Finite Population Inference for Causal Parameters

Wen Wei Loh

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Thomas S. Richardson, Chair

Mathias Drton

Michael G. Hudgens

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Finite Population Inference for Causal Parameters

Wen Wei Loh

Chair of the Supervisory Committee:
Prof. Thomas S. Richardson
Statistics

Randomized experiments are often employed to determine whether a treatment X has a causal effect on an outcome Y. Under the Neyman-Rubin causal model with binary X and Y, each patient is characterized by two binary potential outcomes, leading to four possible response types. In a finite population, the set of individuals of each response type is regarded as fixed over hypothetical rerandomizations, so that individuals are sampled without replacement. The resulting observed-data likelihood, which we term the Neyman-Rubin-Copas (NRC) likelihood, is a convolution of multivariate hypergeometric probabilities. We will derive results for the NRC likelihood that may be used to facilitate calculation of the generalized likelihood ratio (GLR) in more complicated finite population settings. A key finding is that the maximum likelihood under the 'Neyman' null (where the population average causal effect is zero) is always attained by the population in which the 'Fisher' null holds (where the individual causal effect is zero).

Next, we consider the setting where treatment X is no longer randomized, but there is an instrumental variable Z that is randomized. For example, patients in a randomized controlled trial may choose not to adhere to their randomly assigned treatment Z, possibly due to side-effects. In such randomized experiments with noncompliance, scientific interest is often in testing whether the treatment exposure X has an effect on the final outcome Y, among the subset of 'Compliers' who take the treatment only if assigned to do so and would not if assigned not to do so. We propose a finite population significance test of the 'Fisher' null hypothesis among the principal stratum of

'Compliers', using the GLR test statistic under an extended Neyman-Rubin-Copas likelihood that accounts for the noncompliance. New methods that improve the computational efficiency when evaluating the exact p-values are described.

We then extend the randomization-based significance tests using the GLR to construct an exact confidence interval for the Complier Average Causal Effect (CACE). The procedure is illustrated with a small toy example. Finally, we propose a GLR test statistic for a significance test of the 'Fisher' null under the noncompliance setting where we allow for a direct effect of Z on Y.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation and thanks to Thomas Richardson, Mathias Drton, Elena Erosheva, Michael Hudgens, Marina Meila, Adrian Raftery, Ellen Chan Reynolds, Ken Rice, Rekha Thomas and Jon Wakefield.

DEDICATION

In the order of my meeting them, this dissertation is dedicated to:

Loh Si Lan Nee Foo

Loh Kee Wan

Loh Wenbo

Ren Dongning

Chapter 1

LIKELIHOOD ANALYSIS FOR THE FINITE POPULATION NEYMAN-RUBIN BINARY CAUSAL MODEL

1.1 Introduction

Under the Neyman-Rubin causal model with binary treatment and outcome, each patient is characterized by two binary potential outcomes, leading to four response types. In a finite population, individuals are sampled without replacement over hypothetical rerandomizations from a fixed set of individuals characterized by the number of each response type. Under complete randomization where the size of each treatment group remains fixed, the observed data consists of a two-by-two table with fixed row sums. The resulting observed-data likelihood is a convolution of multivariate hypergeometric probabilities.

In this chapter we present two finite population results relating to this likelihood. First restricting attention to perfectly balanced designs with equal treatment group sizes, we show that, depending on the observed data, the maximum likelihood is attained by a population in which either one, two or three types are present, but never only by a population with all four response types. We conjecture that this property also holds for unbalanced designs, and confirm this by brute-force calculations for all populations up to size 200. Next we show that for any observed two-by-two dataset, the maximum likelihood under the 'Neyman' null (where the population average causal effect is zero) is always attained by the population in which the 'Fisher' null holds (where the individual causal effect is zero). These results facilitate the construction of exact significance tests using the generalized likelihood ratio (GLR) in more complex finite population settings.

Consider a randomized study with N individuals where n are randomly chosen to undergo treatment $X = 1$. The remaining $N - n$ individuals are allocated to a different treatment $X = 0$.

Among the n individuals chosen for the $X = 1$ group, n_{11} were observed to have the outcome $Y = 1$, whereas n_{10} had the same $Y = 1$ outcome among those in the $X = 0$ group. The total number of individuals with the observed $Y = 1$ outcome is also denoted $s \equiv n_{1+}$ for brevity. The usual summary of the observed dataset $\boldsymbol{n}$ is in the form of a two-by-two table such as Table 1.1.

n_{yx}	$Y=1$	$Y=0$	Row
$X=1$	n_{11}	n_{01}	$n \equiv n_{+1}$
$X=0$	n_{10}	n_{00}	$N-n$
Column	$s \equiv n_{1+}$	n_{0+}	N

Table 1.1: Observed two-by-two dataset $\boldsymbol{n}$ where each entry is n_{yx}.

	$r_Y=A$	$r_Y=B$	$r_Y=C$	$r_Y=D$	Row
$X=1$	m_{A1}	m_{B1}	m_{C1}	m_{D1}	n
$X=0$	m_{A0}	m_{B0}	m_{C0}	m_{D0}	$N-n$
Column	m_{A+}	m_{B+}	m_{C+}	m_{D+}	N

Table 1.2: Complete table $\boldsymbol{m}$ where m_{ij} is the number of individuals with $r_Y = i$ in group $X = j$.

