p_{10} + q_1(1-p_{11}))(p_{com_{AR}} + p_{com_{NR}})p_c \\ &+ \, 2p_{01}p_{10}p_{com_{TR}}p_c \,+\, 2p_{01}p_{11}p_{com_{TR}}(p_{com_{AR}} + p_{com_{NR}}) \,-\, 2p_{10}p_{11}(p_{com_{AR}} + p_{com_{NR}})p_c \left. \right\}. \end{aligned}$$

Rearranging to group like terms gives

$$\begin{aligned} &= \frac{1}{n\,p_c^4} \left\{ \underbrace{q_1(1-q_1)p_c^2 \,-\, 2p_{10}(1-q_1)p_c^2}_{(a)} \quad \underbrace{-\,2q_1 p_{01} p_{com_{TR}} p_c \,+\, 2p_{01}p_{10}p_{com_{TR}}p_c}_{(b)} \right. \\ &+ \underbrace{p_{10}(1-p_{10})p_c^2}_{(c)} \quad \underbrace{-2(-p_{10} + q_1(1-p_{11}))(p_{com_{AR}} + p_{com_{NR}})p_c \,-\, 2p_{10}p_{11}(p_{com_{AR}} + p_{com_{NR}})p_c}_{(d)} \\ &+ \, p_{01}(1-p_{01})p_{com_{TR}}^2 \,+\, p_{11}(1-p_{11})(p_{com_{AR}} + p_{com_{NR}})^2 \,+\, 2p_{01}p_{11}p_{com_{TR}}(p_{com_{AR}} + p_{com_{NR}}) \left. \right\}. \end{aligned} \tag{3.29}$$

We now simplify (a), (b), (c) and (d). First, for (a) we note that as $q_1 = p_{com_{AR}} + p_{never_{AR}}$ and $p_{never_{AR}} = p_{10}$, we have $q_1 - 2p_{10} = p_{com_{AR}} + p_{10} - 2p_{10} = p_{com_{AR}} - p_{10}$. So

$$(a) = q_1(1-q_1)p_c^2 \,-\, 2p_{10}(1-q_1)p_c^2$$

$$= (q_1 - 2p_{10})(1 - q_1)p_c^2$$
$$= (p_{com_{AR}} - p_{10})(1 - q_1)p_c^2.$$

For (b), we use the same substitution for q_1, noting that $q_1 - p_{10} = p_{com_{AR}}$.

$$\begin{aligned}(b) &= -2q_1 p_{01} p_{com_{TR}} p_c + 2p_{01} p_{10} p_{com_{TR}} p_c \\ &= -2(q_1 - p_{10}) p_{01} p_{com_{TR}} p_c \\ &= -2p_{com_{AR}} p_{01} p_{com_{TR}} p_c \\ &= -2p_{01} p_{com_{AR}} p_{com_{TR}} p_c\end{aligned}$$

Then, using $q_1 = p_{com_{AR}} + p_{10}$,

$$\begin{aligned}(a) + (b) + (c) &= (p_{com_{AR}} - p_{10})(1 - q_1)p_c^2 - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c + p_{10}(1 - p_{10})p_c^2 \\ &= (p_{com_{AR}} - p_{10})(1 - p_{com_{AR}} - p_{10})p_c^2 - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c + p_{10}(1 - p_{10})p_c^2 \\ &= \left[(p_{com_{AR}} - p_{10})(1 - p_{com_{AR}} - p_{10}) + p_{10}(1 - p_{10}) \right] p_c^2 - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c \\ &= \left[p_{com_{AR}} - p_{com_{AR}}^2 - p_{10} p_{com_{AR}} - p_{10} + p_{10} p_{com_{AR}} + p_{10}^2 + p_{10} - p_{10}^2 \right] p_c^2 \\ &\quad - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c \\ &= \left[p_{com_{AR}} - p_{com_{AR}}^2 \right] p_c^2 - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c \\ &= p_{com_{AR}}(1 - p_{com_{AR}})p_c^2 - 2p_{01} p_{com_{AR}} p_{com_{TR}} p_c\end{aligned}$$

.

Finally, in (d), we use $p_{10} - q_1 = -p_{com_{AR}}$ so that.

$$\begin{aligned}(d) &= -2(-p_{10} + q_1(1 - p_{11}))(p_{com_{AR}} + p_{com_{NR}})p_c - 2p_{10} p_{11}(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= 2(p_{10} - q_1(1 - p_{11}) - p_{10} p_{11})(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= 2(p_{10} - q_1 + q_1 p_{11} - p_{10} p_{11})(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= 2(p_{10} - q_1 - p_{11}(p_{10} - q_1)(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= 2(1 - p_{11})(p_{10} - q_1)(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= 2(1 - p_{11})(-p_{com_{AR}})(p_{com_{AR}} + p_{com_{NR}})p_c \\ &= -2(1 - p_{11})p_{com_{AR}}(p_{com_{AR}} + p_{com_{NR}})p_c\end{aligned}$$

We are ready to substitute $(a)+(b)+(c)$ and (d) into Equation 3.29. Before doing so we notice that all terms with q_1 and p_{10} have canceled. Since $p_{01} = p_{com_{NR}}$ and $p_{11} = p_{com_{AR}} + p_{com_{TR}}$,

all of the terms in Equation 3.29 consist of the fractions of complier types: $p_{com_{AR}}$, $p_{com_{NR}}$ and $p_{com_{TR}}$. For ease of notation, in this proof only, we shall denote these three quantities as A, N and T where

$$\begin{aligned} A &= p_{com_{AR}} \\ N &= p_{com_{NR}} \\ T &= p_{com_{TR}} \end{aligned}$$

so that

$$\begin{aligned} p_{01} &= N \\ p_{11} &= A + T \\ p_c &= A + N + T \end{aligned}$$

We may apply these equalities to represent Equation 3.29 in terms of just n, A, N and T. That is:

$$\begin{aligned} &\tfrac{1}{n}\nabla \mathrm{h}(\mathbf{p})' \left[\beta \mathbf{\Sigma_{pop}} - \mathbf{\Sigma_{gap}}\right] \nabla \mathrm{h}(\mathbf{p}) \\ &= \tfrac{1}{n\,(A+N+T)^4} \Big\{ \underbrace{A(1-A)(A+N+T)^2 \;-\; 2ANT(A+N+T)}_{(a)+(b)+(c)} \\ &\quad \underbrace{-\,2A(1-A-T)(A+N)(A+N+T)}_{(d)} \\ &\quad + N(1-N)T^2 \;+\; (A+T)(1-A-T)(A+N)^2 \;+\; 2NT(A+T)(A+N) \Big\}. \end{aligned}$$

Collecting the first three terms along $A(A+N+T)$ and the last two along $(A+T)(1-A-T)$ gives

$$\begin{aligned} &= \tfrac{1}{n\,(A+N+T)^4} \Big\{ [\,(1-A)(A+N+T) - 2NT - 2(1-A-T)(A+N)\,]A(A+N+T) \\ &\quad + N(1-N)T^2 \;+\; [\,(1-A-T)(A+N) + 2NT](A+T)(A+N) \Big\} \\ &= \tfrac{1}{n\,(A+N+T)^4} \Big\{ [A + N + T - A^2 - AN - AT - 2NT \\ &\quad - 2A + 2A^2 + 2AT - 2N + 2AN + 2NT]A(A+N+T) \end{aligned}$$

$$+ N(1-N)T^2 + [A - A^2 - AT + N - AN - NT + 2NT](A+T)(A+N)\Big\}$$

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{[A^2 - A + AN + AT - N + T]A(A+N+T)$$

$$+ N(1-N)T^2 + [-(A^2 - A + AN + AT - N) + NT](A+T)(A+N)\Big\}.$$

Let $\alpha = A^2 - A + AN + AT - N$ so that

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{(\alpha+T)A(A+N+T)$$

$$+ N(1-N)T^2 + (-\alpha + NT)(A+T)(A+N)\Big\}.$$

Now we expand terms to collect along $A(A+T)$.

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{(\alpha+T)A(A+T) + (\alpha+T)AN$$

$$+ N(1-N)T^2 + (-\alpha + NT)A(A+T) + (-\alpha + NT)N(A+T)\Big\}$$

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{(T+NT)A(A+T) + ANT + \alpha AN$$

$$+ N(1-N)T^2 + (-\alpha + NT)N(A+T)\Big\}$$

and collect along α

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{(T+NT)A(A+T) + ANT$$

$$+ N(1-N)T^2 + N^2T(A+T) + \alpha(AN - AN - NT)\Big\}$$

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{(T+NT)A(A+T) + ANT$$

$$+ N(1-N)T^2 + N^2T(A+T) - \alpha NT\Big\}$$

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{A^2T + A^2NT + AT^2 + ANT^2 + ANT$$

$$+ NT^2 - N^2T^2 + N^2T^2 + AN^2T - \alpha NT\Big\}$$

the N^2T^2 terms cancel and we expand the α in the last term so that

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{A^2T + A^2NT + AT^2 + ANT^2 + ANT + NT^2 + AN^2T$$

$$- (A^2 - A + AN + AT - N)NT\Big\}$$

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{A^2T + A^2NT + AT^2 + ANT^2 + ANT + NT^2 + AN^2T$$
$$- A^2NT + ANT - AN^2T - ANT^2 + N^2T\Big\}.$$

The A^2NT, AN^2T and ANT^2 terms all cancel to give

$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{A^2T + AT^2 + 2ANT + NT^2 + N^2T\Big\}$$
$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{T(A^2 + AT + 2AN + NT + N^2)\Big\}$$
$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{T[\,(A+N)^2 + AT + NT]\Big\}$$
$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{T[\,(A+N)^2 + T(A+N)\,]\Big\}$$
$$= \tfrac{1}{n\,(A+N+T)^4}\Big\{T[\,(A+N)(A+N+T)\,]\Big\}$$
$$= \frac{T(A+N)}{n\,(A+N+T)^3}$$

and returning to our original notation we have

$$= \frac{p_{comTR}(p_{comAR} + p_{comNR})}{n(p_{comAR} + p_{comNR} + p_{comTR})^3}$$

□

Chapter 4

Understanding Behavioral Types

In this chapter we provide precise definitions of the terms *behavioral type* and *restriction* introduced in the Chapter 2. These two definitions apply when treatment assigned, treatment received, and outcomes take on a discrete, finite number of values. We show that in experimental settings with these conditions we may understand each subject of the sample as belonging to one of a finite number of distinct types. As was seen in section 2.2, treatment effects are simply proportions of these distinct behavioral types. After presenting the definitions we demonstrate how the concepts are applied in three separate examples.

4.1 Formal Definitions

For an experiment, let $\mathcal{Z}$ be the space of possible assignments to treatment and $\mathcal{D}$ be the space of possible treatment received. Usually we have $\mathcal{Z} = \mathcal{D}$. Let $\mathcal{Y}$ be the space of possible outcomes. We now introduce our first formal definition.

Definition 4.1.1. A *behavioral type*, f, is simply a function, $f : \mathcal{Z} \mapsto \mathcal{D} \times \mathcal{Y}$, or $f(z) = (d, y)$.

We saw a number of examples of behavioral types in Chapter 2 such as the *comTR*, which receive the treatment assigned and vote only if assigned to the treatment group. With a finite number of assigned and received treatments it is helpful to visualize a behavioral type, such as the *comTR*, with the two dimensional array

z \ d(z)	0	1
0	0	
1		1

Moving across the first row, the treatment assignment is $z = 0$. The column in the first row with a value indicates the treatment received for $z = 0$, or $d = 0$. The value inside

of that cell represents the outcome, $y = 0$. Similarly, if $z = 1$ then $d = 1$ and $y = 1$. As before, a behavioral type describes a non-probabilistic response to a stimulus. Viewing the couplet (d, y) as a joint response makes it clear that treatment received is an outcome, not a covariate. Rosenbaum (2006) discusses this in his presentation of bivariate random outcomes. Thus, a unique behavioral type is simply a unique function of z within the space of all possible functions.

Now, let $\mathscr{F}$ be the space of all possible behavioral types, or functions f with domain $\mathcal{Z}$ and range $\mathcal{D} \times \mathcal{Y}$. There are $|\mathcal{Z}|$ possible values for z and for each z there are at most $|\mathcal{D}| \times |\mathcal{Y}|$ possible values of (d, y) where $||$ represents the number of elements in a set. Thus, there are at most $(|\mathcal{D}|\,|\mathcal{Y}|)^{|\mathcal{Z}|}$ possible functions f which make up $\mathscr{F}$. If this number is large, we may have difficulty estimating the fraction of behavioral types from a sample. This brings us to our second definition which reduces the number of types to a more manageable amount.

Definition 4.1.2. A *restriction* is a limit on the possible range of the functions in $\mathscr{F}$ in the context of a certain experiment.

A restriction reduces the set of possible behavioral types to a family $\mathscr{F}'$ where $\mathscr{F}' \subseteq \mathscr{F}$. For example, the *exclusion restriction* of section 2.1.1 was a restriction of this type. As we shall show in the next section, many of the assumptions of the last chapter can also be shown to be restrictions as defined above. With just these two basic definitions, in the context of the potential outcomes framework, one can arrive at a new perspective on an estimation problem. To demonstrate the wide applicability of these concepts and insights they provide to an estimation problem, we take a closer look at a few specific experimental protocols.

4.2 Examples

We now provide examples of three different experimental designs to show how the definitions of behavioral types and restrictions are applied. In each case we describe how restrictions apply, how this leads to a manageable number of behavioral types, and how these behavioral types relate to the treatment effects of interest. We continue to limit ourselves to binary outcome variables.

4.2.1 Experiments with a single treatment, with noncompliance to treatment assignment

In Section 2.2 we showed the setting with a control group and one level of treatment, with noncompliance, results in an estimation problem of five behavioral types. We now formalize this using the above definitions.

Here the assignment, treatment received and outcome are all binary variables so $\mathcal{Z} = \mathcal{D} = \mathcal{Y} = \{0, 1\}$. Each of the two values for z has 2^2 possible values of $f(z) = (d, y)$ so there are a total of $(2^2)^2 = 16$ members of $\mathscr{F}$. These are shown in Figure 4.1 below.

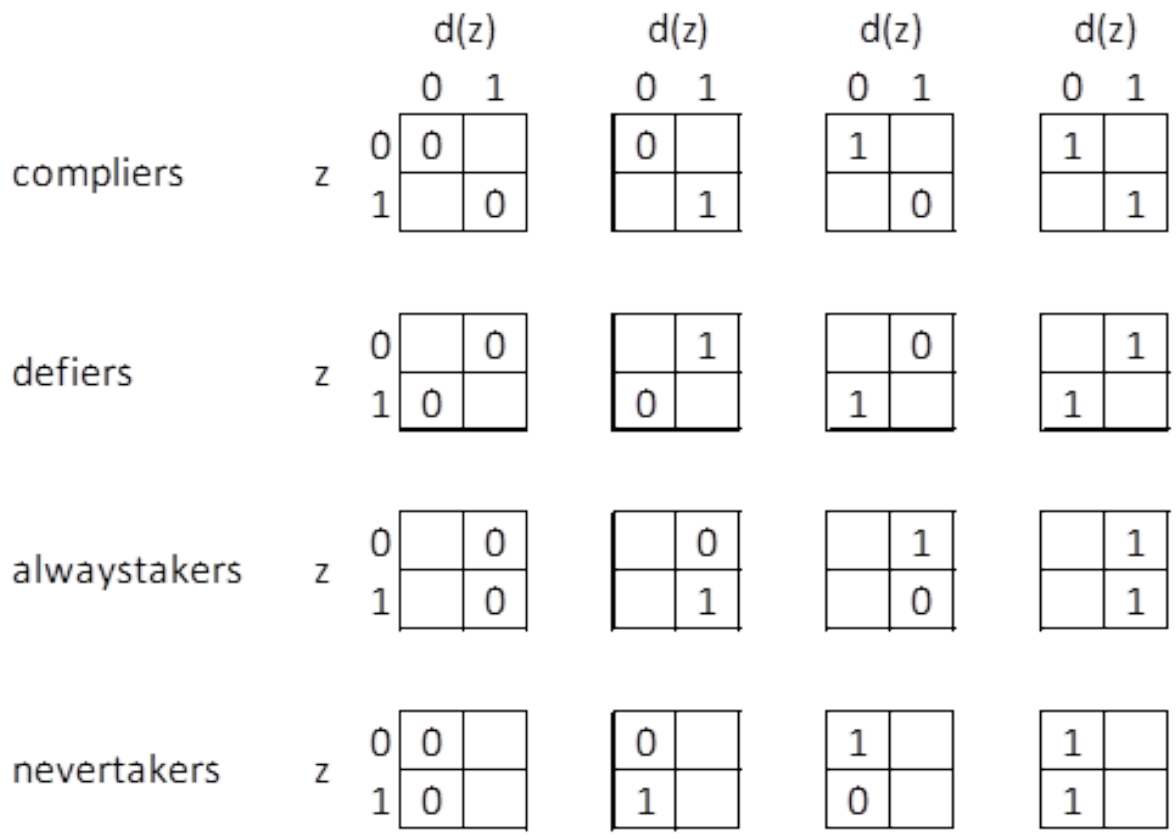


Figure 4.1: The 16 possible behavioral types in $\mathscr{F}$ with one level of treatment and noncompliance.

We group each of the four compliance types introduced in Section 2.1 together. The the compliers, where $d(z) = z$, are followed by the defiers with $d(z) = 1 - z$. The alwaystakers have $d(z) = 1$ and nevertakers are last with $d(z) = 0$.

We now see how assumptions 2, 3, 4 and 6 of chapter 2 act as restrictions leading to five behavioral types in $\mathscr{F}'$. Assumption 2, the exlusion restriction, requires outcomes to depend on treatment received rather than assignment, graphically, means that we cannot have any functions with different values of y within the same column. This eliminates the middle two behavioral types of the alwaystakers and nevertakers in Figure 4.1. Assumption 3, no defiers, rules out the defier types. Assumption 4, no alwaystakers, excludes the two remaining alwaystakers. And finally, Assumption 6, monotonicity, removes the complier behavioral type in the third column, which responds only when no treatment is received. The five remaining behavioral types in $\mathscr{F}'$ are the $comAR$, $comNR$, $comTR$, $nevAR$, and $nevNR$ of section 2.2. We summarize this in Figure 4.2 below.

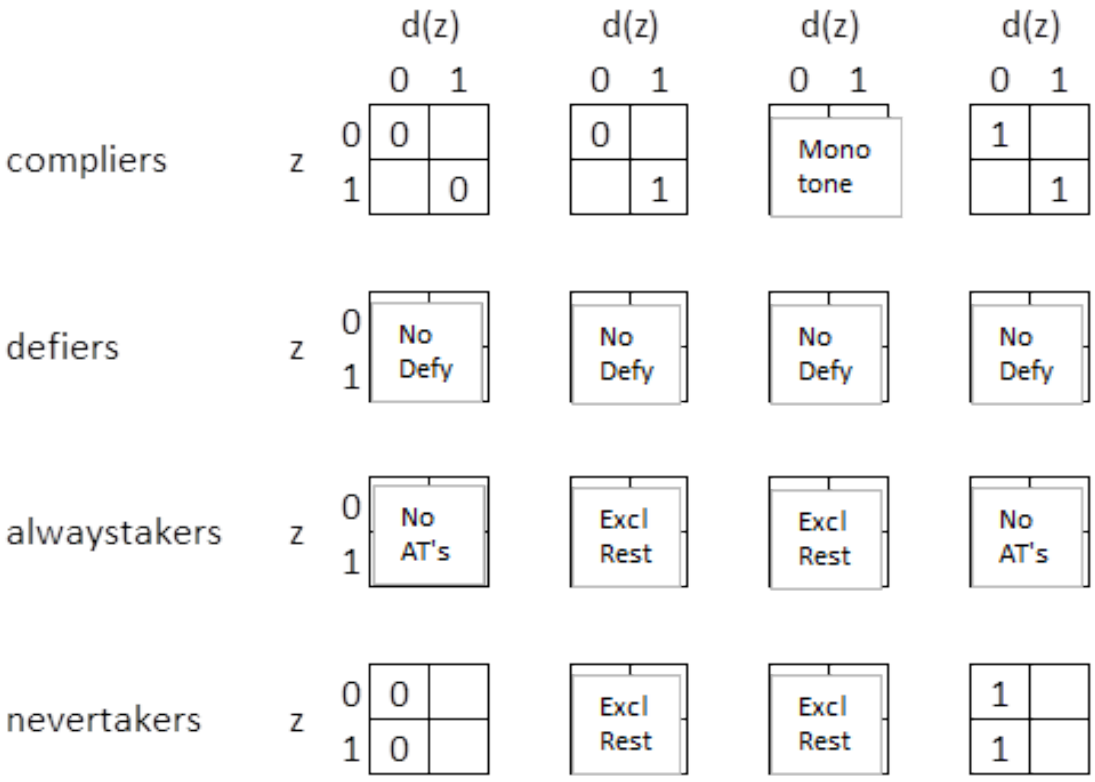


Figure 4.2: How the assumptions of Section 2.1 and 2.2 are viewed as restrictions. The labels indicate how each assumption eliminates certain behavioral types from Figure 4.1.

Section 2.2 lays out how the behavioral types appear in the observed control and treatment data, how treatment effects of interest are thought of as proportions of the different behavioral types and how they are estimated from the data. Again, this is a commonly used design which has been used in a number of applications, particularly in estimating the impact of get-out-the-vote campaigns.

4.2.2 k levels of ordered treatment, with unknown compliance

We now turn to experiments where subjects are assigned to a control group or to one of k levels of *ordered treatment* in the sense that we know ahead of time which treatments are more likely to lead to a positive response of the subjects. We begin by assuming compliance to treatment is unknown (though it's also applies for perfect compliance) and the effects we estimate are attributed to assignment.

Distinct Behavioral Types

Here $\mathcal{Z} = \{0, 1, ..., k\}$ where 0 denotes the control, 1 the weakest treatment and k the strongest treatment. Since the range of f is $\mathcal{Y} = \{0, 1\}$, $\mathscr{F}$ contains 2^{k+1} distinct behavioral types. We reduce the number of behavioral types substantially with the restriction of *monotonicity* introduced in section 2.2.3. As the range of f is only one variable, the outcome, the assumption states that for any two treatments z_1 and z_2 such that $z_1 \leq z_2$, we have $f(z_1) \leq f(z_2)$. As we see in Figure 4.3 below, this restriction implies there are only $k + 2$ different behavioral types in $\mathscr{F}'$

z	0	1	2	...	k-1	k
f(z)	1	1	1	1	1	1

z	0	1	2	...	k-1	k
f(z)	0	1	1	1	1	1

z	0	1	2	...	k-1	k
f(z)	0	0	1	1	1	1

...

z	0	1	2	...	k-1	k
f(z)	0	0	0	0	1	1

z	0	1	2	...	k-1	k
f(z)	0	0	0	0	0	1

z	0	1	2	...	k-1	k
f(z)	0	0	0	0	0	0

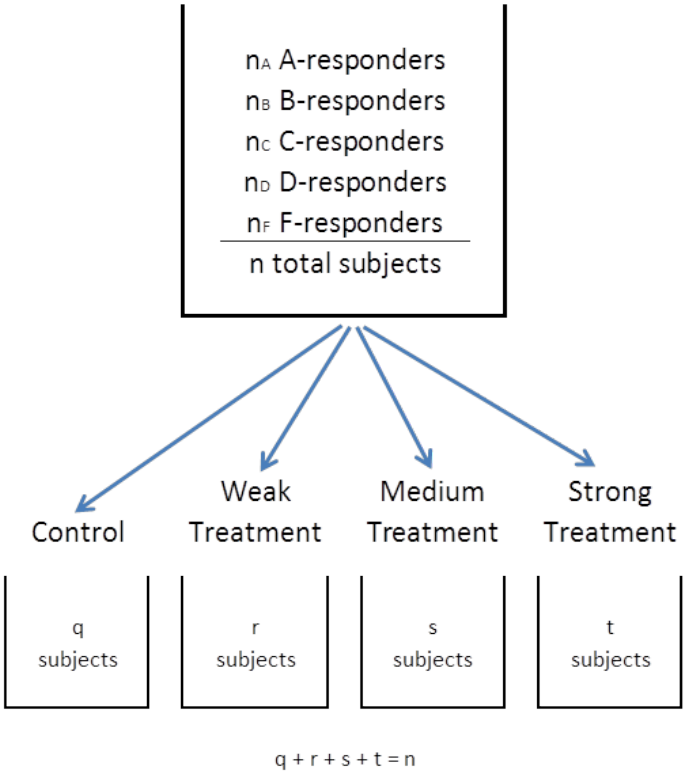
Figure 4.3: Unique behavioral types when treatment is ordered and we assume the Monotonicity.

In this design, every behavioral type is associated with a threshold level of treatment which must be met to trigger a response. The top row has the *always responds* and the bottom row has the *never responds* similar to the behavioral types described in 2.2.

Sampling Model and Observed Data

For ease of exposition we demonstrate this with $k = 3$ though our description holds for an arbitrary value of k.

Let $\mathcal{Z}$ be {control, *weak*, *medium*, *strong* } z As before $z = 0$ denotes the control and $z = 1, 2, 3$ correspond to the *weak*, *medium* and *strong* treatments, respectively. We have five behavioral types. The *A-responders* always respond. *B-responders* respond if assigned to any of the treatments but not to the control. *C-responders* respond only if assigned to the medium or strong treatment. *D-responders* only respond to the strong treatment. And *F-responders* never respond. As was discussed in section **??**, we assume the subjects in the sample are randomly assigned from a finite population. We can imagine data being generated by what Freedman et al. (1998) call a box model with n_A, n_B, n_C, n_D and n_F tickets for each behavioral type such that $n_A + n_B + n_C + n_D + n_F = n$. A ticket for each type resembles the representations in Figure 4.3 with the value of the outcome for each of the treatment assignments. Let q, r, s and t be the number of subjects assigned to the control and weak, medium and strong treatments, respectively so that $q + r + s + t = n$. Returning to the box model framework, we randomly draw from the box of tickets, without replacement, the appropriate number of tickets assigned to each treatment level.



