On Causal Inference for Ordinal Outcomes

On Causal Inference for Ordinal Outcomes

A dissertation presented

by

Jiannan Lu

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

July 2015

Dissertation Advisors: Author:
Professors Tirthankar Dasgupta and Joseph K. Blitzstein Jiannan Lu

On Causal Inference for Ordinal Outcomes

Abstract

This dissertation studies the problem of causal inference for ordinal outcomes. Chapter 1 focuses on the sharp null hypothesis of no treatment effect on all experimental units, and develops a systematic procedure for closed-form construction of sequences of alternative hypotheses in increasing orders of their departures from the sharp null hypothesis. The resulted construction procedure helps assessing the powers of randomization tests with ordinal outcomes. Chapter 2 proposes two new causal parameters, i.e., the probabilities that the treatment is beneficial and strictly beneficial for the experimental units, and derives their sharp bounds using only the marginal distributions, without imposing any assumptions on the joint distribution of the potential outcomes. Chapter 3 generalizes the framework in Chapter 2 to address noncompliance.

Contents

Citations to Previously Published Work

Chapters 1 and 2 of this dissertation are based on the following papers:

1. Lu, J., Ding, P. and Dasgupta, T. (2015) Construction of Alternative Hypotheses for Evaluation of Randomization Tests with Ordinal Outcomes. arXiv:1505.04629

2. Lu, J., Ding, P. and Dasgupta, T. (2015) Sharp Bounds of Causal Effects on Ordinal Outcomes. arXiv:1507.01542

I am the main contributor and lead author of the above papers, Dr. Peng Ding and professor Tirthankar Dasgupta contributed to them, and professor Joseph K. Blitzstein helped edit their early versions.

Acknowledgments

First and foremost, I would like to express my utmost gratitude to my advisors, professors Tirthankar Dasgupta and Joseph K. Blitzstein. Without their trust, guidance and encouragement, I never would have finished this incredible journey. They are not only my mentors in Statistics, but also my role models in life.

I owe tremendously to professor Luke W. Miratrix for his generous help in serving as my committee member, my friend and collaborator Dr. Peng Ding for all his valuable inputs, and the faculty and staff members of the Harvard Statistics Department for their inspiring courses, seminars, and kind assistance.

I want to thank my friends and colleagues at Harvard for the memories and camaraderie, Jie Lai for her constant patience, understanding and care over all these years, and my parents for their unwavering love and support.

To my family.

Chapter 1

Construction of Alternative Hypotheses for Evaluating Randomization Tests with Ordinal Outcomes

Assessing the powers of randomization tests for finite populations has been recognized as a difficult task by several researchers, because the construction of "alternative" finite populations requires specification of a large number of potential outcomes that increases with the number of experimental units and thus is often referred to as "a thankless task" by experts. For ordinal outcomes, we develop a systematic procedure for closed-form construction of sequences of alternative hypotheses in increasing orders of their departures from the sharp null hypothesis of zero treatment effect on each experimental unit. Our construction procedure helps assessing the powers of randomization tests with ordinal data in randomized treatment-control studies and facilitates the comparison of different test statistics. Also, the results, as extensions of two by two tables, provide a way of quantifying the amount of information of association contained in the marginal distributions of general contingency tables.

1.1 Introduction

Introduced by Fisher (1935), randomization tests are useful tools for causal inference, because they assess the statistical significance of estimated treatment effects without making any assumptions about the underlying distributions of the data. Early theories on randomization tests were developed by Pitman (1938), and Kempthorne (1952), which showed that many statistical procedures can be viewed as approximations of randomization tests. To quote Bradley (1968), "[a] corresponding parametric test is valid only to the extent that it results in the same statistical decision [as the randomization test]." A crucial advantage of randomization tests is their abilities to handle non-standard (e.g., ordinal) outcomes, however there appears to be limited research on how to assess the powers of randomization tests for ordinal outcomes.

The potential outcomes framework (Neyman, 1923; Rubin, 1974) makes randomization tests easy to interpret. However, it does not naturally permit the assessment of their powers, which requires constructing alternatives to the sharp null hypothesis of zero treatment effect. The existing literature (e.g., Lehmann, 1975; Rosenbaum, 2010) accessed the powers of randomization tests by invoking infinite-population models. However, under certain circumstances we may prefer finite-population inference over infinite-population inference. For example, sometimes we can not view the experimental units as a random sample of a hypothetical infinite-population. For a comparison of finite-population and infinite-population inference, see Reichardt and Gollob (1999). In this chapter, we first construct alternatives to the sharp null hypothesis for ordinal outcomes within the finite-population framework, and later discuss how to incorporate our methodology into the infinite-population framework. However, constructing finite-population alternatives requires specifying the potential outcomes for all experimental units, and is thus considered "a thankless task" by experts (Rosenbaum, 2010). We demonstrate that this task is feasible for ordinal outcomes, and our finite-population construction procedure facilitates the study of the powers of randomization tests.

The chapter proceeds as follows. Section 1.2 reviews randomization tests of the sharp null hypothesis for ordinal outcomes. Section 1.3 introduces two measures quantifying departures from the sharp null hypothesis, discusses their relationship to the powers of randomization tests, and proposes a systematic procedure to construct alternative hypotheses in closed forms. Section 1.4 reports the results of a simulation study that demonstrates how to use the proposed construction procedure to assess the powers of randomization tests. Section 1.6 concludes.

1.2 Randomization Tests for Ordinal Outcomes

1.2.1 Potential Outcomes, Sharp Null Hypothesis and Randomization Test

We consider a completely randomized experiment with N units, a binary treatment and an ordinal outcome with J categories labeled as $0, \ldots, J-1$, where 0 and $J-1$ are the "worst" and "best" categories respectively. Under the Stable Unit Treatment Value Assumption (Rubin, 1980) that there is only one version of the treatment and no interference among units, we define the pair $\{Y_i(1), Y_i(0)\}$ as the potential outcomes of the ith unit under treatment and control. Let

$$p_{kl} = \text{pr}\left\{Y_i(1) = k, Y_i(0) = l\right\} = \#\left\{i : Y_i(1) = k, Y_i(0) = l\right\}/N$$

be the joint probability of potential outcomes k and l, under treatment and control. The probability "pr(·)" is defined for the finite-population. The $J \times J$ probability matrix $\boldsymbol{P} = (p_{kl})_{0 \leq k,l \leq J-1}$, which summarizes the joint distribution of the potential outcomes, plays a crucial role in our later construction of the alternative hypotheses. Let

$$p_{k+} = \sum_{l'=0}^{J-1} p_{kl'}, \quad p_{+l} = \sum_{k'=0}^{J-1} p_{k'l} \quad (k, l = 0, 1, \ldots, J-1).$$

The vectors $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$ characterize the marginal distributions of the potential outcomes under treatment and control.

Using the potential outcomes, we express the sharp null hypothesis as $Y_i(0) = Y_i(1)$ for all i. Under the sharp null hypothesis, the probability matrix $\boldsymbol{P}$ is diagonal with $p_{j+} = p_{jj} = p_{+j}$, for all j. To test the sharp null hypothesis, we use data from completely randomized experiments with N_1 units assigned to treatment. For the ith unit, we denote its treatment indicator as

W_i, and its observed outcome is consequently $Y_i^{\text{obs}} = W_i Y_i(1) + (1 - W_i) Y_i(0)$. For each j, let n_{0j} and n_{1j} respectively represent the number of units exposed to control and treatment with observed outcome j. Given the observed data, we first choose a suitable test statistic, typically a "measure of extremeness" (Brillinger et al., 1978), and obtain a p-value by comparing the test statistic's observed value to its randomization distribution.

1.3 Characterization and Construction of Alternative Hypotheses

To evaluate the powers of randomization tests, we need to construct alternatives to the sharp null hypothesis. A probability matrix $\boldsymbol{P}$ can violate the sharp null hypothesis in the following two distinct ways:

1. different marginal probabilities, i.e., $\boldsymbol{p}_1 \neq \boldsymbol{p}_0$;
2. identical marginal probabilities and non-zero off-diagonal elements, that is, $\boldsymbol{p}_1 = \boldsymbol{p}_0$ and $\sum_j p_{jj} < 1$.

For example, consider the following probability matrices

$$\boldsymbol{P}_1 = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{P}_2 = \begin{pmatrix} 0 & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & 0 \end{pmatrix}, \quad \boldsymbol{P}_3 = \begin{pmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{pmatrix}, \tag{1.1}$$

all of which violate the sharp null hypothesis. In particular, $\boldsymbol{P}_1$ has different marginal probabilities, $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ have identical marginal probabilities and non-zero off-diagonal elements. Inspired by the above observations, we construct alternative hypotheses following a three-step procedure:

1. Invoke an assumption regarding the marginal probabilities to reduce the number of possible alternative hypotheses;
2. Introduce two measures quantifying violations of the sharp null hypothesis;

3. Construct a sequence of alternative hypotheses by varying these measures.

1.3.1 Reducing the Number of Possible Alternative Hypotheses

To evaluate the powers of randomization tests, we create a sequence of probability matrices of increasing violations from the sharp null hypothesis. However, we can violate the sharp null hypothesis in many ways, making the problem intractable. To make it somewhat tractable we impose the following restriction on $\boldsymbol{P}$.

Assumption 1.1. (Stochastic Dominance) For all $j = 1, \ldots, J-1$, $\sum_{k=j}^{J-1} p_{k+} \geq \sum_{l=j}^{J-1} p_{+l}$.

As an illustration, consider the three probability matrices in (1.1). The stochastic dominance assumption excludes $\boldsymbol{P}_1$. Besides the advantage of reducing alternative hypotheses, in applied research the stochastic dominance pattern occurs frequently (e.g., Bradley et al., 1962; Bajorski and Petkau, 1999), because the treatment is often beneficial on the population level. In fact, stochastic dominance is termed "positive distributional causal effect" in Ju and Geng (2010). Because of the aforementioned technical convenience and the practical importance, we first focus on those marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ that satisfies the stochastic dominance assumption, and then discuss the general marginal probabilities.

1.3.2 Measures of Departures from the Sharp Null Hypothesis

In this section we introduce two measures quantifying violations of the sharp null hypothesis. Because the sharp null hypothesis is violated when $\boldsymbol{p}_1 \neq \boldsymbol{p}_0$, we use Hellinger distance τ_{HD} to quantify difference of the marginal probabilities. Other choices include Kullback–Leibler divergence and total variance distance. Under the sharp null hypothesis $\tau_{HD} = 0$, and therefore large τ_{HD} implies severe violations of the sharp null hypothesis. However, τ_{HD} relies solely on the marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$, implying the need of additional measures to access the violations of the sharp null hypothesis. For example, the probability matrices $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ in (1.1) violate the sharp null hypothesis, albeit $\tau_{HD} = 0$. To address this issue we use Cohen's Kappa (Cohen, 1960):

$$\kappa = \left\{ \mathrm{tr}(\boldsymbol{P}) - \boldsymbol{p}_1^{\mathrm{T}} \boldsymbol{p}_0 \right\} / \left(1 - \boldsymbol{p}_1^{\mathrm{T}} \boldsymbol{p}_0 \right), \tag{1.2}$$

where $\mathrm{tr}(\cdot)$ is the trace function. Cohen's kappa κ relies on the probability matrix $\boldsymbol{P}$, and under the sharp null hypothesis $\kappa = 1$ because $\boldsymbol{P}$ is diagonal.

1.3.3 Construction of Alternative Hypotheses

Having introduced the two measures, we construct a sequence of probability matrices with different Hellinger distances and Cohen's kappas, by the following two-step procedure:

1. construct a sequence of marginal probabilities, in increasing order of τ_{HD};
2. for each fixed marginal probabilities, construct a sequence of probability matrices in increasing order of κ. Construction of such a sequence involves the following sub-steps:
   (a) minimize and maximize κ subject to the following constraints:
   $$\sum_{k'=0}^{J-1} p_{k'l} = p_{+l}, \quad \sum_{l'=0}^{J-1} p_{kl'} = p_{k+}, \quad p_{kl} \geq 0 \quad (k, l = 0, 1, \ldots, J-1);$$
   (b) use a convex combination of the minimizer and maximizer to construct probability matrices with intermediate values of κ.

Step 1 accesses the impact of τ_{HD} on the powers of randomization tests, and step 2 further accesses the impact of κ on the powers of randomization tests. For fixed marginal probabilities, sub-step (a) studies the two extreme cases of "most" and "least" violations of the sharp null hypothesis, and sub-step (b) addresses the "in between" cases. Therefore, this procedure provides a relatively complete picture of violations of the sharp null hypothesis.

For given marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_)$, the minimization problem in the above procedure is somewhat intuitive. Consider the probability matrix $\boldsymbol{P}_I$ with independent potential outcomes, i.e., $p_{kl} = p_{k+}p_{+l}$ for all k, l. If we are not interested in distributions with negatively associated potential outcomes, $\boldsymbol{P}_I$ minimizes κ as zero. The maximization problem is, however, non-trivial. The following theorem provides the maximum value of κ, and the maximizer itself. For simplicity we restrict the maximizer to be lower triangular, because for any marginal probabilities satisfying the stochastic dominance assumption, there exists a corresponding probability matrix that is lower triangular. This is a special case of Strassen's theorem (Strassen, 1965; Lindvall, 1992), and was utilized in Rosenbaum (2001).

Theorem 1.1. For any $J \geq 2$, given marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ satisfying the stochastic dominance assumption, there exists a lower triangular probability matrix $\boldsymbol{P}_+$ achieving the upper bound of κ, i.e.,

$$\kappa\left(\boldsymbol{P}_+\right) = \left\{\sum_{k=0}^{J-1} \min\left(p_{k+}, p_{+k}\right) - \boldsymbol{p}_1^{\mathrm{T}} \boldsymbol{p}_0\right\} / \left(1 - \boldsymbol{p}_1^{\mathrm{T}} \boldsymbol{p}_0\right). \tag{1.3}$$

Proof of Theorem 1.1. The proof consists of two parts. First, for all $j = 0, \ldots, J-1$, the diagonal element p_{jj} of the probability matrix $\boldsymbol{P}$ cannot be greater than either p_{j+} or p_{+j}, i.e., $p_{jj} \leq \min\left(p_{j+}, p_{+j}\right)$. Consequently,

$$\operatorname{tr}(\boldsymbol{P}) \leq \sum_{k=0}^{J-1} \min\left(p_{k+}, p_{+k}\right), \tag{1.4}$$

substituting which in the right hand side of (1.2) yields the right hand side of (1.3). Therefore, it is an upper bound of κ. Second, in Appendix A.1 we construct a lower triangular probability matrix $\boldsymbol{P}_+$, with marginal probabilities $\boldsymbol{p}_1 = \left(p_{0+}, \ldots, p_{J-1,+}\right)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = \left(p_{+0}, \ldots, p_{+,J-1}\right)^{\mathrm{T}}$, that attains the upper bound. □

In the above proof of Theorem 1.1, we suggest one way to construct the maximizer $\boldsymbol{P}_+$. Next we discuss the uniqueness of $\boldsymbol{P}_+$. By restricting $\boldsymbol{P}_+$ to be lower triangular and its $(j+1)$th diagonal element p_{jj} to be $\min\left(p_{j+}, p_{+j}\right)$, what remain to be determined are the $(J-1)J/2$ off-diagonal elements. Note that there are $2J-3$ constraints associated with them. The equality

$$\frac{(J-1)J}{2} = 2J - 3$$

holds if and only if $J = 2$ or 3.

The case with $J = 2$ corresponds to binary outcomes, which occur frequently in both methodology and applied research. For a recent discussion of finite population inference for binary data, see Ding and Dasgupta (2015). The following corollary provides the maximizer under $J = 2$. Although it is a special case of Theorem 1.1, we provide a direct proof to rigorously show the uniqueness of the maximizer.

Corollary 1.1. For $J = 2$, given marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ that satisfy the stochastic dominance assumption, the following matrix is the unique maximizer of κ:

$$\boldsymbol{P}_+ = \begin{pmatrix} p_{0+} & 0 \\ p_{1+} - p_{+1} & p_{+1} \end{pmatrix}. \tag{1.5}$$

Proof of Corollary 1.1. Because $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ satisfy the stochastic dominance assumption, we have $p_{0+} \leq p_{+0}$ and $p_{1+} \geq p_{+1}$, implying that the diagonal elements of the maximizer are $p_{00} = p_{0+}$ and $p_{11} = p_{+1}$. Because the row sums of the maximizer are $\boldsymbol{p}_1$, we uniquely determine the entries of the maximizer, as shown in (1.5). The maximizer has nonnegative entries because $p_{1+} \geq p_{+1}$, and its column sums are $\boldsymbol{p}_0$ because $p_{0+} + p_{1+} - p_{+1} = p_{+0}$. The proof is complete. □

The case with $J = 3$ corresponds to three-level outcomes, which is also important in practice. For example, in a clinical trial we can describe the status of a patient as "deterioration," "no change" or "improvement" (Bajorski and Petkau, 1999). The following corollary provides the maximizer for $J = 3$. Again, we provide a direct proof.

Corollary 1.2. For $J = 3$, given marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ that satisfy the stochastic dominance assumption, the following matrix is the unique maximizer of κ:

$$\boldsymbol{P}_+ = \begin{pmatrix} p_{0+} & 0 & 0 \\ p_{1+} - \min(p_{+1}, p_{1+}) & \min(p_{+1}, p_{1+}) & 0 \\ p_{2+} - p_{+2} - \{p_{+1} - \min(p_{+1}, p_{1+})\} & p_{+1} - \min(p_{+1}, p_{1+}) & p_{+2} \end{pmatrix}. \tag{1.6}$$

Proof of Corollary 1.2. Because $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ satisfy the stochastic dominance assumption, we have $p_{0+} \leq p_{+0}$ and $p_{2+} \geq p_{+2}$, which implies that the diagonal elements of the maximizer are $p_{00} = p_{0+}$, $p_{11} = \min(p_{+1}, p_{1+})$, and $p_{22} = p_{+2}$. First, because the first row sum and third column sum are respectively p_{0+} and p_{+2}, the maximizer is in the following form:

$$\boldsymbol{P}_{+} = \begin{pmatrix} p_{0+} & 0 & 0 \\ ? & \min(p_{+1}, p_{1+}) & 0 \\ ? & ? & p_{+2} \end{pmatrix},$$

where "?" denotes an entry yet to be determined. Second, because the second row sum and column sum are respectively p_{1+} and p_{+1}, the maximizer is in the following form:

$$\boldsymbol{P}_{+} = \begin{pmatrix} p_{0+} & 0 & 0 \\ p_{1+} - \min(p_{+1}, p_{1+}) & \min(p_{+1}, p_{1+}) & 0 \\ ? & p_{\cdot,1} - \min(p_{+1}, p_{1+}) & p_{+2} \end{pmatrix}.$$

Third, because the third row sum is p_{2+}, we uniquely determine the maximizer, as in (1.6). Fourth, $\boldsymbol{P}_{+}$ has nonnegative entries, because

$$p_{2+} - p_{+2} - \{p_{+1} - \min(p_{+1}, p_{1+})\} = \min(p_{2+} - p_{+2}, p_{+0} - p_{0+}) \geq 0.$$

Finally, $\boldsymbol{P}_{+}$ has row sums $\boldsymbol{p}_1$, and column sums $\boldsymbol{p}_0$, because

$$p_{0+} + p_{1+} - \min(p_{+1}, p_{1+}) + p_{2+} - p_{+2} - \{p_{+1} - \min(p_{+1}, p_{1+})\} = p_{+0}.$$

The proof is complete. □

We end this section by discussing how to construct probability matrices with intermediate values of κ. Given the minimizer $\boldsymbol{P_I}$ and maximizer $\boldsymbol{P}_{+}$, let $\boldsymbol{P}_{\lambda} = \lambda \boldsymbol{P_I} + (1-\lambda)\boldsymbol{P}_{+}$. We view $\lambda \in [0, 1]$ as a sensitivity parameter, because we cannot estimate it from the observed data. The resulting probability matrices have the same marginal probabilities as $\boldsymbol{P_I}$ and $\boldsymbol{P}_{+}$, and subsequently the same Hellinger distances. However, they have different κ depending on λ because $\kappa(\boldsymbol{P}_{\lambda}) = (1-\lambda)\kappa(\boldsymbol{P}_{+})$. To complete the construction procedure, note that any entry of a well-defined joint distribution matrix $\boldsymbol{P}$ should be multiples of $1/N$. The marginals of $\boldsymbol{P}_{\lambda}$ are multiples of $1/N$ by definition, but its entries $p_{kl}(\lambda)$ may not be. We propose a calibration step to address this issue. We define

$$\tilde{p}_{kl}(\lambda) = \begin{cases} \frac{\lfloor N p_{kl}(\lambda) \rfloor}{N} & \text{if } k \neq l, \\ p_{+l} - \sum_{k' \neq l} \frac{\lfloor N p_{k'l}(\lambda) \rfloor}{N} & \text{if } k = l, \end{cases}$$

where $\lfloor \cdot \rfloor$ is the floor function. By definition, the column sums of $\widetilde{\boldsymbol{P}}_\lambda$ are $\boldsymbol{p}_0$. Let $\Lambda(\boldsymbol{p}_1, \boldsymbol{p}_0)$ denote the set containing all λ's such that the row sums of $\widetilde{\boldsymbol{P}}_\lambda$ are $\boldsymbol{p}_1$, and our constructed sequence of alternative hypotheses are therefore $\{\widetilde{\boldsymbol{P}}_\lambda\}_{\lambda \in \Lambda(\boldsymbol{p}_1, \boldsymbol{p}_0)}$. In practice, we can use a grid search to obtain an approximation of $\Lambda(\boldsymbol{p}_1, \boldsymbol{p}_0)$.