Such randomized experiments are often employed in order to determine whether the treatment X has a causal effect on the outcome Y. This problem may be formulated in terms of the potential outcome framework, also referred to as the Neyman-Rubin causal framework [Spława-Neyman et al., 1990, Rubin, 1974]. Neyman introduced the potential outcome model in the context of a fixed population of agricultural plots, where any given plot has, for each treatment (variety), a fixed but unknown potential yield. Plots are then sampled without replacement from the fixed population and assigned to a treatment group. Since each plot is assigned to only one treatment, exactly one of the potential yields is observed: this is the 'fundamental problem of causal inference'.

Copas [1973] subsequently analyzed the likelihood under such a randomization-based distribution in the context of a binary treatment and outcome on a set of individuals. Rubin [1974] discusses estimating the average causal effect using the set of all possible randomizations (allocations) in a *completely randomized* experiment where the number of subjects in each treatment group is fixed.

For $x = 1, 0$, denote $Y(x)$ as the outcome that an individual would have if, possibly counter to fact, assigned to $X = x$. The potential outcomes $Y(x)$ are then linked to the observed outcomes

via the causal consistency axiom (see for example [Pearl, 2010]):

$$Y = XY(1) + (1-X)Y(0). \tag{1.1}$$

Each individual may be characterized in terms of the pair of values for their potential outcomes under each treatment compared $\{Y(x) : x = 1, 0\}$. There are thus four possible response types r_Y that the population may be divided into: (A) those who would always have an observed outcome of $Y = 1$ regardless of the treatment they were given, $Y(x) \equiv 1$; (B) those who would have an observed outcome of $Y = 1$ only if given $X = 1$, $Y(x) \equiv x$; (C) those who would have an observed outcome of $Y = 1$ only if given $X = 0$, $Y(x) \equiv 1 - x$; (D) those who would never have an observed outcome of $Y = 1$, $Y(x) \equiv 0$. If $Y = 1$ denotes a good or desirable outcome and $X = 1$ an active treatment, the response types (A, B, C, D) are typically termed (Always Recover, Helped, Hurt, Never Recover) respectively.

The number of individuals of each response type in each treatment group may be summarized with a two-by-four table such as Table 1.2, which we henceforth refer to as a *complete* table and denote as $\boldsymbol{m} \equiv (m_{A1}, m_{A0}, m_{B1}, \ldots, m_{D0})$. Since for any individual we only get to observe exactly one of the potential outcomes $Y(x)$, the exact cell counts in $\boldsymbol{m}$ are not in general directly observable from $\boldsymbol{n}$. The set of all possible $\boldsymbol{m}$ compatible with the observed quantities $\boldsymbol{n}$ is the subset of the eight-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^8$ that satisfies the linear constraints of the observed quantities $\boldsymbol{n}$:

$$\mathcal{N}(\boldsymbol{n}) = \left\{ \boldsymbol{m} : m_{ij} \geq 0, i = A, B, C, D, j = 1, 0; \begin{array}{ll} m_{A1} + m_{B1} = n_{11}, & m_{C1} + m_{D1} = n_{01}, \\ m_{A0} + m_{C0} = n_{10}, & m_{B0} + m_{D0} = n_{00} \end{array} \right\}. \tag{1.2}$$

This is the same set of linear constraints in [Copas, 1973, Equation 6].

1.2 Likelihood Under a Fixed Population

A key feature of the potential outcome framework under the finite population approach is that the set of individuals and the values of their potential outcomes are regarded as *fixed over hypothetical rerandomizations*. Differences between results over hypothetical reallocations arise *only* due to

different assignments of this fixed set of individuals to either $X=1$ or $X=0$. A finite population is uniquely described by the column totals $\boldsymbol{t}$ in Table 1.2, which we denote as

$$\boldsymbol{t} \equiv (t_A, t_B, t_C, t_D) \equiv (m_{A+}, m_{B+}, m_{C+}, m_{D+}). \tag{1.3}$$

The appropriate sampling scheme for randomization-based inference is thus one where individuals are sampled without replacement [Spława-Neyman et al., 1990]. Randomization of treatment implies that

$$\Pr\left(X=1|\{Y(1), Y(0)\}\right) = \Pr\left(X=1\right). \tag{1.4}$$

Under complete randomization, the row totals n and $N-n$, respectively, the size of the $X=1$ and $X=0$ groups, are also fixed. Hence the randomization distribution for the complete table $\boldsymbol{m}$ is the multiple hypergeometric distribution (see for example [Lehmann, 2006, Appendix 7B]):

$$\Pr\left(\boldsymbol{m}|\boldsymbol{t}, n\right) = \binom{N}{n, N-n}^{-1} \prod_{i\in\{A,B,C,D\}} \binom{t_i}{m_{i1}, m_{i0}}; \tag{1.5}$$

where we choose to write the binomial coefficient as:

$$\binom{a+c}{a, c} = \frac{(a+c)!}{a!c!}.$$

In general, there may be multiple complete tables $\boldsymbol{m} \in \mathcal{N}(\boldsymbol{n})$ that are compatible with the observed data $\boldsymbol{n}$ and have the same column totals $\boldsymbol{t}$. For a given population $\boldsymbol{t}$, define the subset of $\mathcal{N}(\boldsymbol{n})$ that satisfies both the linear constraints of the observed dataset $\boldsymbol{n}$, and the population-specific linear constraints determined by $\boldsymbol{t}$, as:

$$\mathcal{M}(\boldsymbol{t}; \boldsymbol{n}) = \left\{ \boldsymbol{m} \in \mathcal{N}(\boldsymbol{n}) : \begin{array}{r} m_{B1} + m_{C0} = s - t_A \\ m_{B1} + m_{B0} = t_B \\ m_{C1} + m_{C0} = t_C \end{array} \right\} \subseteq \mathcal{N}(\boldsymbol{n}). \tag{1.6}$$