Let Q_A, Q_B, Q_C, Q_D and Q_F denote the random number of each behavioral type that appear in the control group. Similarly, define $R_A, ..., R_F$ for those assigned to the weak treatment, $S_A, ..., S_F$ for the medium treatment and $T_A, ..., T_F$ for the strong treatment.

The eight observed data points are

Q_A, those in the control who respond.
$Q_{BCDF} \equiv Q_B + Q_C + Q_D + Q_F$, those in the control who don't respond.
$R_{AB} \equiv R_A + R_B$, those in the weak treatment who respond.
$R_{CDF} \equiv R_C + R_D + R_F$, those in the weak treatment who don't respond.
$S_{ABC} \equiv S_A + S_B + S_C$, those in the medium treatment who respond.
$S_{DF} \equiv S_D + S_F$, those in the medium treatment who don't respond.
$T_{ABCD} \equiv T_A + T_B + T_C + T_D$, those in the strong treatment who respond.
T_F, those in the strong treatment who don't respond.

And the table of observed values is the following:

	Control	Weak Treatment	Medium Treatment	Strong Treatment
No response	Q_{BCDF}	R_{CDF}	S_{DF}	T_F
Responds	Q_A	R_{AB}	S_{ABC}	T_{ABCD}
Total	q	r	s	t

Parameter Estimation

Since q, r, s and t are known by the experimental design, the model has four parameters n_A, n_B, n_C and n_D or, more conveniently, $p_A = \frac{n_A}{n}, ..., p_D = \frac{n_D}{n}$ (where $p_F = 1 - p_A - p_B - p_C - p_D$). Individually, each of Q_A, R_{AB}, S_{ABC} and T_{ABCD} will follow hypergeometric distributions with $\mathbb{E}(Q_A) = q\,p_A$, $\mathbb{E}(R_{AB}) = r(p_A + p_B)$, etc. We have the following unbiased estimates of the parameters:

$$\widehat{p}_A = \frac{Q_A}{q}$$
$$\widehat{p}_B = \frac{R_{AB}}{r} - \widehat{p}_A$$
$$\widehat{p}_C = \frac{S_{ABC}}{s} - \widehat{p}_A - \widehat{p}_B$$
$$\widehat{p}_D = \frac{T_{ABCD}}{t} - \widehat{p}_A - \widehat{p}_B - \widehat{p}_C$$

Treatment Effects

Using the attributable effects notation from Section 3.6, where treatment received is not relevant, let w_{iz} be potential outcome of the response of subject i if assigned to one of the four treatments z. Again, Let Z denote the random value of treatment assignment so that

the random response is $W_i = \sum_{z=0}^{3} w_{iz} \mathbb{1}(Z = z)$. Then the effect of treatment z, beyond the control, is

$$effect_z \equiv \mathbb{E}(W \mid Z = z) - \mathbb{E}(W \mid Z = 0) = \frac{1}{n}\sum_{i=1}^{n} w_{iz} - \frac{1}{n}\sum_{i=1}^{n} w_{i0}$$

So the treatment effects of interest may be understood as the fractions of behavioral types in the group of n subjects. That is,

- $effect_3 = p_B + p_C + p_D$, the increase in the response rate due to the strong treatment
- $effect_2 = p_B + p_C$, the increase in the response rate due to the medium treatment
- $effect_1 = p_B$, the increase in the response rate due to the weak treatment, which is also the marginal increase in the response rate from the control to the weak treatment
- p_C is the marginal increase in the response rate from the weak to the medium treatment
- p_D is the marginal increase in the response rate from the medium to the strong treatment

Applications

This design is commonly used when the treatment(s) are received in the mail and treatment received is not known, resulting in the measurement of intention-to-treat effects. Cotterill, John, and Richardson (2010) conduct such an experiment with letters promising increasing levels of rewards for participation in a book donation drive. And Gerber, Green, and Larimer (2008), Sinclair, McConnell, and Green (2012) and Citrin, Green, and Levy (2014) study the impacts of various mailers in GOTV campaigns. We examine the Gerber, Cotterill and Sinclair experiments in Chapter 5.

4.2.3 GOTV experiment in households with two voters, allowing for noncompliance

For our last example, we consider get-out-the-vote experiments consisting of households with exactly two voters, who may or may not comply with the assigned treatment. Unlike the first two examples, which are more general, we now describe a specific design motivated for a real-world application. Though not as broadly applicable, we include this experimental setup because it highlights a number of interesting features of using behavioral types such as how restrictions whittle thousands of behavioral types to a workable amount and how

complications may arise in parameter estimation. Also, as it turns out, this example yields interpretations into *interference* or indirect, *spill over* effects, the ability of a treatment to influence the outcomes of others who interact with the subject. This topic has recently received considerable attention in the statistics literature (for recent reviews see VanderWeele, Tchetgen, and Halloran, 2014; Aronow, Samii, et al., 2017).

Here, we assume the campaign employs only one form of outreach (e.g. an in-person visit) and may try to reach either of the two voters with individual outreach efforts. While the campaign views both voters in the same manner, we allow only one of the voters be a subject in the experiment. From the point of view of a subject in the study, they may be assigned to the control group or to one of three different ordered treatments. This is analogous to example 4.2.2 so we use similar terminology. The strong treatment occurs when both the subject and the other voter receive the outreach (we call this level of treatment *both*). Under this treatment, the subject is encouraged to vote directly by the outreach, and also indirectly if the attempt to contact the other voter in the household is successful. The medium treatment corresponds with only the subject receiving the treatment themselves which we denote *self*. And the weak treatment is when the outreach effort is only directed at the other voter, so that any increase in voting turnout is due to interactions between the two voters which we shall refer to as *other*.

In practice, an experimental design which designates only one of the two voters as part of the experiment when the other is receiving treatments also, might seem uneconomical as we are squandering half of the voters by keeping them outside of the experiment. In this example we do this to meet the SUTVA assumption from Chapter 2.

Distinct Behavioral Types

As before, we begin with determining the size of $\mathscr{F}$. The assigned treatments $\mathcal{Z}$={control, other, self, both}, ordered from least to greatest. The space of received treatments, $\mathcal{D}$, is the same as $\mathcal{Z}$ and $\mathcal{Y}$ is the voting outcome space of {0,1}. We may represent a behavioral type with the following four-by-four array.

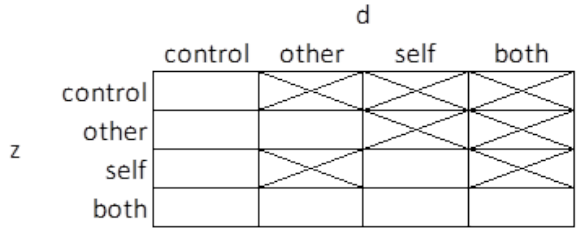
z \ d	control	other	self	both
control				
other				
self				
both				

For each value of z, there are $4 \times 2 = 8$ possible values of $f(z) = (d, y)$. With 4 values of z there are $8^4 = 4096$ possible distinct functions f, or distinct behavioral types in $\mathscr{F}$. Fortunately, there are four restrictions we may employ. These are similar to the assumptions

found in section 4.2.1 but are extended to more levels of treatment. The first two restrictions pertain to treatment received, d, while the last two limit the outcome values, y.

Restriction 1. *A voter cannot receive a GOTV contact if the campaign does not attempt to reach them.*

This applies equally to the subjects and the other voters so that, for example, a subject who is assigned $self$ may only receive the $self$ or *control* treatment depending on whether the subject complies. Since the other voter is not be contacted there is no way for the subject to receive *both* or *other*. Restriction 1 is similar to Assumptions 3 and 4 of Chapter 2 , there are no defiers or alwaystakers. In terms of our visual representation, it states that certain values of d are not possible for $f(z)$. We show the restriction below by crossing out the columns of d that are no longer valid for each value of z.



With this restriction, for $z = control$ the only possible values for (d,y) are (*control*,0) and (*control*,1), or 2 possible values. Similarly for $z = other$ or $z = self$ there are 4 possible values for (d,y) while for $z = both$ there are $2^4 = 16$ possible values. Thus the restriction has reduced the total number of possible functions to $2 \times 4 \times 4 \times 16 = 512$ different behavioral types in $\mathscr{F}'$.

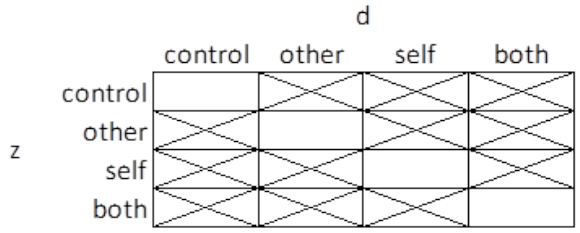
Restriction 2. *Each voter within the same household complies with the assigned individual treatment without regard to the assignment or treatment received of the other voter.*

Let us first be clear that this should not be confused with SUTVA, as the "units" applies to subjects which are, by our design, voters in different households. Also SUTVA addresses outcome values while this restriction addresses compliance. But the restriction touches upon a similar notion. Recall that campaigns, in this setting, conduct outreach directly to the individual. This restriction states that whether we are referring to the subject or to the other voter, whether one received the treatment does not rely on the outreach to or treatment received of the second voter. Or, in math, the behavioral type function $f_i()$ for voter i is only a function of z_i and does not depend on the z, d, or y value for the second voter.

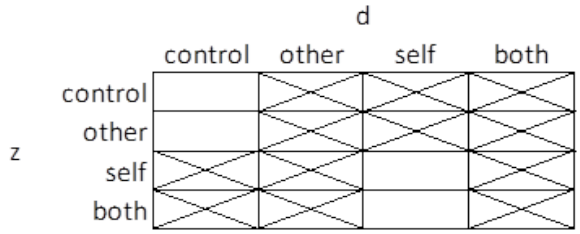
Relating to the visual representation, once the value of d is known for the bottom row, when $z = both$, we know the received value for all other rows. Because there are four possible

receipt values to being assigned *both* there are four possible "compliance clusters" for the behavioral types. For example, if a subject is assigned *both*, they receive the outreach while the other voter does not comply, so the subject receives the treatment of $self$. This implies that if assigned $self$, the subject complies with the outreach and the other voter is not contacted so the subject again receives the treatment $self$. Along the same lines if assigned *other* the subject receives *control*. Of course, if assigned *control* the subject receives *control* by Restriction 1.

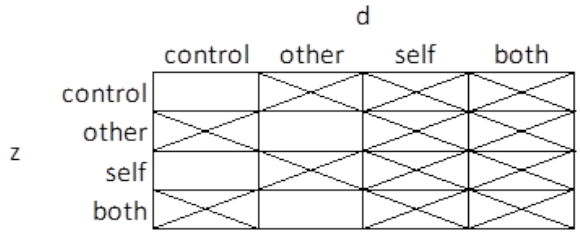
If $d = both$ when $z = both$ then both the subject and the other voter comply perfectly with the individual outreach, so that $d = z$ for all values of z. We call this cluster of behavioral types *perfect compliers*. Their four-by-four array is:



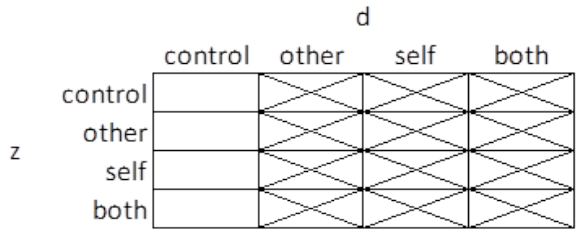
If $d = self$ when $z = both$ this indicates the subject complies with the outreach but the other voter does not and we refer to this cluster of behavioral types as *self compliers*, with the following four-by-four representation.



If $d = other$ when $z = both$, only the other voter complies with treatment and these *other compliers* have arrays of this form.



And *nevertakers* with $d = control$ for every value of z have the simplest array structure, with only one column available, as below.



As seen from the visual representation, we have completely determined the value of d, in the output, (d, y), of $f()$. Each of the compliance clusters has 4 unfilled boxes which contain a value for y, either 0 or 1. Therefore, each of the four compliance clusters has $2^4 = 16$ possibilities for the range, or this restriction has reduced the number of behavioral types in $\mathscr{F}'$ to $4 \times 16 = 64$.

One noteworthy feature about Restriction 2 is that we can, to an extent, test the assumption. Taking the subjects who are assigned *both* we can observe the fraction who are perfect or self compliers as they are the subjects complying with the outreach, receiving treatments of *both* or $self$. If Restriction 2 holds, this fraction should equal the fraction of perfect or self compliers who are assigned to $self$ with any difference between being due to chance. If we observe a significant difference between the two fractions it implies that outreach to one of the voters impacts the likelihood of the other voter in the household complying with the outreach.

As mentioned earlier, the first two restrictions only limited the treatment received, d. The next two restrictions are concerned with y. In fact, both are assumptions we have seen before.

Restriction 3. *The outcome, y, only depends on the treatment received, not the treatment assigned.*

This is the same as Assumption 2 in Chapter 2. As we have discussed, graphically, this means that the values of y, within a column, must be the same. So perfect compliers with four columns available to y still have $2^4 = 16$ possible behavioral types. Self compliers and other compliers, with only two columns open, each must have $2^2 = 4$ possible behavioral types. And nevertakers only have 1 column open to them which must be all 1's or all 0's so there are 2 types. This now gives a total of $16 + 4 + 4 + 2 = 26$ possible behavioral types in $\mathscr{F}'$.

Restriction 4. *Monotonicity.*

We use this in the same manner as described in section 4.2.2 to describe voters as A,B,C,D or F voters where treatment now depends on what is actually received. A-voters always

vote. B-voters vote under any of the three treatments, but not in *control*. C-voters vote if they receive *both* or $self$ but not if they receive *other* or *control*. D-voters only vote if the receive *both*. F-voters never vote. Graphically, as we move from left to right a voting response only occurs if d meets or exceeds a threshold. We see how this is realized for each of the four compliance clusters.

For perfect compliers there are a total of 5 behavioral types, one for each A,B,C,D and F voter: perfect complier A-voter ($perfect_A$), perfect complier B-voter ($perfect_B$), perfect complier C-voter ($perfect_C$), perfect complier D-voter ($perfect_D$), and perfect complier F-voter ($perfect_F$). These are shown below.

perfect $_A$ z

z \ d	control	other	self	both
control	1			
other		1		
self			1	
both				1

perfect $_D$ z

z \ d	control	other	self	both
control	0			
other		0		
self			0	
both				1

perfect $_B$ z

z \ d	control	other	self	both
control	0			
other		1		
self			1	
both				1

perfect $_F$ z

z \ d	control	other	self	both
control	0			
other		0		
self			0	
both				0

perfect $_C$ z

z \ d	control	other	self	both
control	0			
other		0		
self			1	
both				1

Self compliers only receive 2 of the possible assignments, $self$ and *control*, so B-voters act just like C-voters and D-voters act just like F-voters, resulting 3 distinct behavioral types of voters which we call: self complier A-voter ($self_A$), self complier B or C-voter ($self_{BC}$), and the self complier D or F-voter ($self_{DF}$). Each representation is shown below.

self $_A$ z

z \ d	control	other	self	both
control	1			
other	1			
self			1	
both			1	

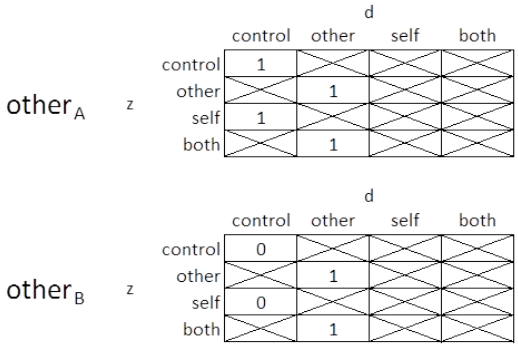
self $_{DF}$ z

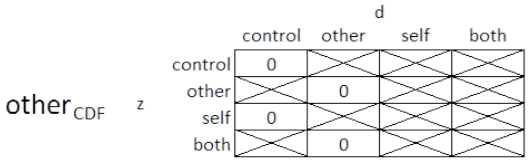
z \ d	control	other	self	both
control	0			
other	0			
self			0	
both			0	

self $_{BC}$ z

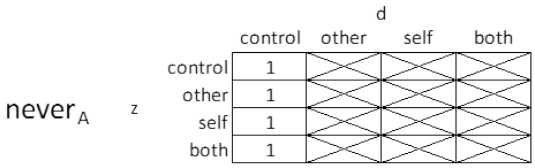
z \ d	control	other	self	both
control	0			
other	0			
self			1	
both			1	

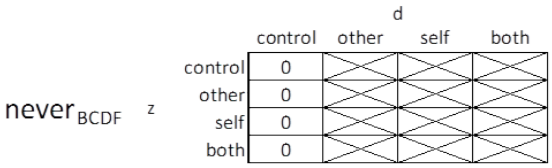
Other compliers, also only receive 2 of the possible assignments, *other* or *control*. In this cluster, C, D and F-voters have identical behavior so there are 3 types of voters: other complier A-voter ($other_A$), other complier, B-voter ($other_B$) and the other complier, C, D or F-voter ($other_{CDF}$). These are shown below.





Nevertakers, never receive any of the three treatments so consist of 2 types of voters: A voters ($never_A$) , and B, C, D, F voters, which are indistinguishable from each other, form their own behavioral type ($never_{BCDF}$).





After beginning with thousands of possible functions in $\mathscr{F}$, we have reduced $\mathscr{F}'$ to just $5 + 3 + 3 + 2 = 13$ behavioral types! This makes the estimation problem much more manageable.

Finally, we emphasize the key idea that while the treatment received of the subject does not depend on the other voter, *the behavioral type for the subject is a function for the treatment received of both the subject and the other voter in the household*. This subtlety may not be readily apparent but it should be clear from this discussion that it underlies how we are able to conceptualize the behavioral types and restrictions in this setting with two voters.

Sampling Model and Observed Data

Similar to the example from 4.2.2 we have a model represented with n total tickets with a certain number belonging to of each of 13 distinct behavioral types. As in 4.2.2 there are 4 possible assignments. The number drawn to the *control*, *other*, $self$ and *both* are fixed in advance as q, r, s and t, respectively such that $q+r+s+t = n$. We show this in Figure 4.4.

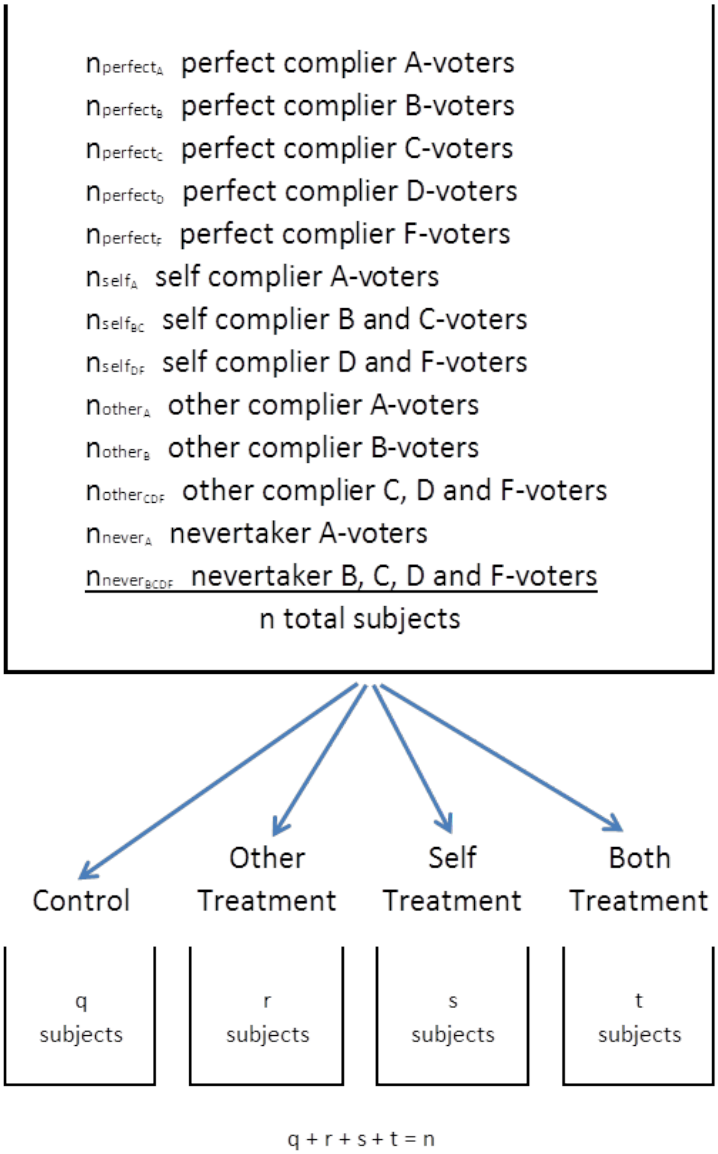


Figure 4.4: Sampling model for GOTV experiment with two-voter households allowing for noncompliance.

As we can observe how each subject complies with the assigned treatment, observed data appears as an extension of the table shown in section 2.2. The treatment received possibilities depend on the treatment assignment so that *self* and *other* treatments only have two possible received treatments while the *both* assignment has all the possible received treatments. Staying with the notation of 4.2.2 we let the Q_{yd} represent the random total of observations of the subjects assigned to the control who respond with voting value y and receive treatment d, R_{yd} for those assigned to the *other* treatment, S_{yd} for *self* and T_{yd} for

both. The observed results appear as this two by nine table.

Treat Assigned	Control	Other		Self		Both			
Treat Received	Control	Control	Other	Control	Self	Control	Other	Self	Both
Didn't Vote	$Q_{0\,control}$	$R_{0\,control}$	$R_{0\,other}$	$S_{0\,control}$	$S_{0\,self}$	$T_{0\,control}$	$T_{0\,other}$	$T_{0\,self}$	$T_{0\,both}$
Voted	$Q_{1\,control}$	$R_{1\,control}$	$R_{1\,other}$	$S_{1\,control}$	$S_{1\,self}$	$T_{1\,control}$	$T_{1\,other}$	$T_{1\,self}$	$T_{1\,both}$
Total	q	r		s		t			

And we can now place the 13 behavioral types according to which cell they appear in the table of observed data.

Treat Assigned	Control	Other		Self		Both			
Treat Received	Control	Control	Other	Control	Self	Control	Other	Self	Both
Didn't Vote	$perfect_F$		$perfect_F$		$perfect_F$				$perfect_F$
	$self_{DF}$	$self_{DF}$			$self_{DF}$			$self_{DF}$	
	$other_{CDF}$		$other_{CDF}$	$other_{CDF}$			$other_{CDF}$		
	$never_{BCDF}$	$never_{BCDF}$		$never_{BCDF}$		$never_{BCDF}$			
	$perfect_D$		$perfect_D$		$perfect_D$				
	$perfect_C$		$perfect_C$						
	$self_{BC}$	$self_{BC}$							
	$perfect_B$								
	$other_B$								
Voted									$perfect_D$
					$perfect_C$				$perfect_C$
					$self_{BC}$			$self_{BC}$	
			$perfect_B$		$perfect_B$				$perfect_B$
			$other_B$	$other_B$			$other_B$		
	$perfect_A$		$perfect_A$		$perfect_A$				$perfect_A$
	$self_A$	$self_A$			$self_A$			$self_A$	
	$other_A$		$other_A$	$other_A$			$other_A$		
	$never_A$	$never_A$		$never_A$		$never_A$			

Parameter Estimation

The assignment totals q, r, s and t are fixed, so the table of observed results has 1 degree of freedom for the subjects assigned to *control*, 3 degrees of freedom each for those assigned to *other* and $self$ and 7 degrees of freedom for those assigned to *both*. The $1 + 3 + 3 + 7 = 14$ degrees of freedom for the 12 unknown parameters leads to an overdetermined system of linear equations. This allows for multiple estimators of the same parameter, which may be combined to reduce variance. In this section we do not delve into choosing optimal estimators but simply show the parameters can indeed be determined from the observations.

We begin with the five behavioral types which appear exclusively in their own cells of the observed data.

$$\begin{aligned}
\widehat{p}_{perfect_F} &= T_{0\,both}/t \\
\widehat{p}_{self_{DF}} &= T_{0\,self}/t \\
\widehat{p}_{other_{CDF}} &= T_{0\,other}/t \\
\widehat{p}_{never_{BCDF}} &= T_{0\,control}/t \\
\widehat{p}_{never_A} &= T_{1\,control}/t
\end{aligned}$$

We build on these immediate estimates with

$$\widehat{p}_{perfect_D} = S_{0\,self}/s - \widehat{p}_{perfect_F} - \widehat{p}_{self_{DF}}$$

and use successive estimates in the equations for the remaining parameters such as the following.

$$\begin{aligned}
\widehat{p}_{perfect_C} &= R_{0\,other}/r - \widehat{p}_{perfect_D} - \widehat{p}_{other_{CDF}} - \widehat{p}_{perfect_F} \\
\widehat{p}_{self_{BC}} &= R_{0\,control}/r - \widehat{p}_{never_{BCDF}} - \widehat{p}_{self_{DF}} \\
\widehat{p}_{self_A} &= T_{1\,self}/t - \widehat{p}_{self_{BC}}
\end{aligned}$$

For the last estimate we could also use $\widehat{p}_{self_A} = R_{1\,control}/r - \widehat{p}_{never_A}$ or some combination of the two. As the linear equations are overdetermined, the final four parameters may be solved by a number of different systems of linear equations such as

$$\begin{aligned}
\widehat{p}_{perfect_B} + \widehat{p}_{other_B} + \widehat{p}_{perfect_A} + \widehat{p}_{other_A} &= R_{0\,control}/r \\
\widehat{p}_{perfect_B} + \widehat{p}_{perfect_A} &= T_{1\,both}/t - \widehat{p}_{perfect_D} - \widehat{p}_{perfect_C} \\
\widehat{p}_{perfect_A} + \widehat{p}_{other_A} &= Q_{1\,control}/q - \widehat{p}_{self_A} - \widehat{p}_{never_A} \\
\widehat{p}_{perfect_B} + \widehat{p}_{other_B} &= Q_{0\,control}/q - \widehat{p}_{self_{BC}} - \widehat{p}_{perfect_C} - \ldots - \widehat{p}_{perfect_F}
\end{aligned}$$

where the quantities on the right hand side of the equations are observations from the data or parameter estimates that have already been solved.