1.4 A Simulation Study

We demonstrate how the above construction procedure facilitates the study the powers of randomization tests. The test statistic we use is the squared Mann–Whitney U-statistic (Agresti, 2002):

$$U^2 = \frac{1}{4N_1^2(N-N_1)^2} \left[\sum_{k=0}^{J-1} \sum_{l=0}^{J-1} n_{1k} n_{0l} \{ I(k > l) - I(k < l) \} \right]^2.$$

Another commonly-used test statistic for categorical data is the χ^2-statistic. However, we do not use it here, because it does not utilize the order information and therefore is less powerful.

Although closed-form expressions of the powers of randomization test using the U^2 statistic are difficult to obtain, numerical calculations by the Monte Carlo method are straightforward, once we determine the alternative hypothesis $\boldsymbol{P}$:

1. under $\boldsymbol{P}$ generate 2×10^5 independent treatment assignments and obtain the corresponding observed data sets;

2. for each observed data set calculate the p-value of the randomization test using the observed value of U^2 and its simulated null distribution;

3. approximate the power of the U^2 statistic as the proportion of the p-values that are smaller than the significance level $\alpha = 0.05$.

In this simulation study, we construct alternative hypotheses using the following four sets of marginals, two with $J = 2$ and two with $J = 3$:

1. $\boldsymbol{p}_1 = (3/10, 7/10)^{\mathrm{T}}$, $\boldsymbol{p}_0 = (3/5, 2/5)^{\mathrm{T}}$, $\tau_{HD} = 0.216$;

2. $\boldsymbol{p}_1 = (1/2, 1/2)^{\mathrm{T}}$, $\boldsymbol{p}_0 = (4/5, 1/5)^{\mathrm{T}}$, $\tau_{HD} = 0.227$;

3. $\boldsymbol{p}_1 = (1/4, 1/4, 1/2)^{\mathrm{T}}$, $\boldsymbol{p}_0 = (2/5, 2/5, 1/5)^{\mathrm{T}}$, $\tau_{HD} = 0.227$;

4. $\boldsymbol{p}_1 = (9/40, 9/40, 11/20)^{\mathrm{T}}$, $\boldsymbol{p}_0 = (2/5, 2/5, 1/5)^{\mathrm{T}}$, $\tau_{HD} = 0.261$.

For each case, we let the sample sizes $N = 120, 160, 240$, and the sensitivity parameters $\lambda = 0, 1/4, 1/2, 3/4, 1$. We then construct the probability matrices, which share the same marginals. For each probability matrix $\widetilde{\boldsymbol{P}}_\lambda$, we use the aforementioned Monte Carlo procedure to calculate the powers. On the one hand, different cases allow us to study the impact of τ_{HD} on the powers. On the other hand, within each case we can study the impact of κ on the powers. The simulation results are summarized in Figure 1.1, from which we draw the following conclusions. For all fixed sample sizes, the power functions of Case 2 dominate those of Case 1, and the power functions of Case 4 dominate those of Case 3. Therefore, for fixed J the power increases as the Hellinger distance increases. Furthermore, for fixed marginals and sample size, the power increases as κ decreases, or equivalently as λ increases. However, this dependence becomes weaker as the sample size increases, because the power converges to one.

We can use the demonstrated methodology to compare the power functions of different test statistics, and also to determine sample sizes that guarantee a pre-specified power. For instance, in Case 3, we cannot guarantee a power of 0.95 with a sample of size 120, but we can with a sample of size 160.

In summary, for a finite population, the power of the randomization test using U^2 depends on the marginal difference of the potential outcomes as well as the association between them. In particular, the power increases as the marginal difference increases, and given the marginals fixed, the power increases as the association decreases. Furthermore, the power converges to one as the sample size increases, for any case with non-zero marginal difference. The above conclusions appear to confirm our intuition, because it should be easier to reject the sharp

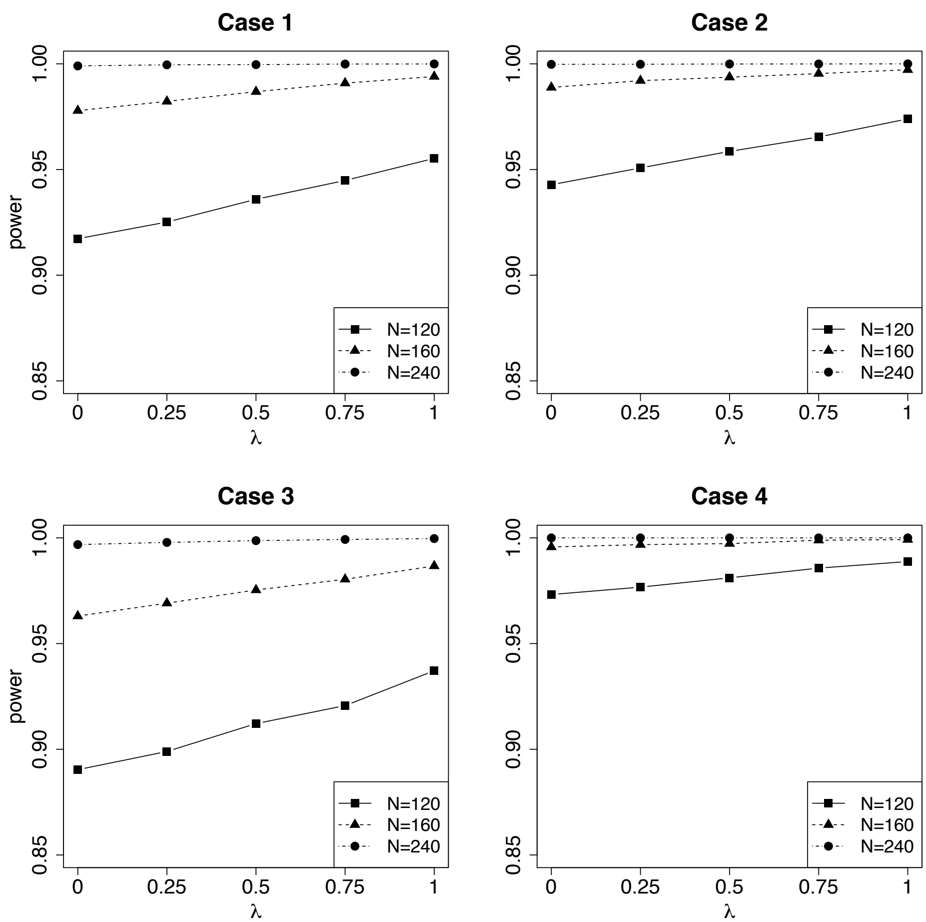

Figure 1.1: *Statistical Powers of Randomization Tests Using* U^2.

null hypothesis given larger differences between the marginals, and given the marginals fixed, it should be easier to reject the sharp null hypothesis given smaller associations between the potential outcomes. Our findings conform to Plackett (1977)'s and Chernoff (2004)'s results about the classical 2×2 tables: the marginals of the contingency tables contain limited amount of information about the association with finite samples, which becomes negligible asymptotically.

1.5 Sharp Bounds of the Proportion of Units with Zero Treatment Effects

For any categorical outcomes, the proportion of units with zero treatment effects is

$$\eta_0 = \text{pr}\{Y_i(1) = Y_i(0)\} = \sum_{j=0}^{J-1} p_{jj},$$

In this section, we bound η_0 using only the marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \dots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \dots, p_{+,J-1})^{\mathrm{T}}$. We first state a fundamental lemma, which not only plays a central role in our later proof of the main theorem, but also is of independent interest.

Lemma 1.1. Assume $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ are nonnegative constants such that

$$\sum_{k=1}^{n} a_k = \sum_{l=1}^{n} b_l.$$

(a) There exists an $n \times n$ matrix $\boldsymbol{P}_n = (p_{kl})_{1 \le k,l \le n}$ with nonnegative elements and the following row sums, column sums and diagonal elements:

$$\sum_{l'=1}^{n} p_{kl'} = a_k, \quad \sum_{k'=1}^{n} p_{k'l} = b_l, \quad p_{jj} = \min(a_j, b_j) \quad (k, l, j = 1, \dots, n).$$

(b) If $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ further satisfy the following condition:

$$a_j + b_j \le 1, \quad (j = 0, \dots, J-1),$$

then there exists an $n \times n$ matrix $\boldsymbol{Q} = (q_{kl})_{1 \le k,l \le n}$ with nonnegative elements and the following row sums, column sums and diagonal elements:

$$\sum_{l'=1}^{n} q_{kl'} = a_k, \quad \sum_{k'=1}^{n} q_{k'l} = b_l, \quad q_{jj} = 0 \quad (k, l, j = 1, \dots, n).$$

The proof of Lemma 1.1 is in Appendix A.2. With the help of the Lemma, we derive the sharp bounds of the probability of the sharp null hypothesis, and construct explicitly the probability matrices that attain these bounds.

Theorem 1.2. The sharp bounds of η_0 are

$$\sum_{j=0}^{J-1} \max(p_{j+} + p_{+j} - 1, 0) \le \eta_0 \le \sum_{j=0}^{J-1} \min(p_{j+}, p_{+j}).$$

Proof of Theorem 1.2. We first prove the lower bound, which follows directly from

$$p_{jj} \ge \max\left(p_{jj} - \sum_{k\neq j}\sum_{l\neq j} p_{kl}, 0\right) = \max(p_{j+} + p_{+j} - 1, 0).$$

We need to show that the above bound is indeed attainable, i.e., there exists a probability matrix $\boldsymbol{P}$ satisfying the following two conditions:

(C1) the marginal probabilities are $(p_{0+}, p_{1+}, \dots, p_{J-1,+})$ and $(p_{+0}, p_{+1}, \dots, p_{+,J-1})$;

(C2) the corresponding η_0 is equal to the lower bound.

Because we have

$$\sum_{i=0}^{J-1} (p_{j+} + p_{+j}) = \sum_{i=0}^{J-1} p_{j+} + \sum_{i=0}^{J-1} p_{+j} = 2,$$

there exists at most one j^* such that $p_{j^*+} + p_{+j^*} - 1 > 0$. Therefore we discuss the following two cases. If such j^* exists, we construct the matrix $\boldsymbol{P}_L$ as:

$$p_{kl} = \begin{cases} 0, & \text{if } k \neq j^*, l \neq j^*, \\ p_{k+}, & \text{if } k \neq j^*, l = j^*, \\ p_{+l}, & \text{if } k = j^*, l \neq j^*, \\ p_{j^*+} + p_{+j^*} - 1, & \text{if } k = j^*, l = j^*. \end{cases}$$

The resulting probability matrix $\boldsymbol{P}_L$ for attaining the lower bound has none zero elements only in the j^*-th column and j^*-th row. Otherwise, Lemma 1.1(b) allows us to construct a probability matrix with the above marginal probabilities and zero diagonal elements.

We then prove the upper bound, which follows directly from $p_{jj} \leq \min(p_{j+}, p_{+j})$. Lemma 1.1(a) further guarantees that the upper bound is indeed attainable. □

1.6 Discussions

We propose a method to construct sequences of finite populations of ordinal outcomes in increasing orders of their departures from the sharp null hypothesis of no treatment effect. Such a construction is useful for a systematic evaluation of the powers of randomization tests. The key idea is to introduce two measures quantifying departures from the sharp null hypothesis and study the relationship between them, as well as their impact on the powers of randomization tests. Our results show that we can retrieve only limited amount of information regarding the association between the potential outcomes from their marginal distributions, and therefore extend Plackett (1977)'s and Chernoff (2004)'s discussions on 2×2 tables by quantifying the information about association in the marginal distributions of general contingency tables. Our methodology can easily be incorporated into the infinite-population framework, by assuming that the potential outcomes $\{Y_i(1), Y_i(0)\}$ are i.i.d. from the constructed probability matrices.

There are multiple future directions based on our work. First, although we adopt a numerical approach, it is possible to derive the distribution of the U^2 statistic under the sharp null hypothesis using asymptotic theory. Second, we can derive the maximizer for marginal probabilities that do not satisfy the stochastic dominance assumption. Third, we can incorporate covariate information to further improve our current framework for ordinal outcomes.

Chapter 2

Sharp Bounds of Causal Effects on Ordinal Outcomes

Under the potential outcomes framework, causal effects are defined as comparisons between the treatment and control potential outcomes. However, the average causal effect, generally the parameter of interest, is not well defined for ordinal outcomes. To address this problem, we propose two new causal parameters, i.e., the probabilities that the treatment is beneficial and strictly beneficial for the experimental units, which are well defined for any outcomes and of particular interest for ordinal outcomes. These two new causal parameters, though of scientific importance and interest, depend on the association between the potential outcomes and therefore without any further assumptions are not identifiable from the observed data. For ordinal outcomes, we derive sharp bounds of the two new causal parameters using only the marginal distributions, without imposing any assumptions on the joint distribution of the potential outcomes.

2.1 Introduction

The potential outcomes framework (Neyman, 1923; Rubin, 1974) permits defining causal effects as comparisons between the potential outcomes under treatment and control. The average causal effect, generally the parameter of interest ever since the seminal work of Neyman (1923), is not applicable to ordinal outcomes, because average outcomes themselves are not well defined. Ordinal outcomes are common in applied research, but discussions about them in the causal inference literature are very limited. Rosenbaum (2001) discussed causal inference for ordinal outcomes under the monotonicity assumption that the treatment is beneficial for all units. Cheng (2009) and Agresti (2010) discussed various causal parameters under the assumption of independent potential outcomes. Volfovsky et al. (2015) exploited a Bayesian strategy, which involved a full parametric model on the joint values of the potential outcomes.

For ordinal outcomes, we propose two new causal parameters measuring the probabilities that the treatment is beneficial and strictly beneficial for the experimental units, which play important roles in decision and policy making for randomized evaluations with ordinal outcomes. Because these two causal parameters depend on the association between the treatment and control potential outcomes, they are generally not identifiable from the observed data. Without imposing any assumptions about the underlying distributions or the association between the potential outcomes, we sharply bound them by using the marginal distributions of the potential outcomes. We sharpen the bounds when covariates are available, and further demonstrate that our nonparametric bounds on causal effects can be incorporated into existing parametric modeling frameworks.

The chapter proceeds as follows. Section 3.2 sets up the theoretical framework of causal inference for ordinal outcomes, and propose two causal parameters that are better measures of causal effects and are of practical importance. Section 2.3 shows the sharp bounds of the causal parameters. Section 2.4 presents some numerical and real examples to demonstrate the advantages of our theoretical results. In Section 2.5, we further sharpen the bounds by using pretreatment covariates, and derive sharp bounds for another causal measure. We conclude with a discussion in Section 2.6, and provide the technical details in Appendix B.

2.2 Causal Inference for Ordinal Outcomes

2.2.1 Potential Outcomes

We consider a finite population with N units, a binary treatment, and an ordinal outcome with J categories labeled $0, \ldots, J-1$, where 0 and $J-1$ respectively represent the worst and best categories. Under the Stable Unit Treatment Value Assumption (Rubin, 1980) that there is only one version of the treatment and no interference among units, we define the pair $\{Y_i(1), Y_i(0)\}$ as the potential outcomes of the ith unit under treatment and control, respectively. We let

$$p_{kl} = \mathrm{pr}\{Y_i(1) = k, Y_i(0) = l\} \quad (k, l = 0, \ldots, J-1)$$

denote the proportion of units whose potential outcome is k under treatment and l under control. Here, the probability "pr(·)" can be defined for either a finite population of N units, or for a super population. The $J \times J$ probability matrix $\boldsymbol{P} = (p_{kl})_{0 \le k,l \le J-1}$ summarizes the joint distribution of the potential outcomes. We let the row and column sums of $\boldsymbol{P}$ be

$$p_{k+} = \sum_{l'=0}^{J-1} p_{kl'}, \quad p_{+l} = \sum_{k'=0}^{J-1} p_{k'l} \quad (k, l = 0, 1, \ldots, J-1)$$

The vectors $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$ characterize the marginal distributions of the potential outcomes under treatment and control, respectively.

2.2.2 Causal Parameters for Ordinal Outcomes

Any causal parameter is a function of the probability matrix $\boldsymbol{P}$. Unfortunately, the average causal effect is not well defined for ordinal outcomes. Instead, we can use the distributional causal effects (cf. Ju and Geng, 2010)

$$\Delta_j = \Pr\{Y_i(1) \ge j\} - \Pr\{Y_i(0) \ge j\} = \sum_{k \ge j} p_{k+} - \sum_{l \ge j} p_{+l} \quad (j = 0, \ldots, J-1) \qquad (2.1)$$

to measure the difference between the marginal distributions of potential outcomes at different levels of j. However, unless Δ_j's have the same sign for all j, it is difficult to decide whether the treatment or the control is preferable based only on the distributional causal effects.

Example 2.1. Let $\boldsymbol{p}_1 = (1/5, 3/5, 1/5)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (2/5, 1/5, 2/5)^{\mathrm{T}}$, with $\Delta_0 = 0$, $\Delta_1 = 1/5$ and $\Delta_2 = -1/5$. The treatment is beneficial at level 1, but not at level 2. In this case, distributional causal effects do not provide straightforward guidance for decision making.

Volfovsky et al. (2015) studied the conditional medians

$$m_j = \mathrm{med}\left\{Y_i(1) \mid Y_i(0) = j\right\} \quad (j = 0, \ldots, J-1), \tag{2.2}$$

which is a set containing all values of k such that

$$\sum_{k'=0}^{k} p_{k'j} \geq \frac{p_{+j}}{2}, \quad \sum_{k'=k}^{J-1} p_{k'j} \geq \frac{p_{+j}}{2}.$$

Therefore, the conditional medians may not be unique, and they are only well defined for j with $p_{+j} > 0$. Moreover, they are not direct measures of the treatment effect itself.

We propose two new causal parameters that measure the probabilities that the treatment is beneficial and strictly beneficial for the experimental units:

$$\tau = \Pr\left\{Y_i(1) \geq Y_i(0)\right\} = \sum\sum_{k \geq l} p_{kl}, \quad \eta = \Pr\left\{Y_i(1) > Y_i(0)\right\} = \sum\sum_{k > l} p_{kl}. \tag{2.3}$$

The causal parameters τ and η are measures of causal effects that are well defined for any types of outcomes and of particular interest to ordinal outcomes.

Example 2.2. Consider the following probability matrix:

$$\boldsymbol{P} = \begin{pmatrix} 0 & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{1}{6} & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \end{pmatrix}.$$

In this case, m_0 is not well defined, m_1 is 1, and $m_2 = \{0, 1, 2\}$. However, we have $\tau = 2/3$ and $\eta = 1/3$, i.e., 2/3 of the population benefit from the treatment, and 1/3 strictly benefit.

2.3 Sharp Bounds on Causal Effects for Ordinal Outcomes

The definitions of τ and η involve the association between the individual potential outcomes $Y_i(1)$ and $Y_i(0)$. Because we can never jointly measure them, any observed data do not provide full information about their association, rendering the causal parameters τ and η not identifiable. To partially circumvent this difficulty, we focus on the sharp bounds of τ and η, which are the minimal and maximal values of τ and η under the constraints of the following marginal distributions:

$$\sum_{l'=0}^{J-1} p_{kl'} = p_{k+}, \quad \sum_{k'=0}^{J-1} p_{k'l} = p_{+l}, \quad p_{kl} \geq 0 \quad (k, l = 0, \dots, J-1). \tag{2.4}$$

Therefore, the sharp bounds depend only on the marginal distributions, but not on the joint distribution of potential outcomes. Finding the sharp bounds is equivalent to solving linear programming problems, because the objective functions in (2.3) and the constraints in (2.4) are all linear. Fortunately, through some rigorous mathematical arguments, we can derive closed-form solutions of the above linear programming problems. We first state a fundamental lemma, which plays a central role in our later proofs and is also of independent interest.