For a complete table $\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$, each element in $\boldsymbol{m}$ is just a linear function of the observed data $\boldsymbol{n}$, the column totals $\boldsymbol{t}$, and one of the cell counts m_{ij}. For example, the entries in Table 1.2 may be written in terms of $\boldsymbol{n}, \boldsymbol{t}$ and m_{B1} as in Table 1.3. Hence the complete tables in $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ may be indexed by the one-dimensional set of possible values for m_{B1}:

$$\left\{m_{B1}^{\min}, m_{B1}^{\min}+1, \ldots, m_{B1}^{\max}\right\}; \tag{1.7}$$
$$m_{B1}^{\min} = \max\left\{s-(t_A+t_C), n_{11}-t_A, t_B-n_{00}, 0\right\},$$
$$m_{B1}^{\max} = \min\left\{n_{01}-(t_A+t_C)+s, n_{11}, t_B, s-t_A\right\}.$$

	$r_Y=A$	$r_Y=B$	$r_Y=C$	$r_Y=D$	Row
$X=1$	$n_{11}-m_{B1}$	m_{B1}	$(t_A+t_C)-(s-m_{B1})$	$n_{01}-(t_A+t_C)+(s-m_{B1})$	n
$X=0$	$t_A-(n_{11}-m_{B1})$	t_B-m_{B1}	$s-t_A-m_{B1}$	$n_{00}-t_B+m_{B1}$	$N-n$
Column	t_A	t_B	t_C	$N-(t_A+t_B+t_C)$	N

Table 1.3: Complete table $\boldsymbol{m}$ of the four response types in Table 1.2 expressed in terms of the observed dataset $\boldsymbol{n}$, population $\boldsymbol{t}$ and a single cell count m_{B1}.

The likelihood given the observed dataset $\boldsymbol{n}$ of a particular population $\boldsymbol{t}$ is thus the total probability of all complete tables $\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})$ [Copas, 1973]:

$$\Pr(\boldsymbol{n}|\boldsymbol{t},n) = \sum_{\boldsymbol{m}\,\in\,\mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \Pr(\boldsymbol{m}|\boldsymbol{t},n) = \binom{N}{n, N-n}^{-1} h(\boldsymbol{t}),$$
$$\text{where } h(\boldsymbol{t}) = \sum_{\boldsymbol{m}\,\in\,\mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \left\{ \prod_{i\in\{A,B,C,D\}} \binom{t_i}{m_{i1}, m_{i0}} \right\}$$
$$= \sum_{m_{B1}\,=\,m_{B1}^{\min}}^{m_{B1}^{\max}} \left\{ \prod_{i\in\{A,B,C,D\}} \binom{t_i}{m_{i1}(\boldsymbol{t},\boldsymbol{n},m_{B1}), m_{i0}(\boldsymbol{t},\boldsymbol{n},m_{B1})} \right\}, \tag{1.8}$$

and the entries m_{ij} in the complete table $\boldsymbol{m}$ are functions of $\boldsymbol{t}$, $\boldsymbol{n}$ and m_{B1} as in Table 1.3. We shall henceforth refer to the likelihood function $h(\boldsymbol{t})$ as the *Neyman-Rubin-Copas likelihood*.

Note that if we knew the cell counts in the complete table $\boldsymbol{m}$, we could obtain the exact values of the parameters $\boldsymbol{t}$ simply by finding the column sums. In contrast, neither $\boldsymbol{m}$ nor $\boldsymbol{t}$ are determined from $\boldsymbol{n}$. Let $T(\boldsymbol{n})$ be the parameter space of possible values for $\boldsymbol{t}$ given the observed dataset $\boldsymbol{n}$.

Since $t_D \equiv N - (t_A + t_B + t_C)$ in a finite population of size N, it is sufficient to specify values for (t_A, t_B, t_C). $T(\boldsymbol{n})$ is then the intersection of the three-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^3$, and the convex polyhedron described with the following inequalities:

$$0 \leq t_A \leq n_{11} + n_{10}, \qquad 0 \leq t_B \leq n_{11} + n_{00}, \qquad 0 \leq t_C \leq n_{01} + n_{10},$$
$$n_{11} \leq t_A + t_B \leq N - n_{01}, \quad n_{10} \leq t_A + t_C \leq N - n_{00}, \quad n_{11} + n_{10} \leq t_A + t_B + t_C \leq N. \tag{1.9}$$

The inequalities were obtained using `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015].