Spillover Treatment Effects

There are numerous treatment effects which we could estimate, including *att* and *itt* for each of the three levels of treatment. Just as statisticians, political scientists have also shown interest in spillover effects, or, in the context of GOTV experiments, the impact of turn out on the subject, when the other voter is contacted. See Sinclair, Rogowski, Bass,

Harrington, et al. (2011) for an overview of the GOTV research. By our definitions, the $perfect_B$ and $other_B$ behavioral types turn out to vote when only the other voter in the household is called. Thus, for the experimental sample we have

$$itt_{spillover} = \frac{n_{perfect_B} + n_{other_B}}{n}.$$

As the perfect compliers and self compliers are the subjects who receive the treatment we may also estimate the *att* for the indirect treatment as

$$att_{spillover} = \frac{n_{perfect_B} + n_{other_B}}{n_{perfect\,complier} + n_{self\,complier}}.$$

Applications

We know of no applications of this specific experimental design. However, as mentioned earlier, this example is motivated by the possibility of carrying out such an experiment, based on the authors work with modest sized GOTV drives. In such campaigns, when the outreach capabilities are limited, campaign workers may only be able to attempt contact with a fraction of the targeted voters. The design of this example requires that a large portion of the voting population not be contacted and thus may be a useful experiment for campaigns with limited resources that are interested in measuring spillover effects.

4.3 Discussion

In this chapter, we provide precise definitions for the terms *behavioral type* and *restriction* which may be applied to any experiment where the possible treatment assigned, the possible treatments received, and the responses are all categorical. We show that while the initial number of distinct behavioral types may be large, the restrictions can provide enough limitations to the treatment received and response variables to reduce the number of types to an estimable amount. Through three illustrative examples we show how the model parameters, and the proportion of the behavioral types in the experimental sample, may be identified from the contingency table summarizing the observed results. For each of the three designs we demonstrate how commonly measured treatment effects for many experiments with binary outcomes may be expressed as functions of the proportion of behavioral types. For the remainder of this thesis we concern ourselves with inference about these proportions via testing, confidence intervals, and multi-parameter confidence regions.

Chapter 5

Applications for Single Parameter Inference

In this chapter we apply the behavioral-types approach to four social science field experiments with multiple degrees of treatment. We begin with an experiment that follows the example in Section 4.2.2, having k degrees of ordered treatment and unknown compliance, which leads to fairly unambiguous findings. We follow this with another experiment in the mold of example 4.2.2 but the conclusions we draw are slightly different from those of the authors. In our third application we tackle a more complicated design and our approach yields much stronger evidence of a spillover treatment effect. The fourth application is a recent experiment tracking compliance and though we don't have the data in hand, we outline how it might be evaluated. The experiment also differs from the first three as the treatments are partially ordered, departing from the restriction of strict monotonicity. For each application we describe the experimental design with an overview of the author's methods and conclusions. We then conduct our own analysis from the point of view of behavioral types and compare the results. The fifth section summarizes our general approach and also highlights the nuances specific to each application. We postpone most of the detailed variance calculations until the sixth section. In this chapter we only concern ourselves with single parameter inference. We discuss confidence regions for multi-parameter inference in Chapter 6.

5.1 Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

Motivated to understand how social pressure can increase voter participation, Gerber et al. (2008) design a series of four GOTV postcards which progressively escalate the degree of social pressure exerted to encourage voting. The mildest of the treatments begins "*DO YOUR CIVIC DUTY–VOTE*" and contains a short message appealing to the voters sense

of social obligation. The same sentence is also repeated in the three other types of mailers. The pressure increases in the second type of postcard which emphasizes "*YOU ARE BEING STUDIED!*", noting that researchers will be monitoring voter's participation via public records (the authors refer to this as the "Hawthorne" treatment). The "Self" treatment includes a similar message about voters being studied but leads with "*WHO VOTES IS PUBLIC INFORMATION*" and boosts the social pressure by printing the names of each of the voters within the household and whether they had voted in the prior two elections, including an empty box for the upcoming election, which the researchers claim they will fill in and resend after votes have been tallied. The strongest pressure is exerted by the "Neighbors" treatment which spurs voters with "*WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?*" and lists the voting history of the household members followed by the voting history of their neighbors, again with a blank box for the upcoming election and a promise to send an updated mailing. That is, hinting that one's neighbors will learn whether they voted. The overlapping nature of the messaging, keeping elements of weaker treatments in the stronger treatments, is in accord with an ordered series of treatments and supports the monotonicity assumption that a voter motivated to vote by a weaker treatment is also motivated by the stronger ones.

The study was conducted in Michigan prior to the August 2006 primary, where most electoral activity focused on two key races to decide the Republican nominees for state governor and a United States senator. Researchers removed individuals from the subject pool who were unlikely to be impacted by the treatments, such as those with bad addresses, likely Democratic registrants and those who did not vote in the high turnout November 2004 election. Additionally, subjects were grouped into blocks of 18 neighboring households where, within each, 10 were assigned to control and two were assigned to each of the four treatments. Certain households were removed that did not fit into the design, such as those in sparsely populated areas where voters may not know their neighbors, or all those in apartment buildings as neighboring households could not be clearly identified. The resulting sample contained approximately 100,000 households in the control group and 20,000 in each of the four treatment groups. In total 180,002 households representing 344,084 voters were included in the study.

Analysis and Conclusions by Experimenters

We argue, shortly, that in order to apply the behavioral-types approach directly, it is necessary to restrict ourselves to single voter households. However, for comparison, we begin with the author's analysis for voters in all households. Here are the main observations of the experiment.

	Experimental Group				
	Control	Civic Duty	Being Studied (Hawthorne)	Self Vote History	Neighbors Vote History
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
Number of Individuals	191,243	38,218	38,204	38,218	38,201

Table 5.1: Voting rates of the control and four postcard treatments in the Social Pressure experiment.

Table 5.1 indicates that each of the treatments correspond with a higher rate of voting than the control, and perhaps even a significant difference from each other. In order to estimate the effects of the mailers, the authors perform a linear regression with an indicator of voting as the response variable and the assigned treatment as the independent variable of interest. They specify three separate models two of which employ covariates of prior voting history and block identifiers to check the robustness of the treatment assignment estimates. Though the analysis is performed at the individual level, most subjects live in households with more than one voter. Gerber, Green and Larimer account for the dependency by using clustered standard errors, with each household as a cluster. Table 5.2 contains the results of their analysis. The first row shows the estimates from direct calculation of the observations, i.e., subtracting the voting rate of the treatment minus that of the control. The second and third row show the linear regression estimates and standard errors using the specification without covariates. The results are nearly identical when the covariates are included.

	Treatment			
	Civic Duty	Being Studied (Hawthorne)	Self Vote History	Neighbors Vote History
Effects, Direct Calculation	1.8%	2.5%	4.8%	8.1%
Effects, Linear Regression	1.8%	2.6%	4.9%	8.1%
SE, Linear Regression	0.3%	0.3%	0.3%	0.3%

Table 5.2: Effects estimated from Social Pressure experiment.

The author's conclude, with strong evidence, that each of the effects are not only significantly different from zero but also large enough in magnitude to be a cost effective method for getting out the vote.

Replicating Experimenters Results for One-voter Households

As we saw from example 4.2.2, modeling under the potential outcomes framework becomes more elaborate when there are two or more voters within the same household, as the outcome

of one voter may depend on the treatment assignment of another. For now, we avoid this complication entirely by limiting our analysis to households with only one voter (in Section 5.3 we tackle the issue directly by addressing the spillover effect between voters from the same household). The individual level data of the subjects, analysis code, and other documentation relevant to the experiment are available in the Data Archive of Yale University's Institution for Social and Policy Studies which houses the data for many recent Get Out The Vote field experiments published in the literature. We simply drop the subjects from multi-voter households from the records and repeat the authors' analysis using their code. The luxury of such a large sample size, which still includes 47,836 single-voter households, allows us to do this and still retain meaningful findings. Table 5.3, below, shows the voting rates of single voter households under the different treatments.

	Experimental Group				
	Control	Civic Duty	Being Studied (Hawthorne)	Self Vote History	Neighbors Vote History
Percentage Voting	33.1%	35.4%	37.0%	40.0%	42.3%
Number of Individuals	26,481	5,398	5,281	5,310	5,364

Table 5.3: Voting rates for those in households with only one voter.

Comparing Tables 5.1 and 5.3 we see a similar pattern of substantial increases in voter turnout accompanying the increased degree of social pressure. Subjects in one-voter households appear to vote at rates 3 to 5 percentage points higher than that of the overall sample, perhaps due to characteristics of one-voter households which correlate with a higher tendency to vote. The estimated effects are also stronger, which is seen more clearly when we repeat the regression analysis in Table 5.4 below. The overall treatment effects are higher than what was observed in Table 5.2 but with the smaller sample size, the standard errors are much larger. Our next task is to reproduce Table 5.4 from the perspective of behavioral types.

Analysis with Behavioral Types for One-voter Households

We now follow the blueprint laid out in Example 4.2.2 when there are k treatments of ordered intensity with unknown compliance. Here we have a control group and $k = 4$ ordered treatments. Given the messaging of the mailers, the assumption of monotonicity seems reasonable. That is, there are $k + 2 = 6$ distinct behavioral types, which we refer to as A, B, C, D, E and *F-voters*. Their voting outcomes are shown in Table 5.5.

	Treatment			
	Civic Duty	Being Studied (Hawthorne)	Self Vote History	Neighbors Vote History
Effects, Direct Calculation	2.32%	3.92%	6.92%	9.20%
Effects, Linear Regression	2.32%	3.92%	6.91%	9.20%
SE, linear regression	0.71%	0.72%	0.72%	0.72%

Table 5.4: Effects estimated from Social Pressure experiment, restricted to households with one voter.

Treatment	Votes	Doesn't Vote
Control	A-voters	B,C,D,E,F-voters
Civic Duty	A,B-voters	C,D,E,F-voters
Being Studied	A,B,C-voters	D,E,F-voters
Self Vote History	A,B,C,D-voters	E,F-voters
Neighbors Vote History	A,B,C,D,E-voters	F-voters

Table 5.5: Response of behavioral types in the Social Pressure experiment.

Still within the framework of 4.2.2 the effects from the first two rows of Table 5.4 are the sample percentages of certain behavioral types. The estimate of the "Civic Duty" effect is the estimated percentage of B-voters ($\widehat{p_B}$), while the estimated "Neighbors" effect is the percentage of B, C, D and E-voters ($\widehat{p_B} + \widehat{p_C} + \widehat{p_D} + \widehat{p_E}$). The parameters are estimated as described in 4.2.2 and shown in Table 5.6. Their standard errors result from a variance calculation similar to that of Section 3.2, the details of which are found at the end of the chapter in Section 5.6.

Parameter	p_A	p_B	p_C	p_D	p_E	p_F
Estimate	33.06%	2.32%	1.60%	3.0%	2.28%	57.74%
SE	0.19%	0.67%	0.90%	0.92%	0.92%	0.64%

Table 5.6: Parameter estimates when applying the behavioral-types approach to one-voter households in the Social Pressure experiment.

As was described in the example from Section 4.2.2, within each treatment group the total number of votes follows a hypergeometric distribution, with a covariance structure across the treatment groups. Here, the finite population size is 47,836 and the number of draws, the number within each treatment, is at least 5,000. Though the distribution is discrete, with the observed response rates in Table 5.3 between 33%–43%, and more than 1,000 counts in

each cell of the contingency table of treatment and outcomes, the observed total votes is well approximated by a multivariate normal distribution, as are the parameter estimates which are linear combinations of the observed totals. Table 5.6 indicates that, with the exception of $\widehat{p_C}$, the parameter estimates are highly significant. Table 5.7 gives the estimates of the treatment effects using the behavioral-types approach. These are comparable to the second and third line of Table 5.4, with slightly smaller standard errors.

	Treatment			
	Civic Duty	Being Studied (Hawthorne)	Self Vote History	Neighbors Vote History
Effects, Behavioral Types	2.32%	3.92%	6.92%	9.20%
SE, Behavioral Types	0.68%	0.69%	0.69%	0.69%

Table 5.7: Effects estimated from Social Pressure experiment, restricted to households with 1 voter.

In summary, our analysis matches well to that of Gerber, Green and Larimer, indicating substantial and highly significant findings but with a different perspective on whether the measurement is of a treatment effect or a percentage of behavioral types in the sample. The conclusions are quite strong in this experiment due to the differences in the response rates to the different treatments. The remaining applications exhibit how the behavioral-types approach compares when the results are more ambiguous.

5.2 The Impact of a Pledge Request and the Promise of Publicity: A Randomized Controlled Trial of Charitable Donations

In this application Cotterill, John, and Richardson (2013) examine how making a commitment of a charitable donation and having that act recognized publicly, impacts the likelihood a household donates. Asking households to make a pledge is a tactic commonly used by charitable or civic engagement organizations to encourage donations or engage in civic action, and as found in the Social Pressure experiment, publicly recognizing acts deemed as beneficial to one's community can strongly influence behavior.

Set in the town of Manchester, United Kingdom, investigators collaborated with a local charity collecting books for school libraries in South Africa during a week-long book donation drive in 2010. The treatment assignments consist of a control and two treatments to examine different sorts of enticement. The control group received an initial letter informing

them of the book drive and asking for donations. The "Pledge" group received the same letter with an additional sentence "*Please pledge to donate a second hand book (by postcard, email or phone)*" and also included an addressed pledge postcard for the subjects to return. The letter in the stronger treatment,"Pledge and Publicity", began with a letter identical to that of the "Pledge" treatment, with the same pledge card, but tacked on one more sentence "*A list of everyone who donates a book will be displayed locally.*" For all subjects, a follow-up letter was sent four weeks later reminding them of the book collection and indicating the drop off locations. As with the initial letter, the follow up letter for the "Pledge" group contained an additional request for a pledge, while the "Pledge and Publicity" group added a final message promising to publish the names of those who donate. Additionally, households who made a pledge were thanked, and reminded of their commitment. Thus, the complete treatment effect is due to both letters, with the content of the second letter conditional on the response to the first letter. The second mailing also included a plastic bag to use for donated books. Each bag contained a unique identifier number allowing the researchers to track who had given.

To compile a list of subject households, researchers identified all residential addresses in two electoral wards one of which was relatively poor while the other was relatively wealthy. Households were approximately split between the two wards. The 11,812 households chosen for the study were then randomly divided into three nearly equal treatment assignments with 3,937 to control, 3,937 to "Pledge" and 3,938 to "Pledge and Publicity". In the second mailing, residents were informed of six locations, three in each ward, where books could be dropped off. Books left without a book bag or any other identifier were recorded as an anonymous donation.

Analysis and Conclusions by Experimenters

The main observations of the experiment are shown below, copied from Table 2 of the article.

	Experimental Group		
	Control	Pledge	Pledge and Publicity
Percentage Donating	7.3%	8.2%	8.9%
Standard Error	0.42%	0.44%	0.45%
Number of Individuals	3,937	3,937	3,938

Table 5.8: Effects estimated from Cotterill et al. experiment.

The first row shows the response rate increases monotonically with the strength of the treatment. Though not stated explicitly, standard errors are in line with a sampling model where subjects are chosen from an infinite population so that the number of donations in each of

the three groups are deemed as independent. The authors begin their analysis by conducting two-tailed z-tests and conclude that the "Pledge" group does not differ significantly from the control (p-value over .05) while the "Pledge and Publicity" group does (p-value below .05). In order to control for various demographic variables between neighborhoods, such as age and income distribution, the authors perform a complementary log-log regression, a form of binary regression with link function such that $\Pr(donate) = 1-\exp[-\exp(\sum_{j=1}^{p} x_j\eta_j)]$ where η_j represents the coefficients of the covariates x_j. Descriptions of the model can be found in (McCullagh and Nelder, 1989, p. 108). The regression analysis supports the findings that the "Pledge and Publicity" treatment has a significant effect, with p-value below .05, while the pledge only group does not.

Analysis with Behavioral Types

As with the single voter households in the Social Pressure experiment, we proceed along the lines of Example 4.2.2 with a control group and $k = 2$ degrees of ordered treatment. Given the layered arrangement of the letters, the assumption of monotonicity seems reasonable. There are $k + 2 = 4$ distinct behavioral types which we refer to as A, B, C and *F-donors*. Their donation outcomes follow this scheme.

Treatment	Donates	Doesn't Donate
Control	A-donors	B,C,F-donors
Pledge	A,B-donors	C,F-donors
Pledge and Publicity	A,B,C-donors	F-donors

In this setting, the donation rates of Table 5.8 are the sample percentages of certain behavioral types. The "Pledge" effect is the percentage of B-donors (p_B), while the "Pledge and Publicity" effect is the percentage of B and C-donors ($p_B + p_C$). The parameters and effects are shown in Table 5.9 and the calculation of the standard error is again postponed to the end of this chapter.

Parameter (Effect)	p_A	p_B (Pledge)	p_C	p_F	$p_B + p_C$ (Pledge and Publicity)
Estimate	7.3%	0.9%	0.7%	91.9%	1.6%
SE	0.34%	0.55%	0.57%	0.37%	0.56%

Table 5.9: Parameter estimates when applying the behavioral-types approach to the Cotterill et al. experiment. The treatment effect of the "Pledge" is p_B while the effect of "Pledge and Publicity" is $p_B + p_C$.

Here, neither of the estimates for the main parameters of interest, $\widehat{p_B}$ and $\widehat{p_C}$, are significantly different from zero but their sum, $\widehat{p_B} + \widehat{p_C}$ certainly appears to be so. Assuming normality,

a null hypothesis claiming the "Pledge and publicity" effect is zero would be rejected with a p-value of 0.002, which appears to be a much stronger conclusion than those found by the authors of the study (though part of the difference may be due to the assumption of dependency among the treatment groups). However, employing these point estimates and standard errors does not necessarily yield an accurately sized confidence region. In Chapter 6 we discuss how, when parameter values are close to zero, as they are here, confidence regions and p-values computed from these standard errors may be somewhat too small.

As an alternative, we test that hypothesis that $p_B = p_C = 0$ with the Fisher Exact test. If we combine the results for the two treatments, under the null, there are only A and F-donors; one type donates and the other does not, regardless of assignment. Any difference observed in the rate of donating between the control and combined treatments is due to the random allocation of the two behavioral types to the treatment groups. Under the null hypothesis, the number donating in the control will follow a hypergeometric distribution. The exact test results in a p-value of 0.018, substantial evidence against the null. Overall, we agree with the authors, that the effect of the "Pledge and Publicity" is significant. But from the perspective of behavioral types the conclusion is more nuanced: there is strong evidence that either B or C-donors exist, but it's not clear if both or just one of the two is found in the sample.

5.3 Detecting Spillover Effects: Design and Analysis of Multilevel Experiments

Motivated to understand how spillover effects impact voter turnout, Sinclair et al. (2012) design an elaborate GOTV mailer experiment to explore how voters within the same household, or residing on the same block, influence each other. Unlike the previous two experiments we explored, where treatment was assigned to a household, here treatment is assigned at the individual level by directly addressing one voter in a household. The authors mail the "Self" treatment from the Social Pressure experiment, which begins with "*WHO VOTES IS PUBLIC INFORMATION*" with only the name and vote history of the addressee. Treatment assignments are randomized at three different levels. First, nine-digit zip codes are assigned to one of four groups: control, one household in zip mailed, half of households mailed, or all of households mailed. Second, the appropriate number of households within each zip code are assigned to receive the mailer. Third, in households with more than one voter, one is randomly selected to be the addressee.

We will describe the exact assignment procedure in more detail below but first we establish the treatments effects which can be estimated from the various treatment assignments. From the perspective of an individual subject in the study, they can be assigned to one of the following nine possible experimental conditions.

1. Control, with no mailings sent to anyone in their zip code
2. One household in zip mailed, not to the subject's household
3. One household in zip mailed, sent to other voter in subject's household
4. One household in zip mailed, sent to the subject
5. Half of households in zip mailed, not to the subject's household.
6. Half of households in zip mailed, sent to other voter in subject's household
7. Half of households in zip mailed, sent to the subject
8. All of households in zip mailed, sent to other voter in subject's household
9. All of households in zip mailed, sent to the subject

By design, individuals in single-voter households cannot be assigned to the third, sixth or eight treatment. For our purposes, we argue that this order represents a series of ordered treatments, the higher numbered corresponding with the stronger treatments.

The design allows us to measure a number of treatment effects. For example, the spillover effect between voters in the same household, conditional on no neighbors mailed, is found by subtracting the voting rate of those in assignment 1 from those in assignment 3. Similarly, subtracting the response rate of those assigned to treatment 5 from those assigned to treatment 6 also provides a within-household spillover effect, but in the context of half of one's neighbors receiving the mailing. And the spillover effect from half of one's neighbors receiving the mailing is found by subtracting the voting rate of group 1 from group 5.

Sinclair et al. apply this design to a GOTV outreach during the April 2009 special election in the Illinois 5th Congressional District. The previous elected official, Rahm Emanuel, had resigned to become the White House Chief of Staff and few resources were invested in the campaign by any political party as a lopsided Democratic victory was anticipated. The non-competitive nature of the race made it an attractive setting for an empirical application as voting outcomes were unlikely to be obscured by other forms of campaign outreach. Eligible voters for the study were those who had registered before 2006. Eligible households consisted of those with between one and three eligible voters. And eligible zip codes needed to have at least two eligible two-voter households. In total, there were 71,127 eligible individuals, 47,851 eligible households and 4,897 eligible zip codes.

The assignment to treatment group proceeded as follows. First, a two-voter household from each zip code was chosen at random to be what was called the *core* household. Then one quarter of zip codes were assigned to a one-mailer zip code, one quarter of zips to half of households treated, one quarter to all households treated and one quarter were assigned to control where no households were sent the mailer. Next, the appropriate number of houses in the zip code were randomly selected to receive the postcard. Core households not assigned

to control were always sent the mailer so in zip codes assigned to the 1-mailer treatment, the core household is the lone one treated. In zip codes where half the households receive the mailer, half of the non-core households received the mailer. If the number of non-core households was odd, the number receiving treatment was rounded down (so in zip codes with seven or eight households the core and three non-core households, or four in all, are treated). Finally, one voter within the house was selected as the subject. The procedure for assignment to treatment is summarized in Figure 5.1.

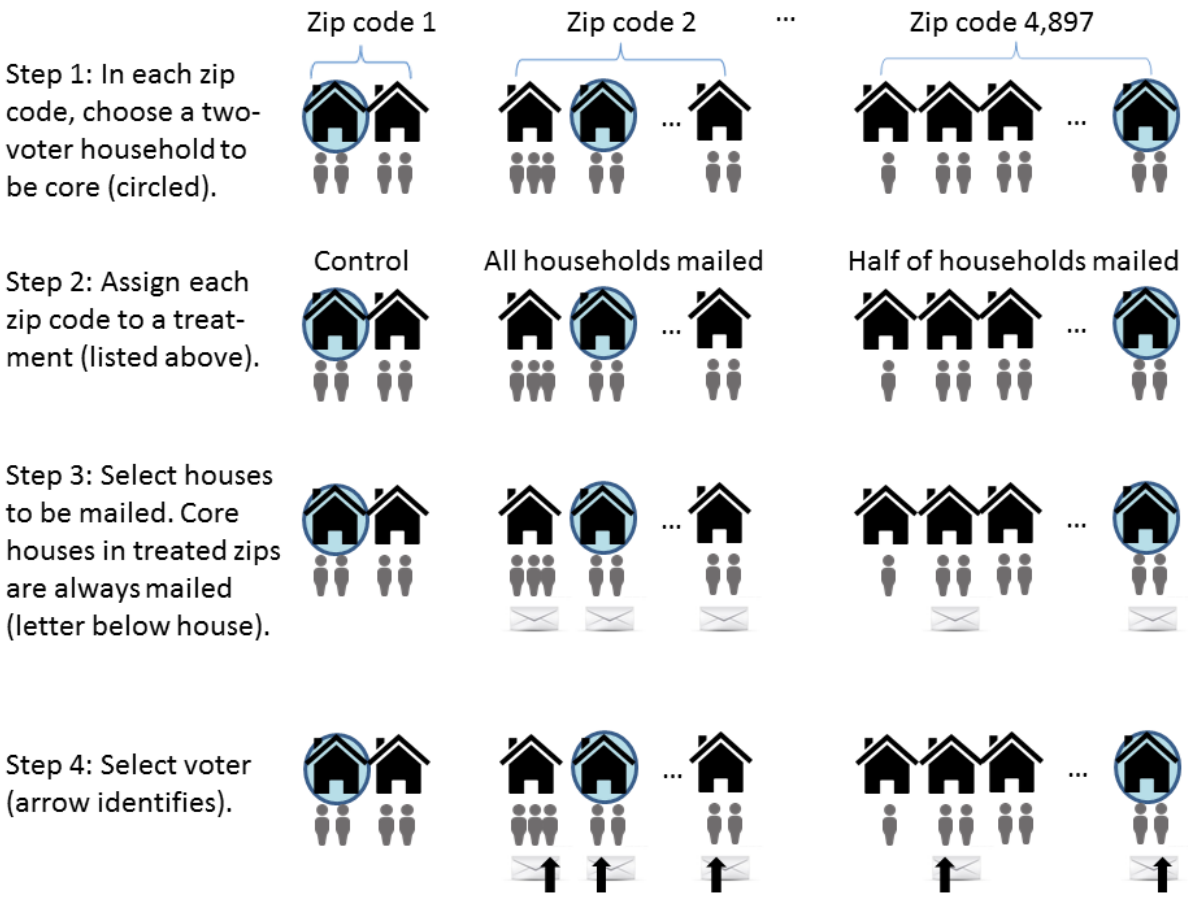


Figure 5.1: Randomization to treatment procedure in Sinclair et al. experiment.