Lemma 2.1. Assume that $(x_0, \dots, x_{n-1})$ and $(y_0, \dots, y_{n-1})$ are nonnegative constants.

(a) If $\sum_{r=s}^{n-1} x_r \geq \sum_{r=s}^{n-1} y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ lower triangular matrix $\boldsymbol{A}_n = (a_{kl})_{0 \leq k,l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} a_{kl'} \leq x_k, \quad \sum_{k'=0}^{n-1} a_{k'l} = y_l \quad (k, l = 0, \dots, n-1). \tag{2.5}$$

(b) If $\sum_{r=s}^{n-1} x_r \leq \sum_{r=s}^{n-1} y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ upper triangular matrix $\boldsymbol{B}_n = (b_{kl})_{0 \leq k,l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} b_{kl'} = x_k, \quad \sum_{k'=0}^{n-1} b_{k'l} \leq y_l \quad (k, l = 0, \dots, n-1). \tag{2.6}$$

(c) If $\sum_{r=0}^{s} x_r \leq \sum_{r=0}^{s} y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ lower triangular matrix $\boldsymbol{C}_n = (c_{kl})_{0\le k,l\le n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} c_{kl'} = x_k, \quad \sum_{k'=0}^{n-1} c_{k'l} \le y_l \quad (k,l = 0,\dots,n-1). \tag{2.7}$$

(d) If $\sum_{r=0}^{s} x_r \ge \sum_{r=0}^{s} y_r$ for all $s = 0,\dots,n-1$, there exists an $n\times n$ upper triangular matrix $\boldsymbol{D}_n = (d_{kl})_{0\le k,l\le n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} d_{kl'} \le x_k, \quad \sum_{k'=0}^{n-1} d_{k'l} = y_l \quad (k,l = 0,\dots,n-1). \tag{2.8}$$

(e) If we further assume $\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, the above inequalities in (2.5)–(2.8) all reduce to equalities, i.e., the matrices $\boldsymbol{A}_n$, $\boldsymbol{B}_n$, $\boldsymbol{C}_n$ and $\boldsymbol{D}_n$ have $(x_0,\dots,x_{n-1})$ and $(y_0,\dots,y_{n-1})$ as their row and column sums.

The proof of Lemma 2.1 is in Appendix B.1. With the help of the lemma, we derive the sharp bounds of the causal parameters of interest, and construct explicitly the probability matrices that attain these bounds. First we state a theorem on the sharp bounds of τ, which is the foundation for the remaining theorems and corollaries.

Theorem 2.1. Fix marginal probabilities $\boldsymbol{p}_1 = (p_{0+},\dots,p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0},\dots,p_{+,J-1})^{\mathrm{T}}$,

(a) the sharp upper bound of τ is $\tau_U = 1 + \min_{0\le j\le J-1} \Delta_j$.

(b) the sharp lower bound of τ is $\tau_L = \max_{0\le j\le J-1} (p_{+j} + \Delta_j)$,

Proof of Theorem 2.1(a). For $j = 0,1,\dots,J-1$, we can bound τ from above by

$$\begin{aligned}
\tau &= \sum\sum_{k\ge l} p_{kl} = 1 - \sum\sum_{k<l} p_{kl} \\
&\le 1 - \sum_{k<j}\sum_{l\ge j} p_{kl} = 1 - \left(\sum_{k=0}^{J-1}\sum_{l\ge j} p_{kl} - \sum_{k\ge j}\sum_{l\ge j} p_{kl}\right) && (2.9)\\
&\le 1 - \left(\sum_{k=0}^{J-1}\sum_{l\ge j} p_{kl} - \sum_{k\ge j}\sum_{l=1}^{J-1} p_{kl}\right) = 1 - \left(\sum_{l\ge j} p_{+l} - \sum_{k\ge j} p_{k+}\right) && (2.10)\\
&= 1 + \Delta_j.
\end{aligned}$$

Because the above inequality holds for all j, we can bound τ from above by

$$\tau \leq 1 + \min_{0 \leq j \leq J-1} \Delta_j. \tag{2.11}$$

In Appendix B.2.1 we construct a probability matrix with row sums $\boldsymbol{p}_1$ and column sums $\boldsymbol{p}_0$, that attains the upper bound in (2.11). □

Proof of Theorem 2.1(b). For $j = 0, 1, \ldots, J-1$, we can bound τ from below by

$$\begin{aligned}
\tau &= \sum\sum_{k \geq l} p_{kl} \\
&\geq \sum_{k \geq j} \sum_{l \leq j} p_{kl} = \sum_{k \geq j} \sum_{l=0}^{J-1} p_{kl} - \sum_{k \geq j} \sum_{l > j} p_{kl} && (2.12)\\
&\geq \sum_{k \geq j} \sum_{l=0}^{J-1} p_{kl} - \sum_{k=0}^{J-1} \sum_{l > j} p_{kl} = \sum_{k \geq j} p_{k+} - \sum_{l > j} p_{+l} && (2.13)\\
&= p_{+j} + \Delta_j.
\end{aligned}$$

Because the above inequality hold for all j, we can bound τ from below by

$$\tau \geq \max_{0 \leq j \leq J-1} (p_{+j} + \Delta_j). \tag{2.14}$$

In Appendix B.2.2 we construct a probability matrix with row sums $\boldsymbol{p}_1$ and column sums $\boldsymbol{p}_0$, that attains the lower bound in (2.14). □

The bounds in Theorem 2.1 are very closely related to the distributional causal effects in (2.1), and we can respectively interpret them as conservative and optimistic estimates of the probability that the treatment is beneficial. Furthermore, the following corollary demonstrates that the sharp upper bound τ_U is related to the stochastic dominance assumption, i.e., $\Delta_j \geq 0$ for all j.

Corollary 2.1. The causal parameter $\tau_U = 1$, if and only if the marginal probabilities $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ satisfy the stochastic dominance assumption.

Proof of Corollary 2.1. By Theorem 2.1, $\tau = 1$ if and only if $\min_{0 \leq j \leq J-1} \Delta_j = 0$. Because

$\Delta_0 = 0$, this is equivalent to

$$\Delta_j \geq 0 \quad (j = 0, \ldots, J-1),$$

i.e., the stochastic dominance assumption holds. □

The above Corollary 2.1 implies that for any marginal distributions $(p_{0+}, \ldots, p_{J-1,+})$ and $(p_{+0}, \ldots, p_{+,J-1})$ satisfying the stochastic dominance assumption, there exists a lower triangular probability matrix P that corresponds to a population satisfying the monotonicity assumption, i.e., $Y_i(1) \geq Y_i(0)$ for all i. Strassen (1965) and Rosenbaum (2001) demonstrated this result, and Theorem 2.1 extends it without imposing the stochastic dominance assumption. Moreover, Theorem 2.1 also justifies using $\min_{0 \leq j \leq J-1} \Delta_j$ as a measure of the deviation from the stochastic dominance assumption (Scharfstein et al., 2004).

Next we consider the sharp bounds of η. Recognizing the equivalent form

$$\eta = \text{pr}\left\{Y_i(1) > Y_i(0)\right\} = 1 - \text{pr}\left\{Y_i(0) \geq Y_i(1)\right\},$$

we can derive bounds for $\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$ by switching the treatment and control labels and applying Theorem 2.1.

Theorem 2.2. The sharp lower and upper bounds of η are

$$\eta_L = \max_{0 \leq j \leq J-1} \Delta_j, \quad \eta_U = 1 + \min_{0 \leq j \leq J-1} \left(\Delta_j - p_{j+}\right). \tag{2.15}$$

Proof of Theorem 2.2. Because $\eta = 1 - \text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$, its lower bound is one minus the upper bound of $\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$. By switching the treatment and control labels, we can bound $\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$ from the above by

$$\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\} \leq 1 - \max_{0 \leq j \leq J-1} \Delta_j,$$

which implies that $\eta_L = \max_{0 \leq j \leq J-1} \Delta_j$.

Similarly, the upper bound of η is one minus the lower bound of $\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$. By switching the treatment and control labels, we can bound $\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\}$ from below by

$$\text{pr}\left\{Y_i(0) \geq Y_i(1)\right\} \geq \max_{0\leq j\leq J-1}\left(p_{j+} - \Delta_j\right),$$

which implies that $\eta_U = 1 + \min_{0\leq j\leq J-1}\left(\Delta_j - p_{j+}\right).$ □

In the proofs of Theorem 2.1 and 2.2, we construct the joint distributions that achieve the lower and upper bounds of τ and η respectively, which correspond to negatively associated and positively associated potential outcomes. They are both extreme scenarios. In practice, we may also be interested in the case with independent potential outcomes (Rubin, 1978; Cheng, 2009; Agresti, 2010; Ding and Dasgupta, 2015), i.e., $p_{kl} = p_{k+}p_{+l}$ for all k and l. With independent potential outcomes, we can identify τ and η by the marginal distributions.

Theorem 2.3. With independent potential outcomes,

$$\tau_I = \sum\sum_{k\geq l} p_{k+}p_{+l}, \quad \eta_I = \sum\sum_{k>l} p_{k+}p_{+l}.$$

Furthermore, $\tau_L \leq \tau_I \leq \tau_U$ and $\eta_L \leq \eta_I \leq \eta_U$.

Proof of Theorem 2.3. With independent potential outcomes, the probability matrix P has elements $p_{kl} = p_{k+}p_{+l}$ for k and l. We obtain τ_I and η_I by their definitions. Obviously, they are between their lower and upper bounds, i.e., $\tau_L \leq \tau_I \leq \tau_U$ and $\eta_L \leq \eta_I \leq \eta_U$. □

To further demonstrate the results in Theorems 2.1 and 2.2, we consider two special cases with $J = 2$ and 3. As mentioned in Chapter 1, they are important cases in both methodology and applied research.

Corollary 2.2. When $J = 2$, the sharp lower and upper bounds of τ are

$$\tau_L = \max\left(p_{+0}, p_{1+}\right), \quad \tau_U = \min\left(1, p_{0+} + p_{+1}\right),$$

and the sharp lower and upper bounds of η are

$$\eta_L = \max\left(0, p_{1+} - p_{+1}\right), \quad \eta_U = \min\left(p_{1+}, p_{+0}\right).$$

The above Corollary 2.2 agrees with our intuitions. Because $\tau = p_{00} + p_{10} + p_{11}$, it is not smaller than the first column sum and the second row sum of the probability matrix P. Similarly, because $\eta = p_{10}$, it is not larger than the second row sum or first column sum of the probability matrix P.

Corollary 2.3. When $J = 3$, the sharp lower and upper bounds of τ are

$$\tau_L = \max\left(p_{+0}, 1 - p_{0+} - p_{+2}, p_{2+}\right), \quad \tau_U = \min\left(1, p_{+0} + p_{1+} + p_{2+}, p_{+0} + p_{+1} + p_{2+}\right),$$

and the sharp lower and upper bounds of η are

$$\eta_L = \max\left(0, p_{+0} - p_{0+}, p_{2+} - p_{+2}\right), \quad \eta_U = \min\left(p_{1+} + p_{2+}, p_{+0} + p_{2+}, p_{+0} + p_{+1}\right).$$

Example 2.3. The marginal probabilities $\boldsymbol{p}_1 = (1/5, 3/5, 1/5)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (2/5, 1/5, 2/5)^{\mathrm{T}}$ do not satisfy the stochastic dominance assumption, because $\Delta_0 = 0$, $\Delta_1 = 1/5$ and $\Delta_2 = -1/5$. By Theorems 2.1 and 2.3, we have $\tau_L = 2/5$, $\tau_I = 16/25$, and $\tau_U = 4/5$. The joint distributions corresponding to negatively associated, independent, and positively associated potential outcomes achieving these values are respectively

$$\boldsymbol{P}_L^{\tau} = \begin{pmatrix} 0 & \frac{1}{5} & 0 \\ \frac{1}{5} & 0 & \frac{2}{5} \\ \frac{2}{5} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{P}_I^{\tau} = \begin{pmatrix} \frac{2}{25} & \frac{1}{25} & \frac{2}{25} \\ \frac{6}{25} & \frac{3}{25} & \frac{6}{25} \\ \frac{2}{25} & \frac{1}{25} & \frac{2}{25} \end{pmatrix}, \quad \boldsymbol{P}_U^{\tau} = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & \frac{1}{5} \end{pmatrix}. \tag{2.16}$$

Similarly, by Theorems 2.2 and 2.3, we have $\eta_L = 1/5$, $\eta_I = 9/25$, and $\eta_U = 3/5$.

Example 2.4. The marginal probabilities $\boldsymbol{p}_1 = (1/5, 1/5, 3/5)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (3/5, 1/5, 1/5)^{\mathrm{T}}$ satisfy the stochastic dominance assumption, because $\Delta_0 = 0$, $\Delta_1 = 2/5$ and $\Delta_2 = 2/5$. By Theorems 2.1 and 2.3, we have $\tau_L = 3/5$, $\tau_I = 22/25$, and $\tau_U = 1$. The joint distributions corresponding to negatively associated, independent, and positively associated potential outcomes achieving these values are respectively

$$\boldsymbol{P}_L^{\tau} = \begin{pmatrix} 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \\ \frac{3}{5} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{P}_I^{\tau} = \begin{pmatrix} \frac{3}{25} & \frac{1}{25} & \frac{1}{25} \\ \frac{3}{25} & \frac{1}{25} & \frac{1}{25} \\ \frac{9}{25} & \frac{3}{25} & \frac{3}{25} \end{pmatrix}, \quad \boldsymbol{P}_U^{\tau} = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ \frac{2}{5} & 0 & \frac{1}{5} \end{pmatrix}. \tag{2.17}$$

Similarly, by Theorems 2.2 and 2.3, we have $\eta_L = 2/5$, $\eta_I = 3/5$, and $\eta_U = 4/5$.

2.4 Simulated and Real Examples

2.4.1 Simulated Examples

We obtain point estimators of the bounds on τ and η based on data from completely randomized experiments by replacing p_{k+}, p_{+l}, and Δ_j with their sample analogues. We use the bootstrap method in Horowitz and Manski (2000) to obtain a confidence interval for the bounds. For computational details of some other bootstrap methods, see Cheng and Small (2006) and Yang and Small (2015).

To save space in the main text, we focus only on τ and its bounds in Theorem 2.1. We choose the sample size to be 200, and consider four cases with different probability matrices. Cases 1 and 2 correspond to the independent and positively associated potential outcomes in (2.16), which share the same marginal distribution but do not satisfy the stochastic dominance assumption. Cases 3 and 4 correspond to the independent and positively associated potential outcomes in (2.17), which share the same marginal distribution and satisfy the stochastic dominance assumption. Columns 2–4 of Table 3.3 summarize the true values of τ, τ_L and τ_U, for all four cases. For cases 1 and 3 with independent potential outcomes, we have $\tau_L < \tau < \tau_U$. For cases 2 and 4 with positively associated potential outcomes, we have $\tau = \tau_U$.

For each case, we independently draw 5000 treatment assignments from a balanced completely randomized experiment. For each observed data set, we calculate point estimates of τ_L and τ_U, and construct a confidence interval for the bounds (τ_L, τ_U) that covers the bounds themselves at least 95% times. In columns 5–8 of Table 3.3, we report the biases and standard errors of the point estimators $\widehat{\tau}_L$ and $\widehat{\tau}_U$; in columns 9 and 10 of Table 3.3, we report the coverage rates of the intervals on the bounds (τ_L, τ_U) and the true parameter τ. Table 3.3 shows that the point estimators have small biases and standard errors, and the confidence intervals achieve reasonable coverage rates on the bounds (τ_L, τ_U) although they over-cover the true parameter τ.

Table 2.1: *Numerical Examples*

case	τ	τ_L	τ_U	bias($\widehat{\tau}_L$)	se($\widehat{\tau}_L$)	bias($\widehat{\tau}_U$)	se($\widehat{\tau}_U$)	coverage of (τ_L, τ_U)	coverage of τ
1	0.640	0.400	0.800	0.016	0.037	0.000	0.045	0.987	1.000
2	0.800	0.400	0.800	0.013	0.043	-0.001	0.057	0.957	0.974
3	0.880	0.600	1.000	0.026	0.030	0.000	0.000	0.967	1.000
4	1.000	0.600	1.000	0.025	0.031	0.000	0.000	0.960	1.000

2.4.2 Real Example

We use the taste-testing experiment data in Bradley et al. (1962) to demonstrate the estimation of bounds of our new causal parameters and the subsequent inferences drawn. In this experiment, the outcome of interest Y is ordinal, taking values from terrible with $Y = 0$ to excellent with $Y = 4$. The treatment has five unordered levels, labeled A, B, C, D and E, respectively. We summarize the data in Table 2.2.

Table 2.2: *A Taste-Testing Experiment*

	Outcome Categories					
Treatment	0	1	2	3	4	row sum
A	9	5	9	13	4	40
B	7	3	10	20	4	44
C	14	13	6	7	0	40
D	11	15	3	5	8	42
E	0	2	10	30	2	44

For illustration purpose, we consider only treatment E versus treatment C, and treatment E versus treatment D, and summarize the results in Table 2.3. They lead to several conclusions. First, treatment E stochastically dominates treatment C. Furthermore, the estimated lower bound of τ is 0.779 and the estimated lower bound of η is 0.630, suggesting that treatment E is indeed better than treatment C. Second, treatment E and treatment D do not stochastically dominate each other. However, the estimated lower bound of τ is 0.645 and the estimated lower bound of η is 0.574, suggesting that treatment E is better than treatment D. Therefore, our new causal parameters τ and η are useful for decision making, especially when the stochastic dominance assumption does not hold.

Table 2.3: *Comparisons of three treatments*

	$\widehat{\tau}_L$	$\widehat{\tau}_I$	$\widehat{\tau}_U$	CI for (τ_L, τ_U)	$\widehat{\eta}_L$	$\widehat{\eta}_I$	$\widehat{\eta}_U$	CI for (η_L, η_U)
E vs C	0.779	0.945	1.000	(0.673, 1.000)	0.630	0.777	0.870	(0.480 1.000)
E vs D	0.645	0.782	0.855	(0.495, 1.000)	0.574	0.660	0.736	(0.423, 0.886)

2.5 Extensions

2.5.1 Tighter Bounds by Using Covariates

With covariates, we can further sharpen the bounds on the causal pamameters (Lee, 2009; Long and Hudgens, 2013; Mealli and Pacini, 2013). Without loss of generality, we focus only on the bounds of the causal parameter τ. Within each level of the pretreatment covariates $\boldsymbol{X} = \boldsymbol{x}$, we can define the conditional probability that the treatment is beneficial, i.e.,

$$\tau\left(\boldsymbol{x}\right) = \text{pr}\{Y_i(1) \geq Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}\},$$

and obtain its conditional sharp upper and lower bounds $\tau_L\left(\boldsymbol{x}\right)$ and $\tau_U\left(\boldsymbol{x}\right)$. If we average the conditional bounds over the distributions of the covariates $F\left(\boldsymbol{x}\right)$, then the bounds for $\tau = \int \tau\left(\boldsymbol{x}\right) F\left(d\boldsymbol{x}\right)$ become

$$\tau_L' = \int \tau_L\left(\boldsymbol{x}\right) F\left(d\boldsymbol{x}\right), \quad \tau_U' = \int \tau_U\left(\boldsymbol{x}\right) F\left(d\boldsymbol{x}\right). \tag{2.18}$$

Theorem 2.4. The adjusted bounds are tighter than the unadjusted ones.

Proof of Theorem 2.4. The proof follows the same logic as Lee (2009). Because any value of τ within the covariate adjusted bounds $[\tau_L', \tau_U']$ must be compatible with the distributions of $\{Y(1), \boldsymbol{X}\}$ and $\{Y(0), \boldsymbol{X}\}$, it must also be compatible with the distributions of $Y(1)$ and $Y(0)$ by discarding $\boldsymbol{X}$. Therefore, any value of τ within the adjusted bounds $[\tau_L', \tau_U']$ must also be within the unadjusted bounds $[\tau_L, \tau_U]$. Consequently, the adjusted bounds are tighter, i.e., $[\tau_L', \tau_U'] \subset [\tau_L, \tau_U]$. □

Example 2.5. Consider a population consisting of two sub-populations of equal sizes, labeled by a binary covariate $\boldsymbol{X} \in \{0, 1\}$. Assume that the potential outcomes of sub-populations $\boldsymbol{X} = 1$ and $\boldsymbol{X} = 0$ are the independent potential outcomes in Example 2.3 and 2.4. Some simple algebra gives the following probability matrix

$$\boldsymbol{P} = \begin{pmatrix} \frac{1}{10} & \frac{1}{25} & \frac{3}{50} \\ \frac{9}{50} & \frac{2}{25} & \frac{7}{50} \\ \frac{11}{50} & \frac{2}{25} & \frac{1}{10} \end{pmatrix},$$

and consequently $\tau = 19/25$. Moreover, the corresponding marginal probabilities are $\boldsymbol{p}_1 = (1/5, 2/5, 2/5)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (1/2, 1/5, 3/10)^{\mathrm{T}}$. Therefore, if we apply Theorem 2.1 using only the above marginal distributions, we obtain that $\tau_L = 1/2$ and $\tau_U = 1$. However, if we first obtain the bounds for the two sub-populations and then average over them, we obtain the following sharper covariate adjusted bounds:

$$\tau_L' = \frac{1}{2}\tau_L(1) + \frac{1}{2}\tau_L(0) = \frac{1}{2}, \quad \tau_U' = \frac{1}{2}\tau_U(1) + \frac{1}{2}\tau_U(0) = \frac{9}{10}.$$

2.5.2 Sharp Bounds of the Conditional Medians

In spite of the limitations mentioned earlier, the conditional medians in (2.2) provide useful information about the causal effects on ordinal outcomes. The following theorem presents their sharp bounds.