1.3 Global Maximum Likelihood

Our motivation to study the maximum of the likelihood function, or equivalently, $h(\boldsymbol{t})$ stems from the calculation of the *generalized likelihood ratio* (GLR) test statistic; see for example [Perlman and Wu, 1999]. For a null hypothesis H_0, denote the subset of $T(\boldsymbol{n})$ containing the population column totals $\boldsymbol{t}$ compatible with H_0 as $T^0(\boldsymbol{n})$. The GLR is defined as:

$$\lambda(\boldsymbol{n}) \equiv \frac{\max\limits_{\boldsymbol{t} \in T^0(\boldsymbol{n})} \Pr(\boldsymbol{n}|\boldsymbol{t}, n)}{\max\limits_{\boldsymbol{t} \in T(\boldsymbol{n})} \Pr(\boldsymbol{n}|\boldsymbol{t}, n)} = \frac{\max\limits_{\boldsymbol{t} \in T^0(\boldsymbol{n})} h(\boldsymbol{t})}{\max\limits_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t})}. \tag{1.10}$$

In subsequent chapters, we seek to extend the GLR to more complicated finite population settings with nuisance parameters, where finding an appropriate test statistic is difficult.

Denote the global maximum likelihood given an observed dataset $\boldsymbol{n}$ as:

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}). \tag{1.11}$$

As we will see, the parameters $\boldsymbol{t}$ that we wish to maximize over are not point-identified asymptotically. Naïvely, finding $\hat{h}$ requires calculating all possible likelihood values $h(\boldsymbol{t})$ for all $\boldsymbol{t} \in T(\boldsymbol{n})$. In general, $h(\boldsymbol{t})$ is the convolution of multivariate hypergeometric probabilities for $m_{B1}^{\max} - m_{B1}^{\min} + 1$ different complete tables. But if at least one of the column totals in $\boldsymbol{t}$ is zero, then from Equation (1.7), we have that $m_{B1}^{\min} = m_{B1}^{\max}$. $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ then consists of a single complete table, so that calculating the likelihood function does not require summing over multiple complete tables:

$$\min_i t_i = 0 \quad \Rightarrow \quad h(\boldsymbol{t}) = \prod_{i \in \{A,B,C,D\}} \binom{t_i}{m_{i1}, m_{i0}}.$$

Thus if we know that the global maximum likelihood $\hat{h}$ is attained by at least one population with three or fewer types, equivalently that $\hat{h}$ is never achieved by just populations that have all four types, then it would be easier to find $\hat{h}$. In this case we need only calculate a single multivariate hypergeometric probability, and not sums over sets of probabilities, to evaluate a candidate $\boldsymbol{t}$ while searching for $\hat{h}$.

1.3.1 Bounds for the population totals for individuals with non-zero individual causal effects

An alternative parametrization to (t_A, t_B, t_C) given a finite population of size N is (μ_1, μ_0, β), where:

$$\mu_1 \equiv (t_A + t_B)/N,\ \mu_0 \equiv (t_A + t_C)/N,\ \beta \equiv (t_B + t_C)/N. \tag{1.12}$$

(In [Copas, 1973], μ_x is m_x for $x = 0, 1$.) The (marginal) probability of observing $Y = 1$ if everyone in the population had been assigned to $X = x$ is just $\mu_x \equiv \Pr(Y(x) = 1)$. Note that the 'Fisher' null of zero individual causal effects corresponds to $(\mu_1 = \mu_0, \beta = 0)$, whereas the 'Neyman' null of equal average potential outcomes corresponds to just $(\mu_1 = \mu_0)$.

Given an observed table $\boldsymbol{n}$, it follows from the description of the parameter space $T(\boldsymbol{n})$ described in (1.9) that the set of possible values for (μ_1, μ_0) is the intersection of the discrete lattice $\{0/N, 1/N, \ldots, N/N\}^2$ and the subset of the unit square defined by the following inequalities:

$$\frac{n_{1x}}{N} \leq \mu_x \leq 1 - \frac{n_{0x}}{N};\ x = 0, 1. \tag{1.13}$$

For a fixed value of $\boldsymbol{\mu} \equiv (\mu_1, \mu_0)$ that satisfies (1.13), β is variation-dependent on $\boldsymbol{\mu}$ and is determined by the Fréchet bounds:

$$\beta^-(\boldsymbol{\mu}) \equiv |\mu_1 - \mu_0| \leq \beta \leq \min(\mu_1 + \mu_0, 2 - (\mu_1 + \mu_0)) \equiv \beta^+(\boldsymbol{\mu}). \tag{1.14}$$

The set of possible values for β is thus the intersection of the discrete set $\{0/N, 1/N, \ldots, N/N\}$ and the interval $[\beta^-(\boldsymbol{\mu}), \beta^+(\boldsymbol{\mu})]$ as described in Equation (1.14). Lemma 1 below describes the column totals $\boldsymbol{t}$ corresponding to the endpoints of the bounds.

Lemma 1. *Denote the population characterized by the proportions of the four response types for a given value of* $(\boldsymbol{\mu}, \beta)$ *as:*

$$\boldsymbol{\pi}(\boldsymbol{\mu},\beta) \equiv \left\{ \pi_i, i=A,B,C,D : \begin{array}{ll} \pi_A \equiv (\mu_1+\mu_0-\beta)/2, & \pi_B \equiv (\mu_1-\mu_0+\beta)/2, \\ \pi_C \equiv (\mu_0-\mu_1+\beta)/2, & \pi_D \equiv 1-(\mu_1+\mu_0+\beta)/2 \end{array} \right\}. \quad (1.15)$$