After the election, public records were used to track the voting outcomes. The authors were able to determine the responses of 64,445 of the subjects, about 90%, of the original sample.

Analysis and Conclusions by Experimenters

The authors complete three separate analyses for the one, two and three-voter households. We focus on just the two-voter households which, as we shall soon demonstrate, have a sample sufficiently large enough to yield interesting results when viewed from the perspective of behavioral types. For the two-voter households, Sinclair et al. use the linear regression model

$$\begin{aligned} Y_i = \alpha + \beta \mathbb{1}(\text{Mailed Directly})_i + \gamma \mathbb{1}(\text{Mailed to Other Voter in Household})_i \\ + \delta_1 \mathbb{1}(\text{1 Other HH, Not Sent to Subject})_i + \delta_2 \mathbb{1}(\text{Half HH})_i + \delta_3 \mathbb{1}(\text{All HH})_i \\ + \epsilon \mathbb{1}(\text{Core HH})_i + \sum_{j=1}^{m} \zeta_j \mathbb{1}(\text{Zip Configuration } j)_i + u_i \end{aligned} \tag{5.1}$$

where the first five indicator covariates jointly provide an economical representation of the nine separate treatments. For example, the second treatment, where one household in the zip code is mailed, not addressed to the subject's household would have

$$\mathbb{1}(\text{1 Other HH, Not Sent to Subject}) = 1$$

with the other four treatment indicators set to 0 while the fourth treatment, where one household in the zip code is mailed, addressed to the subject would have

$$\mathbb{1}(\text{Mailed Directly}) = 1$$

with the other four treatment indicators set to 0. The binary variable for core household was included to account for the core and non-core assigned to treatment with different probabilities. The $\mathbb{1}(\text{Zip Configuration 1}), ..., \mathbb{1}(\text{Zip Configuration } m)$ represent m possible zip code configurations. Each configuration j represents one of the m possible configurations of the number one, two and three-voter households in the zip code. For example, all individuals living in zip codes with exactly four one-voter households, five two-voter households and one three-voter household would have the same values for $\mathbb{1}(\text{Zip Configuration 1}), ..., \mathbb{1}(\text{Zip Configuration } m)$. This controls for the impact of different household sizes as one could imagine zips with just a few households could have a different tendency to vote than ones with many households. And as we saw in Section 5.1 when we restricted ourselves to single voter households, voting rates do vary by household size. The coefficients $\zeta_1, ..., \zeta_m$ represent a control for each configuration. There were nearly 300 unique configurations. Additionally, the authors duplicate the model with and without indicator variables of individual turnout for the 2008, 2006, 2004, 2002 and 2000 November elections.

The authors arrive at the following estimates for treatment and spillover effects. We show the estimates without the controls for vote history which, as we have seen in the two previous applications, has little change on the estimated coefficients.

Treatment	Estimate	95% Confidence Interval
Mailed Directly (β)	3.6%	(1.9%, 5.2%)
Mailed to Other Voter in Household (γ)	1.2%	(-0.5%, 3.0%)
Mailed to One Other Household in Zip (δ_1)	0.5%	(-1.6%, 2.8%)
Mailed to Half of Households in Zip (δ_2)	0.6%	(-1.3%, 2.6%)
Mailed to All Households in Zip (δ_3)	1.0%	(-1.2%, 3.0%)

Table 5.10: Regression estimates of treatment and spillover effects in the Sinclair et al. experiment for two-voter households.

Table 5.10 indicates strong evidence for a direct impact of being sent the postcard, near significant evidence for a within-household spillover effect and little evidence for spillover from neighbors. Confidence intervals were computed via a bootstrap. The authors conclude a strong significant direct effect confirms the Social Pressure experiment's findings of the impact of the "Self" mailer and that there is just below significant evidence of a within-household indirect effect. We focus on the within-household spillover effect and show that a behavioral-types approach, while it does not provide a confidence interval, can employ hypothesis testing to provide even stronger evidence of the indirect spillover effect.

Analysis with Behavioral Types

We begin by showing the different behavioral types. Again our primary interest lies in the spillover effect. First, for simplicity, and to increase power, we group all subjects in households that were not mailed a postcard to form the "Not mailed to house" treatment group. Thus we disregard any effects that may arise due to one, half, or all of neighbors being mailed and consider only three ordered treatments: "Not Mailed to Household", "Mailed to Other Voter in Household" and "Mailed Directly". Assuming monotonicity, we can again follow the blueprint of Example 4.2.2, which gives rise to four behavioral types of voters: *A*, *B*, *C* and *F-voters*. Their voting outcomes are as follows.

Treatment	Votes	Doesn't Vote
Not Mailed to Household	A-voters	B,C,F-voters
Mailed to Other Voter in Household	A,B-voters	C,F-voters
Mailed Directly	A,B,C-voters	F-voters

Here, the percent of B-voters represents the within household spillover effect. As with the Social Pressure experiment, the individual level data is available from the Data Archive of Yale University's Institution for Social and Policy Studies which we use to calculate the voting rates for our aggregated treatment groups. Due to the complex randomization process we omit standard errors, for now but show the voting rates in Table 5.11.

	Experimental Group		
	Not Mailed to Household	Mailed to Other Voter in House	Mailed Directly
Percentage Voting	22.8%	24.4%	26.7%
Number of Individuals	15,632	7,194	7,268

Table 5.11: Response rates for the aggregated treatment groups for two-voter households.

This leads to the following parameter estimates in Table 5.12.

Parameter	p_A	p_B	p_C	p_F
Estimate	22.8%	1.6%	2.3%	73.3%

Table 5.12: Parameter estimates for aggregated treatment groups for two-voter households.

In terms of the treatment effects, the impact of the Mailed Directly treatment is the percent of B and C-voters, or β of Equation 5.1 is comparable to $p_B + p_C$. Comparing to the estimated treatment effects of Table 5.10, the linear model yields $\widehat{\beta} = 3.6\%$ while our approach gives $\widehat{p}_B + \widehat{p}_C = 3.9\%$. Continuing the comparison of treatment effects, the spillover effect in Table 5.10 of $\widehat{\gamma} = 1.2\%$, is the percent of B-voters, and Table 5.12 puts this slightly higher at 1.6%. Computing the standard errors, however, is not as straightforward as in the first two examples of the chapter. First, individuals receiving the "Mailed Directly" and "Mailed to Other Voter in Household" treatments live in the same address which violates the SUTVA (which was the motivation for restricting our analysis to one voter households in the Social Pressure experiment).

In fact, a closer look at the random quantities involved in the calculation of $\widehat{p}_B$ brings to light the complications in computing a standard error for $\widehat{p}_B$. To see this more clearly we represent Table 5.11 in terms of the random totals appearing in a contingency table of results. We use the notation of Example 4.2.2 where the subjects who vote under the "Not Mailed to Household" treatment, the A-voters, are denoted by Q_A. The total who don't vote if assigned this treatment, the B, C and F voters are Q_{BCF}. Similarly, the voters and non-voters assigned to "Mailed to Other Voter in Household" will be R_{AB} and R_{CF}, respectively. In the "Mailed Directly" treatment we observe S_{ABC} voters and S_F non-voters. The contingency table of counts by treatment and voting outcome appears as Table 5.13 where Q, R and S are now random quantities that depend on the zip code level assignment and the configuration of household sizes on the block. We estimate the percent of *B-voters*, the statistic of interest, with

$$\widehat{p}_B = \frac{R_{AB}}{R} - \frac{Q_A}{Q}.$$

	Experimental Group		
	Not Mailed to Household	Mailed to Other Voter in House	Mailed Directly
Didn't Vote	Q_{BCF}	R_{CF}	S_F
Voted	Q_A	R_{AB}	S_{ABC}
Total	Q	R	S

Table 5.13: Random variables within the contingency table of counts by treatment and voting outcomes.

but the covariance of R_{AB}/R and Q_A/Q is much harder to evaluate, as each house receiving the mailer has two voters with dependent outcomes. To fully proceed with the calculation of standard errors, and incorporate this dependency between the two voters of the household, we cannot view the behavioral types from the perspective of an individual, we must view them as a particular type of household consisting of a pair of voters. Each *household behavioral type* depends on the outcomes of the two voters. For example, we denote a household behavioral type as AC if it consists of one A-voter and one C-voter, the order of the pair is irrelevant. Thus there are 10 possible unordered pairs or 10 unique household behavioral types: AA, AB, AC, AF, BB, BC, BF, CC, CF and FF. Finding the proportion of each, $p_{AA}, ..., p_{FF}$, is the estimation goal.

If we were to follow the procedure of the first two applications, once the proportions of household behavioral types is known, this can be combined with the sampling design to solve the covariance structure of $(Q_A, Q_{BCF}, R_{AB}, R_{CF}, S_{ABC}, S_F)$ to determine the variance of $\widehat{p}_B$. Here, however, we encounter an identifiability issue as there are nine parameters (since the proportion of each of the 10 behavioral types must sum to 1) but, it can be shown, there are only seven possible estimating equations. The authors do not encounter this identifiability problem as the direct impact of a mailer, or the spillover effect from the other in the household are viewed as additive effects of a linear model. Complicating the calculation of the variance even more, because of the randomizing scheme, assignment is not analogous to pulling households randomly from an infinite or finite population. Assignment is intertwined with the household configuration of each nine-digit zip code as, for example, core households are assigned to treatment with different probability and households residing in zip codes with an odd number of houses are more likely to receive the treatment than if in a zip with an even number of houses. In conclusion, though we can estimate $\widehat{p}_B$ there is no clear way to calculate it's variance for in this experimental design.

Testing the Hypothesis of No Spillover Effect

Fortunately, we can take an easier path to assessing the significance of $\widehat{p}_B$ via hypothesis testing. Under the null hypothesis of no spillover effect, that is, no B-voters, the individual behavioral types act accordingly:

Treatment	Votes	Doesn't Vote
Not Mailed to Household	*A-voters*	*~~B~~,C,F-voters*
Mailed to Other Voter in Household	*A,~~B~~-voters*	*C,F-voters*
Mailed Directly	*A,~~B~~,C-voters*	*F-voters*

so that R_{AB}/R and Q_A/Q represent the fraction of A-voters in the control and weaker treatment and their difference, the test statistic $\widehat{p}_B$, varies around zero. We restrict ourselves to the two-voter households with complete data, where neither vote outcomes are missing. This gives 28,182 voters in 14,091 households. The observed $\widehat{p}_B$, from the complete data sample is still 1.6%, as it was in Table 5.12, and we shall use it as a test statistic, computing a p-value by simulating it's distribution under the null hypothesis.

However, simulating the distribution of $\widehat{p}_B$, assuming no B-voters requires additional assumptions about the A-voters within two-voter households. Under the null hypothesis of no B-voters, the distributions of R_{AB} and Q_A, and thus $\widehat{p}_B$, only depend on three aggregated groupings of the 10 household behavioral types: households with two A-voters (AA), households with one A-voter and households with no A-voters. Stated another way, in order to simulate the distribution of $\widehat{p}_B$ under the null, we must also specify the percent of households with two, one or no A-voters among the 14,091 two-voter households in the study. We argue that we can use observations from the control to provide a range of possible proportions for the three household behavioral types. We test the null hypothesis of no B-voters, repeatedly, over this range of possible proportions and show that in each case, the null hypothesis is rejected.

To estimate the proportion of these three household behavioral types, we observe the control as only A-voters will vote in this treatment condition. By counting the number of votes within each household we count the number of A-voters in each household and are able to detect each of the three aggregated household behavioral types. With a total of 7,310 households assigned to the control, we can estimate the proportion of each type and use it in our simulation. Table 5.14 shows the estimated percent of each type assigned to control. We see that A-voters are strongly grouped within the two-voter households. Over 22% of the subjects are A-voters but they are far from evenly distributed as over 70% of the households contain no A-voters. To understand the range of error, the final column of Table 5.14 lists a 95% confidence interval computed as if the 7,310 households in the control were drawn without replacement as a simple random sample (SRS) from the 14,091 households.

Household Type	Observed in Control	Percent	95% Confidence Interval (SRS)
Two A-voters	1,137	15.6%	(14.9,16.3%)
One A-voter	1,029	14.1%	(13.4,14.8%)
No A-voters	5,144	70.3%	
Total	7,310	100%	

Table 5.14: Household types observed in the control, when restricting to two-voter households with complete data. The confidence interval is calculated assuming the control is derived from a simple random sample.

By the complex randomization, the control group is not drawn from an SRS. The sampling procedure more closely resembles a type of cluster sample which would have a larger confidence region than an SRS of similar size (Lohr, 2009). But the bounds do provide a sense of the possible proportions of each of the three behavioral types. Once the proportion of the three household behavioral types is set, it can be used in the simulation to test the hypothesis of no B-voters.

One final concern is how clustered together the two A and one A-voter households are to each other. Simply assigning them randomly among the 14,091 households may not accurately reflect how they could be grouped within certain blocks or certain neighborhoods. We simulated under two extreme assumptions : no clustering, that is, random placement of the three types of households on the same block and, full clustering, where two A-voter houses are, to the degree allowable, only neighboring other two A-voter households, and the same for the one A-voter households and households with no A-voters. With the fraction of two, one and no A-voter households set, along with the clustering assumption, we take the raw data, with the zip codes and households of each voter and simulate the randomization process for all of the one, two and three-voter households, collect the results of the 14,091 two-voter households representing 28,182 voters and calculate the test statistic $\widehat{p}_B$. We repeat the simulation 40,000 times to give the distribution of $\widehat{p}_B$ under the null hypothesis. To recap, we proceed with the algorithm:

1. Set the two assumptions for the within-household distribution of A-voters.

   (i) The proportion of two, one and no A-voter households among the 14,091 two-voter households

   (ii) Clustering of A-voters or randomly dispersed.

2. Distribute the three household behavioraly types among the 14,091 two-voter households according to the assumptions from Step 1.

3. Using all of the households, including the one and three voter households, and those without complete data, randomly assign the zip codes, households and voters to treatment as described earlier and shown in Figure 5.1.
4. Construct the contingency table of observations from the 14,091 two-voter households as in Table 5.13.
5. Calculate $\widehat{p}_B$ from the contingency table.
6. Repeat Steps 2 – 5 until 40,000 simulations of $\widehat{p}_B$ are obtained.

We ran simulations under the various assumptions in Step 1 under the hypothesis of no B-voters to arrive at the following p-values for the observed $\widehat{p}_B$ of 1.6%. Each row represents a separate set of assumptions.

Assumptions on distribution of A-voters			
HH with 2 A-voters	HH with 1 A-voter	Same Type Clustering	p-value
15.6%	14.1%	No	.0087
15.6%	14.1%	Yes	.0084
19.5%	18.1%	Yes	.0133

Table 5.15: Hypothesis tests specifications and p-values under the null of no B-voters.

We see that even under a range of different assumptions the p-value are smaller than those found by Sinclair et al. We conclude that, at least for two-voter households with complete voting data, there is strong evidence for a within-household spillover effect.

5.4 When Does Increasing Mobilization Effort Increase Turnout? New Theory and Evidence from a Field Experiment on Reminder Calls

Many GOTV experiments measure the impact of just a single intervention, such as a specific mailer or phone message, to isolate and quantify a single effect. In the application in Section 5.3, researchers chose a political race expected to receive little attention, hoping to minimize other electoral influence. Commonly though, citizens are inundated with election mailers, door knocks and advertisements when voting in districts with heavily resourced campaigns. In such circumstances a more practical aim is to measure the marginal impact of additional contacts to voters already reached. Does an extra phone call matter? Political scientists have approached this question by studying reminder calls, those made close to the election date after initial outreach attempts, but the studies have differed on whether reminder calls

are effective. Green and Gerber (2001) find little evidence that multiple contacts increase turnout beyond the initial attempt, while Michelson, García Bedolla, and McConnell (2009) find convincing evidence that, for those stating an intention to vote, follow-up calls can have a substantial impact.

In a recently posted paper Gerber, Huber, Fang, and Reardon (2016) attempt to measure the reminder call effect and also uncover the underlying reasons why a reminder call would increase turnout. Is it the timing of the call, close to the election? Or the combination of a follow-up call after the voter has already confirmed an intention to vote from an earlier contact? Does the voter perceive the reminder as an act of kindness from the canvasser and reciprocates by voting? The experimental design allows the authors to answer such queries. We highlight this application because, unlike our previous ones, it incorporates compliance to treatment. As we saw from Sections 2.2 and 4.2.3, the dimension of compliance leads to a larger number of behavioral types and estimators, based on ratios of random variables, with more complex variance structure. Additionally, it gives us the opportunity to analyze effects under a partial ordering of treatments when the relationship between some of the treatments is ambiguous. Though we don't have all of the data available to us, we demonstrate how to draw conclusions if all of the information were in hand.

Gerber et al. devise an experiment with six treatment conditions. The first treatment includes a nonpartisan *early call*, 22-25 days before the election, with no other attempts at contact. The second treatment consists of a *late call*, 5-7 days before the election, which reminds subjects of the upcoming vote. The third treatment administers an early and a late call. In all instances, late calls occur regardless of whether an earlier call was successful and the script of the late call does not refer to, nor depend on an earlier attempt. The separation of the early and late call interventions, by design, allows for the testing of numerous hypothesis about the reasons for the impact of the reminder calls. The next two treatments include an extra interaction in the early call, where the canvasser asks the voter if they would like a reminder call. The fourth treatment consists of an early call, this time with the offer of a reminder call, but there is no late call. The fifth treatment is similar to the fourth treatment but includes a late call, again regardless of whether the voter requests one or not. The sixth experimental group is a control with no outreach attempts. The six experimental conditions are summarized below.

1. Early call only
2. Late call only
3. Early and late call
4. Early call, only, with offer of reminder
5. Early with offer of reminder and late call

6. Control

The inclusion of the offer of a reminder call allows the authors to analyze the impact of the call conditional on the interactions between the caller and voter. The authors theorize that the act of the canvasser making a reminder call, or perhaps just offering to make the follow up call, indebts the subject in such a way to reciprocate the courtesy by carrying out the canvasser's request. Much of the paper centers on the underlying reasons reminder calls may increase turnout. We mention this to explain the motivation for the inclusion of the offer of a reminder call. For our analysis though, we do not concern ourselves with the testing of the theories as to *why* reminder calls may have an effect and instead concentrate on the initial question of measuring the magnitude of the effect. Again, our interest lies mainly in applying the behavioral-types approach to an experiment with noncompliance and partial ordering of the multiple treatments.

The experiment was held in Colorado during the November 2014 midterm election. The Colorado Civic Engagement Roundtable implemented the experiment targeting African Americans, Latinos, young voters between the age of 17-34 and unmarried women. They excluded voters who voted in all of the last four elections as they were expected to turnout with or without contact. Also excluded were long term registrants who didn't vote in the 2012 presidential election as they were unlikely to vote in the lower turnout midterm race. This led to a pool of 225,717 eligible voters. The study participants were whittled further by removing those without a valid phone number or without a valid state ID–needed to match voters to administrative data on voter turnout. Finally, just one voter was chosen in households with multiple voters to minimize subject-to-subject interactions and violations of the SUTVA. The ultimate sample numbered 139,153 registered voters. Each subject was then assigned independently to a treatment according to the following probabilities. The first three treatments with an early call, a late call or both calls, without an offer of a followup, were assigned with a probability of 0.125. The fourth and fifth treatments, which offered a reminder call were assigned with a probability of 0.25 each. The control condition was assigned with a probability of 0.125. This differs slightly from the first two applications where the total in each treatment condition is fixed before the assignment process. As we shall show, even with a random total in each treatment, we will be able to estimate parameters from the table of observations, similar to the first three applications. Though this does lead to a more complicated variance calculation which is presented in Section 5.6.

Analysis and Conclusions by Experimenters

The paper contains a number of separate investigations but begins with most straightforward, measuring an ITT effect for a reminder call after an early call. To measure the effect they subset their sample to the 104,674 subjects assigned to the four treatments with an early call: the first, third, fourth and fifth treatments. This is not an ATT effect -compliance is not used– the authors characterize it as an ITT effect, conditional on being assigned an early

call. They use ordinary least squares assuming the model

$$Y_i = \alpha + \beta \mathbb{1}(\text{Assigned Late Call})_i + \sum_j \gamma_j X_{ji} + u_i \tag{5.2}$$

where Y is the usual voting indicator, β is the primary parameter of interest, and X_j represent covariates for age, gender, race and past voting history. As in the other applications presented in the chapter, they find the extra predictor variables have little impact on the results. They arrive at an estimate for β of -0.1% with an SE of 0.3% and conclude there is little evidence for an average effect over the entire sample.

The authors make two choices in their analysis worth noting. First, they omit including the offer of a reminder into the model, effectively combining the four treatments into two. Also, they do not try to estimate the impact of the early call. By restricting to just those assigned an early call, groups three and five become a quasi-control group to determine the impact of the reminder call. Curiously, throughout the paper, they do not use the second treatment group, with just the late call, nor do they use the control group in any of their investigations. Thus, no separate estimates of the effect of an early call is reported. Nor do they report an estimate of a stand-alone "late call effect", the influence of a late call is always within the context of an earlier call attempt.

Next they determine the effect of being assigned a reminder call, for those who received the early call. This is closer to, but still different from, the ATT which would focus on those who receive the reminder call. Instead they estimate an ITT effect, a reminder call conditional on receipt of an early call. Again, the separate interventions of the early and late call allow for this type of subgroup analysis because the assignment of the late call is independent of whether a subject receives an early call. Since the compliance rate for those receiving the early call is just a little higher than 20%, this reduces the sample size to 22,120 and they estimate this conditional ITT effect again with OLS using Equation 5.2. This results in an estimate of 1.2% with a standard error of 0.6%. This is the primary finding of the study.

The authors deepen their investigation of the reminder call by further subsetting the data to test four hypotheses pertaining to the underlying reasons of *why* the calls could be effective, all of which utilize a variable for "offer of a reminder call" in the modeling. And they conclude the paper with constructing models which control for certain covariates, to explore heterogeneous effects such as the impact of reminder calls on those who state they intend to vote. We won't delve further into these, and focus our application of the behavioral-types approach on ITT and ATT effects for reminder calls, similar to the first two estimates provided in the study.

Analysis with Behavioral Types

Gerber et al. focus on ITT effects which do not depend on compliance. Though we do not have access to the data, we describe how to examine the treatment effects if we had a summary table of the aggregated compliance and voting totals for each treatment group. We estimate the same conditional ITT effects with the behavioral-types approach, and also estimate an ATT effect.

In our approach, the aim is to completely determine the fraction of each of the distinct behavioral types in the sample. To do so, we use the observations from all six of the treatment conditions. As can be seen in Equation 5.2 the offer of a follow up is not included in the model used by the authors to measure the ITT effect for a remainder call after an early call nor to measure the ITT effect of a reminder call conditional on receipt of an early call. The authors do include a variable for the offer of a follow up in their explorations of why follow up calls have impact. But since we won't look into this, restricting ourselves to the first two estimates, we follow the authors lead, disregarding the offer of a follow up as a separate treatment, combining the first and fourth groups into one "early call" treatment and the third and fifth group into a "both early and late" treatment. Along with the control and late calls only group, this gives four separate treatment conditions for the experiment.

We can now describe the different behavioral types. As was shown in Sections 2.2, 4.2.1 and 4.2.3, when compliance is observed we may separate the compliance behavior from the response behavior, and we find it more illuminating to begin with the compliance. Similar to Example 4.2.3 with a control and three degrees of treatment, in this experiment there are four flavors of compliance behavior. *Perfect compliers* receive whichever treatment is assigned. *Nevertakers* never pick up the phone and receive "no call" (control) each time. *Early compliers* only comply with an early call. They receive the early call but not the late call, so receive the early treatment if assigned to early or both and receive no call if assigned to the control or late call. In the same manner, *Late compliers* receive just the late call, thus receive the late treatment if assigned to late or both and receive no call if assigned to the control or early call.

The response behavior presents a new complication as it is not clear which of the early or late calls is the stronger treatment so the restriction of monotonicity may not apply. If there was a clear order, then the distinct behavioral types would be analogous to the 13 types in Example 4.2.3. Could the late call be a stronger treatment than the early call? Nickerson (2007) argues that, on average, calls made closer to the election date are more effective in a study where voters were subject to multiple outreach efforts. However, a GOTV study by Panagopoulos (2011) tests this and finds little evidence that later calls increase turnout more than earlier ones. In any case, it seems plausible that there could exist sizable fractions of the sample who vote if called early but not if called late–perhaps someone who would mail an absentee ballot early on but would not be able to vote at a polling station–

and also individuals who vote if called late but not if called early–where the reminder is more effective closer to election day. Instead of a strict ordering of treatments we consider a partial ordering where the control is the weakest treatment condition and receiving both calls is the strongest. In between, the early and late calls are not comparable but we allow types that have voting outcomes changed by one of the timings of the call but not the other. This results in six different response behaviors. The first four are as before. *A-voters* vote and *F-voters* do not, no matter the treatment received. *B-voters* vote when receiving the early or late call or both. And *D-voters* vote if and only if they receive both calls. The difference is in the *C-voters* which now has two distinct subtypes: *C-early voters (C_E)* vote if they receive an early call, but not if they receive a late call; *C-late voters (C_L)* vote if they receive a late call, but not for an early one. The behavior of the six voting types to each received treatment are summarized below.