Theorem 2.5. If $p_{+j} > 0$, the sharp lower and upper bounds of m_j are the minimum and maximum values of the following set:

$$M_j = \left\{ 0 \le k \le J-1 : \quad \sum_{k' \le k} p_{k'+} \ge \frac{p_{+j}}{2}, \quad \sum_{k' \ge k} p_{k'+} \ge \frac{p_{+j}}{2} \right\}.$$

If we further assume monotonicity, i.e., $Y_i(1) \ge Y_i(0)$ for all i, the sharp lower and upper bounds of m_j are the minimum and maximum values of the set $M_j \cap \{j, j+1, \ldots, J-1\}$.

Proof of Theorem 2.5. To make sure that the theorem is meaningful, we first prove that M_j is not empty for all $j = 0, 1, \ldots, J-1$. We let

$$k_{\min} = \min \left\{ k : \sum_{k' \le k} p_{k'+} \ge \frac{1}{2} \right\}$$

be the minimum value of the medians of $Y_i(1)$'s. Clearly $k_{\min}$ is well defined, and satisfies the condition that

$$\sum_{k' \le k_{\min}} p_{k'+} \ge \frac{1}{2} \ge \frac{p_{+j}}{2}.$$

We prove that $\sum_{k' \ge k_{\min}} p_{k'+} \ge 1/2$. If not, we must have that

$$\sum_{k' \le k_{\min}-1} p_{k'+} = 1 - \sum_{k' \ge k_{\min}} p_{k'+} > 1 - \frac{1}{2} = \frac{1}{2},$$

which contradicts the definition of $k_{\min}$. Therefore, $k_{\min}$ must satisfy

$$\sum_{k' \ge k_{\min}} p_{k'+} \ge \frac{1}{2} \ge \frac{p_{+j}}{2}.$$

We conclude that $k_{\min} \in M_j$, and therefore M_j is not empty.

Now we prove Theorem 2.5 itself. For any number $k \in M_j$, we prove that there exists a probability matrix $\boldsymbol{P}$, with fixed marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$, such that the conditional median m_j can be k. Let $\boldsymbol{C}_j = (p_{0j}, ..., p_{J-1,j})^{\mathrm{T}}$ be the jth column of the probability matrix $\boldsymbol{P}$. We first construct $\boldsymbol{C}_j$, with nonnegative elements, such that the sum of its elements is p_{+j}, and

$$\sum_{k' \le k} p_{k'j} \ge \frac{p_{+j}}{2}, \quad \sum_{k' \ge k} p_{k'j} \ge \frac{p_{+j}}{2}. \tag{2.19}$$

Because $\sum_{k' \le k} p_{k'+} \ge \sum_{k' \le k} p_{k'j}$ and $\sum_{k' \ge k} p_{k'+} \ge \sum_{k' \ge k} p_{k'j}$, a necessary condition for the existence of such $\boldsymbol{C}_j$ is

$$\sum_{k' \le k} p_{k'+} \ge \frac{p_{+j}}{2}, \quad \sum_{k' \ge k} p_{k'+} \ge \frac{p_{+j}}{2}. \tag{2.20}$$

We now show that (2.20) is also sufficient for the existence of such $\boldsymbol{C}_j$. To do this, we discuss the following four cases according to the values of $\sum_{k' \le k} p_{k'+}$ and $\sum_{k' \ge k} p_{k'+}$:

(1) $\sum_{k' \le k} p_{k'+} < 1/2$ and $\sum_{k' \ge k} p_{k'+} < 1/2$. This is a impossible scenario, because

$$\sum_{k' \le k} p_{k'+} + \sum_{k' \ge k} p_{k'+} \ge \sum_{k'=0}^{J-1} p_{k'+} = 1.$$

(2) $\sum_{k' \le k} p_{k'+} \ge 1/2$ and $\sum_{k' \ge k} p_{k'+} \ge 1/2$. We construct $\boldsymbol{C}_j$ by letting

$$p_{k'j} = p_{k'+} p_{+j} \quad (k' = 0, \ldots, J-1),$$

and (2.19) holds trivially.

(3) $\sum_{k' \le k} p_{k'+} \ge 1/2$ and $\sum_{k' \ge k} p_{k+} < 1/2$. We construct $\boldsymbol{C}_j$ in the following way:

   (a) If $k' < k$, let $p_{k'j} = p_{k'+}p_{+j} \Big/ \left(2\sum_{k'<k} p_{k'+}\right)$.

   (b) If $k' \ge k$, let $p_{k'j} = p_{k'+}p_{+j} \Big/ \left(2\sum_{k' \ge k} p_{k'+}\right)$.

   Clearly $\boldsymbol{C}_j$ has nonnegative entries, and their sum is p_{+j}, because

$$\sum_{k'=0}^{J-1} p_{k'j} = \sum_{k'<k} p_{k'j} + \sum_{k' \ge k} p_{k'j} = \frac{p_{+j}}{2} + \frac{p_{+j}}{2} = p_{+j}.$$

   Furthermore, (2.19) holds, because

$$\sum_{k' \le k} p_{k'j} \ge \sum_{k'<k} p_{k'j} = \frac{p_{+j}}{2}, \quad \sum_{k' \ge k} p_{k'j} = \frac{p_{+j}}{2}.$$

(4) $\sum_{k' \le k} p_{k'+} < 1/2$ and $\sum_{k' \ge k} p_{k'+} \ge 1/2$. Similar to case ((3)) above, we construct $\boldsymbol{C}_j$ in the following way:

   (a) If $k' \le k$, let $p_{k'j} = p_{k'+}p_{+j} \Big/ \left(2\sum_{k' \le k} p_{k'+}\right)$.

   (b) If $k' > k$, let $p_{k'+}p_{+j} \Big/ \left(2\sum_{k'>k} p_{k'+}\right)$.

To summarize, we have constructed the jth column of $\boldsymbol{P}$ that satisfies (2.19). To finish the construction of the probability matrix $\boldsymbol{P}$, we let

$$p_{k'l} = (p_{k'+} - p_{k'j}) \times \frac{p_{+l}}{1 - p_{+j}} \ge 0 \quad (l \ne j).$$

It is obvious that the constructed probability matrix $\boldsymbol{P}$ has marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$. By the construction of $\boldsymbol{C}_j$, m_j can be k for any $k \in M_j$. □

Example 2.6. First, if $\boldsymbol{p}_1 = (1/2, 1/4, 1/4)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (2/3, 1/6, 1/6)^{\mathrm{T}}$, then the bounds of m_1 are $0 \le m_1 \le 2$. Therefore, the marginal distributions provide no information for m_1.

Second, if $\boldsymbol{p}_1 = (1/6, 1/6, 2/3)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (1/4, 1/2, 1/4)^{\mathrm{T}}$, then the bounds of m_1 are $1 \le m_1 \le 2$. Therefore, the marginal distributions provide partial information for m_1.

Third, if $\boldsymbol{p}_1 = (1/6, 2/3, 1/6)^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (1/4, 1/2, 1/4)^{\mathrm{T}}$, then the bounds of m_1 shrink to a point $m_1 = 1$. Therefore, the marginal distributions completely determine m_1.

In the first case, the marginal distributions satisfy the stochastic dominance assumption. If we further assume monotonicity, we can improve the bounds of m_1 by $1 \leq m_1 \leq 2$.

2.6 Discussions

For ordinal outcomes, we have discussed various causal parameters and their sharp bounds by using only the marginal distributions of the potential outcomes. Our causal parameters τ and η are closely related to the relative treatment effect previously studied under the assumption of independent potential outcomes (Agresti, 2010):

$$\alpha = \mathrm{pr}\left\{Y_i(1) > Y_i(0)\right\} + \frac{1}{2}\mathrm{pr}\left\{Y_i(1) = Y_i(0)\right\}.$$

This relative treatment effect α and our new ones have a simple algebraic relationship, i.e., $\alpha = (\tau + \eta)/2$. Therefore, our newly proposed causal parameters τ and η determine α, and consequently our bounds also provide information for α.

Example 2.5 has illustrated the role of a binary covariate in improving bounds on the causal parameters. In practice, with general and high dimensional covariates, we can invoke parametric or nonparametric models for ordinal outcomes (Agresti, 2010) to estimate the marginal distributions of the potential outcomes, and then apply the formulas for bounds on the causal effects. Our theoretical results are closely related to the probability structure of the potential outcomes, and thus can be incorporated into different statistical inference frameworks and statistical models.

We have discussed causal inference for ordinal outcomes with a single treatment. For factorial experiments, Dasgupta et al. (2015) discussed causal inference of average treatment effects for continuous outcomes. In the future, we will explore factorial experiments with ordinal outcomes.

Chapter 3

Causal Inference of Ordinal Outcomes with Noncompliance

Noncompliance is an important issue in both methodology and applied causal inference research. The principal stratification framework addresses this issue by defining subgroup causal effects, based on the joint value of the potential outcomes of treatment received. In this chapter, we extend the results in Chapter 2, by studying the probabilities that the treatment is beneficial and strictly beneficial for compliers, i.e., the experimental units complying to whichever treatment assigned. For ordinal outcomes, we derive the sharp bounds of these two causal parameters using the marginal distributions of potential outcomes for compliers. To identify such marginal distributions, we invoke two classical assumptions in the causal inference literature, namely monotonicity and exclusion restriction. We tighten the bounds using pretreatment covariates, and demonstrate our results by numerical and real examples.

3.1 Introduction

Under the potential outcomes framework (Neyman, 1923; Rubin, 1974), we can define causal effects as comparisons between the potential outcomes under treatment and control. In applied research ordinal outcomes are common (e.g., Bruce et al., 2004). However, average outcomes themselves are not well defined for ordinal outcomes, rendering the average causal effect, generally the parameter of interest, not applicable. Realizing this salient feature of ordinal outcomes, Chapter 2 proposed two new causal parameters measuring the probabilities that the treatment is beneficial and strictly beneficial for the experimental units. These two causal parameters are well defined for any outcomes, and of particular interest for ordinal outcomes. Unfortunately, however, Chapter 2 did not address the noncompliance issue, which is also common in applied research. For instance, in clinical trials some patients may not comply with their assigned treatment, due to fear of potential side effects. Although noncompliance itself has been extensively studied (e.g., Sommer and Zeger, 1991; Baker and Lindeman, 1994; Robins and Greenland, 1994; Angrist et al., 1996; Hirano et al., 2000; Frangakis and Rubin, 2002), there appears to be very limited discussions about causal inference of ordinal outcomes in the presence of noncompliance, to our best knowledge. Cheng (2009) discussed various causal parameters under the assumptions of one-sided noncompliance and independent potential outcomes, and Baker (2011) generalized her results to two-sided noncompliance.

In this chapter, we generalize the framework in Chapter 2 to address noncompliance. We invoke the principal stratification framework (Angrist et al., 1996; Frangakis and Rubin, 2002), which permits defining subgroup causal effects based on the principal stratification variable, i.e., the joint value of the potential outcomes of treatment received. Because the principal stratification variable is unaffected by the treatment, inference conditioning on it yields valid causal interpretations. For ordinal outcomes, we study the probabilities that the treatment is beneficial and strictly beneficial for the principal strata, i.e., subgroups defined by the principal stratification variable. Under the classical monotonicity and exclusion restriction assumptions (Angrist et al., 1996), we can either pointly identify or derive sharp bound for our desired causal parameters using the marginal distributions of the potential outcomes for the principal strata. We further tighten the bounds when pretreatment covariates are available, and propose an EM algorithm to estimate the tighter bounds in practice.

The chapter proceeds as follows. Section 3.2 sets up the theoretical framework of causal inference for ordinal outcomes with noncompliance, and introduces the causal parameters of interest. Section 3.3 derives the sharp bounds of the causal parameters, and Section 3.4 discusses the inference of the bounds. Sections 3.5 and 3.6 present some numerical and real examples to demonstrate the advantages of our theoretical results. Section 3.7 concludes. Web Appendix A provides the proofs of the lemmas, theorems and corollaries in Section 3.3, and Web Appendix B provides the technical details in Section 3.4.

3.2 Basic Framework

3.2.1 Potential Outcomes and Principal Stratification

We consider a completely randomized experiment with N units drawn independently from a hypothetical super-population, and a binary treatment. For the ith unit, we let $\boldsymbol{X}_i$ be the pretreatment covariates, Z_i be the treatment assignment indicator with $Z_i = 1$ for active treatment and $Z_i = 0$ for control, D_i^{obs} be the treatment receipt indicator with $D_i^{\text{obs}} = 1$ for active treatment and $D_i^{\text{obs}} = 0$ for control, and Y_i^{obs} be the outcome of interest that is ordinal, with J categories labeled $0, \ldots, J-1$, where 0 and $J-1$ respectively represent the worst and best categories. Noncompliance occurs when there exists i such that $Z_i \neq D_i^{\text{obs}}$.

We invoke the potential outcomes framework to define causal effects. Under the Stable Unit Treatment Value Assumption (Rubin, 1980), there is only one version of the treatment and no interference among units. For the ith unit, we define the pair $\{D_i(1), D_i(0)\}$ as the potential values of treatment received and $\{Y_i(1), Y_i(0)\}$ as the potential values of outcome of interest, under treatment and control respectively. We adopt the principal stratification framework (Angrist et al., 1996; Frangakis and Rubin, 2002). For the ith unit, we define the principal stratification variable U_i as the joint value of $\{D_i(1), D_i(0)\}$. To be specific,

- let $U_i = a$, if $D_i(1) = 1$ and $D_i(0) = 1$;

- let $U_i = c$, if $D_i(1) = 1$ and $D_i(0) = 0$;
- let $U_i = d$, if $D_i(1) = 0$ and $D_i(0) = 1$;
- let $U_i = n$, if $D_i(1) = 0$ and $D_i(0) = 0$.

The four principal strata, i.e., subgroups defined by U, are sometimes referred to as "always-takers," "compliers," "defiers" and "never-takers" (Angrist et al., 1996). For principal stratum u, let $\pi_u = \text{pr}(U = u)$ represent the probabilities of the principal strata, and $u_{kl} = \text{pr}\{Y(1) = k, Y(0) = l \mid U = u\}$ denote the probability of having potential outcome k under treatment and potential outcome l under control. The $J \times J$ probability matrix $\{u_{kl}\}_{0 \le k,l \le J-1}$ summarizes the joint distribution of the potential outcomes for principal stratum u. We let

$$u_{k+} = \sum_{l'=0}^{J-1} u_{kl'}, \quad u_{+l} = \sum_{k'=0}^{J-1} u_{k'l} \quad (k, l = 0, 1, \ldots, J-1).$$

The vectors $(u_{0+}, \ldots, u_{J-1,+})$ and $(u_{+0}, \ldots, u_{+,J-1})$ characterize the marginal distributions of the potential outcomes under treatment and control. By law of total probability, let

$$p_{kl} = \sum_u \pi_u u_{kl}, \quad p_{k+} = \sum_u \pi_u u_{k+}, \quad p_{+l} = \sum_u \pi_u u_{+l},$$

characterize the joint and marginal distributions of the potential outcomes under treatment and control, for the super-population.

3.2.2 Causal Parameters for Ordinal Outcomes

For ordinal outcomes, because the average itself is not well defined, Chapter 2 studied the distributional causal effects (cf. Ju and Geng, 2010)

$$\Delta_j = \text{pr}\{Y(1) \ge j\} - \text{pr}\{Y(0) \ge j\} = \sum_{k \ge j} p_{k+} - \sum_{l \ge j} p_{+l} \quad (j = 0, \ldots, J-1),$$

and proposed two new causal parameters, i.e., the probabilities that the treatment is beneficial and strictly beneficial for the population:

$$\tau = \mathrm{pr}\left\{Y_i(1) \geq Y_i(0)\right\} = \sum\sum_{k \geq l} p_{kl}, \quad \eta = \mathrm{pr}\left\{Y_i(1) > Y_i(0)\right\} = \sum\sum_{k > l} p_{kl}. \tag{3.1}$$

With noncompliance, because the principal stratification variable is unaffected by the treatment, inference conditioning on it yields valid causal interpretations. Therefore, in this chapter focus on the "principal stratification counterparts" of the above causal parameters, i.e., the distributional causal effects for principal stratum u :

$$\Delta_{u,j} = \mathrm{pr}\left\{Y(1) \geq j \mid U = u\right\} - \mathrm{pr}\left\{Y(0) \geq j \mid U = u\right\} = \sum_{k \geq j} u_{k+} - \sum_{l \geq j} u_{+l},$$

as well as the probabilities that the treatment is beneficial and strictly beneficial for principal stratum u :

$$\tau_u = \mathrm{pr}\left\{Y_i(1) \geq Y_i(0) \mid U = u\right\} = \sum\sum_{k \geq l} u_{kl}, \quad \eta_u = \mathrm{pr}\left\{Y_i(1) > Y_i(0) \mid U = u\right\} = \sum\sum_{k > l} u_{kl}.$$

Under the monotonicity and exclusion restriction assumptions which we formally define later, we can either pointly identify or bound the above causal parameters. The definition of point identification is that, given the distribution of $(\boldsymbol{X}, Z, D^{\mathrm{obs}}, Y^{\mathrm{obs}})$, we can uniquely determine the value of the parameter of interest.

3.3 Sharp Bounds on Causal Effects among Compliers

3.3.1 Main Results

The definitions of the causal parameters τ and η involve the association between the potential outcomes $\{Y_i(1), Y_i(0)\}$, and those of τ_u and η_u further involve the principal stratification variable U_i, which is defined by $\{D_i(1), D_i(0)\}$. Unfortunately, we can never jointly measure $\{Y_i(1), Y_i(0)\}$ or $\{D_i(1), D_i(0)\}$, rendering the causal parameters unidentifiable. To partially circumvent this difficulty, Chapter 2 derived the sharp bounds of τ and η, which are defined as the minimal and maximal values of the objective functions (3.1) under the constraints of the marginal probabilities:

$$\sum_{l'=0}^{J-1} p_{kl'} = p_{k+}, \quad \sum_{k'=0}^{J-1} p_{k'l} = p_{+l}, \quad p_{kl} \geq 0 \quad (k, l = 0, \ldots, J-1).$$

To recall, we summarize their main results in the following lemma.

Lemma 3.1. The sharp lower and upper bounds of τ are

$$\tau_L = \max_{0 \leq j \leq J-1} \{p_{+j} + \Delta_j\}, \quad \tau_U = 1 + \min_{0 \leq j \leq J-1} \Delta_j,$$

and the sharp lower and upper bounds of η are

$$\eta_L = \max_{0 \leq j \leq J-1} \Delta_j, \quad \eta_U = 1 + \min_{0 \leq j \leq J-1} \{\Delta_j - p_{j+}\}.$$

If the marginal probabilities of potential outcomes for principal stratum u are identifiable, we can accordingly sharply bound the causal parameters τ_u and η_u in the same way. To make the desired marginal probabilities identifiable, we make the following assumptions.

Assumption 3.1 (Monotonicity)**.** $D_i(1) \geq D_i(0)$ for all $i = 1, \ldots, N$.

The monotonicity assumption rules out defiers. Within the treatment group, those receiving treatment are compliers or always-takers, and those receiving control are never-takers. Within the control group, those receiving treatment are always-takers, and those receiving control are compliers or never-takers. We can identify the probabilities of all principal strata as

$$\pi_a = \mathrm{pr}(D^{\mathrm{obs}} = 1 \mid Z = 0), \quad \pi_n = \mathrm{pr}(D^{\mathrm{obs}} = 0 \mid Z = 1), \quad \pi_c = 1 - \pi_a - \pi_n. \tag{3.2}$$

Assumption 3.2 (Exclusion Restriction)**.** $Y_i(1) = Y_i(0)$ for all i such that $U_i = a$ or n.