Let the populations corresponding to the bounds $\beta^-(\boldsymbol{\mu})$ *and* $\beta^+(\boldsymbol{\mu})$ *as described in Equation* (1.14) *for a given value of* $\boldsymbol{\mu}$ *be* $\boldsymbol{\pi}^-(\boldsymbol{\mu}) \equiv \boldsymbol{\pi}(\boldsymbol{\mu}, \beta^-(\boldsymbol{\mu}))$ *and* $\boldsymbol{\pi}^+(\boldsymbol{\mu}) \equiv \boldsymbol{\pi}(\boldsymbol{\mu}, \beta^+(\boldsymbol{\mu}))$ *respectively. There are then at most three response types in each of these populations i.e.*

$$\min_i \pi_i^-(\boldsymbol{\mu}) = \min_i \pi_i^+(\boldsymbol{\mu}) = 0.$$

Proof. First consider the population $\boldsymbol{\pi}^-(\boldsymbol{\mu})$ where $\beta = |\mu_1 - \mu_0|$:

$$\boldsymbol{\pi}^-(\boldsymbol{\mu}) = \begin{cases} (\mu_1, 0, \mu_0 - \mu_1, 1 - \mu_0) & , \mu_1 \leq \mu_0; \\ (\mu_0, \mu_1 - \mu_0, 0, 1 - \mu_1) & , \mu_1 > \mu_0. \end{cases} \quad (1.16)$$

Next consider the population $\boldsymbol{\pi}^+(\boldsymbol{\mu})$ where $\beta = \min(\mu_1 + \mu_0, 2 - (\mu_1 + \mu_0))$:

$$\boldsymbol{\pi}^+(\boldsymbol{\mu}) = \begin{cases} (0, \mu_1, \mu_0, 1 - (\mu_1 + \mu_0)) & , \mu_1 + \mu_0 \leq 1; \\ (\mu_1 + \mu_0 - 1, 1 - \mu_0, 1 - \mu_1, 0) & , \mu_1 + \mu_0 > 1. \end{cases} \quad (1.17)$$

There is thus at least one response type absent in each population i.e.

$$\min_i \pi_i^-(\boldsymbol{\mu}) = \min_i \pi_i^+(\boldsymbol{\mu}) = 0.$$

Furthermore, if $\mu_1 = \mu_0$ so that $\beta^-(\boldsymbol{\mu}) = 0$, then $\boldsymbol{\pi}^-(\boldsymbol{\mu}) = (\mu_1, 0, 0, 1 - \mu_1)$. Or if $\mu_1 + \mu_0 = 1$ so that $\beta^+(\boldsymbol{\mu}) = 1$, then $\boldsymbol{\pi}^+(\boldsymbol{\mu}) = (0, \mu_1, \mu_0, 0)$. □

Note that the bounds in Equation (1.14) may also be obtained from Equation (1.15) together with the requirement that $\pi_i \geq 0$.

1.3.2 Asymptotic results

In order to present asymptotic results in the finite population context, we consider the setting where the population characterized by some true value of $\boldsymbol{t}$ is a member of a hypothetical infinite sequence of populations $\boldsymbol{t}^{(k)} \equiv \{kt_i, i = A, B, C, D\}$ where k is a positive integer. Each element in $\boldsymbol{t}^{(k)}$ is k times that in $\boldsymbol{t}$, so that the finite population size is $N^{(k)} = kN$. All the populations $\boldsymbol{t}^{(k)}$ have the same true value of the parameters $(\boldsymbol{\mu}^{(k)}, \beta^{(k)}) = (\boldsymbol{\mu}, \beta)$ as defined in Eq. (1.12), as well as the same fixed proportions of the four response types $\boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}$ as defined in Eq. (1.15).

Given an observed table $\boldsymbol{n}^{(k)}$ that is compatible with $\boldsymbol{t}^{(k)}$ and has $n^{(k)} = kn$ individuals assigned to the $X = 1$ group, denote the observed proportions of individuals with $Y = 1$ in each $X = x$ group as $p_x^{(k)} \equiv n_{1x}^{(k)} / \sum_{y=0}^{1} n_{yx}^{(k)}$. Then as $k \to \infty$, the observed proportions $\widehat{\boldsymbol{\mu}}^{(k)} \equiv (p_1^{(k)}, p_0^{(k)})$ converge in probability to the true value of $\boldsymbol{\mu}$:

$$\lim_{k\to\infty} \widehat{\mu}_x^{(k)} = \mu_x, \ x = 0, 1. \tag{1.18}$$

In contrast, β is partially identified with support determined by $\widehat{\boldsymbol{\mu}}^{(k)}$ as described in Eq. (1.14). As $k \to \infty$, the true value of β is contained within the bounds determined by $\boldsymbol{\mu}$ with probability 1:

$$\lim_{k\to\infty} \Pr\left(\beta^-(\widehat{\boldsymbol{\mu}}^{(k)}) \le \beta \le \beta^+(\widehat{\boldsymbol{\mu}}^{(k)})\right) = \Pr\left(\beta^-(\boldsymbol{\mu}) \le \beta \le \beta^+(\boldsymbol{\mu})\right) = 1. \tag{1.19}$$

Proposition 1. *As $k \to \infty$, the population column totals $\hat{\boldsymbol{t}}^{(k)}$ that attain the global maximum likelihood contain at most three response types i.e.*

$$\min_i \hat{t}_i^{(k)} = 0.$$

Proof. From [Copas, 1973], as $k \to \infty$, the observed proportions $\widehat{\boldsymbol{\mu}}^{(k)}$ have a bivariate normal distribution with mean $\boldsymbol{\mu}$ and correlation strictly increasing with β. The likelihood is then maximized at $\widehat{\boldsymbol{\mu}}^{(k)}$ and $\widehat{\beta}^{(k)}$ being either of the bounds $\beta^-(\widehat{\boldsymbol{\mu}}^{(k)})$ or $\beta^+(\widehat{\boldsymbol{\mu}}^{(k)})$.