Treatment Received	Votes	Doesn't Vote
No Call (Control)	A-voters	B,C,D,E,F-voters
Early Call	A,B,C_E-voters	C_L,D,F-voters
Late Call	A,B,C_L,-voters	C_E,D,F-voters
Both Calls	A,B,C_E,C_L,D-voters	F-voters

We combine the compliance and response behaviors to enumerate the distinct behavioral types intrinsic to the experiment. Perfect compliers have a distinct behavioral type for each of the six voting types ($perfect_A$, ..., $perfect_F$). Nevertakers consist of just two voting types: A-voters ($never_A$) and all other voting types which act in the same way as they don't receive either call ($never_{B-F}$). For early compliers, who don't receive late calls, A-voters form a distinct behavioral type. And ($early_A$) and B and C_E-voters are indistinguishable, as the early call is the strongest treatment received, thus forming a single behavioral type ($early_{BC_E}$). Furthermore, C_L, D and F-voters do not receive calls and nor do they vote in any treatment and can also be combined as ($early_{C_L DF}$). Using the same reasoning, late compliers also include three behavioral types ($late_A$, $late_{BC_L}$ and $late_{C_E DF}$). Tables 5.16 to 5.20 show the compliance and response values for each assignment, for each of the behavioral types. In total, there are $6+2+3+3=14$ distinct behavioral types, whose proportions must sum to 1, yielding an estimation problem with 13 parameters.

Assigned Treatment	Received Treatment	Response $perfect_A$	Response $perfect_B$	Response $perfect_D$	Response $perfect_F$
Control	No call	Votes	Doesn't vote	Doesn't vote	Doesn't vote
Early call	Early call	Votes	Votes	Doesn't vote	Doesn't vote
Late call	Late call	Votes	Votes	Doesn't vote	Doesn't vote
Both calls	Both calls	Votes	Votes	Votes	Doesn't vote

Table 5.16: Compliance and response to each assigned treatment for the A, B, D and F types of perfect compliers.

Assigned Treatment	Received Treatment	Response $perfect_{C_E}$	Response $perfect_{C_L}$
Control	No call	Doesn't vote	Doesn't vote
Early call	Early call	Votes	Doesn't vote
Late call	Late call	Doesn't vote	Votes
Both calls	Both calls	Votes	Votes

Table 5.17: Compliance and response to each assigned treatment for the C-early and C-late types of perfect compliers.

Assigned Treatment	Received Treatment	Response $early_A$	Response $early_{BC_E}$	Response $early_{C_L DF}$
Control	No call	Votes	Doesn't vote	Doesn't vote
Early call	Early call	Votes	Votes	Doesn't vote
Late call	No call	Votes	Doesn't vote	Doesn't vote
Both calls	Early call	Votes	Votes	Doesn't vote

Table 5.18: Compliance and response to each assigned treatment for the three types of early compliers.

Assigned Treatment	Received Treatment	Response $late_{A}$	Response $late_{BC_L}$	Response $late_{C_E DF}$
Control	No call	Votes	Doesn't vote	Doesn't vote
Early call	No call	Votes	Doesn't vote	Doesn't vote
Late call	Late call	Votes	Votes	Doesn't vote
Both calls	Late call	Votes	Votes	Doesn't vote

Table 5.19: Compliance and response to each assigned treatment for the three types of late compliers.

Assigned Treatment	Received Treatment	Response $never_{A}$	Response $never_{B-F}$
Control	No call	Votes	Doesn't vote
Early call	No call	Votes	Doesn't vote
Late call	No call	Votes	Doesn't vote
Both calls	No call	Votes	Doesn't vote

Table 5.20: Compliance and response to each assigned treatment for the two types of nevertakers.

With the distinct behavioral types established, the next task is estimate the proportions of each type from the data, thus identifying the parameters of the model. Using the notation introduced in Section 4.2.3 we let Q_{yd} represent the random total of observations of the subjects assigned to the control who respond with voting value y and comply with treatment d, R_{yd} for those assigned to the *early* treatment,S_{yd} for *late* and T_{yd} for *both*. The aggregated voting outcomes appear in the two by nine table of observed results.

	Assigned Control	Assigned Early Call		Assigned Late Call		Assigned Both Calls			
	Received	Received		Received		Received			
	No Call	No Call	Early	No Call	Late	No Call	Early	Late	Both
Didn't Vote	$Q_{0\,no\,call}$	$R_{0\,no\,call}$	$R_{0\,early}$	$S_{0\,no\,call}$	$S_{0\,late}$	$T_{0\,no\,call}$	$T_{0\,early}$	$T_{0\,late}$	$T_{0\,both}$
Voted	$Q_{1\,no\,call}$	$R_{1\,no\,call}$	$R_{1\,early}$	$S_{1\,no\,call}$	$S_{1\,late}$	$T_{1\,no\,call}$	$T_{1\,early}$	$T_{1\,late}$	$T_{1\,both}$
Total	Q	R		S		T			

Since they are not fixed by design, the totals in the treatment groups, Q, R, S and T, are random quantities so, in this study, the table of observations has 17 degrees of freedom (if the number assigned to each treatment was set in advance there would be $1+3+3+7=14$ degrees of freedom, still sufficient for estimating the 13 parameters). We now place the behavioral types according to the cell in which they appear in the table of observed data. This is shown in Table 5.21.

	Assigned Control	Assigned Early Call		Assigned Late Call		Assigned Both Calls			
	Received	Received		Received		Received			
	No Call	No Call	Early	No Call	Late	No Call	Early	Late	Both
Didn't Vote	$perfect_{F}$		$perfect_{F}$		$perfect_{F}$				$perfect_{F}$
	$late_{C_E DF}$	$late_{C_E DF}$			$late_{C_E DF}$			$late_{C_E DF}$	
	$early_{C_L DF}$		$early_{C_L DF}$	$early_{C_L DF}$			$early_{C_L DF}$		
	$never_{B-F}$	$never_{B-F}$		$never_{B-F}$		$never_{B-F}$			
	$perfect_{D}$		$perfect_{D}$		$perfect_{D}$				
	$perfect_{C_E}$				$perfect_{C_E}$				
	$perfect_{C_L}$		$perfect_{C_L}$						
	$late_{BC_L}$	$late_{BC_L}$							
	$perfect_{B}$								
	$early_{BC_E}$								
Voted									$perfect_{D}$
					$perfect_{C_L}$				$perfect_{C_L}$
			$perfect_{C_E}$						$perfect_{C_E}$
					$late_{BC_L}$			$late_{BC_L}$	
			$perfect_{B}$		$perfect_{B}$				$perfect_{B}$
			$early_{BC_E}$	$early_{BC_E}$			$early_{BC_E}$		
	$perfect_{A}$		$perfect_{A}$		$perfect_{A}$				$perfect_{A}$
	$late_{A}$	$late_{A}$			$late_{A}$			$late_{A}$	
	$early_{A}$		$early_{A}$	$early_{A}$			$early_{A}$		
	$never_{A}$	$never_{A}$		$never_{A}$		$never_{A}$			

Table 5.21: Location of the 14 behavioral types in the table of observations for the reminder call experiment.

Each cell of table of observations provides a linear equation in the estimation problem. For example, there are only three behavioral types which, if assigned to the early call treatment, do not comply with treatment (receiving the the control treatment, that is, no call) and do not vote: $late_{BC_L}$, $late_{C_E DF}$ and $never_{B-F}$. Thus the fraction of voters assigned to the early

call who neither comply nor vote provide an estimate of the fraction of the three behavioral types.

$$R_{0\,no\,call}/R = \widehat{p}_{late_{C_E DF}} + \widehat{p}_{late_{BC_L}} + \widehat{p}_{never_{B-F}}$$

In total, this gives an overdetermined system of 17 equations (again 14 if the number in each treatment group is fixed) and 13 parameters. We refer the readers to other discussions of such estimation problems such as a recent piece by Awange and Paláncz (2016).

Once the parameters have been estimated, the plug-in principle is used to calculate the variances and standard errors which are found in Section 5.6. And we arrive at our ultimate aim, using the fractions of the behavioral types to estimate the treatment effects of interest.

The first estimate: ITT effect of a reminder call after an early GOTV attempt

We turn our attention to the interpreting the effect as a combination of behavioral types. Gerber et al. characterize the intention-to-treat effect of the reminder call as the marginal increase in voter turnout for those assigned to both calls, compared to those assigned to the early call.

$$itt_{reminder} \equiv itt_{both} - itt_{early} \tag{5.3}$$

We reach this in two ways, first by describing both terms on the right hand side of 5.3 in terms of behavioral types, and second by a more intuitive approach. The first path is more methodical, but also more accessible as it utilizes the well understood intention-to-treat effect. Setting the total number of subjects in the sample to n, where $n = Q + R + S + T$, the *itt* for the early call is the difference between those who vote if assigned to the early call and those who vote if assigned to control.

$$itt_{early} = \frac{\#(\text{vote if assigned early})}{n} - \frac{\#(\text{vote if assigned control})}{n}$$

We separate subjects who vote if assigned to the early treatment into two behavioral groups: vote if assigned to control versus don't vote if assigned to control but do vote if assigned to the early call.

$$itt_{early} = \left[\frac{\#(\text{vote if assigned control}) + \#(\text{vote if assigned early, don't vote if assigned control})}{n}\right] - \frac{\#(\text{vote if assigned control})}{n}$$

$$= \frac{\#(\text{vote if assigned early, don't vote if assigned control})}{n}$$

and from Table 5.21 we identify the specific behavioral types which vote if assigned early, but not if assigned to the control so that

$$itt_{early} = \frac{\#(early_{BC_E} \text{ or } perfect_B \text{ or } perfect_{C_E})}{n}$$

$$= p_{early_{BC_E}} + p_{perfect_B} + p_{perfect_{C_E}} \tag{5.4}$$

The intention-to-treat effect for both calls is evaluated in a like manner to uncover its connection to the behavioral types. Here, we recognize that those who vote if assigned to both calls may be divided into those who vote, or don't vote under the control condition.

$$itt_{both} = \frac{\#(\text{vote if assigned both})}{n} - \frac{\#(\text{vote if assigned control})}{n}$$

$$= \left[\frac{\#(\text{vote if assigned control}) + \#(\text{vote if assigned both, don't vote if assigned control})}{n}\right] - \frac{\#(\text{vote if assigned control})}{n}$$

$$= \frac{\#(\text{vote if assigned both, don't vote if assigned control})}{n}$$

$$= \frac{\#(early_{BC_E} \text{ or } perfect_B \text{ or } perfect_{C_E} \text{ or } perfect_{C_L} \text{ or } perfect_D \text{ or } late_{BC_L})}{n}$$

$$= p_{early_{BC_E}} + p_{perfect_B} + p_{perfect_{C_E}} + p_{perfect_{C_L}} + p_{perfect_D} + p_{late_{BC_L}} \tag{5.5}$$

We substitute Equations 5.4 and 5.5 into 5.3 and reveal the behavioral types making up $itt_{reminder}$.

$$\begin{aligned} itt_{reminder} &= itt_{both} - itt_{early} \\ &= p_{perfect_{C_L}} + p_{perfect_D} + p_{late_{BC_L}} \end{aligned} \tag{5.6}$$

We can also arrive at 5.6 via a more intuitive, and immediate path. As $itt_{reminder}$ is the increase in the percentage of voters from the early call to both calls, the effect consists of the three behavioral types in Table 5.21 which don't vote if assigned to early but do vote if assigned to both calls: $perfect_{C_L}$, $perfect_D$ and $late_{BC_L}$.

There are two clear estimates of $itt_{reminder}$, the first is directly from the table of observations, the difference between the voting rate of the early and both calls treatments.

$$\widehat{itt}_{reminder} = \frac{T_{1\,no\,call} + T_{1\,early} + T_{1\,late} + T_{1\,both}}{T} - \frac{R_{1\,no\,call} + R_{1\,early}}{R} \tag{5.7}$$

The variance of this estimate, which now has random quantities in the numerator and denominator, is discussed in Section 5.6. However we may also use another estimate of the effect from the estimated percentage of behavioral types

$$\widehat{itt}_{reminder} = \widehat{p}_{perfect_{C_L}} + \widehat{p}_{perfect_D} + \widehat{p}_{late_{BC_L}} \tag{5.8}$$

where the fraction of behavioral types are solved from the overdetermined system of linear equations. It should be clear that while these estimates should be close, they aren't the same. Equation 5.7 may provide a less accurate estimate as it is a linear combination of the equations which lead to the behavioral types in Equation 5.8. But computing the variance of the $\widehat{itt}_{reminder}$ under Equation 5.8 is more difficult because one must solve the covariance structure of the $\widehat{p}_{type}$ which is not so easy when they are solutions to an overdetermined system of linear equations. For these reasons we use the estimate in Equation 5.7 for our analysis.

The second estimate: ITT effect of a reminder call conditional on receiving the early call

The second estimate of the paper is the $itt_{reminder}$ restricted to those who received the first call. We use what we've learned from the first estimate, modifying our evaluation so that it is conditional on receipt of the first call.

$$\begin{aligned} &itt_{reminder} \mid \text{receive early} \\ &\quad = itt_{both} \mid \text{receive early} - itt_{early} \mid \text{receive early} \\ &\quad = \frac{\#(\text{vote if assigned both, don't vote in control, receive early})}{\#(\text{receive early})} \\ &\qquad - \frac{\#(\text{vote if assigned early, don't vote in control, receive early})}{\#(\text{receive early})} \end{aligned}$$

where the numerators may be evaluated in the manner of those in Equations 5.4 and 5.5 with $late_{BC_L}$ excluded as they don't receive the early treatment so that

$$= \frac{\#(early_{BC_E} \text{ or } perfect_B \text{ or } perfect_{C_E} \text{ or } perfect_{C_L} \text{ or } perfect_D)}{\#(\text{early or perfect compliers})}$$

$$- \frac{\#(early_{BC_E} \text{ or } perfect_B \text{ or } perfect_{C_E})}{\#(\text{early or perfect compliers})}$$

$$= \frac{\#(perfect_{C_L} \text{ or } perfect_D)}{\#(\text{early or perfect compliers})}$$

$$= \frac{p_{perfect_{C_L}} + p_{perfect_D}}{p_{early} + p_{perfect}} \tag{5.9}$$

where p_{early} represents the fraction of all three of the early compliers and $p_{perfect}$ represents the fraction of all six of the perfect compliers. As with $\widehat{itt}_{reminder}$, we have two immediately available options from Equation 5.9 for estimating the parameter. We may use the $\widehat{p}_{type}$ from the system of equations but that presents the same complicated variance calculation. Again, we take the approach of estimating from the table of observations where the numerator of 5.9 is estimated as

$$\widehat{p}_{perfect_{C_L}} + \widehat{p}_{perfect_D} = \frac{T_{1\,early} + T_{1\,both}}{T} - \frac{R_{1\,early}}{R} \tag{5.10}$$

and the denominator is estimated from the subjects who receive treatment if assigned to the early or both calls condition, i.e.

$$\widehat{p}_{early} + \widehat{p}_{perfect} = \frac{R_{0\,early} + R_{1\,early} + T_{0\,early} + T_{1\,early} + T_{0\,both} + T_{1\,both}}{R + T}$$

which gives the estimate of

$$\widehat{itt}_{reminder} \mid \text{receive early} = \frac{\frac{T_{1\,early}+T_{1\,both}}{T} - \frac{R_{1\,early}}{R}}{\frac{R_{0\,early}+R_{1\,early}+T_{0\,early}+T_{1\,early}+T_{0\,both}+T_{1\,both}}{R+T}}.$$

A third estimate: ATT of a reminder call

We can further the authors analysis to provide an estimate of the ATT effect of the reminder call. This is the treatment effect on those who comply with the entire treatment protocol of receiving both calls, i.e., the effect on the perfect compliers. Restricting ourselves to those treated, the effect of interest is the marginal increase in voting for those who are assigned and received both calls, above the turnout for the perfect compliers that are assigned only the early call.

$$att_{reminder} = \frac{\#(\text{vote if assigned both, don't vote if assigned early, receive both if assigned both})}{\#(\text{receive both if assigned both})}$$

$$= \frac{\#(\,perfect_{C_L} \text{ or } perfect_D)}{\#(\text{perfect compliers})}$$

$$= \frac{p_{perfect_{C_L}} + p_{perfect_D}}{p_{perfect}} \tag{5.11}$$

For $\widehat{att}_{reminder}$ the numerator is estimated via Equation 5.10 and the denominator is the fraction of subjects who receive both calls if assigned both calls.

$$\widehat{att}_{reminder} = \frac{\frac{T_{1\,early}+T_{1\,both}}{T} - \frac{R_{1\,early}}{R}}{\frac{T_{0\,early}+T_{1\,early}+T_{0\,both}+T_{1\,both}}{T}} \tag{5.12}$$

We conclude by noting a familiar identity for the ATT parameter. From Equation 5.11

$$\begin{aligned} att_{reminder} &= \frac{p_{perfect_{C_L}} + p_{perfect_D}}{p_{perfect}} \\ &= \frac{\frac{p_{perfect_{C_L}} + p_{perfect_D}}{p_{early}+p_{perfect}}}{\frac{p_{perfect}}{p_{early}+p_{perfect}}} \\ &= \frac{itt_{reminder} \mid \text{receive early}}{(\text{compliance rate for late call given receive early})} \end{aligned} \tag{5.13}$$

so that Equation 5.13 is the analogue of Equation 2.7, $att = itt/(\text{compliance rate})$.

5.5 Discussion

We included four applications in this chapter to exhibit the use of the behavioral-types approach to estimate causal effects under a variety of experimental settings with multiple levels of treatment. A behavioral-types approach is well suited to multi-treatment experiments because it distills these often complex designs into an estimation problem of a manageable number of types. When compared to linear modeling, the common method of the studies, the behavioral types framework may not lead to different estimates but often leads to a different standard error, changing the significance of the conclusions. The first two applications adhere to the widely applicable Intention-to-Treat analysis of strictly ordered treatments, described in Example 4.2.2. The last two applications illustrate the nuances in the analysis

brought on by more elaborate experimental designs. The application in 5.3 demonstrates how hypothesis testing may be able to verify the existence of an effect even when the distinct behavioral types are not identifiable. Additionally the application displays how simulations can be used to measure significance when randomization schemes are complex. In 5.4 the experimental design results in a partially ordered set of treatments. It also presents the challenge of an overdetermined system of equations to identify the parameters. And 5.4 shows how seemingly minor details in the design, such as whether the total assigned to treatment is fixed or random, changes the number of rows in the system of estimating equations.

As shown in Chapter 4, a design with more distinct treatments and compliance outcomes leads to more distinct behavioral types. And, more complicated randomization schemes result in more complicated calculations of significance. Despite these differences among the applications, there are common steps in our behavioral types analysis. In each, we proceeded along the following steps.

1. Identify the distinct behavioral types. These are intrinsically tied to experimental design. Each behavioral type can be located in the table of observations, as in Table 5.21, indicating the cell in which the behavioral type lands for each treatment assignment.

2. Represent the treatment effects of interest in terms of the behavioral types.

3. Determine how to estimate the distinct behavioral types, and thus treatment effects, from the table of observations. In some cases, when parameters are not identifiable, we may still be able to detect the presence of certain effects via hypothesis testing.

4. Use the sampling process of the experiment to calculate either the variances of the estimates or, if hypothesis testing, the distribution of the test statistic. In complex randomization designs, simulations maybe be helpful when analytical calculations are not possible.

At this point, we have only concerned ourselves with single parameter inference. We discuss confidence regions for multi-parameter inference in Chapter 6.

5.6 Variance Calculations

This section describes how we arrive at the standard errors in Sections 5.1 and 5.2. As discussed in Chapter 3, the asymptotic normality of the estimators is established by Theorem 5 of Li and Ding (2017).

Social Pressure Experiment

To determine the variance of $\widehat{\mathbf{p}}$ we note that it can be represented as a transformation of the observations listed as a column vector.

$$\widehat{\mathbf{p}} = \mathbf{\Psi} \cdot [Q_A, R_{AB}, S_{ABC}, T_{ABCD}, U_{ABCDE}]'$$

where

$$\mathbf{\Psi} = \begin{pmatrix} \frac{1}{q} & 0 & 0 & 0 & 0 \\ -\frac{1}{q} & \frac{1}{r} & 0 & 0 & 0 \\ 0 & -\frac{1}{r} & \frac{1}{s} & 0 & 0 \\ 0 & 0 & -\frac{1}{s} & \frac{1}{t} & 0 \\ 0 & 0 & 0 & -\frac{1}{t} & \frac{1}{u} \end{pmatrix},$$

so the variance of $\widehat{\mathbf{p}}$ will be $\mathbf{\Psi}\text{Var}([Q_A, R_{AB}, S_{ABC}, T_{ABCD}, U_{ABCDE}])\mathbf{\Psi}'$. To find the covariance matrix of the observations, the diagonals terms are from the hypergeometric distribution. That is,

$$\begin{aligned} \text{Var}(Q_A) &= \frac{q(n-q)n_A(n-n_A)}{n^2(n-1)} \\ &= \frac{q(n-q)p_A(1-p_A)}{n-1} \end{aligned}$$

and similarly

$$\begin{aligned} \text{Var}(R_{AB}) &= \frac{r(n-r)(p_A+p_B)(1-p_A-p_B)}{n-1} \\ \text{Var}(S_{ABC}) &= \frac{s(n-s)(p_A+p_B+p_C)(1-p_A-p_B-p_C)}{n-1} \\ \text{Var}(T_{ABCD}) &= \frac{t(n-t)(p_A+p_B+p_C+p_D)(1-p_A-p_B-p_C-p_D)}{n-1} \\ \text{Var}(U_{ABCDE}) &= \frac{u(n-u)(p_A+p_B+p_C+p_D+p_E)(1-p_A-p_B-p_C-p_D-p_E)}{n-1} \end{aligned}$$

We calculate the covariance terms in the same manner as in Section 3.2. For example, consider $\mathrm{Cov}(R_{AB}, T_{ABCD})$. Since R_{AB} is $R_A + R_B$ and T_{ABCD} is $T_A + T_B + T_C + T_D$, we have

$$\begin{aligned}\mathrm{Cov}(R_{AB}, T_{ABCD}) =& \mathrm{Cov}(R_A, T_A) + \mathrm{Cov}(R_A, T_B) + \mathrm{Cov}(R_A, T_C) + \mathrm{Cov}(R_A, T_D) + \\ & \mathrm{Cov}(R_B, T_A) + \mathrm{Cov}(R_B, T_B) + \mathrm{Cov}(R_B, T_C) + \mathrm{Cov}(R_B, T_D). \quad (5.14)\end{aligned}$$

As it turns out, the covariance formulas from Equations 3.4 and 3.5 of section 3.2 apply here as well so that

$$\begin{aligned}\mathrm{Cov}(R_{type\,i}, T_{type\,i}) &= -\frac{r\,t\,n_{type\,i}(n - n_{type\,i})}{n^2(n-1)} \\ &= -\frac{r\,t\,p_{type\,i}(1 - p_{type\,i})}{n-1}\,, \quad (5.15)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(R_{type\,i}, T_{type\,j}) &= \frac{r\,t\,n_{type\,i}\,n_{type\,j}}{n^2(n-1)} \\ &= \frac{r\,t\,p_{type\,i}\,p_{type\,j}}{n-1} \quad \text{for } i \neq j. \quad (5.16)\end{aligned}$$

We may substitute these into Equation 5.14 so that

$$\begin{aligned}\mathrm{Cov}(R_{AB}, T_{ABCD}) =& \frac{rt}{n-1}\,\{\, p_A(1-p_A) + p_A p_B + p_A p_C + p_A p_D + \\ & p_A p_B + p_B(1-p_B) + p_B p_C + p_B p_D \,\}. \quad (5.17)\end{aligned}$$

The other nine covariance entries in $\mathrm{Var}([Q_A, R_{AB}, S_{ABC}, T_{ABCD}, U_{ABCDE}])'$ are calculated in the same manner as (5.17) and listed below.

$$\mathrm{Cov}(Q_A, R_{AB}) = \frac{qr}{n-1}\,\{p_A(1-p_A) + p_A p_B\}$$

$$\mathrm{Cov}(Q_A, S_{ABC}) = \frac{qs}{n-1}\,\{p_A(1-p_A) + p_A p_B + p_A p_C\}$$

$$\mathrm{Cov}(Q_A, T_{ABCD}) = \frac{qt}{n-1}\,\{p_A(1-p_A) + p_A p_B + p_A p_C + p_A p_D\}$$

$$\begin{aligned}
\mathrm{Cov}(Q_A, U_{ABCDE}) =& \frac{qu}{n-1}\{p_A(1-p_A) + p_Ap_B + p_Ap_C + p_Ap_D + p_Ap_E\} \\
\mathrm{Cov}(R_{AB}, S_{ABC}) =& \frac{rs}{n-1}\{\ p_A(1-p_A) + p_Ap_B + p_Ap_C + \\
& \qquad p_Ap_B + p_B(1-p_B) + p_Bp_C\ \} \\
\mathrm{Cov}(R_{AB}, T_{ABCD}) =& \text{ see above} \\
\mathrm{Cov}(R_{AB}, U_{ABCDE}) =& \frac{ru}{n-1}\{\ p_A(1-p_A) + p_Ap_B + p_Ap_C + p_Ap_D + p_Ap_E + \\
& \qquad p_Ap_B + p_B(1-p_B) + p_Bp_C + p_Bp_D + p_Bp_E\ \} \\
\mathrm{Cov}(S_{ABC}, T_{ABCD}) =& \frac{st}{n-1}\{\ p_A(1-p_A) + p_Ap_B + p_Ap_C + p_Ap_D + \\
& \qquad p_Ap_B + p_B(1-p_B) + p_Bp_C + p_Bp_D + \\
& \qquad p_Ap_C + p_Bp_C + p_C(1-p_C) + p_Cp_D\ \} \\
\mathrm{Cov}(S_{ABC}, U_{ABCDE}) =& \frac{su}{n-1}\{\ p_A(1-p_A) + p_Ap_B + p_Ap_C + p_Ap_D + p_Ap_E + \\
& \qquad p_Ap_B + p_B(1-p_B) + p_Bp_C + p_Bp_D + p_Bp_E + \\
& \qquad p_Ap_C + p_Bp_C + p_C(1-p_C) + p_Cp_D + p_Cp_E\ \} \\
\mathrm{Cov}(T_{ABCD}, U_{ABCDE}) =& \frac{tu}{n-1}\{\ p_A(1-p_A) + p_Ap_B + p_Ap_C + p_Ap_D + p_Ap_E + \\
& \qquad p_Ap_B + p_B(1-p_B) + p_Bp_C + p_Bp_D + p_Bp_E + \\
& \qquad p_Ap_C + p_Bp_C + p_C(1-p_C) + p_Cp_D + p_Cp_E + \\
& \qquad p_Ap_D + p_Bp_D + p_Cp_D + p_D(1-p_D) + p_Dp_E\ \}
\end{aligned}$$

We have all the formulas needed for the variance as

$$\mathrm{Var}(\widehat{\mathbf{p}}) = \mathbf{\Psi}\mathrm{Var}([Q_A, R_{AB}, S_{ABC}, T_{ABCD}, U_{ABCDE}])\mathbf{\Psi}'.$$

For the standard errors we use the plug-in principle substituting $\widehat{\mathbf{p}}$ for $\mathbf{p}$.