Under the monotonicity and exclusion restriction assumptions, from the distribution of $(Z, D^{\mathrm{obs}}, Y^{\mathrm{obs}})$, we can identify the marginal probabilities of potential outcomes for all the principal strata. First, for the always-takers, under the monotonicity assumption we have

$$\begin{aligned} a_{+j} &= \mathrm{pr}\{Y(0) = j \mid U = a\} \\ &= \mathrm{pr}\{Y^{\mathrm{obs}} = j \mid Z = 0, D^{\mathrm{obs}} = 1\}, \end{aligned} \tag{3.3}$$

and under the exclusion restriction assumption we have

$$\begin{aligned} a_{j+} &= \text{pr}\{Y(1) = j \mid U = a\} \\ &= \text{pr}\{Y(0) = j \mid U = a\} = a_{+j}. \end{aligned} \tag{3.4}$$

Second, for the never-takers, under the monotonicity assumption we have

$$\begin{aligned} n_{j+} &= \text{pr}\{Y(1) = j \mid U = n\} \\ &= \text{pr}\{Y^{\text{obs}} = j \mid Z = 1, D^{\text{obs}} = 0\}, \end{aligned} \tag{3.5}$$

and under the exclusion restriction assumption we have

$$\begin{aligned} n_{+j} &= \text{pr}\{Y(0) = j \mid U = n\} \\ &= \text{pr}\{Y(1) = j \mid U = n\} = n_{j+}. \end{aligned} \tag{3.6}$$

Third, by the law of total probability, we have

$$\begin{aligned} p_{k+} &= \text{pr}\{Y(1) = k\} \\ &= \sum_{u=a,c,n} \text{pr}\{U = u\} \, \text{pr}\{Y(1) = k \mid U = u\} \\ &= \pi_a a_{k+} + \pi_c c_{k+} + \pi_n n_{k+} \end{aligned}$$

and

$$\begin{aligned} p_{+l} &= \text{pr}\{Y(0) = l\} \\ &= \sum_{u=a,c,n} \text{pr}\{U = u\} \, \text{pr}\{Y(0) = l \mid U = u\} \\ &= \pi_a a_{+l} + \pi_c c_{+l} + \pi_n n_{+l} \end{aligned}$$

Therefore, for the compliers we have

$$\begin{aligned} c_{k+} &= \text{pr}\{Y(1) = k \mid U = c\} \\ &= (p_{k+} - a_{k+}\pi_a - n_{k+}\pi_n) / \pi_c, \end{aligned} \tag{3.7}$$

and

$$
\begin{aligned}
c_{+l} &= \mathrm{pr}\left\{Y(0)=l \mid U=c\right\} \\
&= \left(p_{+l} - a_{+l}\pi_a - n_{+l}\pi_n\right)/\pi_c. \qquad (3.8)
\end{aligned}
$$

Furthermore, under the monotonicity and exclusion assumptions, we can pointly identify or sharply bound the causal parameters τ_u and η_u. To be more specific, for the always-takers and never-takers we have $\tau_u = 1$ and $\eta_u = 0$. For the compliers, we can identify τ_u and η_u using the marginal probabilities for the compliers, which are identifiable by (3.7) and (3.8).

Theorem 3.1. The sharp lower and upper bounds of τ_c are

$$
\tau_{c,L} = \max_{0\le j\le J-1}\left(c_{+j} + \Delta_{c,j}\right), \quad \tau_{c,U} = 1 + \min_{0\le j\le J-1}\Delta_{c,j}, \qquad (3.9)
$$

and the sharp lower and upper bounds of η_c are

$$
\eta_{c,L} = \max_{0\le j\le J-1}\Delta_{c,j}, \quad \eta_{c,U} = 1 + \min_{0\le j\le J-1}\left(\Delta_{c,j} - c_{j+}\right). \qquad (3.10)
$$

Additionally, we can establish the following relationships between the causal parameters τ and τ_c, and between η and η_c.

Lemma 3.2. The following equalities holds:

$$
\tau_c = \tau/\pi_c - \left(1-\pi_c\right)/\pi_c, \quad \eta_c = \eta/\pi_c.
$$

Theorem 3.1 and Lemma 3.2 imply two sets of new bounds for τ_c and η_c, which do not involve the marginal probabilities of potential outcomes for compliers, and therefore easy to compute.

Corollary 3.1. We can bound τ_c from below and above using

$$
\tau_{c,L}^{\mathrm{N}} = \tau_L/\pi_c - \left(1-\pi_c\right)/\pi_c, \quad \tau_{c,U}^{\mathrm{N}} = \tau_U/\pi_c - \left(1-\pi_c\right)/\pi_c,
$$

and bound η_c from below and above using

$$
\eta_{c,L}^{\mathrm{N}} = \eta_L/\pi_c, \quad \eta_{c,U}^{\mathrm{N}} = \eta_U/\pi_c.
$$

Moreover, these new bounds and those in Theorem 3.1 satisfy the following:

$$\tau_{c,L} \geq \tau_{c,L}^{\mathrm{N}}, \quad \tau_{c,U} = \tau_{c,U}^{\mathrm{N}}; \quad \eta_{c,L} = \eta_{c,L}^{\mathrm{N}}, \quad \eta_{c,U} \leq \eta_{c,U}^{\mathrm{N}}.$$

Theorem 3.1 and Lemma 3.2 also imply two new sets of bounds for τ and η, tighter than those in Lemma 3.1.

Corollary 3.2. We can bound τ from below and above using

$$\tau_L^{\mathrm{PS}} = \pi_c \tau_{c,L} + 1 - \pi_c, \quad \tau_U^{\mathrm{PS}} = \pi_c \tau_{c,U} + 1 - \pi_c,$$

and bound η from below and above using

$$\eta_L^{\mathrm{PS}} = \pi_c \eta_{c,L}, \quad \eta_U^{\mathrm{PS}} = \pi_c \eta_{c,U}.$$

Moreover, these new bounds and those in Lemma 3.1 satisfy the following:

$$\tau_L \leq \tau_L^{\mathrm{PS}}, \quad \tau_U = \tau_U^{\mathrm{PS}}; \quad \eta_L = \eta_L^{\mathrm{PS}}, \quad \eta_U \geq \eta_U^{\mathrm{PS}}.$$

There are two reasons that we can obtain tighter bounds. First, we use the principal stratification variable U as a pretreatment variable. Second, the monotonicity and exclusion restriction assumptions further restrict the probability structure of the potential outcomes. It is worth noting that the new lower bound of τ_c is larger than the old lower bound, but the two upper bounds are identical. Similarly, the new upper bound of η_c is larger than the old upper bound, but the two lower bounds are identical. This "partial improvement" phenomenon is observed in Grilli and Mealli (2008) and later pointed out and explained in Long and Hudgens (2013). We use the following example to further compare the "principal stratification" adjusted bounds and the unadjusted bounds in Corollary 3.2.

Example 3.1. Consider a balanced completely randomized experiment with a binary treatment and a three-level outcome. For the distribution of the observed data $(Z, D^{\mathrm{obs}}, Y^{\mathrm{obs}})$, first assume that

$$\mathrm{pr}(D^{\mathrm{obs}} = 1 \mid Z = 1) = 3/4, \quad \mathrm{pr}(D^{\mathrm{obs}} = 1 \mid Z = 0) = 1/4.$$

Next, we report the distribution of Y^{obs} given (Z, D^{obs}) in column 4 of Table 3.1.

Having specified the distribution of the observed data $(Z, D^{\text{obs}}, Y^{\text{obs}})$, we use it to identify the probabilities of all the principal strata, and the marginal probabilities of potential outcomes for the whole population and all the principal strata. First, by (3.2) we can identify the probabilities of the principal strata as

$$\pi_a = 1/4, \quad \pi_n = 1/4, \quad \pi_c = 1/2.$$

Second, by law of total probability we can identify the marginal probabilities of potential outcomes for the whole population. To be more specific,

$$\begin{aligned}\text{pr}\{Y(z) = j\} &= \text{pr}(Y^{\text{obs}} = j \mid Z = z) \\ &= \sum_{d=0}^{1} \text{pr}(D^{\text{obs}} = d \mid Z = z)\text{pr}(Y^{\text{obs}} = j \mid Z = z, D^{\text{obs}} = d)\end{aligned}$$

for $z = 0$ and 1. We report the marginal probabilities of potential outcomes for the whole population, in column 5 of Table 3.1. Third, by (3.7) and (3.8) we recover the marginal probabilities of potential outcomes for compliers, and we report them in column 7 of Table 3.1. With the above identification results, we can obtain the adjusted and unadjusted bounds. To save space for the main text, we only focus on τ. On the one hand, using the marginal probabilities of potential outcomes for compliers, we obtain the sharp bounds of τ_c, which by Corollary 3.2 imply the "principal stratification" adjusted bounds of τ. We report them in columns 8–9 of Table 3.1. On the other hand, using the marginal probabilities of potential outcomes for the whole population, by Lemma 3.1 we obtain the unadjusted bounds of τ. We report them in columns 10–11 of Table 3.1. The unadjusted upper bound is identical to the adjusted one, and the adjusted lower bound improves upon the unadjusted one.

Analogous to Chapter 2, the probability distributions that achieve the lower and upper bounds in (3.9) and (3.10) correspond to negatively associated and positively associated potential outcomes for compliers. In practice, we may also be interested in the case with independent potential outcomes (Rubin, 1978; Agresti, 2010), where $c_{kl} = c_{k+}c_{+l}$ for all k, l. For instance, Cheng (2009) implicitly assumed independent potential outcomes when discussing various causal parameters. With independent potential outcomes, we can identify τ_c and η_c by the marginal probabilities of potential outcomes for compliers.

Theorem 3.2. With independent potential outcomes for compliers,

$$\tau_{c,I} = \sum\sum_{k \geq l} c_{k+}c_{+l}, \quad \eta_{c,I} = \sum\sum_{k>l} c_{k+}c_{+l}.$$

Furthermore, $\tau_{c,L} \leq \tau_{c,I} \leq \tau_{c,U}$ and $\eta_{c,L} \leq \eta_{c,I} \leq \eta_{c,U}$.

3.3.2 Tighter Bounds by Covariate Adjustment

With pretreatment covariates, we can further sharpen the unadjusted bounds, i.e., the bounds that do not utilize the pretreatment covariate information. In the existing literature, a few examples are Lee (2009); Long and Hudgens (2013); Mealli and Pacini (2013). To save space for the main text, we focus only on τ_c. Within each level of the pretreatment covariates $\boldsymbol{X} = \boldsymbol{x}$, we can define let the probabilities of the principal strata as

$$\pi_u(\boldsymbol{x}) = \text{pr}(U = u \mid \boldsymbol{X} = \boldsymbol{x}) \quad (u = a, c, n); \tag{3.11}$$

the marginal probabilities of the potential outcomes for principal strata u as

$$u_{k+}(\boldsymbol{x}) = \text{pr}\{Y(1) = k \mid U = u, \boldsymbol{X} = \boldsymbol{x}\}, \quad u_{+l}(\boldsymbol{x}) = \text{pr}\{Y(0) = l \mid U = u, \boldsymbol{X} = \boldsymbol{x}\}; \tag{3.12}$$

and the probability that the treatment is beneficial for a complier as

$$\tau_c(\boldsymbol{x}) = \text{pr}\{Y(1) \geq Y(0) \mid U = c, \boldsymbol{X} = \boldsymbol{x}\}.$$

Let $F_c(\boldsymbol{x})$ and $F(\boldsymbol{x})$ respectively denote the cumulative distribution functions of the covariates for compliers and the whole population, respectively. By Bayes rule, we have

$$dF_c(\boldsymbol{x}) = \frac{\pi_c(\boldsymbol{x})\, dF(\boldsymbol{x})}{\int \pi_c(\boldsymbol{x})\, dF(\boldsymbol{x})},$$

which implies that

$$\tau_c = \int \tau_c(\boldsymbol{x})\, dF_c(\boldsymbol{x}) = \frac{\int \tau_c(\boldsymbol{x})\, \pi_c(\boldsymbol{x})\, dF(\boldsymbol{x})}{\int \pi_c(\boldsymbol{x})\, dF(\boldsymbol{x})}.$$

Consequently, if we obtain the sharp upper and lower bounds of $\tau_c(\boldsymbol{x})$, denoted as $\tau_{c,L}(\boldsymbol{x})$ and $\tau_{c,U}(\boldsymbol{x})$ respectively, the bounds for τ_c become

$$\tau'_{c,L} = \frac{\int \tau_{c,L}(\boldsymbol{x})\,\pi_c(\boldsymbol{x})\,dF(\boldsymbol{x})}{\int \pi_c(\boldsymbol{x})\,dF(\boldsymbol{x})}, \quad \tau'_{c,U} = \frac{\int \tau_{c,U}(\boldsymbol{x})\,\pi_c(\boldsymbol{x})\,dF(\boldsymbol{x})}{\int \pi_c(\boldsymbol{x})\,dF(\boldsymbol{x})}. \tag{3.13}$$

Theorem 3.3. For any pretreatment covariates $\boldsymbol{X}$, the covariate adjusted bounds in (3.13) are tighter than the unadjusted bounds in (3.9).

Example 3.2. Consider a three-level outcome of interest $Y^{\mathrm{obs}} \in \{0,1,2\}$, a population consisting of two equal-sized sub-populations labeled by a binary covariate $\boldsymbol{X} \in \{0,1\}$, and a balanced completely randomized experiment. For the treatment received D^{obs}, assume that for $\boldsymbol{x} = 0,1$ we have

$$\mathrm{pr}(D^{\mathrm{obs}} = 1 \mid Z = 1, \boldsymbol{X} = \boldsymbol{x}) = 2/3, \quad \mathrm{pr}(D^{\mathrm{obs}} = 1 \mid Z = 0, \boldsymbol{X} = \boldsymbol{x}) = 1/3.$$

We consider four cases with different conditional distributions of the observed outcome Y^{obs} given $(\boldsymbol{X}, Z, D^{\mathrm{obs}})$, in column 5 of Table 3.2. For fixed distribution of the observed data $(\boldsymbol{X}, Z, D^{\mathrm{obs}}, Y^{\mathrm{obs}})$, under the monotonicity and exclusion restriction assumptions we can recover the latent distributions of potential outcomes within principal strata. To be more specific, first by (3.2) we have for $\boldsymbol{x} = 0,1$

$$\pi_a(\boldsymbol{x}) = 1/3, \quad \pi_c(\boldsymbol{x}) = 1/3, \quad \pi_n(\boldsymbol{x}) = 1/3.$$

Next, for all four cases and $\boldsymbol{x} = 0,1$, by (3.3) and (3.4) we have

$$a_{0+}(\boldsymbol{x}) = a_{+0}(\boldsymbol{x}) = 2/5, \quad a_{1+}(\boldsymbol{x}) = a_{+1}(\boldsymbol{x}) = 2/5, \quad a_{2+}(\boldsymbol{x}) = a_{+2}(\boldsymbol{x}) = 1/5.$$

Then, by (3.6) and (3.5) we have for all four cases

$$n_{0+}(\boldsymbol{x}) = n_{+0}(\boldsymbol{x}) = 1/5, \quad n_{1+}(\boldsymbol{x}) = n_{+1}(\boldsymbol{x}) = 2/5, \quad n_{2+}(\boldsymbol{x}) = n_{+2}(\boldsymbol{x}) = 2/5.$$

Finally, by (3.7) and (3.8) we recover the distributions of potential outcomes for compliers within the two sub-populations in column 6 of Table 3.2.

Using the above identification results, we obtain the unadjusted and covariate adjusted bounds of τ_c in columns 7–10 of Table 3.2. For all four cases, the covariate adjusted bounds are tighter than the unadjusted ones. In Case A both the lower and upper bounds are improved by covariate adjustment; in Case B only the lower bound is improved; in Case C only the upper bound is improved; in Case D both covariate adjusted bounds are the same as the unadjusted ones.

3.4 Inference of the Bounds

In practice, we can never measure the true values of the marginal probabilities for compliers, and therefore need to estimate them and the bounds on τ_c and η_c. In this section we discuss the inference of the bounds for completely randomized experiments. We first summarize the observed data $\{\boldsymbol{X}_i, Z_i, D_i^{\text{obs}}, Y_i^{\text{obs}}\}_{i=1}^N$ by letting

$$n_{zdj}^{\text{obs}} = \#\{i : Z_i = z, D_i^{\text{obs}} = d, Y_i^{\text{obs}} = j\} \quad (z, d = 0, 1; j = 0, \dots, J-1)$$

be the number of units with observed outcome $Y_i^{\text{obs}} = j$, that are assigned to treatment arm z and receive d. Furthermore, let

$$n_{zd+} = \sum_{j=0}^{J-1} n_{zdj}^{\text{obs}}, \quad n_{++j}^{\text{obs}} = \sum_{z=0}^{1} \sum_{d=0}^{1} n_{zdj}^{\text{obs}},$$

be the number of units that are assigned to treatment z and receive d; and the number of units with observed outcome $Y_i^{\text{obs}} = j$, respectively.

3.4.1 Simple Moment Estimators

For the unadjusted bounds in (3.9) and (3.10), we first estimate the probability of compliers by $\hat{\pi}_c = 1 - N_{01}/N_0 - N_{10}/N_1$, and then obtain the point estimators of the bounds by estimating the marginal probabilities of potential outcomes for the compliers:

$$\hat{c}_{k+} = (n_{11k}^{\text{obs}}/N_1 - n_{01k}^{\text{obs}}/N_0)/\hat{\pi}_c, \quad \hat{c}_{+l} = (n_{00l}^{\text{obs}}/N_0 - n_{10l}^{\text{obs}}/N_1)/\hat{\pi}_c.$$

Or more efficiently, we can use the EM algorithm (Dempster et al., 1977) to estimate π_c, c_{k+} and c_{+l}. For the relevant computation details see Baker (2011).