It then follows from Lemma 1 that the corresponding population proportions $\widehat{\boldsymbol{\pi}}^{(k)} \equiv \boldsymbol{\pi}(\widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\beta}^{(k)})$ will have at most three response types i.e.

$$\min_i \widehat{\pi}_i^{(k)} = 0.$$

As $k \to \infty$, the column totals $\hat{\boldsymbol{t}}^{(k)}$ that attain the global maximum likelihood will thus have at most three response types as well i.e.

$$\Pr\left(\min_i \hat{t}_i^{(k)} \equiv \min_i \left\{N^{(k)} \cdot \widehat{\pi}_i^{(k)}\right\} = 0\right) \to 1.$$

□

1.4 Finite Population Examples

For a given observed table $\boldsymbol{n}$, let $\widehat{\boldsymbol{\mu}} \equiv (p_1, p_0)$ be the observed proportions, and the corresponding bounds for β as described in Eq. (1.14) be $\widehat{\beta}^- \equiv \beta^-(\widehat{\boldsymbol{\mu}})$ and $\widehat{\beta}^+ \equiv \beta^+(\widehat{\boldsymbol{\mu}})$. From Lemma 1, denote the respective populations as $\widehat{\boldsymbol{\pi}}^- \equiv \boldsymbol{\pi}(\widehat{\boldsymbol{\mu}}, \widehat{\beta}^-)$ and $\widehat{\boldsymbol{\pi}}^+ \equiv \boldsymbol{\pi}(\widehat{\boldsymbol{\mu}}, \widehat{\beta}^+)$, so that the column totals are just $\hat{\boldsymbol{t}}^- \equiv N \cdot \widehat{\boldsymbol{\pi}}^-$ and $\hat{\boldsymbol{t}}^+ \equiv N \cdot \widehat{\boldsymbol{\pi}}^+$. If both $\hat{\boldsymbol{t}}^- \in T(\boldsymbol{n})$ and $\hat{\boldsymbol{t}}^+ \in T(\boldsymbol{n})$ are column totals compatible with the observed $\boldsymbol{n}$, let $h^* \equiv \max\left\{h(\hat{\boldsymbol{t}}^-), h(\hat{\boldsymbol{t}}^+)\right\}$. Following the asymptotic result in Proposition 1, one may then use h^* to approximate the global maximum likelihood $\hat{h}$. However, there may be a substantial difference between h^* and $\hat{h}$ in finite populations, as the following examples show. Consider the toy datasets:

$$\text{I}: \quad n_{11} = 3, n_{01} = 2, n_{10} = 2, n_{00} = 3; \qquad \text{III}: \quad n_{11} = 4, n_{01} = 16, n_{10} = 9, n_{00} = 11;$$
$$\text{II}: \quad n_{11} = 2, n_{01} = 3, n_{10} = 1, n_{00} = 4; \qquad \text{IV}: \quad n_{11} = 5, n_{01} = 15, n_{10} = 9, n_{00} = 11.$$

Datasets I and II are a tenth of the observed quantities in the numerical examples from [Copas, 1973]. The profile likelihood as a function of β, where we fix a particular value of $\beta = b$, and maximize over all populations $\boldsymbol{t} \in T(\boldsymbol{n})$ subject to $(t_B + t_C)/N = b$, takes the following form:

$$\hat{h}_\beta(b) = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) \,:\, (t_B + t_C)/N = b\}} h(\boldsymbol{t}). \tag{1.20}$$

The global maximum likelihood solution for β is then:

$$\widehat{\beta} = \arg\max_b \; \hat{h}_\beta(b).$$

A plot of $\hat{h}_\beta(b)$ as a function of b for each dataset is shown in Figure 1.1. The respective values of $\widehat{\beta}^-$ and $\widehat{\beta}^+$ are indicated by the blue vertical lines. For example in dataset IV, $\widehat{\boldsymbol{\mu}} = (0.25, 0.45)$,

so that $\widehat{\beta}^- = 0.2$ and $\widehat{\beta}^+ = 0.7$. To contrast against a larger population, the profile likelihood for a larger version of each dataset (ten times the respective observed quantities) is drawn as a red solid line. The larger versions of datasets I and II are thus the same numerical examples from [Copas, 1973]: for the larger version of dataset I, $\widehat{\beta} = \widehat{\beta}^+ = 1$, whereas for the larger version of dataset II, $\widehat{\beta} = \widehat{\beta}^- = 0.2$. For the larger version of each dataset, the profile likelihood is indeed bimodal with the (local) modes at $\widehat{\beta}^-$ and $\widehat{\beta}^+$.

For the smaller observed datasets, the profile likelihood (empty circles) have local modes not only at $\widehat{\beta}^-$ and $\widehat{\beta}^+$, but at $\beta = 0$ as well. In fact for datasets II and IV, the global maximum solutions are $\widehat{\beta} = 0$! This suggests that in small populations, $\widehat{\beta}$ may not lie within the range $\left[\widehat{\beta}^-, \widehat{\beta}^+\right]$: there is information on β based on other values of $\boldsymbol{\mu}$ besides $\widehat{\boldsymbol{\mu}}$ that informs where the global maximum likelihood is attained.