Book Donation Experiment

For this experiment $\text{Var}([Q_A, R_{AB}, S_{ABC})$ is simply the upper left three by three submatrix of $\text{Var}([Q_A, R_{AB}, S_{ABC}, T_{ABCD}, U_{ABCDE}])$, that is consisting of the first three rows and first three columns. Similarly to transform the observations to $\widehat{\mathbf{p}}$ we apply the matrix $\mathbf{\Psi}[1:3, 1:3]$ which equals the first three rows and first three columns of $\mathbf{\Psi}$. Thus

$$\text{Var}(\widehat{\mathbf{p}}) = \mathbf{\Psi}[1:3, 1:3]\text{Var}([Q_A, R_{AB}, S_{ABC}])\mathbf{\Psi}[1:3, 1:3]'$$

and we again use the plug-in principle substituting $\widehat{\mathbf{p}}$ for $\mathbf{p}$.

Chapter 6

Confidence Regions for Multi-parameter Inference

In Chapter 5 we demonstrated how to calculate confidence intervals for each individual parameter but postponed concerns of joint significance. In general, the more parameters estimated the higher the chance of erroneous conclusions. In our inference problem the parameter estimates, the fraction of each behavioral type, sum to one so they are negatively correlated. With this relationship, if one estimate is above is above it's true value, another estimate is more likely to be below it's true value. Simply reporting the individual confidence intervals may misrepresent the joint variation of the two estimates. We address this concern by constructing multiparameter confidence regions. We restrict ourselves to the simplest case when assignment is done at the individual level, without replacement with complete, or unknown, compliance and extend the work of the single parameter inference to find confidence regions for two of the applications in the previous chapter. We discuss how more complex experiments may be addressed in the final section.

The broader subject of multi-parameter estimation and multiple testing has been an area of research in statistics for decades (see Miller, 1981; Shaffer, 1995) and has seen renewed interest due to problems arising in genetics and bioinformatics (see Goeman and Solari, 2014). In settings with multiple parameters, questions of joint significance, based on the joint distribution of the estimates, naturally arise. In our experiments the highly correlated parameter estimates lead to highly dependent hypothesis tests and such settings can lead to overly conservative tests (Perneger, 1998; Fiedler, Kutzner, and Krueger, 2012) We do not aim to develop optimal simultaneous confidence intervals or confidence regions for the parameters. Instead, we propose using confidence regions as a tool to understand the joint variation of the parameter estimates. Confidence regions still come with their own challenges. We show how well understood methods to develop regions often do not attain the desired nominal coverage rates and how this difference may be corrected. And we show how useful conclusions can be drawn from the examination of these regions.

We proceed as follows. In the first section we discuss the normal approximation and show,

despite the theoretical benefits, the approximation results in regions with poor coverage rates. In section two we show how bootstrap approaches provide a pragmatic way to overcome the shortcomings of the normal approximation regions. In section three we refine our approach via the double bootstrap, first proposed by Beran (1987), which offers a self-correcting method to arrive at appropriately sized regions. We also examine the regions for two of the applications from the previous chapter. In section four we present other methods to construct confidence sets and briefly discuss the advantages and challenges with each. Finally we draw conclusions and discuss areas for further exploration.

6.1 Constructing Regions via the Normal Approximation

We return to the experimental design with k levels of ordered treatments where compliance is either complete or unknown as described in Example 4.2.2 of Chapter 4. This design leads to an estimation problem with $l = k + 1$ free parameters. It is also the design of two of the experiments from the previous chapter: the Social Pressure experiment, restricted to single-voter households, of Gerber, Green and Larimer, described in Section 5.1, and the Book Donation experiment of Cotterill, John and Richardson, described Section 5.2. Here our estimate is $\widehat{\mathbf{p}} = (\widehat{p_A}, \widehat{p_B}, ..., \widehat{p_F})$, the fraction of behavioral types of all experimental subjects which has an expectation of $\mathbf{p} = (p_A, p_B, ..., p_F)$ and a variance of $\boldsymbol{\Sigma}/n$. If $\widehat{\mathbf{p}}$ is well approximated by the normal distribution, that is, if $\widehat{\mathbf{p}} \sim \mathrm{N}(\mathbf{p}, \boldsymbol{\Sigma}/n)$ then we may use the T^2 statistic of Hotelling (1931) defined as

$$T^2 = (\widehat{\mathbf{p}} - \mathbf{p})' \frac{1}{n} \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\mathbf{p}} - \mathbf{p}),$$

where $\widehat{\boldsymbol{\Sigma}}/n$ is the estimated covariance of $\widehat{\mathbf{p}}$ as described in Section 5.6. A rescaling gives

$$T^2 \frac{1}{l} \left(\frac{n-l}{n-1} \right) \sim \mathrm{F}(l, n-l). \tag{6.1}$$

The elliptical region determined by the 1-α quantile of the F distribution,

$$\left\{ \mathbf{p} : (\widehat{\mathbf{p}} - \mathbf{p})' \frac{1}{n} \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\mathbf{p}} - \mathbf{p}) \frac{1}{l} \left(\frac{n-l}{n-1} \right) \leq F_{l,n-l(1-\alpha)} \right\}, \tag{6.2}$$

forms a $1 - \alpha$ level confidence region for $\mathbf{p}$. Furthermore, under the key assumption of normality of $\widehat{\mathbf{p}}$, $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\boldsymbol{\Sigma}}^{-1}/n (\widehat{\mathbf{p}} - \mathbf{p})$ is the test statistic of a uniformly most powerful (UMP) test for $\mathbf{p}$ so that the region described by 6.2 corresponds with an optimal confidence region for the parameter (see Anderson, 1984, Ch 5). As n increases, $lF_{l,n-l}$ converges in distribution to a χ^2_l random variable. For large n, we may more elegantly approximate the confidence region as

$$\left\{ \mathbf{p} : (\widehat{\mathbf{p}} - \mathbf{p})' \frac{1}{n} \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\mathbf{p}} - \mathbf{p}) \leq \chi^2_{l,(1-\alpha)} \right\}. \tag{6.3}$$

This returns us to directly using Hotelling's T^2 statistic (as we moved l to the right hand side and $n-l/n-1goesto1$). In the applications we evaluate, $n = 47,836$ and $l = 5$ for the Social Pressure experiment while $n = 11,812$ and $l = 3$ for the Book Donation study. For both experiments, if $\widehat{\mathbf{p}}$ follows a normal distribution, the confidence region is well approximated by the region in 6.3. For smaller n, or larger k, the region in 6.2 is more appropriate.

One final consideration for this estimation problem is proportions of each behavioral type are known to lie within $[0, 1]$ and proportions will sum to one. In contrast, normal random variables, and their estimates, are unbounded and could be negative. We incorporate these parameter constraints, by taking the additional step of truncating the confidence region outside these bounds. Truncation, for any confidence set, does not change the coverage rate if the parameters truly are subject to the constraints.

The assumption of normality is central to the validity of the region. As described in Example 4.2.2, the underlying data generating process is a multivariate hypergeometric distribution with multiple draws. Thus, the normality assumption relies on the degree to which $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\mathbf{\Sigma}}^{-1}/n (\widehat{\mathbf{p}} - \mathbf{p})$ is distributed as an χ^2_l random variable. To test this we conduct simulations of the Social Pressure and Book Donation experiments. In both simulation settings we choose the true parameters to be the same as the estimates of parameters found in Tables 5.6 and 5.9. The sample sizes of each of the experimental groups is set to the sample sizes of the original studies and we simulate each experiment 10,000 times resulting in an estimate $\widehat{\mathbf{p}}$ and confidence region of $\mathbf{p}$ for each simulation. With the simulations in hand, we may then examine the distribution of the $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\mathbf{\Sigma}}^{-1}/n (\widehat{\mathbf{p}} - \mathbf{p})$ used to form the region and calculate how often the confidence region actually covers the true parameter vector.

Figure 6.1 shows the resulting quantile-quantile plot of $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\mathbf{\Sigma}}^{-1}/n (\widehat{\mathbf{p}} - \mathbf{p})$ to the theoretical χ^2_l distribution. For both experiments, the actual distribution of the statistic tends to slightly higher values than those of the theoretical one, if the underlying $\widehat{\mathbf{p}}$ was a normal random variable. This indicates the critical value for the region, $\chi^2_{l(1-\alpha)}$, will be too small and lead to regions that are too small and have under coverage of the true parameters. The distribution of the statistic from the Book Donation experiment appears to diverge more strongly from normality than that of the Social Pressure experiment. However, for the Social Pressure experiment $l = 5$ while for the Book Donation $l = 3$ so this may just be due to the lower degrees of freedom.

The impact of the non-normality may also be seen in Table 6.1 which shows the actual coverage rates of the true parameter for 99%, 95% and 90% coverage regions for the 10,000 simulations. As expected from Figure 6.1, the coverages rates are significantly below the the target, or nominal, rate.

6.2 Constructing Regions via the Bootstrap

Similar to regions constructed via normal approximation, bootstrap regions consist of the elliptical sphere

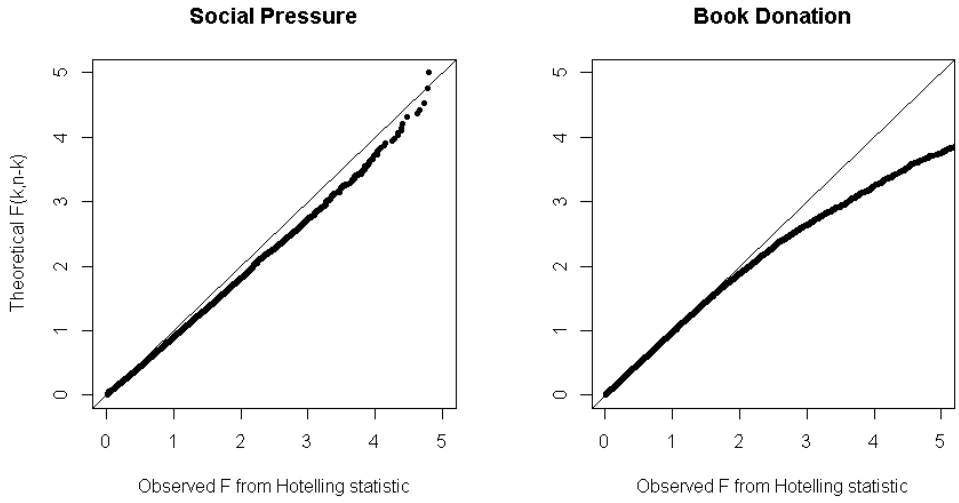


Figure 6.1: Comparing $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\mathbf{\Sigma}}^{-1}/n \, (\widehat{\mathbf{p}} - \mathbf{p})$ to the theoretical χ^2_l distribution.

	Social Pressure		Book Donation	
	Coverage Rate	Std. Error	Coverage Rate	Std. Error
99% region	97.7%	0.15%	97.3%	0.16%
95% region	90.9%	0.29%	92.4%	0.27%
90% region	83.3%	0.37%	87.3%	0.33%

Table 6.1: Coverage rates, and standard error for coverage rates, for normal approximation confidence regions at 99%, 95% and 90% coverage levels, based on 10,000 simulations.

$$\left\{ \mathbf{p} : (\widehat{\mathbf{p}} - \mathbf{p})' \frac{1}{n} \widehat{\mathbf{\Sigma}}^{-1} (\widehat{\mathbf{p}} - \mathbf{p}) \leq t^{2*}_{(1-\alpha)} \right\}, \tag{6.4}$$

where the critical value $t^{2*}_{(1-\alpha)}$ is determined by the bootstrap distribution (see Efron and Tibshirani, 1993; Davison and Hinkley, 1997) instead of a critical value of the F or χ^2 distribution. The distribution of the bootstrap random variable t^{2*}, is found by creating a series of replicates of the data. We take the Book Donation experiment as an example, with observed data of $(Q_A, Q_{BCF}, R_{AB}, R_{CF}, S_{ABC}, S_F)$. We will describe shortly, in detail, how the bootstrap replicates are created. The bootstrap procedure is to find b bootstrap replicates of $(\widehat{\mathbf{p}} - \mathbf{p})' \widehat{\mathbf{\Sigma}}^{-1}/n \, (\widehat{\mathbf{p}} - \mathbf{p})$ according to Algorithm 1.

Following the work of Hall (1992, p160), and also described by Davison and Hinkley (1997, p231), the elliptical region derived by finding $t^{2*}_{(1-\alpha)}$ in this manner is called the

Algorithm 1 Bootstrap to construct a $1-\alpha$ confidence region for $\mathbf{p}$.

1. Calculate $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{\Sigma}}$ from the original observed data, $(Q_A, Q_{BCF}, R_{AB}, R_{CF}, S_{ABC}, S_F)$, based on the formulas in 5.6.

2. For $i = 1, 2, ..., b$

   (a) Replicate the data, denote it as $(Q^*_A, Q^*_{BCF}, R^*_{AB}, R^*_{CF}, S^*_{ABC}, S^*_F)_i$. We discuss in more detail how to do this shortly.

   (b) From the formulas in 5.6 use $(Q^*_A, Q^*_{BCF}, R^*_{AB}, R^*_{CF}, S^*_{ABC}, S^*_F)_i$ to calculate $\widehat{\mathbf{p}}^*_i$ and $\widehat{\mathbf{\Sigma}}^*_i$.

   (c) Set $t^{2*}_i = (\widehat{\mathbf{p}}^* - \widehat{\mathbf{p}})' \, \widehat{\mathbf{\Sigma}}^{*-1}/n \, (\widehat{\mathbf{p}}^* - \widehat{\mathbf{p}})$

   End loop

3. Set $t^{2*}_{(1-\alpha)}$ to be the $1-\alpha$ quantile of $t^{2*}_1, t^{2*}_2, ..., t^{2*}_b$.

4. Define the confidence region of $\mathbf{p}$ as the set

$$\left\{ \mathbf{p} : (\widehat{\mathbf{p}} - \mathbf{p})' \, \frac{1}{n} \widehat{\mathbf{\Sigma}}^{-1} \, (\widehat{\mathbf{p}} - \mathbf{p}) \leq t^{2*}_{(1-\alpha)} \right\}$$

studentized bootstrap region as it is the multidimensional extension of the *studentized* or *bootstrap-t* method (Efron, 1982, p87). The advantage of the bootstrap comes from using the full empirical distribution of the observed data, instead of just the sufficient statistics for the mean and sample covariance. Rather than assume $(\widehat{\mathbf{p}} - \mathbf{p})' \, \widehat{\mathbf{\Sigma}}^{-1/2}$ is normal, it's distribution may be well approximated by the bootstrap replicates $(\widehat{\mathbf{p}}^* - \widehat{\mathbf{p}})' \, \widehat{\mathbf{\Sigma}}^{*-1/2}$ and the distribution of the replicates may be found via simulation.

The resampling incorporates a number of key features of the experimental setting. First, adhering to the potential outcomes framework, we have a clear data generating process and, though a simple one, a parametric model where the parameter vector is the fraction of behavioral types. Second, the assignment to treatment is done without replacement, which impacts the covariance structure of the joint estimates. We find little in the bootstrap literature on sampling of dependent data of the type in our experiment. Even the aptly named *Resampling Methods for Dependent Data* (Lahiri, 2003, 2013) has little related to RCTs, with complex designs, where assignment is done without replacement. Much of the focus of research on bootstrapping with dependent data has been on times series (Davison, Hinkley, and Young, 2003; Carey, 2005). Third, the data is categorical which is nearly always handled with multinomial sampling scheme though there are notable exceptions such the archaeological work of Lockyear (2013) which employs correspondence analysis and resamples the contingency table with fixed marginal totals.

Bootstrap replicates are created via two main branches of resampling: parametric and nonparametric. In the parametric bootstrap, a parametric distribution is assumed for the observed $\widehat{\mathbf{p}}$. The parameters are estimated and then the full model is used to replicate the data. In contrast, the nonparametric bootstrap makes no modeling assumptions and simply resamples the original observations with replacement. We describe each in more detail and begin with the nonparametric approach as it is more common in the analysis of randomized control trials. Employing the bootstrap for our experiments is a multi-treatment extension of Algorithm 6.1 of Efron and Tibshirani (1993, p47). The nonparametric approach is distribution free as it resamples the data from each experimental group with replacement. As outcomes are binary, this is equivalent to generating each cell in the table of observations from a binomial random variable. In the Book Donation experiment the i^{th} replicate, $(Q^*_A, Q^*_{BCF}, ..., S^*_F)_i$, is generated from the observed $(Q_A, Q_{BCF}, ..., S_F)$ according to Algorithm 2.

Algorithm 2 Nonparametric Bootstrap Replicates

1. Generate Q^*_A from a Binomial(q, Q_A/q) random variable. The total of each experimental group is fixed so $Q^*_{BCF} = q - Q^*_A$.
2. Generate R^*_{AB} from a Binomial(r, R_{AB}/r) random variable with $R^*_{CF} = r - R^*_{AB}$.
3. Generate S^*_{ABC} from a Binomial(s, S_{ABC}/s) random variable with $S^*_F = s - S^*_{ABC}$.

From an examination of the literature this appears to be, by far, the most common way bootstrap replicates are created in RCTs (for example, see Barber and Thompson, 2000; Bachmann, Fairall, Clark, and Mugford, 2007). When creating a bootstrap replicate in this manner the response of each subject is independent of all other subjects in the experiment. This form of sampling is analogous to the infinite population assumption from Section 3.1. One appeal of the nonparametric bootstrap is that it requires no prior knowledge, and thus no assumptions, of the data generating process. In fact, suppose we used a data generating process that was more informed by our behavioral types model according to Algorithm 3, below. Here, we assume each subject belongs to one the behavioral types and the only difference between Algorithm 3 and the framework of Example 4.2.2 of Chapter 4 is the infinite population assumption in Step 1. It can be shown that Algorithm 3 is equivalent to Algorithm 2, that is, the distribution of $(Q^*_A, Q^*_{BCF}, ..., S^*_F)_i$ is identical for both algorithms. Thus, even with it's distribution free features the nonparametric bootstrap still captures many of the features of the underlying model.

In the parametric bootstrap we assume the data is created by a parametric model, in our case the one described in Example 4.2.2 of Chapter 4. We estimate the parameters, $\mathbf{p}$, from the observations and plug $\widehat{\mathbf{p}}$ into the model to generate new replicates. The method for generating replicates via the parametric bootstrap is described in Algorithm 4.

Algorithm 3 Alternative Nonparametric Bootstrap Replicates

1. Use the estimated fraction of behavioral types, $\widehat{\mathbf{p}}$, to generate an IID sample of subjects of size n where each subject distributed as a Multinomial(1,$\widehat{\mathbf{p}}$) random variable.
2. Randomly assign each subject, without replacement, to the control, weak or strong treatments.
3. Tabulate the response totals according to the number of behavioral types in each experimental group to form $(Q_A^*, Q_{BCF}^*, ..., S_F^*)_i$.

Algorithm 4 Parametric Bootstrap Replicates

1. Create an experimental sample of size n where the fraction of each behavioral type exactly matches $\widehat{\mathbf{p}}$.
2. Randomly assign each subject, without replacement, to the control, weak or strong treatments.
3. Tabulate the response totals according to the number of behavioral types in each experimental group to form $(Q_A^*, Q_{BCF}^*, ..., S_F^*)_i$.

We see that Algorithm 4 only differs from Algorithm 3 in the first step. Once again this is exactly the difference between drawing subjects from an infinite population or assigning subjects to treatments from a finite sample. Thus, as we have argued throughout this thesis that the finite sample assumptions are preferred for most randomized studies, Step 1 of Algorithm 4 has a higher fidelity to the data generating process of the behavioral types model. As stated in Efron and Tibshirani (1993, p55-56), "The parametric bootstrap is useful in problems where some knowledge about the form of the underlying population is available, and for comparison to nonparametric analyses." However, as we saw in section 3.4, this may make very little difference in the resulting confidence regions.

To determine which approach we should use for generating the replicates we use the same 10,000 simulations from Section 6.1, this time constructing a nonparametric bootstrap region and a parametric region and recording how often each covers the true parameter vector. Here, the regions are constructed as a result of 40,000 bootstrap replications. The coverage rates are shown in Table 6.2. The coverage probabilities for the nonparametric bootstrap are too high, that is, the regions are too large. Meanwhile, the parametric bootstrap appears to work well, as the coverage rate for the Social Pressure Experiment are close to, and the bounds for coverage rates contain, the nominal rate. Though the coverage probabilities for the Book Donation experiment are higher than the nominal rate they are still much closer than those of the nonparametric bootstrap. In summary, the parametric bootstrap is the

preferred bootstrap procedure and is also a clear improvement on constructing the confidence region from the normal approximation.

	Social Pressure		Book Donation	
	Coverage Rate	Std. Error	Coverage Rate	Std. Error
Nonparametric Bootstrap				
99% region	99.5%	0.07%	99.7%	0.06%
95% region	97.2%	0.17%	98.1%	0.14%
90% region	94.3%	0.23%	95.7%	0.20%
Parametric Bootstrap				
99% region	98.9%	0.10%	99.4%	0.08%
95% region	94.7%	0.22%	96.1%	0.19%
90% region	90.3%	0.30%	90.7%	0.29%

Table 6.2: Coverage rates, and standard error for coverage rates, for parametric and nonparametric bootstrap confidence regions at 99%, 95% and 90% coverage levels, based on 10,000 simulations.

One final question is how many bootstrap replicates are needed to construct the region? For the simulations in Table 6.2 we used 40,000 replicates to determine the cutoff of $t^{2*}_{(1-\alpha)}$. The 40,000 was motivated by the call of Hesterberg (2008) for tens of thousands of bootstrap replicates when constructing one dimensional confidence intervals, challenging the suggestion by Efron and Tibshirani (1993, Ch12) that one or two thousand are sufficient. Table 6.3 shows the coverage probabilities of the regions built from 1,000, 4,000, 10,000 and 40,000 bootstrap replicates. Here, at least for these two experiments, it appears that 1,000 replicates is sufficient. Though the coverage rates for the Book Donation Experiment improve slightly with more replicates, there is no significant difference between the coverage rates using 1,000 replicates and the coverage rates using 40,000 replicates.

In summary, the bootstrap confidence regions are a clear improvement on the normal approximation regions as they correct for any deviations from normality. Also, despite the common appearance in the literature of generating replicates without replacement, as described by Efron, the parametric bootstrap outperforms the nonparametric one and, perhaps, is more in the spirit of the bootstrap as it more closely mimics the data generation process. Yet, for the Book Donation Experiment, even the coverage rates for the parametric bootstrap are still (a little) too high. In the next section we show how performing a two bootstrap process, can improve the coverage levels of the confidence region to reach nominal coverage rates.

	Social Pressure		Book Donation	
	Coverage Rate	Std. Error	Coverage Rate	Std. Error
Parametric Bootstrap with 1,000 replicates				
99% region	98.8%	0.10%	99.5%	0.07%
95% region	94.6%	0.23%	96.2%	0.19%
90% region	89.6%	0.31%	91.3%	0.28%
4,000 replicates				
99% region	99.1%	0.10%	99.5%	0.07%
95% region	95.0%	0.22%	96.1%	0.19%
90% region	90.2%	0.30%	91.3%	0.28%
10,000 replicates				
99% region	99.1%	0.09%	99.6%	0.06%
95% region	95.0%	0.22%	96.1%	0.19%
90% region	90.0%	0.30%	91.4%	0.28%
40,000 replicates				
99% region	98.9%	0.10%	99.4%	0.08%
95% region	94.7%	0.22%	96.1%	0.19%
90% region	90.3%	0.30%	90.7%	0.29%

Table 6.3: Coverage rates, and standard error for coverage rates, for parametric bootstrap confidence regions using 1,000, 4,000, 10,000 and 40,000 bootstrap replicates to determine the region. Some similarities, or differences, may appear to exist solely due to rounding.