3.4.2 Covariate Adjusted Estimators

For the covariate adjusted bounds in (3.13), we obtain their point estimators by invoking parametric models for the principal stratification (parameters denoted as $\boldsymbol{\theta}_{\text{PS}}$), and the marginal probabilities of the potential outcomes (parameters denoted as $\boldsymbol{\theta}_{\text{PO}}$). Given the maximum likelihood estimates of the modeling parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{PS}}, \boldsymbol{\theta}_{\text{PO}})$, first we plug them in (3.11) and (3.12), and for each unit estimate its probability of being a complier, and its marginal probabilities of potential outcomes given that it is a complier. We denote those probabilities as $\widehat{\pi}_c(\boldsymbol{x}_i)$, $\widehat{c}_{k+}(\boldsymbol{x}_i)$ and $\widehat{c}_{+l}(\boldsymbol{x}_i)$, respectively. Next, we use $\{\widehat{c}_{0+}(\boldsymbol{x}_i), \ldots, \widehat{c}_{J-1,+}(\boldsymbol{x}_i)\}$ and $\{\widehat{c}_{+0}(\boldsymbol{x}_i), \ldots, \widehat{c}_{+,J-1}(\boldsymbol{x}_i)\}$ to estimate $\tau_{c,L}(\boldsymbol{x}_i)$ and $\tau_{c,U}(\boldsymbol{x}_i)$, the sharp lower and upper bounds of $\tau_c(\boldsymbol{x}_i)$. We denoted the estimates as $\widehat{\tau}_{c,L}(\boldsymbol{x}_i)$ and $\widehat{\tau}_{c,U}(\boldsymbol{x}_i)$ respectively. Finally, we estimate $\tau'_{c,L}$ and $\tau'_{c,U}$, the covariate adjusted bounds of τ_c, by the finite sample analogous of their expressions in (3.13):

$$\widehat{\tau}'_{c,L} = \frac{\sum_{i=1}^{N} \widehat{\tau}_{c,L}(\boldsymbol{x}_i)\, \widehat{\pi}_c(\boldsymbol{x}_i)}{\sum_{i=1}^{N} \widehat{\pi}_c(\boldsymbol{x}_i)}, \quad \widehat{\tau}'_{c,U} = \frac{\sum_{i=1}^{N} \widehat{\tau}_{c,U}(\boldsymbol{x}_i)\, \widehat{\pi}_c(\boldsymbol{x}_i)}{\sum_{i=1}^{N} \widehat{\pi}_c(\boldsymbol{x}_i)}.$$

For the case where we model the principal stratification by multiple logistic regression and the potential outcomes by proportional odds models, we propose an EM algorithm to obtain the maximum likelihood estimates, which is a nontrivial generalization of Baker (2011). To be specific, let $\boldsymbol{X}_i = \boldsymbol{x}_i$, $Z_i = z_i$, $D_i^{\text{obs}} = d_i$ and $Y_i^{\text{obs}} = y_i$ be the values of the pretreatment covariates, treatment assigned, treatment received and observed outcome of the ith unit. We treat the principal stratification variable U_i as missing data, and denote the realizations of (4) and (5) for the ith unit when evaluated at the true parameter value $\boldsymbol{\theta}$ as $\pi_u(\boldsymbol{x}_i)$, $u_{k+}(\boldsymbol{x}_i)$ and $u_{+l}(\boldsymbol{x}_i)$, and those when evaluated at the tth iteration of the parameter estimate $\boldsymbol{\theta}^{(t)}$ as $\pi_u^{(t)}(\boldsymbol{x}_i)$, $u_{k+}^{(t)}(\boldsymbol{x}_i)$ and $u_{+l}^{(t)}(\boldsymbol{x}_i)$. The EM algorithm proceeds as follows. Given the current tth iteration of the parameter estimate $\boldsymbol{\theta}^{(t)}$, we obtain the updated $(t+1)$th iteration $\boldsymbol{\theta}^{(t+1)}$ as follows:

1. E-Step: obtain the conditional expectation of the complete-data log-likelihood, given observed data and the current parameter estimate $\boldsymbol{\theta}^{(t)}$, by finding the (current) conditional probabilities of the principal stratum u, denoted as $\pi_{u,i}^{(t)}$:

   - for all i such that $z_i = 1$ and $d_i = 1$, let

   $$\pi_{n,i}^{(t)} = 0; \quad \pi_{a,i}^{(t)} = \frac{\pi_u^{(t)}\left(\boldsymbol{x}_i\right) a_{y_i,+}^{(t)}\left(\boldsymbol{x}_i\right)}{\pi_a^{(t)}\left(\boldsymbol{x}_i\right) a_{y_i,+}^{(t)}\left(\boldsymbol{x}_i\right) + \pi_c^{(t)}\left(\boldsymbol{x}_i\right) c_{y_i,+}^{(t)}\left(\boldsymbol{x}_i\right)} \quad \pi_{c,i}^{(t)} = 1 - \pi_{a,i}^{(t)};$$

   - for all i such that $z_i = 1$ and $d_i = 0$, let

   $$\pi_{a,i}^{(t)} = 0, \quad \pi_{c,i}^{(t)} = 0, \quad \pi_{n,i}^{(t)} = 1;$$

   - for all i such that $z_i = 0$ and $d_i = 1$, let

   $$\pi_{a,i}^{(t)} = 1, \quad \pi_{c,i}^{(t)} = 0, \quad \pi_{n,i}^{(t)} = 0;$$

   - for all i such that $z_i = 0$ and $d_i = 0$, let

   $$\pi_{a,i}^{(t)} = 0; \quad \pi_{c,i}^{(t)} = \frac{\pi_u^{(t)}\left(\boldsymbol{x}_i\right) c_{+,y_i}^{(t)}\left(\boldsymbol{x}_i\right)}{\pi_c^{(t)}\left(\boldsymbol{x}_i\right) c_{+,y_i}^{(t)}\left(\boldsymbol{x}_i\right) + \pi_n^{(t)}\left(\boldsymbol{x}_i\right) n_{+,y_i}^{(t)}\left(\boldsymbol{x}_i\right)} \quad \pi_{n,i}^{(t)} = 1 - \pi_{c,i}^{(t)}.$$

2. M-Step: obtain the updated parameter estimate $\boldsymbol{\theta}^{(t+1)}$, by maximizing the conditional expectation with respect to $\boldsymbol{\theta}$. To do this, we adopt the following two-step procedure:

   (a) Obtain $\boldsymbol{\theta}_{\text{PS}}^{(t+1)}$, the updated estimates of the parameters in the model for the principal strata, by fitting the following weighted multinomial logistic regression:

   - for i such that $z_i = 1$ and $d_i = 1$, create two new observations for the regression: one always-taker with weight $\pi_{a,i}^{(t)}$ and one complier with weight $\pi_{c,i}^{(t)}$;
   - for i such that $z_i = 0$ and $d_i = 0$, create two new observations: one complier with weight $\pi_{c,i}^{(t)}$ and one never-taker with weight $\pi_{n,i}^{(t)}$;
   - for i such that $z_i = 1$ and $d_i = 0$, create one never-taker with weight 1;
   - for i such that $z_i = 0$ and $d_i = 1$, create one always-taker with weight 1.

   (b) Similarly, obtain $\boldsymbol{\theta}_{\text{PO}}^{(t+1)}$, the updated estimates of the parameters in the model for the potential outcomes, by fitting weighted proportional odds models:

- for $u = a$, use all i such that $z_i = 1$ and $d_i = 1$ with weight $\pi_{a,i}^{(t)}$, and all i such that $z_i = 0$ and $d_i = 1$ with weight 1;
- for $u = c$, use all i such that $z_i = 1$ and $d_i = 1$ and all i such that $z_i = 0$ and $d_i = 0$ with weight $\pi_{c,i}^{(t)}$;
- for $u = n$, use all i such that $z_i = 1$ and $d_i = 0$ with weight 1, and all i such that $z_i = 0$ and $d_i = 0$ with weight $\pi_{n,i}^{(t)}$.

3.4.3 Confidence Intervals

To quantify the uncertainty associated with the aforementioned point estimators, we can use the bootstrap method in Horowitz and Manski (2000) to obtain the confidence intervals (CI) for the unadjusted and covariate adjusted bounds. For computational details of some other bootstrap methods, see Cheng and Small (2006) and Yang and Small (2015).

3.5 Simulation Studies

To demonstrate the estimators of our bounds and the inference drawn, especially the performance of the EM algorithm in Section 3.4.2, we conduct simulation studies under different model specifications. To save space, we focus only on the causal parameter τ_c.

We consider six simulation cases. Cases 1–3 are indexed by the parameter $\beta \in \{1, 1/2, 0\}$, and Cases 4–6 are indexed by the parameter $\eta \in \{1, 1/2, 0\}$. We postpone the interpretation of β and η until afterwards. For each case, let the pretreatment covariates $\boldsymbol{X} = (1, X_1, X_2)$, where $X_1 \sim N(0,1)$, and $X_2 \sim \text{Bern}(1/2)$. For fixed $\boldsymbol{X} = \boldsymbol{x}$, generate the principal stratification variable U from multiple logistic regression:

$$\pi_u(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\eta}_u^{\mathrm{T}}\boldsymbol{x})}{\sum_{u'} \exp(\boldsymbol{\eta}_{u'}^{\mathrm{T}}\boldsymbol{x})} \quad (u = a, c, n),$$

where $\boldsymbol{\eta}_c = \mathbf{0}$, $\boldsymbol{\eta}_a = (1/2, 1, 0)$ and $\boldsymbol{\eta}_n = (-1/2, 1, 0)$. Furthermore, we generate the potential outcomes from proportional odds models.

1. For always-takers, let $Y_i(1) = Y_i(0)$, and their marginal distributions be

$$\text{logit} \sum_{k \le j} a_{k+}(\boldsymbol{x}) = \text{logit} \sum_{l \le j} a_{+l}(\boldsymbol{x}) = \alpha_{a,j} - 2x_1,$$

   where $\alpha_{a,0} = -1/2$ and $\alpha_{a,1} = 1$.

2. For never-takers, let $Y_i(1) = Y_i(0)$, and their marginal distributions be

$$\text{logit} \sum_{k \le j} n_{k+}(\boldsymbol{x}) = \text{logit} \sum_{l \le j} n_{+l}(\boldsymbol{x}) = \alpha_{n,j},$$

   where $\alpha_{a,0} = -3/2$ and $\alpha_{a,1} = 0$.

3. For compliers, let $Y_i(1)$ and $Y_i(0)$ be independent, $\alpha_{c,0} = -1$, $\alpha_{c,1} = 1/2$, $\gamma_{c,0} = 1/2$ and $\gamma_{c,1} = 2$. For Cases 1–3, let the marginal distributions of the potential outcomes be

$$\text{logit} \sum_{k \le j} c_{k+}(\boldsymbol{x}) = \alpha_{c,j} - 2\beta x_1, \quad \text{logit} \sum_{l \le j} c_{+l}(\boldsymbol{x}) = \gamma_{c,j} + \beta x_1.$$

   For Cases 4–6, let the marginals distributions of the potential outcomes be

$$\text{logit} \sum_{k \le j} c_{k+}(\boldsymbol{x}) = \alpha_{c,j} - 2x_1 - \xi x_2, \quad \text{logit} \sum_{l \le j} c_{+l}(\boldsymbol{x}) = \gamma_{c,j} + x_1 + \xi x_2.$$

For the above six cases, their true values of τ_c, their unadjusted and adjusted bounds are in columns 2, 3–4 and 5–6 of the first three rows of Table 3.3, respectively. For Cases 1–3, as the association between the covariates and potential outcomes (quantified by β) decreases, the covariate adjusted bounds become closer to the unadjusted bounds. For Cases 4–6, the parameter ξ quantifies the association between the binary covariate X_2 and the potential outcomes of compliers.

We conduct inference of the bounds without the binary covariate X_2, which does not affect Cases 1–3 because X_2 is irrelevant in the data generating process, but does affect Cases 4–6. This is purposefully designed to examine the performance of our estimators under correct and incorrect model specifications. For each case, we choose the sample size to be 1000, and independently draw 1000 treatment assignments from a balanced completely randomized experiment. For each observed data set, we first estimate the bounds $\tau_{c,L}$ and $\tau_{c,U}$, and construct a confidence interval for $(\tau_{c,L}, \tau_{c,U})$ that covers the bounds themselves at least 95% times; we then estimate the bounds $\tau'_{c,L}$ and $\tau'_{c,U}$, and construct a confidence interval for $(\tau'_{c,L}, \tau'_{c,U})$ that covers the bounds themselves at least 95% times.

We report the simulation results in Table 3.3, in which columns 7–10 include the biases of the point estimators, and columns 11–14 include the average lengths and coverage rates of the intervals on the bounds $(\tau_{c,L}, \tau_{c,U})$ and $(\tau'_{c,L}, \tau'_{c,U})$. The results lead to several conclusions. First, the point estimators all have small biases. Second, when the pretreatment covariates are associated with the potential outcomes, the confidence intervals of the bounds $(\tau_{c,L}, \tau_{c,U})$ are longer than those of $(\tau'_{c,L}, \tau'_{c,U})$, on average. Third, the confidence intervals for the bounds $(\tau_{c,L}, \tau_{c,U})$ and $(\tau'_{c,L}, \tau'_{c,U})$ achieve reasonable coverage rates. Fourth, our proposed EM algorithm is robust to the missing of the binary covariate.

3.6 Application

We use the data set from Prevention of Suicide in Primary Care Elderly Collaborative Trial (PROSPECT; Bruce et al., 2004) to demonstrate the estimation of bounds of our new causal parameters and the subsequent inferences drawn. In the PROSPECT study, 82 of the 167 patients were randomly assigned to the treatment arm with specialist, and the rest to control with usual care. In the treatment group, 75 patients received treatment, among whom 23 developed major depression symptoms (labeled 0), 17 developed minor symptoms (labeled 1), 32 developed no symptoms (labeled 2); and 10 received control, among whom the numbers of patients with outcome 0, 1 and 2 are respectively 3, 2 and 5. In the control group everyone received control, and the numbers of patients with outcome 0, 1 and 2 are respectively 28, 19 and 38. For the original data set, see Baker (2011).

By design the monotonicity assumption holds, and Cheng (2009) verified the plausibility of exclusion restriction assumption. We apply our framework to this data set. For τ_c, its estimated lower and upper bounds are $\widehat{\tau}_{c,L} = 0.444$ and $\widehat{\tau}_{c,U} = 1.000$, the 95% confidence interval for the bounds $(\tau_{c,L}, \tau_{c,U})$ is $(0.284, 1.000)$, and $\widehat{\tau}_{C,I} = 0.683$. For η_c, its estimated lower and upper bounds are $\widehat{\eta}_{c,L} = 0.014$ and $\widehat{\eta}_{c,U} = 0.560$, the 95% confidence interval for the bounds $(\eta_{c,L}, \eta_{c,U})$ is $(0.000, 0.729)$, and $\widehat{\eta}_{C,I} = 0.328$. Under the assumption of independent potential outcomes, Cheng (2009) claimed that "the beneficial effect of intervention on the two multinomial depression outcomes is not significant," implying a zero treatment effect. Our results suggest that, depending on the unknown probabilistic structure of the potential outcomes, the treatment could either be worse or better than the control.

3.7 Discussions

By invoking the principal stratification framework to address noncompliance, for ordinal outcomes we have discussed various causal parameters and their sharp bounds by using only the marginal distributions of the potential outcomes for principal strata. Our theoretical results can be conveniently incorporated into different statistical inference frameworks and statistical models, including the case with pretreatment covariates.

We have discussed causal inference for ordinal outcomes with noncompliance to a binary treatment only, because of its wide applications. For instance, Cheng and Small (2006) studied the case with three-arm treatment and noncompliance, and Dasgupta et al. (2015) discussed causal inference of average treatment effects for continuous outcomes for factorial experiments. We will explore these directions in the future.

Table 3.1: *Illustration of Tighter Bounds by Principal Stratification*

Case	Z	D^{obs}	$\text{pr}(Y^{\text{obs}} \mid Z, D^{\text{obs}})$	$\text{pr}\{Y(Z)\}$	$\text{pr}\{Y(Z) \mid U = c\}$	τ_L	τ_U	τ_L^{PS}	τ_U^{PS}
A	0	0	$(1/2, 1/4, 1/4)$	$(11/40, 13/40, 2/5)$	$(1/4, 1/4, 1/2)$	39/80	1	33/40	1
		1	$(2/5, 2/5, 1/5)$						
	1	0	$(1/5, 2/5, 2/5)$	$(19/40, 23/80, 19/80)$	$(13/20, 7/40, 7/40)$				
		1	$(3/10, 3/10, 2/5)$						

Table 3.2: *Illustration of Tighter Bounds by Covariate Adjustment*

Case	$\boldsymbol{X}$	Z	D^{obs}	$\text{pr}\left(Y^{\text{obs}} \mid \boldsymbol{X}, Z, D^{\text{obs}}\right)$	$\text{pr}\{Y(Z) \mid \boldsymbol{X}, U = c\}$	$\tau_{c,L}$	$\tau_{c,U}$	$\tau'_{c,L}$	$\tau'_{c,U}$
A	0	0	0	$(1/2, 1/4, 1/4)$	$(4/5, 1/10, 1/10)$	1/2	1	3/5	9/10
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(1/5, 1/5, 3/5)$				
			1	$(3/10, 3/10, 2/5)$					
	1	0	0	$(1/5, 9/20, 7/20)$	$(1/5, 1/2, 3/10)$				
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(2/5, 1/5, 2/5)$				
			1	$(2/5, 3/10, 3/10)$					
B	0	0	0	$(1/5, 1/2, 3/10)$	$(1/5, 3/5, 1/5)$	3/10	4/5	2/5	4/5
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(1/5, 2/5, 2/5)$				
			1	$(3/10, 2/5, 3/10)$					
	1	0	0	$(1/5, 2/5, 2/5)$	$(1/5, 2/5, 2/5)$				
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(3/5, 1/5, 1/5)$				
			1	$(1/2, 3/10, 1/5)$					
C	0	0	0	$(2/5, 3/10, 3/10)$	$(3/5, 1/5, 1/5)$	1/2	1	1/2	9/10
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(1/5, 1/5, 3/5)$				
			1	$(3/10, 3/10, 2/5)$					
	1	0	0	$(3/10, 3/10, 2/5)$	$(2/5, 1/5, 2/5)$				
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(1/5, 3/5, 1/5)$				
			1	$(3/10, 1/2, 1/5)$					
D	0	0	0	$(3/10, 3/10, 2/5)$	$(2/5, 1/5, 2/5)$	3/10	7/10	3/10	7/10
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(3/5, 1/5, 1/5)$				
			1	$(1/2, 3/10, 1/5)$					
	1	0	0	$(1/5, 3/10, 1/2)$	$(1/5, 1/5, 3/5)$				
			1	$(2/5, 2/5, 1/5)$					
		1	0	$(1/5, 2/5, 2/5)$	$(1/5, 3/5, 1/5)$				
			1	$(3/10, 1/2, 1/5)$					

Table 3.3: *Results of Simulation Studies*

Case	τ_c	$\tau_{c,L}$	$\tau_{c,U}$	$\tau'_{c,L}$	$\tau'_{c,U}$	bias($\hat{\tau}_{c,L}$)	bias($\hat{\tau}_{c,U}$)	bias($\hat{\tau}'_{c,L}$)	bias($\hat{\tau}'_{c,U}$)	len. CI. $(\tau_{c,L}, \tau_{c,U})$	len. CI. $(\tau'_{c,L}, \tau'_{c,U})$	cov. CI. $(\tau_{c,L}, \tau_{c,U})$	cov. CI. $(\tau'_{c,L}, \tau'_{c,U})$
1	0.686	0.488	0.970	0.503	0.772	0.002	-0.028	0.008	-0.006	0.658	0.466	0.947	0.968
2	0.770	0.553	1.000	0.563	0.935	0.005	-0.005	0.004	-0.007	0.574	0.530	0.973	0.968
3	0.856	0.622	1.000	0.622	1.000	0.034	-0.000	0.021	-0.001	0.485	0.489	0.958	0.959
4	0.782	0.590	1.000	0.602	0.846	0.000	-0.002	0.005	0.012	0.528	0.436	0.976	0.960
5	0.738	0.542	1.000	0.556	0.817	0.002	-0.016	0.007	-0.003	0.588	0.447	0.966	0.965
6	0.686	0.488	0.970	0.503	0.772	0.002	-0.028	0.008	-0.006	0.658	0.466	0.947	0.968

Bibliography

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition*. Hoboken, New Jersey: John Wiley and Sons.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd Edition*. Hoboken, New Jersey: John Wiley and Sons.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Assoc.*, 91:444–455.

Bajorski, P. and Petkau, J. (1999). Nonparametric two-sample comparison of changes on ordinal responses. *J. Am. Statist. Assoc.*, 94:970–978.

Baker, S. G. (2011). Estimation and inference for the causal effect of receiving treatment on a multinomial outcome: An alternative approach. *Biometrics*, 67:319–323.

Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Stat. Med.*, 13:2269–2278.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Upper Saddle River, New Jersey: Prentice Hall.

Bradley, R. A., Katti, S. K., and Coons, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika*, 27:355–374.

Brillinger, D. R., Jones, L. V., and Tukey, J. W. (1978). Report of the statistical task force to the weather modification advisory board. *The Management of Weather Resources, II: The Role of Statistics on Weather Resources Management*. Government Printing Office, Washington D.C.

Bruce, M., Ten Have, T., Reynolds, C., Katz, I., Schulberg, H., Mulsant, B., Brown, G., McAvay, G., Pearson, J., and Alexopoulos, G. (2004). Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *J. Am. Med. Assoc.*, 291:1081–1091.

Cheng, J. (2009). Estimation and inference for the causal rffect of receiving treatment on a multinomial outcome. *Biometrics*, 65:96–103.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc. B*, 68:815–836.

Chernoff, H. (2004). Information, for testing the equality of two probabilities, from the margins of the 2× 2 table. *J. Stat. Plan. Inference*, 121:209–214.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20:37–46.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Statist. Soc. B*, 77:727–753.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38.

Ding, P. and Dasgupta, T. (2015). A potential tale of two by two tables from completely randomized experiments. *J. Am. Statist. Assoc.*, in press.

Fisher, R. A. (1935). *The Design of Experiments, 1st Edition*. Edinburgh, London: Oliver and Boyd.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21–29.

Grilli, L. and Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *J. Educ. Behav. Stat.*, 33:111–130.

Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X.-H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Statist. Assoc.*, 95:77–84.

Ju, C. and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc. B*, 72:129–142.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: John Wiley and Sons.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.*, 76:1071–1102.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks, 1st Edition*. San Francisco: Holden-Day.

Lindvall, T. (1992). *Lectures on the coupling method*. New York: Wiley.

Long, D. M. and Hudgens, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics*, 69:812–819.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Am. Statist. Assoc.*, 108:1120–1131.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essay des principles. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, Trans.). *Statist. Sci.*, 5:465–472.

Pitman, E. J. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29:322–335.

Plackett, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika*, 64:37–42.

Reichardt, C. S. and Gollob, H. F. (1999). Justifying the use and increasing the power of a t-test for a randomized experiment with a convenient sample. *Psychological Methods*, 4:117–128.

Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for pcp in high- versus low-dose azt treatment arms in an aids randomized trial. *J. Am. Statist. Assoc.*, 89:737–749.

Rosenbaum, P. R. (2001). Effects attributable to treatment: inference in experiments and observational studies within a discrete pivot. *Biometrika*, 88:219–231.

Rosenbaum, P. R. (2010). *Design of Observational Studies.* New York: Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66:688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Stat.*, 6:34–58.