Note the GLR for testing the 'Fisher' null hypothesis that $\beta = 0$ is $\lambda(\boldsymbol{n}) = \hat{h}_\beta(0)/\hat{h}_\beta(\widehat{\beta})$ (which Copas [1973] mentions is also appropriate for comparing the hypotheses $\beta = 0$ and $\beta = \widehat{\beta}$). If we were to approximate the global maximum likelihood in the denominator with h^*, we may obtain a value of the GLR that is greater than 1; for example in dataset II, $\hat{h}_\beta(0)/h^* = 1.3$!

1.5 Finite Population Properties of the Global Maximum Likelihood

We now turn our attention to the global maximum likelihood $\hat{h}$ as defined in Equation (1.11). First, under a perfectly balanced design where $n = N - n$, Theorem 1 states that $\hat{h}$ is attained by at least one population with three or fewer types. The proof is provided in Appendix A.

Theorem 1. *For an observed table $\boldsymbol{n}$, denote the global maximum likelihood as:*

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}).$$

If $\boldsymbol{n}$ is perfectly balanced so that $n = N - n$, then there exists a population $\hat{\boldsymbol{t}}$ where $h(\hat{\boldsymbol{t}}) = \hat{h}$ and

$$\min_i \hat{t}_i = 0.$$

Next, we use brute-force computations to find the exact maximum likelihood $\hat{h}$ for all possible two-by-two observed datasets $\boldsymbol{n}$ where $N \leq 200$. Our findings confirm the following proposition.

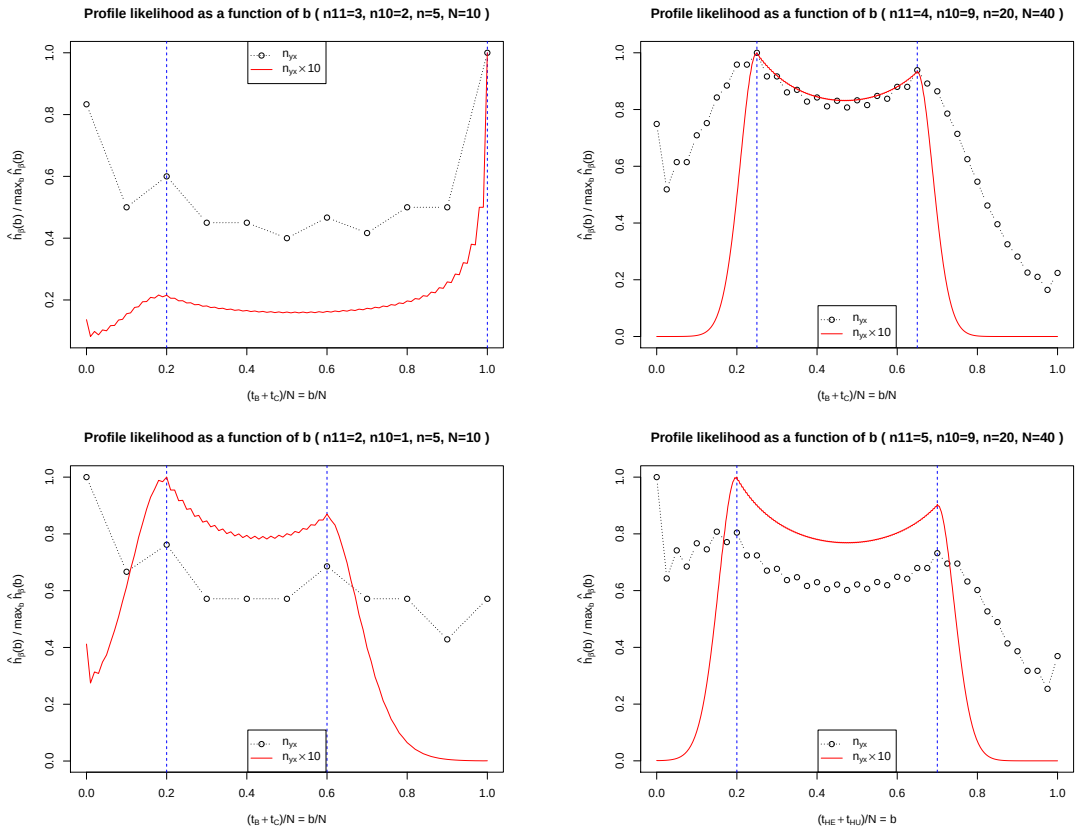

Figure 1.1: Plots of the profile likelihood $\hat{h}_\beta(b)$ against b for datasets I,II,III and IV respectively; the values on the vertical axis have been rescaled as $\hat{h}_\beta(b)/\max_b \hat{h}_\beta(b)$, which is also the GLR for comparing the hypothesis that $\beta = b$ against the hypothesis that $\beta = \widehat{\beta}$; the red solid lines are for larger versions of (ten times) the observed quantities in each dataset; the blue broken lines are $\widehat{\beta}^-$ and $\widehat{\beta}^+$, the bounds for β under $\widehat{\boldsymbol{\mu}} \equiv (p_1, p_0)$ as defined in Eq. (1.14).