6.3 Improved Coverage Rate with the Double Bootstrap

In the previous section we showed via simulations that even the parametric bootstrap regions for the Book Donation experiment do not achieve the nominal coverage rates. This is common for the bootstrap. For example, for a 95% region, suppose that

$$\Pr_{\mathbf{p}}\left((\widehat{\mathbf{p}}-\mathbf{p})'\,\frac{1}{n}\widehat{\boldsymbol{\Sigma}}^{-1}\,(\widehat{\mathbf{p}}-\mathbf{p})\leq t^{2*}_{(.95)}\right)\neq 0.95,$$

where the subscript $\mathbf{p}$ on Pr specifies that the underlying data generating process follows our behavioral types model with true parameter $\mathbf{p}$. The inequality indicates a bias in the

$t^{2*}_{(.95)}$ found from the bootstrap. However, rather than solve for the bias directly, we take another approach, we adjust the amount of the quantile. That is, instead of using the 95^{th} percentile of t^{2*} we would like to find the quantile, q, such that

$$\Pr_{\mathbf{p}}\left((\widehat{\mathbf{p}}-\mathbf{p})'\,\frac{1}{n}\widehat{\mathbf{\Sigma}}^{-1}\,(\widehat{\mathbf{p}}-\mathbf{p})\leq t^{2*}_{(q)}\right)=0.95\,. \tag{6.5}$$

For the Book Donation experiment the coverage rates are higher than the nominal rates. This indicates the 95% region is too large and we should choose a value for q that is smaller than 0.95. Since $\mathbf{p}$ is unknown we have no way to directly estimate the q in Equation 6.5 either analytically or via simulation. Instead, in the spirit of the bootstrap, we work with $\widehat{\mathbf{p}}$ from our table of observations and we find the quantile $\widehat{q}$ such that

$$\Pr_{\widehat{\mathbf{p}}}\left((\widehat{\mathbf{p}}^{*}-\widehat{\mathbf{p}})'\,\frac{1}{n}\widehat{\mathbf{\Sigma}}^{*-1}\,(\widehat{\mathbf{p}}^{*}-\widehat{\mathbf{p}})\leq t^{2**}_{(\widehat{q})}\right)=0.95\,. \tag{6.6}$$

Here, $\Pr_{\widehat{\mathbf{p}}}()$ is the probability distribution generated by the bootstrap process parameterized by the observed $\widehat{\mathbf{p}}$. The $\widehat{\mathbf{p}}^{*}$ and $\widehat{\mathbf{\Sigma}}^{*}$ are the random estimates generated by the bootstrap process and the t^{2**} are the simulated t^2 that result from a *nested*, or *double bootstrap* for each of the original replicates. This procedure was first proposed by Beran (1987) and an overview can be found in Davison and Hinkley (1997, section 5.6).

More concretely, for each of the bootstrap replicates indexed by $i=1,...,b$, we generate $\widehat{\mathbf{p}}^{*}_{i}$, $\widehat{\mathbf{\Sigma}}^{*}_{i}$ and $t^{2*}_{i}=(\widehat{\mathbf{p}}^{*}_{i}-\widehat{\mathbf{p}})'\,\widehat{\mathbf{\Sigma}}^{*-1}_{i}/n\,(\widehat{\mathbf{p}}^{*}_{i}-\widehat{\mathbf{p}})$. And for each of these bootstrap replicates we generate a series of m double bootstrap replicates by resampling from the model parameterized by $\widehat{\mathbf{p}}^{*}_{i}$ to give $\widehat{\mathbf{p}}^{**}_{ij}$, $\widehat{\mathbf{\Sigma}}^{**}_{ij}$ and t^{2**}_{ij} for $j=1,...,m$ (for a grand total of $b\times m$ simulations). The process of the double bootstrap allows us to simulate the joint distribution of $\widehat{\mathbf{p}}^{*}$, $\widehat{\mathbf{\Sigma}}^{*}$, t^{2*} and t^{2**} and estimate $\widehat{q}$. To see this we note that the left side of the inequality in Equation 6.6, by definition, equals t^{2*} so we are searching for the value of $\widehat{q}$ to satisfy

$$\Pr_{\widehat{\mathbf{p}}}\left(t^{2*}\leq t^{2**}_{(\widehat{q})}\right)=0.95\,. \tag{6.7}$$

At first glance, the probability statement in Equation 6.7 may appear ambiguous and we take a moment to describe it explicitly. In the bootstrap procedure, where the fraction of behavioral types is $\widehat{\mathbf{p}}$, a table of observations is generated, and we call the t-statistic from the first bootstrap replicate t^{2*}. From that original replicate we conduct a double bootstrap and call the first replicate of the double bootstrap t^{2**}. Then t^{2*} and t^{2**} are correlated random variables and from their joint distribution we can evaluate probability statements such as the one in Equation 6.7. In this setting define F^{**} to be the cumulative distribution function of t^{2**} so, by definition, $t^{2**}_{(\widehat{q})}=F^{**-1}(\widehat{q})$. We must find $\widehat{q}$ such that

$$\Pr_{\widehat{\mathbf{p}}}\left(t^{2*}\leq F^{**-1}(\widehat{q})\right)=0.95\,.$$

By the monotonicity of F^{**}, $(t^{2*}\leq F^{**-1}(\widehat{q}))$ is equivalent to $(\,F^{**}(t^{2*})\leq F^{**}(F^{**-1}(\widehat{q}))\,)$. Thus, $\widehat{q}$ must satisfy

$$\Pr_{\widehat{\mathbf{p}}}\left(F^{**}(t^{2*})\leq\widehat{q}\right)=0.95\,. \tag{6.8}$$

From our double bootstrap replicates we have all the components to find $\widehat{q}$ from Equation 6.8. We have b replicates, $t_1^{2*}, ..., t_b^{2*}$, of t^{2*} and for each t_i^{2*} we approximate $F^{**}(t_i^{2*})$ from the m double bootstrap replicates, that is, $F^{**}(t_i^{2*})$ is the fraction of $t_{i1}^{2**}, ..., t_{im}^{2**}$ which are less than t_i^{2*}. This gives b values of $F^{**}(t_1^{2*})$, ...,$F^{**}(t_b^{2*})$ and we choose the 95^{th} percentile to be $\widehat{q}$. We summarize this process in Algorithm 5 replacing the 0.95 confidence level with a general $1-\alpha$.

Algorithm 5 Double Bootstrap to construct a $1-\alpha$ confidence region for $\mathbf{p}$.

---

1. Calculate $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{\Sigma}}$ from the observations.
2. For $i = 1, 2, ..., b$
   (a) Replicate the data according to Algorithm 4.
   (b) Calculate $\widehat{\mathbf{p}}_i^*$, $\widehat{\mathbf{\Sigma}}_i^*$ and t_i^{2*}.
   (c) For $j = 1, 2, ..., m$
      i. Replicate the data, using $\widehat{\mathbf{p}}_i^*$, according to Algorithm 4.
      ii. Calculate $\widehat{\mathbf{p}}_{ij}^{**}$, $\widehat{\mathbf{\Sigma}}_{ij}^{**}$ and t_{ij}^{2**}.

      End loop
   (d) Set $F^{**}(t_i^{2*})$ to the fraction of $t_{i1}^{2**}, ..., t_{im}^{2**}$ which are less than t_i^{2*}.

   End loop
3. Set $\widehat{q}$ to be the $1-\alpha$ quantile of $F^{**}(t_1^{2*}), ..., F^{**}(t_b^{2*})$.
4. Define the confidence region of $\mathbf{p}$ as the set

$$\left\{ \mathbf{p} : (\widehat{\mathbf{p}} - \mathbf{p})' \frac{1}{n} \widehat{\mathbf{\Sigma}}^{-1} (\widehat{\mathbf{p}} - \mathbf{p}) \leq t_{(\widehat{q})}^{2*} \right\}$$

The number of double strap replicates, m, can often be much smaller than b (as argued by Davison and Hinkley, 1997, p. 178). To evaluate the double bootstrap we choose m=4000 and $b = 1000$ and again repeat the simulations from the previous section for the parametric bootstrap. The results are in Table 6.4. Comparing these coverage rates to those for the parametric bootstrap in Table 6.2 we see that the double bootstrap appears to partially correct the bootstrap regions, though their is still difference from the nominal coverage rates. It may be that our choice for the number of replicates, or double bootstrap replicates are too low, or that the bias in the bootstrap procedure is not fully corrected with just one iteration of nested bootstrap adjustments. Though computationally more cumbersome we propose the double bootstrap for our working method and turn to the next step, evaluating

what we learn about the parameters from the confidence regions.

	Book Donation	
	Coverage Rate	Std. Error
Double Bootstrap		
99% region	99.4%	0.08%
95% region	95.8%	0.20%
90% region	90.5%	0.29%

Table 6.4: Coverage rates, and standard error for coverage rates, for double bootstrap confidence regions at 99%, 95% and 90% coverage levels for the Book Donation experiment. Regions constructed with 4,000 bootstrap replicates and 1,000 double (nested) replicates.

6.4 Examining the Confidence Regions

For the Social Pressure Experiment, the confidence region reveals a greater degree of uncertainty for the parameters than found using the the confidence intervals in Section 5.1. Figure 6.2 shows the 95% confidence region for the Social Pressure experiment using the parametric bootstrap with 40,000 bootstrap replicates. To represent the five-dimensional region we project it onto two-dimensional scatter plots for each pair of parameters. The regions are discrete, as the value of each parameters must be a whole number multiple of 1/47836. For these scatter plots we show every 47th point (e.g. 1/47836, 48/47836, 95/47836...) so the representation is approximately discretized to 0.001 (one tenth of 1%). With dotted lines we also show the 95% confidence bounds for each of the single parameter estimates, based on the point estimates and standard error in Table 5.6. The plots show the correlation between the parameter estimates, particularly the strong negative correlation between $\widehat{p}_A$ and $\widehat{p}_B$, between $\widehat{p}_B$ and $\widehat{p}_C$ and between $\widehat{p}_C$ and $\widehat{p}_D$. In general, the strongest correlations are found along the diagonal of the plots in Figure 6.2, that is, between the estimators of the "adjacent" parameters such as $\widehat{p}_A$ and $\widehat{p}_B$. To see why, note that $\widehat{p}_A = Q_A/q$ and $\widehat{p}_B = R_{AB}/r - Q_A/q$ so both contain Q_A. The resulting $\text{Cov}(\widehat{p}_A, \widehat{p}_B)$ is a linear combination of the covariance of the observations but the term with the highest magnitude will be $\text{Cov}(Q_A, -Q_A) = -\text{Var}(Q_A)$, which is negative. The same reasoning applies to the other adjacent estimates which gives the strong relationship along the diagonal. The confidence region also shows the degree to which the one-dimensional confidence intervals understate the variability of the point estimates, as for each parameter, the projected region exceeds the bounds of the confidence intervals. For example, the 95% confidence interval for p_B is [1.01%, 3.63%] while the region projected onto p_B spans [0, 4.62%]. In fact, for every

parameter except p_A, the region projected onto the single parameter includes 0, indicating a far greater degree of uncertainty.

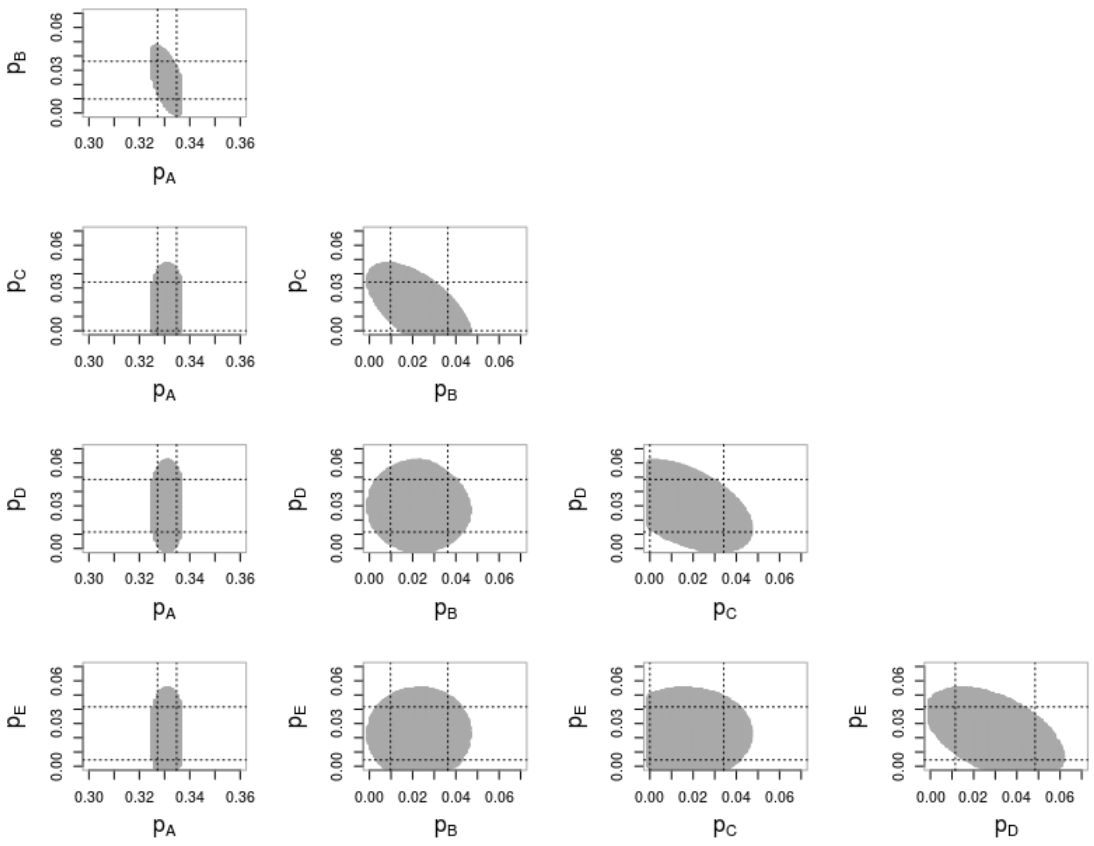


Figure 6.2: Bootstrap confidence region for the Social Pressure experiment. Each plot represents a projection onto a two-dimensional plane for a pair of parameters. The dotted lines indicate the single parameter confidence intervals found in Chapter 5.

The confidence region for the Book Donation experiment also spans a wider range of values for the parameters than those shown by the confidence intervals. Figure 6.3 shows the 95% region derived from the double bootstrap with 4,000 bootstrap replicates and 1,000 double (or nested) replicates. Similar to the other experiment, the three-dimensional region is represented by the two-dimensional projections onto each pair of parameters, shown by the three scatter plots in Figure 6.3. Here, every point of the region is a multiple of 1/11812 and the plots include every second point. Again the 95% region, projected onto the parameters, shows greater upper confidence limits. For example, the projected region for p_B exceeds 2.7%, much wider than the confidence interval upper bound of 2.0%. And again $\widehat{p}_A$ and $\widehat{p}_B$ are correlated as are $\widehat{p}_B$ and $\widehat{p}_C$ (though the linear relationship is somewhat obscured by the truncation). Also, the region fails to confirm one of the conclusions from Section 5.2, that $p_B + p_C > 0$. Though not easy to see from the scatter plot of the region projected along p_B

and p_C, the handful of points close to the origin, such as $(p_B, p_C) = (1/11812, 1/11812)$ lie inside the 95% confidence set. This demonstrates how the confidence regions may lead to different conclusions than those found from the single parameter approaches of Chapter 5.

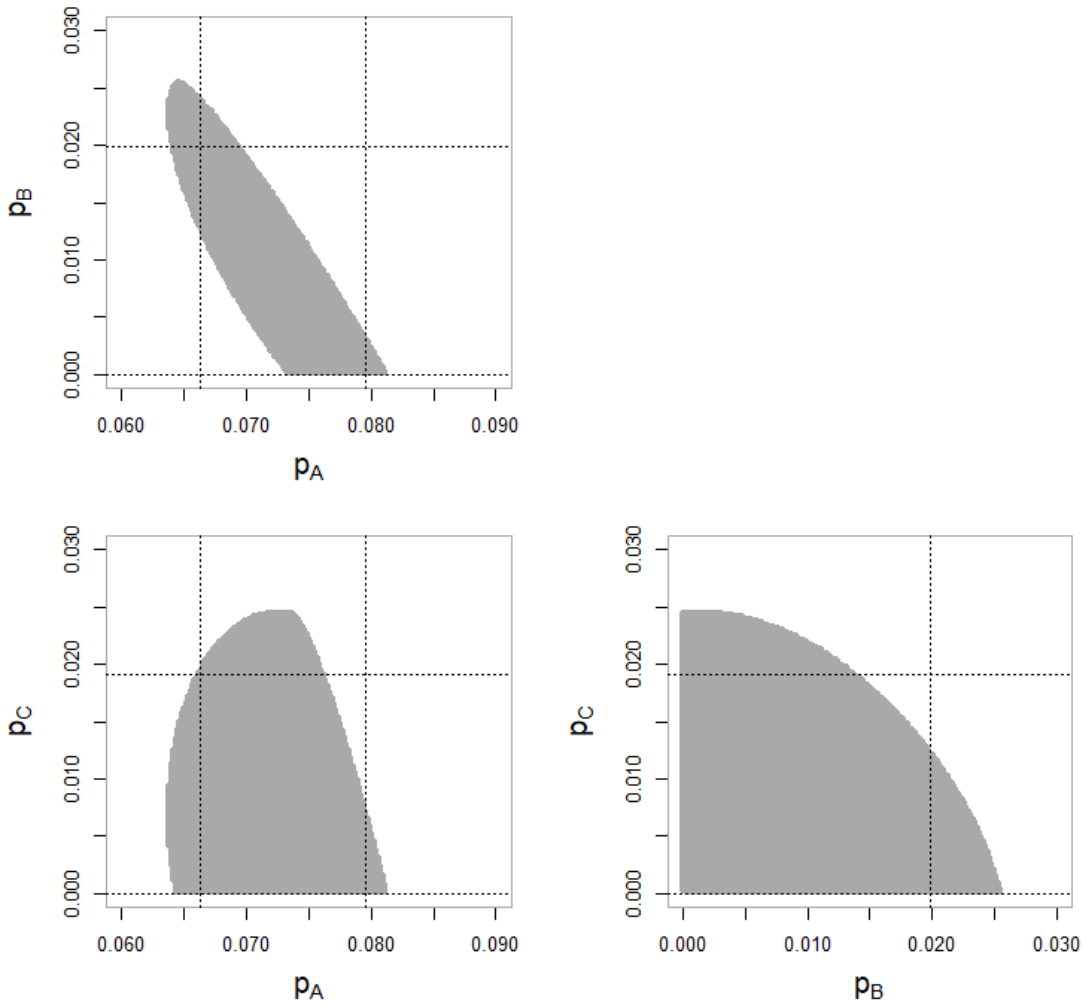


Figure 6.3: Double bootstrap confidence region for the Book Donation experiment. Each plot represents a projection onto a two-dimensional plane for a pair of parameters. The dotted line indicates the single parameter confidence intervals found in Chapter 5.

6.5 Other Approaches to Confidence Regions

In this section we briefly discuss three other approaches we explored in constructing confidence regions. All merit further consideration.

6.5.1 Hypothesis Testing

The region found in Section 6.1, uses the normal approximation. It is part of a more generalized approach based on the duality between confidence regions and hypothesis testing. For any simple hypothesis test with significance of α and null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, accepting the test, i.e. not rejecting the null hypothesis, corresponds with including $\mathbf{p}_0$ in a level $1-\alpha$ confidence region for $\mathbf{p}$. As discussed in Section 6.1, if $\widehat{\mathbf{p}}$ is normally distributed then the test statistic $(\widehat{\mathbf{p}}-\mathbf{p})' \, \widehat{\mathbf{\Sigma}}^{-1}/n \, (\widehat{\mathbf{p}}-\mathbf{p})$ may be used to form a uniformly most powerful test to evaluate whether $\mathbf{p} = \mathbf{p}_0$. The region it forms will be a uniformly most accurate confidence region. However, as we demonstrated in Sections 6.1 and 6.2 the deviations from normality are sufficient enough that the distribution of $(\widehat{\mathbf{p}}-\mathbf{p})' \, \widehat{\mathbf{\Sigma}}^{-1}/n \, (\widehat{\mathbf{p}}-\mathbf{p})$ differs substantially from the χ^2 distribution. However the true distribution of $\widehat{\mathbf{p}}$ is challenging to evaluate analytically.

To demonstrate the challenge we first recall that since the sample sizes of the experimental groups are fixed there is a one-to-one correspondence between $\widehat{\mathbf{p}}$ and the table of observations. As a concrete example, we focus on the Book Donation experiment as it is the simplest case with a control and only two levels of treatment. The table of observations is (Q_A, Q_{BCF}, R_{AB}, R_{CF}, S_{ABC}, S_F) but since the total of each experimental group, q, r, and s are fixed, only the triple (Q_A, R_{AB}, S_{ABC}) is needed for the likelihood. Let q_A, r_{AB}, and s_{ABC} represent the specific realizations of the observed values. In place of $\mathbf{p}$ we use the equivalent parameters of the total of each behavioral type, $\mathbf{n} = n\,\mathbf{p} = (n_A, n_B, n_C, n_D)$. The likelihood function is

$$
\begin{aligned}
&\mathit{likelihood}(\mathbf{p}, \widehat{\mathbf{p}}) \\
&\quad = \mathit{likelihood}(\mathbf{n}, q_A, r_{AB}, s_{ABC}) \\
&\quad = \Pr{}_{\widehat{\mathbf{n}}}(Q_A{=}q_A, R_{AB}{=}r_{AB}, S_{ABC}{=}s_{ABC}) \\
&\quad = \Pr{}_{\widehat{\mathbf{n}}}(Q_A{=}q_A) \Pr{}_{\widehat{\mathbf{n}}}(R_{AB}{=}r_{AB}, S_{ABC}{=}s_{ABC} \mid Q_A{=}q_A) \\
&\quad = \underbrace{\Pr{}_{\widehat{\mathbf{n}}}(Q_A{=}q_A)}_{(a)} \underbrace{\Pr{}_{\widehat{\mathbf{n}}}(R_{AB}{=}r_{AB} \mid Q_A{=}q_A)}_{(b)} \underbrace{\Pr{}_{\widehat{\mathbf{n}}}(S_{ABC}{=}s_{ABC} \mid Q_A{=}q_A, R_{AB}{=}r_{AB})}_{(c)}.
\end{aligned}
\tag{6.9}
$$

To see why the likelihood function is intractable we examine each of the components (a),(b) and (c). First, we see that (a) is found immediately from the hypergeometric. More complicated is (b) where we must further condition on R_A, the number of A-types in the weak treatment, that is

$$
\begin{aligned}
(b) &= \Pr{}_{\widehat{\mathbf{n}}}(R_{AB}{=}r_{AB} \mid Q_A{=}q_A) \\
&= \sum_{i=0}^{min(r_{AB}, n_A)} \Pr{}_{\widehat{\mathbf{n}}}(R_A{=}i, R_B{=}r_{AB}-i \mid Q_A{=}q_A)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=0}^{min(r_{AB},n_A)} \Pr_{\widehat{\mathbf{n}}}(R_A{=}i \mid Q_A{=}q_A) \Pr_{\widehat{\mathbf{n}}}(R_B{=}r_{AB}-i \mid Q_A{=}q_A, R_A{=}i) \\
&= \sum_{i=0}^{min(r_{AB},n_A)} \frac{\binom{n_A-q_A}{i}\binom{n-n_A}{r-i}}{\binom{n-q_A}{r}} \frac{\binom{n_B}{r_{AB}-i}\binom{n-n_B-i}{r-r_{AB}}}{\binom{n-i}{r-i}} \qquad (6.10)
\end{aligned}
$$

where the last substitution holds as both probabilities can also be evaluated from the hypergeometric distribution. With multiple factorials, the summation in Equation 6.10 is challenging to evaluate. Furthermore, (c) is even more complicated involving a double summation. The intractable likelihood hampers our ability to derive a likelihood ratio test, and without such a test there is no obvious optimal test procedure.

Even without optimality properties guaranteed, the distribution of $(\widehat{\mathbf{p}}-\mathbf{p})' \widehat{\boldsymbol{\Sigma}}^{-1}/n\,(\widehat{\mathbf{p}}-\mathbf{p})$ is similar enough to χ^2 that it is still a useful test statistic. In place of a known distribution we may use Monte Carlo simulation to determine the distribution, and hence the critical values of $(\widehat{\mathbf{p}}-\mathbf{p})' \widehat{\boldsymbol{\Sigma}}^{-1}/n\,(\widehat{\mathbf{p}}-\mathbf{p})$ under the null hypothesis $\mathbf{p}=\mathbf{p}_0$. This means that at every possible point in the confidence region we must carry out enough Monte Carlo simulations, say 10,000, to obtain the critical values and evaluate whether the point is in the acceptance region. In a three parameter problem it can be shown there are $(n+1)(n+2)(n+3)/6$ possible points in the parameter space. For the Book Donation experiment, with n=11,812, this is more than 10^{11} points. But we may take advantage of the convexity of the confidence region, and by solving for reasonable values of the parameters, we can test for far fewer points.

Another possible test statistic is the maximum modulus, or maximal deviation, attributed to Tukey (1953). This measures how far each of the components of $\widehat{\mathbf{p}}$ are from a tested $\mathbf{p}_0$, after standardization. Define $\mathbf{p}_z$ to be the difference between $\widehat{\mathbf{p}}$ and $\mathbf{p}_0$ after standardization,

$$\mathbf{p}_z = (\widehat{\mathbf{p}}-\mathbf{p}_0)' \widehat{\boldsymbol{\Sigma}}^{-1/2}/\sqrt{n}.$$

The maximum modulus statistic is the component of $\mathbf{p}_z$ with the largest deviation from 0 or

$$max\left\{|p_{zi}| : i = 1, ..., l\right\}.$$

In our experience, applying the procedures to a number of experiments, neither method dominates the other. The maximum modulus often demonstrates higher power when the significance level is 10% or higher but any advantage diminish at lower sizes. The regions do look quite different with the $(\widehat{\mathbf{p}}-\mathbf{p})' \widehat{\boldsymbol{\Sigma}}^{-1}/n\,(\widehat{\mathbf{p}}-\mathbf{p})$ leading to elliptical regions while the maximum modulus results in confidence sets with sharp, angular boundaries.