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *J. Am. Statist. Assoc.*, 75:591–593.

Scharfstein, D. O., Manski, C. F., and Anthony, J. C. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes: application to the good behavior game trial. *Biometrics*, 60:154–164.

Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Stat. Med.*, 10:45–52.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Stat.*, 36:423–439.

Volfovsky, A., Airoldi, E. M., and Rubin, D. B. (2015). Causal inference for ordinal outcomes. *arXiv:1501.01234*.

Yang, F. and Small, D. S. (2015). Using post-quality of life measurement information in censoring by death problems. *J. R. Statist. Soc. B*, in press.

Appendix A

Supplementary Materials for Chapter 1

A.1 Construction of Maximizer in Theorem 1.1

To complete the proof of Theorem 1.1, we construct a lower triangular probability matrix with fixed row rums $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and column sums $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$, that attains the upper bound of κ in (1.3). We start with the last column and proceed backwards. At any point in the construction, we denote any element in the matrix which we have already filled by $\tilde{p}_{kl}$ and any we have not by p_{kl}, in order to distinguish them.

First, for the last column with index $J-1$, only the last entry needs to be filled, and we set it equal to the corresponding column sum, i.e., $\tilde{p}_{J-1,J-1} = p_{+,J-1}$. Next, for all $r = 1, \ldots, J-1$, given all elements in the last r columns are already filled, we consider the problem of filling in the elements of column with index $l = J-r-1$, as shown in Table A.1. At this point, the already filled elements in the matrix are $\tilde{p}_{kl}$, where $k = 0, \ldots, J-1$ and $l = J-r, \ldots, J-1$.

To fill the column with index $l = J-r-1$, note that all entries for $k < J-r-1$ will be equal to zero. We set the diagonal element with row index $k = l = J-r-1$ equal to the minimum of the corresponding row and column sums, i.e., $\tilde{p}_{J-r-1,J-r-1} = \min\left(p_{J-r-1,+}, p_{+,J-r-1}\right)$. Now the difference $p_{+,J-r-1} - \min(p_{J-r-1,+}, p_{+,J-r-1})$ needs to be distributed over the remaining entries of the column. Note that this difference is zero if $\min\left(p_{J-r-1,+}, p_{+,J-r-1}\right) = p_{+,J-r-1}$. Therefore, for all $k = J-r, \ldots, J-1$, we make the entry $p_{k,J-r-1}$ proportional to the

Table A.1: *Filling in the column with index $l = J - r - 1$ when the last r columns are already filled*

Row index (k)	Column index (l): 0	$\cdots$	$J-r-1$	$J-r$	$\cdots$	$J-1$	Row Sum p_{k+}
0	p_{00}	$\cdots$	0	0	$\cdots$	0	p_{0+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$J-r-1$	$p_{J-r-1,0}$	$\cdots$	$\min(p_{J-r-1,+}, p_{+,J-r-1})$	0	$\cdots$	0	$p_{J-r-1,+}$
$J-r$	$p_{J-r,0}$	$\cdots$	$p_{J-r,J-r-1} =?$	$\tilde{p}_{J-r,J-r}$	$\cdots$	0	$p_{J-r,+}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$J-1$	$p_{J-1,0}$	$\cdots$	$p_{J-1,J-r-1} =?$	$\tilde{p}_{J-1,J-r}$	$\cdots$	$\tilde{p}_{J-1,J-1} = p_{+,J-1}$	$p_{J-1,+}$
Column sum	p_{+0}	$\cdots$	$p_{+,J-r-1}$	$p_{+,J-r}$	$\cdots$	$p_{+,J-1}$	1

"remaining balance" $p_{+,J-r-1} - \min(p_{J-r-1,+}, p_{+,J-r-1})$, where we choose the proportionality constant as

$$\frac{\sum_{l=0}^{J-r-1} p_{kl}}{\sum_{k' \geq J-r} \sum_{l=0}^{J-r-1} p_{k'l}}, \tag{A.1}$$

that is, the ratio of the sum of empty entries in the row with label k and the sum of empty entries in all rows below the one labeled $J - r - 1$. Both the numerator and denominator of (A.1) can be expressed in terms of the given marginal probabilities and the already filled-in entries in the last r columns:

$$\begin{aligned} \sum_{l=0}^{J-r-1} p_{kl} &= p_{k+} - \sum_{l \geq J-r} \tilde{p}_{kl}, \\ \sum_{k' \geq J-r} \sum_{l=0}^{J-r-1} p_{k'l} &= \sum_{k' \geq J-r} \left(p_{k'+} - \sum_{l \geq J-r} \tilde{p}_{k'l} \right), \end{aligned} \tag{A.2}$$

and hence can be computed uniquely. The construction method, (A.1) or (A.2), eventually leads to the following iterative imputation equation for all $k = J - r, \ldots, J - 1:$

$$\tilde{p}_{k,J-r-1} = \frac{p_{k+} - \sum_{l \geq J-r} \tilde{p}_{kl}}{\sum_{k' \geq J-r} \left(p_{k'+} - \sum_{l \geq J-r} \tilde{p}_{k'l} \right)} \left\{ p_{+,J-r-1} - \min(p_{J-r-1,+}, p_{+,J-r-1}) \right\}. \tag{A.3}$$

Having constructed the matrix

$$\boldsymbol{P}_+ = \begin{pmatrix} \tilde{p}_{00} & 0 & \cdots & 0 \\ \tilde{p}_{10} & \tilde{p}_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{p}_{J-1,0} & \tilde{p}_{J-1,1} & \cdots & \tilde{p}_{J-1,J-1} \end{pmatrix},$$

we show that it indeed satisfies (i) $\tilde{p}_{kl} \geq 0$ for all $k, l = 0, \ldots, J-1$, (ii) the equality in condition (1.4), for which a sufficient condition is $\tilde{p}_{kk} = \min(p_{k+}, p_{+k})$ for all $k = 0, \ldots, J-1$, (iii) the vector of column sums is $\boldsymbol{p}_0$, i.e., $\sum_{k=0}^{J-1} \tilde{p}_{kl} = p_{+l}$ for all $l = 0, \ldots, J-1$, and (iv) the vector of row sums is $\boldsymbol{p}_1$, i.e., $\sum_{l=0}^{J-1} \tilde{p}_{kl} = p_{k+}$ for all $k = 0, \ldots, J-1$.

Among (i)–(iv) described above, (i)–(iii) follow directly by the construction of $\boldsymbol{P}_+$ described above. We need only to prove that $\sum_{l=0}^{J-1} \tilde{p}_{kl} = p_{k+}$ for all $k = 0, \ldots, J-1$. To prove this, first note that by Stochastic Dominance, we have $p_{0+} \leq p_{+0}$, implying that $\tilde{p}_{00} = p_{0+}$. Therefore the sum of the first row of $\boldsymbol{P}_+$ is p_{0+}. Now for all $k = 1, \ldots, J-1$, we have from (A.3) by substituting $r = J-1$ (that is, filling up the first column given the last $J-1$),

$$\begin{aligned} \tilde{p}_{k0} &= \frac{p_{k+} - \sum_{l=1}^{J-1} \tilde{p}_{kl}}{\sum_{k=1}^{J-1} \left(p_{k+} - \sum_{l=1}^{J-1} \tilde{p}_{kl} \right)} (p_{+0} - p_{0+}) \\ &= \frac{p_{k+} - \sum_{l=1}^{J-1} \tilde{p}_{kl}}{(1 - p_{0+}) - (1 - \tilde{p}_{00})} (p_{+0} - p_{0+}) \\ &= \frac{p_{k+} - \sum_{l=1}^{J-1} \tilde{p}_{kl}}{p_{+0} - p_{0+}} (p_{+0} - p_{0+}) \\ &= p_{k+} - \sum_{l=1}^{J-1} \tilde{p}_{kl}, \end{aligned}$$

which implies that $\sum_{l=0}^{J-1} \tilde{p}_{kl} = p_{k+}$. The construction is thus complete.

A.2 Proof of Lemma 1.1

Proof of Lemma 1.1(a). We prove by induction. For $n = 1$, let $\boldsymbol{P}_1 = (a_1)$ and Lemma 1.1(a) holds trivially. For $n \geq 2$, assume that Lemma 1.1(a) holds for $n-1$, and we prove it holds for n. The key idea is to first construct the first row and column of the $n \times n$ matrix $\boldsymbol{P}_n$, and use the assumption to construct a $(n-1) \times (n-1)$ matrix $\boldsymbol{P}_{n-1}$, which becomes the remaining part of $\boldsymbol{P}_n$. To construct the first row and column, we discuss the following two cases.

(1) If $a_1 = b_1$, we construct the first row and column of $\boldsymbol{P}$ by letting

$$p_{11} = a_1; \quad p_{k1} = p_{1l} = 0 \quad (k, l = 2, \ldots, n).$$

Because

$$\sum_{k'=2}^{n} a_{k'} = \sum_{l'=2}^{n} b_{l'},$$

we apply Lemma 1.1(a) to $(a_2, \ldots, a_n)$ and $(b_2, \ldots, b_n)$, and obtain a $(n-1) \times (n-1)$ matrix $\boldsymbol{P}_{n-1} = (p_{kl})_{2 \le k,l \le n}$ with nonnegative elements such that

$$\sum_{l'=2}^{n} p_{kl'} = a_k, \quad \sum_{k'=2}^{n} p_{k'l} = b_l, \quad p_{jj} = \min(a_j, b_j) \quad (k, l, j = 2, \ldots, n),$$

Therefore, we let

$$\boldsymbol{P}_n = \begin{pmatrix} p_{11} & \boldsymbol{0}^{\mathrm{T}} \\ 0 & \boldsymbol{P}_{n-1} \end{pmatrix}.$$

(2) If $a_1 \neq b_1$, by symmetry we only need to discuss the case where $a_1 > b_1$. In this case, we construct the first column of $\boldsymbol{P}$ by letting

$$p_{11} = b_1; \quad p_{k1} = 0 \quad (k = 2, \ldots, n).$$

To construct the first row of $\boldsymbol{P}$, note that

$$a_1 - b_1 = \sum_{j=2}^{n} (b_j - a_j) = \sum_{j=2}^{n} (b_j - a_j)_+ - \sum_{j=2}^{n} (b_j - a_j)_- ,$$

where $\{x\}_+ = |x| I_{(x \ge 0)}$ and $\{x\}_- = |x| I_{(x \le 0)}$ are respectively the positive and negative parts of x. By the above equality we have

$$\sum_{j=2}^{n} (b_j - a_j)_+ \ge a_1 - b_1 > 0.$$

Consequently, We let

$$p_{1l} = \left\{ (b_l - a_l) \times \frac{a_1 - b_1}{\sum_{j=2}^{n} (b_j - a_j)_+} \right\}_+ \quad (l = 2, \ldots, n). \tag{A.4}$$

By (A.4) we have

$$p_{1l} \leq (b_l - a_l)_+ \leq b_l \quad (l = 2, \ldots, n),$$

which implies that

$$\min(a_l, b_l) = \min(a_l, b_l - p_{1l}) \quad (l = 2, \ldots, n).$$

Therefore, we apply Lemma 1.1(a) to $(a_2, \ldots, a_n)$ and $(b_2 - p_{12}, \ldots, b_n - p_{1n})$, and obtain a $(n-1) \times (n-1)$ matrix $\boldsymbol{P}_{n-1} = (p_{kl})_{2 \leq k,l \leq n}$ with nonnegative elements such that

$$\sum_{l'=1}^{n} p_{kl'} = a_k, \quad \sum_{k'=1}^{n} p_{k'l} = b_l - p_{1l} \quad (k, l = 2, \ldots, n),$$

and

$$p_{jj} = \min(a_j, b_j - p_{1j}) = \min(a_j, b_j) \quad (j = 2, \ldots, n).$$

Therefore, we let $\boldsymbol{p} = (p_{12}, \ldots, p_{1n})^{\mathrm{T}}$ and

$$\boldsymbol{P}_n = \begin{pmatrix} p_{11} & \boldsymbol{p}^{\mathrm{T}} \\ \boldsymbol{0} & \boldsymbol{P}_{n-1} \end{pmatrix}.$$

Therefore, Lemma 1.1(a) holds for n. □

Proof of Lemma 1.1(b). We sequentially construct the $n \times n$ matrix $\boldsymbol{Q}$. First we let

$$q_{kl} = q_{k+} \times q_{+l} \quad (k, l = 1, \ldots, n),$$

and $\{i_1, \ldots, i_n\}$ be the permutation of $\{1, \ldots, n\}$ such that

$$q_{i_1,i_1} \leq \ldots \leq q_{i_n,i_n}. \tag{A.5}$$

Next, we propose a procedure to sequentially make the above diagonal elements zero, such that at any point all elements of $\boldsymbol{Q}$ are nonnegative, and the row and column sums of $\boldsymbol{Q}$ are

a_k's and b_l's. For all $j = 1, \dots, n-1$, we change the following 2×2 sub-matrix in $\boldsymbol{Q}$:

$$\begin{pmatrix} q_{i_j,i_j} & q_{i_j,i_{j+1}} \\ q_{i_{j+1},i_j} & q_{i_{j+1},i_{j+1}} \end{pmatrix},$$

by subtracting q_{i_j,i_j} from the two diagonal elements, and adding q_{i_j,i_j} to the two off-diagonal elements, resulting

$$\begin{pmatrix} 0 & q_{i_j,i_{j+1}} + q_{i_j,i_j} \\ q_{i_{j+1},i_j} + q_{i_j,i_j} & q_{i_{j+1},i_{j+1}} - q_{i_j,i_j} \end{pmatrix}.$$

By (A.5), all elements of $\boldsymbol{Q}$ remain nonnegative after the above operation. Furthermore, a diagonal element becomes zero. Therefore, eventually we have

$$q_{i_j,i_j} = 0 \quad (j = 1, \dots, n-1); \quad q_{i_n,i_n} = \sum_{j=1}^{n} (-1)^{n-j} a_{i_j} b_{i_j} \geq 0.$$

If $q_{i_n,i_n} = 0$ in the above, we are done with the construction. If $q_{i_n,i_n} > 0$, we need to adjust q_{i_n,i_n} to be zero, and preserve the row and column sums at the same time. The current probability matrix has

$$\sum_{k=1}^{n-1} \sum_{l=1}^{n-1} q_{i_k,i_l} = 1 - a_{i_n} - b_{i_n} + q_{i_n,i_n} \geq q_{i_n,i_n} > 0,$$

but our final probability matrix should have $q_{i_n,i_n} = 0$ and therefore

$$\sum_{k=1}^{n-1} \sum_{l=1}^{n-1} q_{i_k,i_l} = 1 - a_{i_n} - b_{i_n} \geq 0.$$

We multiply p_{kl} by the factor $(1 - a_{i_n} - b_{i_n})/(1 - a_{i_n} - b_{i_n} + q_{i_n,i_n}) \in [0, 1)$ for k and $l = 1, \dots, n-1$, and make the following adjustments to the remaining elements on the final row and column:

$$q_{i_k,i_n} = a_{i_k} - \sum_{l'=1}^{n-1} q_{i_k,i_{l'}} \quad (k = 1, \dots, n-1); \qquad q_{i_n,i_l} = b_{i_l} - \sum_{k'=1}^{n-1} q_{i_{k'},i_l} \quad (l = 1, \dots, n-1).$$

□

Appendix B

Supplementary Materials for Chapter 2

B.1 Proof of Lemma 2.1

Proof of Lemma 2.1(a). We prove by induction. When $n = 1$, we let $\boldsymbol{A}_1 = y_0 \geq 0$, and (2.5) holds because $y_0 \leq x_0$. When $n \geq 2$, suppose Lemma 2.1(a) holds for $n-1$. In particular, for any $(x_1, \ldots, x_{n-1})$ and $(y_1, \ldots, y_{n-1})$ such that $\sum_{r=s}^{n-1} x_r \geq \sum_{r=s}^{n-1} y_r$ for all $s = 1, \ldots, n-1$, there exists a lower triangular matrix $\boldsymbol{A}_{n-1} = (a_{kl})_{1 \leq k,l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=1}^{n-1} a_{kl'} \leq x_k, \quad \sum_{k'=1}^{n-1} a_{k'l} = y_l \quad (k, l = 1, \ldots, n-1). \tag{B.1}$$

To prove that Lemma 2.1(a) holds for n, we let

$$\boldsymbol{A}_n = \begin{pmatrix} a_{00} & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{a} & \boldsymbol{A}_{n-1} \end{pmatrix},$$

where a_{00} and $\boldsymbol{a} = (a_{10}, \ldots, a_{n-1,0})^{\mathrm{T}}$ are defined for two separate cases below.

(1) $y_0 < x_0$. We let $a_{00} = y_0$, and $a_{k0} = 0$ for all $k = 1, \ldots, n-1$. Clearly, $\boldsymbol{A}_n$ has nonnegative elements, and satisfies the row and column sum conditions in (2.5).

(2) $y_0 \geq x_0$. We let $a_{00} = x_0$, and

$$a_{k0} = (y_0 - a_{00}) \frac{x_k - \sum_{l'=1}^{n-1} a_{kl'}}{\sum_{k'=1}^{n-1} \left(x_{k'} - \sum_{l'=1}^{n-1} a_{k'l'} \right)} \geq 0 \quad (k = 1, \ldots, n-1). \tag{B.2}$$

This construction guarantees that the column sums of $\boldsymbol{A}_n$ are y_l's. We need to verify the conditions for the row sums required by (2.5). Because $\boldsymbol{A}_{n-1}$ satisfies (B.1), we have

$$\begin{aligned}\sum_{k'=1}^{n-1}\left(x_{k'}-\sum_{l'=1}^{n-1}a_{k'l'}\right) &= \sum_{k'=1}^{n-1}x_{k'}-\sum_{k'=1}^{n-1}\sum_{l'=1}^{n-1}a_{k'l'}=\sum_{k'=1}^{n-1}x_{k'}-\sum_{l'=1}^{n-1}\sum_{k'=1}^{n-1}a_{k'l'}\\ &= \sum_{k'=1}^{n-1}x_{k'}-\sum_{k'=1}^{n-1}y_{k'}\geq y_0-x_0=y_0-a_{00}>0. \qquad \text{(B.3)}\end{aligned}$$

Formulas (B.2) and (B.3) imply that $a_{k0}\leq x_k-\sum_{l'=1}^{n-1}a_{kl'}$ and therefore $\sum_{l'=0}^{n-1}a_{kl'}\leq x_k$ for $k=1,\ldots,n-1$.

Therefore Lemma 2.1(a) holds for n, and the proof is complete. □

Proof of Lemma 2.1(b). By applying Lemma 2.1(a) to $(y_0,\ldots,y_{n-1})$ and $(x_0,\ldots,x_{n-1})$, we obtain a lower triangular matrix $\widetilde{\boldsymbol{B}_n}=(\tilde{b}_{kl})_{0\leq k,l\leq n-1}$ with nonnegative elements such that

$$\sum_{k'=0}^{n-1}\tilde{b}_{k'l}=x_k,\quad \sum_{l'=0}^{n-1}\tilde{b}_{kl'}\leq y_k \quad (k,l=0,\ldots,n-1).$$

Let $\boldsymbol{B}_n=\widetilde{\boldsymbol{B}_n}^{\mathrm{T}}$, and the proof is complete. □

Proof of Lemma 2.1(c). By applying Lemma 2.1(a) to $(y_{n-1},\ldots,y_0)$ and $(x_{n-1},\ldots,x_0)$, we obtain a lower triangular matrix $\widetilde{\boldsymbol{C}_n}=(\tilde{c}_{kl})_{0\leq k,l\leq n-1}$ with nonnegative elements such that

$$\sum_{k'=0}^{n-1}\tilde{c}_{k'l}=x_{n-l-1},\quad \sum_{l'=0}^{n-1}\tilde{c}_{kl'}\leq y_{n-k-1} \quad (k,l=0,\ldots,n-1).$$

Let $\boldsymbol{C}_n=(\tilde{c}_{n-l-1,n-k-1})_{0\leq k,l\leq n-1}$, and the proof is complete. □

Proof of Lemma 2.1(d). By applying Lemma 2.1(c) to $(y_0,\ldots,y_{n-1})$ and $(x_0,\ldots,x_{n-1})$, we obtain a lower triangular matrix $\widetilde{\boldsymbol{D}_n}=(\tilde{d}_{kl})_{0\leq k,l\leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{k}\tilde{d}_{kl'}=y_k,\quad \sum_{k'=l}^{n-1}\tilde{d}_{k'l}\leq x_k \quad (k,l=0,\ldots,n-1).$$

Let $\boldsymbol{D}_n=\widetilde{\boldsymbol{D}_n}^{\mathrm{T}}$, and the proof is complete. □

Proof of Lemma 2.1(e). In addition to the proof of Lemma 2.1(a), we further need to show that if $\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, the row sums of the constructed matrix $\boldsymbol{A}_n$ are x_k's. In the induction of the proof of Lemma 2.1(a), if we have constructed matrix A_{n-1}, the case with $y_0 < x_0$ would not happen. We consider only the case with $y_0 \geq x_0$. Because the lower triangular matrix A_{n-1} has the column sums y_l's, and $\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, we have

$$\sum_{k'=1}^{n-1} \left(x_{k'} - \sum_{l'=1}^{n-1} a_{k'l'} \right) = \sum_{k'=1}^{n-1} x_{k'} - \sum_{k'=1}^{n-1} y_{k'} = y_0 - x_0 = y_0 - a_{00} > 0.$$

The above formula, coupled with the construction of the first column of $\boldsymbol{A}_n$ in (B.2), gives $a_{k0} = x_k - \sum_{l'=1}^{n-1} a_{kl'}$ and thus $\sum_{l'=0}^{n-1} a_{kl'} = x_k$ for all k. □

This proof of Lemma 2.1 can also serve as a constructive proof of the result in Strassen (1965), which was later utilized by Rosenbaum (2001).

B.2 Construction of Probability Matrices in Theorem 2.1

B.2.1 Probability Matrix for the Upper Bound of τ

To finish the proof of Theorem 2.1(a), we construct a probability matrix $\boldsymbol{P}$, with fixed marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$, to attain the bound in (2.11). To do this, we let

$$j_1 = \min \left\{ 0 \leq j' \leq J-1 : \Delta_{j'} = \min_{0 \leq j \leq J-1} \Delta_j \right\}$$

be the minimum index j that attains the minimum value of Δ_j's. To attain the upper bound (2.11), the equalities in (2.9) and (2.10) must hold, i.e.,

$$\sum_{k<l} \sum p_{kl} = \sum_{k<j_1} \sum_{l \geq j_1} p_{kl}, \quad \sum_{k \geq j_1} \sum_{l \geq j_1} p_{kl} = \sum_{k \geq j_1} \sum_{l=1}^{J-1} p_{kl}. \tag{B.4}$$

If $j_1 = 0$, $\min_{0 \leq j \leq J-1} \Delta_j = \Delta_0 = 0$, implying that $\Delta_j = \sum_{k=j}^{J-1} p_{k+} - \sum_{l=j}^{J-1} p_{+l} \geq 0$ for all j, i.e., the marginal probabilities satisfy the stochastic dominance assumption. According to Lemma 2.1(e), there exists a lower triangular probability matrix $\boldsymbol{P}$ with marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$. Correspondingly, $\tau = 1 + \Delta_0 = 1$.

If $j_1 > 0$, the constraints in (B.4) force some elements of the probability matrix to be zeros. To be more specific, the constraints in (B.4) imply that the probability matrix has the following block structure:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_{\text{tl}} & \boldsymbol{P}_{\text{tr}} \\ \boldsymbol{0} & \boldsymbol{P}_{\text{br}} \end{pmatrix}, \tag{B.5}$$

where the $j_1 \times j_1$ sub-matrix $\boldsymbol{P}_{\text{tl}}$ on top left and the $(J - j_1) \times (J - j_1)$ sub-matrix $\boldsymbol{P}_{\text{br}}$ on bottom right are both lower triangular, and the $j_1 \times (J - j_1)$ sub-matrix $\boldsymbol{P}_{\text{tr}}$ on top right has no restrictions.

Because $\Delta_{j_1} \leq \Delta_j$ for all $j = 0, 1, \ldots, J-1$, we have

$$\sum_{k=j}^{j_1-1} p_{k+} \geq \sum_{l=j}^{j_1-1} p_{+l} \quad (j = 0, \ldots, j_1 - 1); \qquad \sum_{k=j_1}^{j} p_{k+} \leq \sum_{l=j_1}^{j} p_{+l} \quad (j = j_1, \ldots, J-1).$$

Given the above two sets of constraints on the marginal probabilities, we construct the probability matrix $\boldsymbol{P}$ in three steps.

(1) We apply Lemma 2.1(a) to $(p_{0+}, \ldots, p_{j_1-1,+})$ and $(p_{+0}, \ldots, p_{+,j_1-1})$, and obtain a lower triangular matrix $\boldsymbol{P}_{\text{tl}} = (p_{kl})_{0 \leq k,l \leq j_1-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{j_1-1} p_{kl'} \leq p_{k+}, \quad \sum_{k'=0}^{j_1-1} p_{k'l} = p_{+l} \quad (k, l = 0, \ldots, j_1 - 1).$$

(2) We apply Lemma 2.1(c) to $(p_{j_1+}, \ldots, p_{J-1,+})$ and $(p_{+j_1}, \ldots, p_{+,J-1})$, and obtain a lower triangular matrix $\boldsymbol{P}_{\text{br}} = (p_{kl})_{j_1 \leq k,l \leq J-1}$ with nonnegative elements such that

$$\sum_{l'=j_1}^{J-1} p_{kl'} = p_{k+}, \quad \sum_{k'=j_1}^{J-1} p_{k'l} \leq p_{+l} \quad (k, l = j_1, \ldots, J-1).$$

(3) We construct $\boldsymbol{P}_{\text{tr}} = (p_{kl})_{0 \leq k \leq j_1-1, j_1 \leq l \leq J-1}$ by letting

$$p_{kl} = \left(p_{k+} - \sum_{l'=0}^{j_1-1} p_{kl'}\right)\left(p_{+l} - \sum_{k'=j_1}^{J-1} p_{k'l}\right) \geq 0 \quad (k = 0, \ldots, j_1 - 1;\ l = j_1, \ldots, J-1).$$

The constructed probability matrix $\boldsymbol{P}$ has marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\text{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\text{T}}$. What is more, by (B.5) the τ of $\boldsymbol{P}$ is the sum of all the elements in

$\boldsymbol{P}_{\text{tl}}$ and P_{br}, which we construct in the above (1) and (2). Therefore, we have

$$\tau = \sum_{l'=0}^{j_1-1} p_{+l'} + \sum_{k'=j_1}^{J-1} p_{k'+} = 1 + \Delta_{j_1},$$

which implies that the probability matrix $\boldsymbol{P}$ attains the bound (2.11).

B.2.2 Probability Matrix for the Lower Bound of τ

To finish the proof of Theorem 2.1(b), we construct a probability matrix $\boldsymbol{P}$, with fixed marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\text{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\text{T}}$, to attain the bound in (2.14). To do this, we let

$$j_2 = \min \left\{ j' : p_{+j'} + \Delta_{j'} = \max_{0 \le j \le J-1} (p_{+j} + \Delta_j) \right\}$$

be the minimum index j that attains the maximum value of $(p_{+j} + \Delta_j)$'s. To attain the lower bound (2.14), the equalities in (2.12) and (2.13) must hold, i.e.,

$$\sum_{k \ge l} \sum p_{kl} = \sum_{k \ge j_2} \sum_{l \le j_2} p_{kl}, \quad \sum_{k \ge j_2} \sum_{l > j_2} p_{kl} = \sum_{k=0}^{J-1} \sum_{l > j_2} p_{kl}. \tag{B.6}$$

If $j_2 = 0$, from (B.6) we know that the elements in the lower triangular part but not in the first column of the probability matrix $\boldsymbol{P}$ are all zeros, i.e.,

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{p} & \boldsymbol{P}_{\text{tr}} \\ p_{J-1,0} & 0^{\text{T}} \end{pmatrix}, \tag{B.7}$$

where $\boldsymbol{p} = (p_{0,0}, \ldots, p_{J-2,0})^{\text{T}}$, and the $(J-1) \times (J-1)$ sub-matrix P_{tr} on top right is upper triangular. Because $p_{+0} + \Delta_0 \ge p_{+j} + \Delta_j$ for all j, we have

$$\sum_{k=0}^{j} p_{k+} \ge \sum_{l=0}^{j} p_{+,l+1} \quad (j = 0, \ldots, J-2).$$

Applying Lemma 2.1(d) to $(p_{0+}, \ldots, p_{J-2,+})$ and $(p_{+1}, \ldots, p_{+,J-1})$, we obtain an upper triangular matrix $\boldsymbol{P}_{\text{tr}} = (p_{kl})_{0 \le k \le J-2, 1 \le l \le J-1}$ with nonnegative elements such that

$$\sum_{l'=1}^{J-1} p_{kl'} \le p_{k+}, \quad \sum_{k'=0}^{J-2} p_{k'l} = p_{+l} \quad (k = 0, \ldots, J-2;\ l = 1, \ldots, J-1).$$

To complete the construction, let $p_{J-1,0} = p_{J-1,+}$, and

$$p_{k0} = p_{k+} - \sum_{l'=1}^{J-1} p_{kl'} \geq 0 \quad (k = 0, \ldots, J-2).$$

The constructed probability matrix $\boldsymbol{P}$ has marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\mathrm{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\mathrm{T}}$. Moreover, by (B.7) the τ of $\boldsymbol{P}$ is the sum of all the elements in the first column. Therefore $\tau = p_{+0} = p_{+0} + \Delta_0$, which implies that $\boldsymbol{P}$ attains the bound (2.14).

If $j_2 = J-1$, the proof is similar to the above case with $j_2 = 0$.

If $0 < j_2 < J-1$, because the first equality in (B.6) is equivalent to

$$\sum_{k<j_2} \sum_{l \leq k} p_{kl} + \sum_{k \geq j_2} \sum_{l \leq k} p_{kl} = \sum_{k \geq j_2} \sum_{l \leq j_2} p_{kl},$$

the probability matrix $\boldsymbol{P}$ must satisfy the following constraints:

(C1) For all $k = 0, \ldots, j_2 - 1$, $p_{kl} = 0$ for all $l = 0, \ldots, k$.

(C2) For all $k = j_2 + 1, \ldots, J-1$, $p_{kl} = 0$ for all $l = j_2 + 1, \ldots, k$.

Similarly, because the second equality in (B.6) is equivalent to

$$\sum_{k \geq j_2} \sum_{l > j_2} p_{kl} = \sum_{k \geq j_2} \sum_{l > j_2} p_{kl} + \sum_{k < j_2} \sum_{l > j_2} p_{kl},$$

the probability matrix $\boldsymbol{P}$ must further satisfy the following constraint:

(C3) $p_{kl} = 0$, for all $k = 0, \ldots, j_2 - 1$ and $l = j_2 + 1, \ldots, J-1$.

The constraints in (C1), (C2) and (C3) imply that $\boldsymbol{P}$ must have the following block structure:

$$\boldsymbol{P} = \begin{pmatrix} (\boldsymbol{0}, \boldsymbol{P}_{\mathrm{tl}}) & \boldsymbol{0} \\ \boldsymbol{P}_{\mathrm{bl}} & \begin{pmatrix} \boldsymbol{P}_{\mathrm{br}} \\ \boldsymbol{0}^{\mathrm{T}} \end{pmatrix} \end{pmatrix} \tag{B.8}$$

where the $j_2 \times j_2$ sub-matrix $\boldsymbol{P}_{\mathrm{tl}}$ and the $(J - j_1 - 1) \times (J - j_1 - 1)$ sub-matrix $\boldsymbol{P}_{\mathrm{br}}$ are both upper triangular, and the $(J - j_2) \times (j_2 + 1)$ sub-matrix $\boldsymbol{P}_{\mathrm{bl}}$ on bottom left has no restrictions.

Because $p_{+j_2} + \Delta_{j_2} \geq p_{+j} + \Delta_j$ for all j, we have

$$\sum_{k=j}^{j_2-1} p_{k+} \leq \sum_{l=j}^{j_2-1} p_{+,l+1} \quad (j = 0, \ldots, j_2 - 1); \qquad \sum_{k=j_2}^{s} p_{k+} \geq \sum_{l=j_2}^{s} p_{+,l+1} \quad (j = j_2, \ldots, J-2).$$

Given the above two sets of constraints for the marginal probabilities, we construct the probability matrix $\boldsymbol{P}$ in three steps.

(1) We apply Lemma 2.1(b) to $(p_{0+}, \ldots, p_{j_2-1,+})$ and $(p_{+1}, \ldots, p_{+,j_2})$, and obtain an upper triangular matrix $\boldsymbol{P}_{\text{tl}} = (p_{kl})_{0 \leq k \leq j_2-1, 1 \leq l \leq j_2}$ with nonnegative elements such that

$$\sum_{l'=1}^{j_2} p_{kl'} = p_{k+}, \quad \sum_{k'=0}^{j_2-1} p_{k'l} \leq p_{+l} \quad (k = 0, \ldots, j_2 - 1;\ l = 1, \ldots, j_2).$$

(2) We apply Lemma 2.1(d) to $(p_{j_2+}, \ldots, p_{J-2,+})$ and $(p_{+,j_2+1}, \ldots, p_{+,J-1})$, and obtain an upper triangular matrix $\boldsymbol{P}_{\text{br}} = (p_{kl})_{j_2 \leq k \leq J-2, j_2+1 \leq l \leq J-1}$ with nonnegative elements such that

$$\sum_{l'=j_2+1}^{J-1} p_{kl'} \leq p_{k+}, \quad \sum_{k'=j_2}^{J-2} p_{k'l} = p_{+l} \quad (k = j_2, \ldots, J-2;\ l = j_2 + 1, \ldots, J-1).$$

(3) We construct $\boldsymbol{P}_{\text{bl}} = (p_{kl})_{j_2 \leq k \leq J-1, 0 \leq l \leq j_2}$ by letting

$$p_{kl} = \left(p_{k+} - \sum_{l'=j_2+1}^{J-1} p_{kl'} \right) \left(p_{+l} - \sum_{k'=0}^{j_2-1} p_{k'l} \right) \geq 0 \quad (k = j_2, \ldots, J-1;\ l = 0, \ldots, j_2).$$

The constructed probability matrix $\boldsymbol{P}$ has marginal probabilities $\boldsymbol{p}_1 = (p_{0+}, \ldots, p_{J-1,+})^{\text{T}}$ and $\boldsymbol{p}_0 = (p_{+0}, \ldots, p_{+,J-1})^{\text{T}}$. Moreover, by (B.8) the corresponding τ is the sum of all the elements in $\boldsymbol{P}_{\text{bl}}$, which we construct in the above (3). Therefore, we have

$$\tau = 1 - \sum_{k'=0}^{j_2-1} p_{k'+} - \sum_{l'=j_2+1}^{J-1} p_{+l'} = \sum_{k'=j_2}^{J-1} p_{k'+} - \sum_{l'=j_2+1}^{J-1} p_{+l'} = p_{+j_2} + \Delta_{j_2},$$

which implies that $\boldsymbol{P}$ attains the lower bound (2.14).

Appendix C

Supplementary Materials for Chapter 3

Proof of Theorem 3.1. The theorem follows by applying Lemma 3.1 on the marginal probabilities of potential outcomes for compliers, i.e., $\{c_{0+}, \ldots, c_{J-1,+}\}$ and $\{c_{+0}, \ldots, c_{+,J-1}\}$. □

Proof of Lemma 3.2. First we consider τ. Under the monotonicity assumption, by law of total probability we have $\tau = \pi_c \tau_c + \pi_a \tau_a + \pi_n \tau_n$. Under the exclusion restriction assumption, we have $\tau_a = 1$ and $\tau_n = 1$, yielding $\tau = \pi_c \tau_c + 1 - \pi_c$, which implies that $\tau_c = \tau/\pi_c - (1-\pi_c)/\pi_c$. Analogously, we have $\eta = \pi_c \eta_c + (1-\pi_c) \times 0 = \pi_c \eta_c$, which implies that $\eta_c = \eta/\pi_c$. □

Proof of Corollary 3.1. First, the new bounds $\tau^{\mathrm{N}}_{c,L}$, $\tau^{\mathrm{N}}_{c,U}$, $\eta^{\mathrm{N}}_{c,L}$ and $\eta^{\mathrm{N}}_{c,U}$ follow directly from Lemmas 1 and 2. Next, under the monotonicity assumption we have

$$\Delta_j = \pi_a \Delta_{a,j} + \pi_c \Delta_{c,j} + \pi_n \Delta_{n,j} \quad (j = 0, \ldots, J-1).$$

Further under the exclusion restriction assumption, we have $\Delta_{a,j} = 0$ and $\Delta_{n,j} = 0$, and therefore the above equalities reduce to $\Delta_j = \pi_c \Delta_{c,j}$. We use this fact to derive the relationships between the old and new sets of bounds. We first discuss the bounds of τ_c, and then the bounds of η_c.

1. For the upper bounds of τ_c we have

$$\begin{aligned}
\tau_{c,U} &= 1+\min_j \Delta_{c,j} \\
&= \left\{\pi_c+\min_j\left(\pi_c\Delta_{c,j}\right)\right\}/\pi_c \\
&= \left\{1+\min_j\left(\pi_c\Delta_{c,j}\right)\right\}/\pi_c-\left(1-\pi_c\right)/\pi_c \\
&= \left(1+\min_j \Delta_j\right)/\pi_c-\left(1-\pi_c\right)/\pi_c \\
&= \tau^{\mathrm{N}}_{c,U}.
\end{aligned}$$

2. The inequality

$$a_{+j}\pi_a+n_{+j}\pi_n\leq \pi_a+\pi_n$$

implies that

$$p_{+j}-1+\pi_c\leq c_{+j}\pi_c.$$

Consequently, for the lower bounds of τ_c we have

$$\begin{aligned}
\tau_{c,L} &= \max_j\left(c_{+j}+\Delta_{c,j}\right) \\
&\geq \max_j\left\{\left(p_{+j}+\pi_c\Delta_{c,j}-1+\pi_c\right)/\pi_c\right\} \\
&= \left\{\max_j\left(p_{+j}+\Delta_j-1+\pi_c\right)\right\}/\pi_c \\
&= \left\{\max_j\left(p_{+j}+\Delta_j\right)\right\}/\pi_c-\left(1-\pi_c\right)/\pi_c \\
&= \tau^{\mathrm{N}}_{c,L}.
\end{aligned}$$

The two bound are equal if and only if $\left(1-a_{+j}\right)\pi_a=0$ and $\left(1-n_{+j}\right)\pi_n=0$.

3. For the lower bound of η_c we have

$$\eta_{c,L}=\max_j \Delta_{c,j}=\left\{\max_j\left(\pi_c\Delta_{c,j}\right)\right\}/\pi_c=\max_j \Delta_j/\pi_c=\eta^{\mathrm{N}}_{c,L},$$

4. The inequality

$$a_{j+}\pi_a+n_{j+}\pi_n\leq \pi_a+\pi_n.$$

implies that

$$p_{j+} + \pi_c - 1 \leq c_{j+}\pi_c.$$

Consequently, for the upper bound of η_c we have

$$\begin{aligned}
\eta_{c,U} &= 1 + \min_j \left(\Delta_{c,j} - c_{j+}\right) \\
&= \left\{\min_j \left(\pi_c + \pi_c\Delta_{c,j} - c_{j+}\pi_c\right)\right\} / \pi_c \\
&= \left\{1 + \min_j \left(\pi_c - 1 + \Delta_j - c_{j+}\pi_c\right)\right\} / \pi_c \\
&\leq \left\{1 + \min_j \left(\Delta_j - p_{j+}\right)\right\} / \pi_c \\
&= \eta^{\mathrm{N}}_{c,U}.
\end{aligned}$$

The two bounds are equal if and only if $(1 - a_{j+})\,\pi_a = 0$ and $(1 - n_{j+})\,\pi_n = 0$.

□

Proof of Corollary 3.2. The corollary follows directly from Corollary 3.1. □

Proof of Theorem 3.2. With independent potential outcomes for the compleirs, the probability matrix $\boldsymbol{P}_c$ has elements $c_{kl} = c_{k+}c_{+l}$. We obtain $\tau_{c,I}$ and $\eta_{c,I}$ by their definitions, which are between their lower and upper bounds. □

Proof of Theorem 3.3. The proof follows the same logic as Lee (2009). Because any value of τ_c within the covariate adjusted bounds $[\tau'_{c,L}, \tau'_{c,U}]$ must be compatible with the conditional distributions of $\{Y(1), \boldsymbol{X}\}$ and $\{Y(0), \boldsymbol{X}\}$ given that $U = c$, it must also be compatible with the conditional distributions of $Y(1)$ and $Y(0)$ given $U = c$, by discarding $\boldsymbol{X}$. Therefore, any value of τ_c within the adjusted bounds $[\tau'_{c,L}, \tau'_{c,U}]$ must also be within the unadjusted bounds $[\tau_{c,L}, \tau_{c,U}]$. Consequently, the adjusted bounds are tighter, i.e., $[\tau'_{c,L}, \tau'_{c,U}] \subset [\tau_{c,L}, \tau_{c,U}]$. □