One advantage of constructing regions via hypothesis testing is we may employ tests which cater to particular research objectives. A recent example is the work of Benjamini, Madar, and Stark (2013) to create non-centered confidence bounds to maximize the discernment of the sign of an effect, which decreases the confidence bounds. Their assumptions of independence between samples and symmetric distributions for the parameter estimates

does not hold for our setting. However, the work models how more flexible regions could be developed, which in the case of our applications, lead to an enhanced ability to detect non-zero effects. Developing such regions may be a promising area for future inquiry.

6.5.2 A Bayesian Approach

Given the wealth of information from Get-Out-The-Vote experiments we might begin an analysis of the Social Pressure experiment with prior knowledge about the range of possible values for the parameters $\mathbf{p}$. To approach the problem from a Bayesian perspective we translate this knowledge to some prior distribution, $\pi(\mathbf{p})$, which is used to compute the posterior distribution, $g()$, given observed values. To describe the process more clearly we introduce a notational change in this section. We denote the parameter estimates from the observed values as $\widehat{\mathbf{p}}_{obs}$, a fixed number, which we distinguish from the parameter estimates as a random variable which we write as $\widehat{\mathbf{P}}$. We reserve $\widehat{\mathbf{p}}$ as a specific value for the random variable $\widehat{\mathbf{P}}$. Thus, the posterior distribution of $\mathbf{p}$ is

$$g(\mathbf{p}|\widehat{\mathbf{p}}_{obs}) = \frac{\Pr_{\mathbf{p}}(\widehat{\mathbf{P}} = \widehat{\mathbf{p}}_{obs})\,\pi(\mathbf{p})}{\Pr(\widehat{\mathbf{P}} = \widehat{\mathbf{p}}_{obs})}. \tag{6.11}$$

Equation 6.11 shows the posterior distribution is proportional to the likelihood times the prior, as the denominator of the right hand side is considered a scale parameter for the posterior to integrate to 1. We still have the same analytical obstacles with the likelihood which motivated the Monte Carlo simulations in Section 6.5.1. Because the likelihood, $\Pr_{\mathbf{p}}(\widehat{\mathbf{P}} = \widehat{\mathbf{p}})$, is difficult to evaluate in closed form we cannot directly compute the posterior confidence regions. Fortunately, given $\mathbf{p}$ it is easy to simulate $\widehat{\mathbf{P}}$. Combined with the prior, we may simulate $(\mathbf{p}, \widehat{\mathbf{P}})$ from the joint density

$$g(\mathbf{p}, \widehat{\mathbf{p}}) = \Pr_{\mathbf{p}}(\widehat{\mathbf{P}} = \widehat{\mathbf{p}})\pi(\mathbf{p}).$$

After producing simulated values of $(\mathbf{p}, \widehat{\mathbf{P}})$, if there is a sufficient number of values so that $\widehat{\mathbf{P}}$ is"close" to $\widehat{\mathbf{p}}_{obs}$ we can estimate the density $g(\mathbf{p}, \widehat{\mathbf{p}})$ in the neighborhood of $\widehat{\mathbf{p}}_{obs}$ to obtain an approximation of $g(\mathbf{p} \mid \widehat{\mathbf{p}}_{obs})$. This technique is called Approximate Bayesian Computation (ABC) and is based on an idea first proposed by Rubin et al. (1984) though much of it's development has been from applications in the biological sciences (see Beaumont, Zhang, and Balding, 2002; Beaumont, 2010). With the posterior distribution, we can then calculate 1-α *credible regions*, the Bayesian equivalent of confidence regions.

There are a number of challenges to constructing regions via ABC. The first being how close must a simulated $\widehat{\mathbf{P}}$ be to $\widehat{\mathbf{p}}_{obs}$, to be "close enough"? For continuous $\widehat{\mathbf{P}}$ some threshold, for some distance metric, must be specified. Even for discrete $\widehat{\mathbf{P}}$ such as in our applications, where it is a multiple of $1/n$, it may be challenging to obtain enough values at $\widehat{\mathbf{P}} = \widehat{\mathbf{p}}_{obs}$ and using points in the neighborhood of $\widehat{\mathbf{p}}_{obs}$ may be desired. The next question is how to obtain enough simulated values of $(\mathbf{p}, \widehat{\mathbf{P}})$ with $\widehat{\mathbf{P}} = \widehat{\mathbf{p}}_{obs}$ so that there are enough $\mathbf{p}$'s to

estimate $g(\mathbf{p} \mid \widehat{\mathbf{p}}_{obs})$. In our experience, if the sample size of the experiment is small and if the dimension of $\mathbf{p}$ is low, then repeated sampling from the joint distribution may provide a large enough pool to approximate $g()$. For larger n, or for larger dimensions this may not be the case. To obtain more sample points in the neighborhood of $\widehat{\mathbf{P}} = \widehat{\mathbf{p}}_{obs}$ one solution is importance sampling (see Ripley, 2009) which over samples from values of $\mathbf{p}$ likely to lead to values of $\widehat{\mathbf{P}}$ near $\widehat{\mathbf{p}}_{obs}$. Another tack we tried was to use Markov Chain Monte Carlo methods to create a Markov Chain with a stationary distribution equal to $g(\mathbf{p} \mid \widehat{\mathbf{p}}_{obs})$. This is described in more detail by Marjoram, Molitor, Plagnol, and Tavaré (2003).

Another benefit of the ABC technique is that we can apply it to one dimensional parameters as well. Suppose our interest was primarily on p_A. In the same manner used to approximate $g(\mathbf{p} \mid \widehat{\mathbf{p}}_{obs})$ we may use ABC to approximate the marginal distribution for p_A, $g(p_A \mid \widehat{\mathbf{p}}_{obs})$, by estimating the univariate distribution instead of the multivariate one. Recall that the main motivation for this chapter was to account for the joint significance of single value parameter estimates. From a Bayesian perspective this is not a concern as the marginal posterior distribution can be accessed immediately and provides all of the information need to create a level $1 - \alpha$ credible interval.

For all examples we examined, the ABC methods lead to smaller regions and intervals than other methods we have discussed in this chapter and the previous one. The smaller ABC regions were found even using fairly "non-informative" priors. For example, in the Book Donation experiment a non-informative prior might be independent uniform priors on [0,.2] for p_A and [0,.1] for p_B and p_C. Once one accepts the Bayesian view of probability, the intervals are much smaller, but this is more a result of the Bayesian perspective than any particular methodological advantage. We offer no argument either for or against a Bayesian view but we acknowledge that this thesis has a clear frequentist bent because we have presented the behavioral types and the Neyman-Rubin Causal Model as entrenched in frequentist notions of parameters as fixed features of populations. There is nothing inherently frequentist about the potential outcomes framework as Imbens and Rubin (1997) present many of the ideas from their earlier work from a Bayesian perspective.

6.5.3 Confidence regions of optimal expected size

Our difficulties with the hypothesis testing approach of section 6.5.1 stemmed from an inability to find a uniformly most powerful test for a simple hypothesis. An alternative route is to choose a different criteria of optimality and in this section we consider finding confidence regions of smallest expected size. We begin with a review of decision theoretic concepts, building on the Bayesian notions introduced in the previous section, which allow for a discussion of optimality. We continue to frame the concepts in terms in terms of our problem (e.g. our decision rules are choices of confidence region) using notation which differs from most treatments of decision theory.

We define the *least favorable prior* (often called the "least favorable alternative"). Suppose we have a risk function, $Risk(\mathbf{p}, c)$, where $\mathbf{p}$ is the parameter and c, or $c(\widehat{\mathbf{p}})$, is a

confidence region procedure (such as any of the ones described in this chapter) based on the observed $\widehat{\mathbf{p}}$. Let $\mathcal{C}$ represent the set of all possible confidence regions. For any prior distribution $\pi(\mathbf{p})$ there is a *Bayes rule* region, c_π, to minimize the average risk, $\mathbb{E}_\pi\left[Risk(\mathbf{p}, c)\right]$, for all c of $\mathcal{C}$. The *least favorable prior*, π^*, is the prior whose Bayes rule confidence region, c_{π^*}, has greater average risk than the Bayes rule for any other prior. That is,

$$\mathbb{E}_\pi\left[Risk(\mathbf{p}, c_{\pi^*})\right] \geq \mathbb{E}_\pi\left[Risk(\mathbf{p}, c_\pi)\right] \quad \text{for all } \pi.$$

Next, we describe our research question in terms of minimax risk, an optimality criteria which holds in either the Bayesian or frequentist perspectives. Suppose we constrain our parameter $\mathbf{p}$ to some restricted set Θ (in our case $\mathbf{p}$ is a real-valued vector whose components sum is at most 1). For any risk function $Risk(\mathbf{p}, c)$, over the parameter space Θ we can find the *minimax risk*,

$$\inf_{c \in \mathcal{C}} \sup_{\mathbf{p} \in \Theta} Risk(\mathbf{p}, c).$$

A confidence region that attains the minimax risk, a minimax region, will have the "best worst case" risk. There exists a duality between these two concepts: *the Bayes rule confidence region for the least favorable prior is the minimax confidence region.* In many cases it may be very hard to find a minimax estimator but one may be able to find or approximate the least favorable prior. If this is possible, and we can solve for it's Bayes Rule, we will have a minimax estimator.

Evans, Hansen, and Stark (2005) consider the risk function of the expected size (length, area, volume, etc.) of a region, that is

$$Risk(\mathbf{p}, c) = \mathbb{E}_{\mathbf{p}}(\text{ size of } c(\widehat{\mathbf{p}})),$$

where $\widehat{\mathbf{p}}$ is the random quantity that will determine the region. In their work they show that for any prior, the confidence region constructed by a likelihood ratio test has the smallest expected size; that is, *is the* Bayes rule region. Furthermore, suppose that the set of possible priors, Γ, is convex. A least favorable prior, π^*, for Θ may be found and the confidence region constructed with the likelihood ratio test for π^* will be a minimax expected size region over the parameter space Θ. In general, analytically finding π^* is only possible for simple cases. However, if Θ may be approximated by a finite number of points, Schafer, Stark, Evans, and Hansen (2003); Schafer and Stark (2009) develop an algorithm to find π^* and the corresponding minimax expected size region.

The algorithm does not require a closed form solution of the likelihood and makes use of simulation methods to approximate it, so may be applied to our model. We consider this work an important contribution that finds optimal regions for a very useful measure of risk. It also incorporates constraints put upon the parameters. To our knowledge, there have been no new applications of the method though we think it a very broad approach useful for situations, such as our own estimation problem, when typical theoretical restrictions, such as normality assumptions, do not hold.

6.6 Discussion

In this chapter, we show how construction of confidence regions can lead to a more complete understanding of the variation of the estimated parameters, the proporttion of the behavioral types in an experimental sample. We show that even though the normal approximation does not hold, and the likelihood function is challenging to evaluate analytically, confidence regions which achieve their nominal levels can be attained. We propose finding the regions using a double bootstrap approach but we describe other methods, all simulation-based, which merit further exploration. We consider this chapter an initial step towards what could be a much deeper focus of research.

We have only considered the simplest case of experiments with multiple, and strictly ordered, treatments without noncompliance. However, these methods immediately apply to the experiments with noncompliance, such as the ones described in Chapter 2 and in section 5.4. In all cases the objective is the same; we must estimate the fraction of behavioral types. As long as the parameters $\mathbf{p}$ are identifiable they may be estimated by $\widehat{\mathbf{p}}$, and we may construct regions in the same manner.

Chapter 7

Conclusion

In this research we show that for randomized control trials with binary outcomes, each individual in the study may be classified as one of a finite number of distinct *types*. We call these *behavioral types* because they characterize the individual's complete reaction, their measured response and how they receive treatment, to the assignment of each possible experimental group. In this setting, the contingency table that summarizes the observed results is generated by randomly allocating these various behavioral types to the different experimental groups. Since the model is parameterized by the unknown proportions of the different behavioral types, every statistical aspect of the experiment, such as the various average treatment effects, may be written as a function of these proportions. This suggests a different focus for the estimation problem. Instead of finding a particular treatment effect, the ultimate goal can be seen as estimating proportions of behavioral types. With this frame of reference, the effect of a certain treatment will be be most accurately represented as the fraction of the experimental sample for which the treatment has an effect.

While this work was motivated by estimation problems in get-out-the-vote (GOTV) campaigns, the behavioral-types approach may be used in a number of settings, as long as the the experimental design includes a few key features. While we have narrowed our analysis to outcomes that are binary, the results can be easily extended to categorical responses. The more categories for the outcome, the more behavioral types, though this total may be reduced if outcome categories are ordinal. Another common feature is noncompliance to the assigned treatment, which is essential for experiments with human subjects who may choose to adhere to the assigned treatment, or not. When compliance is unknown, or perfect, the estimation problem becomes simpler as multiple behavioral types may be combined into one. The treatment assigned and the treatment received must also be categorical and, again, the estimating problem will be easier if the categories are ordinal. Finally, the experiment should feature some restrictions which limit the treatment received or individual responses. These restrictions may be explicit, such as monotonicity or the exclusion restrictions, or they may be implicit in the experimental design, such as no individuals assigned to control receive the treatment (or, there are no "alwaystakers").

Through the examples and applications in this work we strive to show how the behavioral-types approach is general enough for use in a variety of randomized studies. We introduced many of the concepts by examining the average treatment effect for the treated of Angrist et al. (1996), since this is a well understood work, with explicit assumptions that can be readily recognized as restrictions. All examples, those from Sections 4.2.2, 4.2.3 and the four applications from Chapter 5, pertain to investigations with multiple levels of (mostly ordered) treatment. A behavioral-types approach is well suited to multi-treatment experiments because it distills these often complex designs into an estimation problem of a manageable number of types. In sections 4.2.3 and 5.3 the experimental designs contain violations of the stable unit treatment value assumption which allow estimation of spillover effects. In Sections 4.2.3 and 5.4 we address experiments whose parameters are estimated by an overdetermined systems of equations. In Sections 5.3 we have a complex randomization scheme and even though we are not able to fully identify the model, the partial identification of parameters still results in useful conclusions. In Section 5.4 we show that, even if the orderings of the treatments are ambiguous, we can still proceed with the estimation problem. Each of these examples present their own complications, but since the eventual goal is to find a proportion of types, each inference problem can be understood as one of discrete valued estimation. Once the parameters, meaning the number of distinct behavioral types, are explicitly identified, the inference problem is addressed by (often well) known statistical techniques.

Though a behavioral-types approach shifts the interpretation of the causal effects the experiment is designed to measure, sometimes this has no impact on how we evaluate the data. For example, from a behavioral types perspective, the treatment received is as much of an outcome as the binary response of interest, even though this does not change the statistical analysis. Viewing experimental subjects as belonging to a certain behavioral type suggests, particularly for small sample sizes, an analysis via Fisher randomization inference along the lines of what was described in Section 3.6. In that section, the attributable effect due to treatment is interpreted as the number of complier-if-treated-respond subjects assigned to the treatment group. An analysis of the effects would not be impacted. Other analyses with a behavioral-types approach would lead to the same conclusion. For instance, a hypothesis test of no effect of (perhaps multiple levels of) treatment is equivalent to there being only behavioral types who always respond or never respond. Both interpretations would likely lead to an analysis via a Fisher exact test, resulting in the same observed p-value.

In other settings, the behavioral-types approach leads to meaningful differences in either how the results are viewed or to different conclusions altogether. We saw this in three of the applications in Chapter 5. For example, in Section 5.2, while we agree with the conclusions of the author, we differ in our interpretation of the causal effect of interest and how this relates to the existence of certain kinds of behavioral types. In Section 5.3, taking a behavioral types perspective helps illuminate that a Fisher sharp null hypothesis test can be used to evaluate the existence of a spillover effect. We find stronger evidence for the presence of the indirect effect. And in Section 5.4 we show that understanding the experimental design

through a behavioral types lens reveals treatment-for-the-treated effects may be estimated, thus furthering the analysis.

The methods presented in this research could be improved in a number of ways. While our focus has been on *sample* average treatment effects, there has been much research into *conditional* average treatment effects, that is the average treatment effect for individuals with certain attributes (see Imbens and Rubin, 2015, p269). From a behavioral types view, the approach is the same and the statistical endeavor centers around estimating the fraction of behavioral types for the subset of subjects with certain covariate values. Though we have confined ourselves to randomized controlled trials, the parameter identification exercises in Chapters 2 and 4 pertain to observational studies as well. The lack of randomized assignment to experimental group presents many difficulties and inference under this context is addressed by Rosenbaum (2002). Also, the important restriction of monotonicity may not hold for medical experiments where receiving a treatment cannot be guaranteed to lead to higher outcomes. For example, some cancer patients fare better and some fare worse under chemotherapy than if no treatment was administered. A behavioral-types approach might still be fruitful without monotonicity, though the identification of the parameters will be more complicated.

Furthermore, a number of recent statistical publications address issues material to this work such as inference with noncompliance and outcomes that are discrete and finite. And many apply randomization based inference which do not require large sample sizes. Examples include Keele et al. (2017), Sekhon and Shem-Tov (2017), Ding and Miratrix (2017), Ding and Dasgupta (2016) and Kang, Peck, and Keele (2016). Exploring the connections between these studies, and our own, would surely yield useful insights to incorporate into this work.

Future research could also look further into issues of multiparameter inference and multiple testing. Our exploration of the topic in Chapter 6 is a beginning to this line of inquiry. We illustrate the dependency between the multiple parameter estimates and the degree to which one dimensional intervals overstate the confidence level around parameter estimates. Aside from bootstrap methods, we suggest three alternative routes to constructing confidence regions and all merit further investigation.

Finally, we have aimed this work for practitioners of social science field experiments, including those of GOTV campaigns. We hope our work clarifies the underlying mechanisms at work in experiments with categorical outcomes and adds a new perspective to direct the resulting statistical analysis.

Bibliography

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1984.

Joshua D. Angrist and Guido W. Imbens. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrument variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Peter M Aronow, Cyrus Samii, et al. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

Joseph L Awange and Béla Paláncz. Solutions of overdetermined systems. In *Geospatial Algebraic Computations*, pages 89–112. Springer, 2016.

Max O Bachmann, Lara Fairall, Allan Clark, and Miranda Mugford. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost effectiveness and resource allocation*, 5(1):12, 2007.

Julie A Barber and Simon G Thompson. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in medicine*, 19(23):3219–3236, 2000.

Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406, 2010.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Lisa Garcia Bedolla and Melissa R Michelson. *Mobilizing inclusion: Transforming the electorate through get-out-the-vote campaigns*. Yale University Press, 2012.

Yoav Benjamini, Vered Madar, and Philip B Stark. Simultaneous confidence intervals uniformly more likely to determine signs. *Biometrika*, 100(2):283–300, 2013.

Rudolf Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.

Yvonne MM Bishop, Stephen E Fienberg, and PW Holland. Discrete multivariate analysis: Theo y and practice, 1975.

Jake Bowers and Ben Hansen. Attributing effects to a get-out-the-vote campaign using full matching and randomization inference, April 2005. URL http://www.test.org/doe/.

Vincent J Carey. Resampling methods for dependent data: A review. *Journal of the American Statistical Association*, 100(470):712–713, 2005.

Jack Citrin, Donald P Green, and Morris Levy. The effects of voter id notification on voter turnout: Results from a large-scale field experiment. *Election Law Journal*, 13(2):228–242, 2014.

S Cotterill, P John, and L Richardson. The impact of a pledge campaign and the promise of publicity on charitable giving: a randomised controlled trial of a book donation campaign. In *Randomised Controlled Trials in the Social Sciences Conference, York, September 2010*, 2010.

Sarah Cotterill, Peter John, and Liz Richardson. The impact of a pledge request and the promise of publicity: A randomized controlled trial of charitable donations. *Social Science Quarterly*, 94(1):200–216, 2013.

Anthony C Davison, David V Hinkley, and G Alastair Young. Recent developments in bootstrap methodology. *Statistical Science*, pages 141–157, 2003.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

Peng Ding and Tirthankar Dasgupta. A potential tale of two-by-two tables from completely randomized experiments. *Journal of the American Statistical Association*, 111(513):157–168, 2016.

Peng Ding and Luke Miratrix. Model-free causal inference of binary experimental data. 05 2017.

Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1993.

Steven N Evans, Ben B Hansen, and Philip B Stark. Minimax expected measure confidence sets for restricted location parameters. *Bernoulli*, pages 571–590, 2005.

Klaus Fiedler, Florian Kutzner, and Joachim I Krueger. The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, 7(6):661–669, 2012.

David Freedman, Robert Pisani, and Roger Purves. Statistics (3rd edn), 1998.

Alan S. Gerber and Donald P. Green. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *America Political Science Review*, 94(3):653–663, 2000.

Alan S Gerber, Donald P Green, and Christopher W Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(01):33–48, 2008.

Alan S Gerber, Gregory A Huber, Albert H Fang, and Catlan E Reardon. When does increasing mobilization effort increase turnout? new theory and evidence from a field experiment on reminder calls. *Institution for Social and Policy Studies, Yale University*, 2016.

Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

Donald P Green and Alan S Gerber. Getting out the youth vote: Results from randomized field experiments. *Unpublished report to the Pew Charitable Trusts and Yale Universitys Institute for Social and Policy Studies*, 2001.

Donald P Green and Alan S Gerber. *Get out the vote: How to increase voter turnout*. Brookings Institution Press, 2015.

Donald P Green, Alan S Gerber, and David W Nickerson. Getting out the vote in local elections: results from six door-to-door canvassing experiments. *Journal of Politics*, 65 (4):1083–1096, 2003.

Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 1992.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

Tim Hesterberg. Its time to retire the n ¿= 30 rule, 2008. URL https://ai.google/research/pubs/pub34906.

Joseph Lawson Hodges and Erich L. Lehmann. *Basic Concepts of Probability and Statistics*. Holden-Day, 1964.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–60, 1986.

H Hotelling. The generalization of" student's" ratio. *Annals of Mathematical Statistics*, 1931.

Kousuke Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, 2005.

Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Institution for Social and Policy Studies. Get-out-the-vote research page. URL http://gotv.research.yale.edu/.

NL Johnson, S Kotz, and N Balakrishnan. Discrete multivariate distributions. 1997.

Hyunseung Kang, Laura Peck, and Luke Keele. Inference for instrumental variables: a randomization inference approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2016.

Luke Keele, Dylan Small, and Richard Grieve. Randomization-based instrumental variables methods for binary outcomes with an application to the improvetrial. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2):569–586, 2017.

Soumendra Nath Lahiri. Resampling methods for dependent data. *New York*, 2003.

Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520): 1759–1769, 2017.

Kris Lockyear. Applying bootstrapped correspondence analysis to archaeological data. *Journal of Archaeological Science*, 40(12):4744–4753, 2013.

Sharon Lohr. *Sampling: design and analysis*. Nelson Education, 2009.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328, 2003.

Richard E Matland and Gregg R Murray. An experimental test of mobilization effects in a latino community. *Political Research Quarterly*, 65(1):192–205, 2012.

Peter McCullagh and James A Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.

Melissa R Michelson. Getting out the latino vote: How door-to-door canvassing influences voter turnout in rural central california. *Political Behavior*, 25(3):247–263, 2003.

Melissa R Michelson, Lisa García Bedolla, and Margaret A McConnell. Heeding the call: The effect of targeted two-round phone banks on voter turnout. *The Journal of Politics*, 71(4):1549–1563, 2009.

Rupert G Miller. *Simultaneous statistical inference*. Springer, 1981.

Jerzey Neyman. On the application of probability theory to agricultural experiments. translated and edited by d.m. dabrowska and t.p. speed (1990). 5(4):465–472, 1923.

David W. Nickerson. Quality is job one: Professional and volunteer voter mobilization calls. *American Journal of Political Science*, 51(2):269–282, 2007.

Costas Panagopoulos. Timing is everything? primacy and recency effects in voter mobilization campaigns. *Political Behavior*, 33(1):79–93, 2011.

Thomas V Perneger. Whats wrong with bonferroni adjustments. *BMJ: British Medical Journal*, 316(7139):1236, 1998.

Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

Paul R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88:219–231, 2001.

Paul R Rosenbaum. Observational studies. In *Observational studies*. Springer, 2002.

Paul R Rosenbaum. Comment: the place of death in the quality of life. *Statistical Science*, 21(3):313–316, 2006.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin et al. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

Chad M Schafer and Philip B Stark. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104(487):1080–1089, 2009.

Chad M Schafer, Philip B Stark, Steve Evans, and Ben Hansen. Using what we know: Inference with physical constraints. *Stat. Probl. Part. Phys, Astrophys. Cosmol*, pages 25–34, 2003.

Jasjeet S Sekhon and Yotam Shem-Tov. Inference on a new class of sample average treatment effects. *arXiv preprint arXiv:1708.02140*, 2017.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1): 561–584, 1995.

Betsy Sinclair, John Balz Rogowski, Alex Bass, Jaira Harrington, et al. Design and analysis of experiments in multilevel populations. *Cambridge handbook of experimental political science*, page 906, 2011.

Betsy Sinclair, Margaret McConnell, and Donald P Green. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069, 2012.

Alfred Sommer and Scott L. Zeger. On estimating efficacy from clinical trials. *Statistics in Medicine*, 10:45–52, 1991.

Frederick F Stephan. The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. *The Annals of Mathematical Statistics*, 16(1): 50–61, 1945.

J. W. Tukey. The problem of multiple comparisons. *Unpublished manuscript*, 1953.

Tyler J VanderWeele, Eric J Tchetgen Tchetgen, and M Elizabeth Halloran. Interference and sensitivity analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):687, 2014.

Marko Znidaric. Asymptotic expansion for inverse moments of binomial and poisson distributions. *arXiv preprint math/0511226*, 2005.