Proposition 2. *For an observed table* $\boldsymbol{n}$*, denote the global maximum likelihood as:*

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}).$$

If $N \leq 200$*, where* $N = \sum_{x=0}^{1} \sum_{y=0}^{1} n_{yx}$*, then*

$$\max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) \,:\, h(\boldsymbol{t}) = \hat{h}\}} \mathbb{1}\Big\{\min_i t_i = 0\Big\} = 1, \tag{1.21}$$

where $\mathbb{1}\{E\}$ returns a value of 1 if the condition E is met, and 0 otherwise. In other words, there is at least one population where $h(\hat{\boldsymbol{t}}) = \hat{h}$, and $\hat{\boldsymbol{t}}$ has at most three types.

We have hence shown for observed datasets with sample size $N \leq 200$ (Proposition 2), for perfectly balanced datasets where $n = N - n$ (Theorem 1), and as $k \to \infty$ with $n^{(k)}/N^{(k)}$ fixed (Proposition 1), that there exists some population $\hat{\boldsymbol{t}}$ with at most three types and achieves the maximum likelihood $h(\hat{\boldsymbol{t}}) = \hat{h}$. This provides compelling evidence that the result holds in general, which leads us to the following conjecture.

Conjecture 1 (Never Only Four)**.** *For an observed table $\boldsymbol{n}$, denote the global maximum likelihood as:*

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}).$$

Then there is at least one population $\hat{\boldsymbol{t}}$ where $h(\hat{\boldsymbol{t}}) = \hat{h}$, and $\min_i \hat{t}_i = 0$. The global maximum likelihood $\hat{h}$ may then be found with the multivariate hypergeometric probability of a single complete table $\boldsymbol{m}$:

$$\hat{h} = \prod_{i \in \{A,B,C,D\}} \binom{\hat{t}_i}{m_{i1}, m_{i0}},$$

where the entries m_{ij} are implicitly functions of $\boldsymbol{t}, \boldsymbol{n}$ and one of the cell counts, for example m_{B1} as in Table 1.3.

1.6 Restricted Maximum Likelihood

In the previous section, we conjecture that the maximum likelihood $\hat{h}$ is always attained by at least one population with at most three types. Now we examine whether a similar result holds for the maximum likelihood under restrictions on the parameter space.

Define the Average Causal Effect (ACE) of X on Y as

$$\text{ACE} \equiv \frac{1}{N} \sum_{j=1}^{N} Y_j(1) - Y_j(0), \tag{1.22}$$

where the subscript j indexes the potential outcome $Y(x)$ for the j-th individual.

The difference in the proportions of individuals of types B and C in the population is $\delta \equiv (t_B - t_C)/N \equiv \mu_1 - \mu_0 =$ ACE (Copas [1973] denotes δ as α). The profile likelihood as a function of δ, where we fix a particular value of $\delta = c$, and maximize over all populations $\boldsymbol{t} \in T(\boldsymbol{n})$ subject to $(t_B - t_C)/N = c$, takes the following form:

$$\hat{h}_\delta(c) = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) \,:\, (t_B - t_C)/N = c\}} h(\boldsymbol{t}); \tag{1.23}$$

the global maximum likelihood solution is just:

$$\widehat{\delta} = \arg\max_c \ \hat{h}_\delta(c).$$

We found that the profile likelihood $\hat{h}_\delta(c)$ can be attained only by populations with all four types. For example, in the larger version of dataset I, the profile likelihood $\hat{h}_\delta(0.18) = 0.0411$ is attained by two populations $\boldsymbol{t} = \{(1, 58, 40, 1), (2, 57, 39, 2)\}$. The largest likelihood for a population with three types or fewer and $\delta = 0.18$ was 0.0397 at $\boldsymbol{t} = \{(0, 58, 40, 2), (2, 58, 40, 0)\}$. A possible explanation is that for a fixed value of $\delta \equiv \mu_1 - \mu_0 = c$, the restrictions for μ_1 in (1.9) now apply to $\mu_0 + \delta$, so that the range of possible values for μ_0 are variation-dependent on δ as well:

$$\max(n_{10}, n_{11} - \delta)/N \leq \mu_0 \leq \min(N - n_{01} - \delta, N - n_{00})/N. \tag{1.24}$$

Hence, the most likely value of μ_0 in the profile likelihood $\hat{h}_\delta(c)$ may differ from the maximum likelihood solution in $\hat{h}$.

We return to the four toy datasets analyzed in Figure 1.1. A plot of $\hat{h}_\delta(c)$ as a function of c for each dataset is shown in Figure 1.2. For the larger versions of each dataset, the profile likelihood (red solid line) with respect to δ is clearly unimodal, with the maximum located at $\widehat{\delta} = \widehat{\text{ACE}} = p_1 - p_0$, as indicated by the blue vertical line. However, at smaller population sizes, the profile likelihood (empty circles) is actually bimodal, with local maxima at $\delta = \hat{\delta}$ and $\delta = 0$. For datasets II and IV, the global maximum likelihood $\hat{h}$ is attained at $\delta = 0$, since $\beta \equiv (t_B + t_C)/N = 0$ implies $\delta \equiv (t_B - t_C)/N = 0$.

Conversely, $\delta = 0$ does not necessarily imply that $\beta = 0$, since there are in general multiple populations where $t_B = t_C$ and $t_B + t_C > 0$. In other words, if we assume only a zero population

ACE (the 'Neyman' null), there may still be an equal but non-zero number of individuals of types B and C in the population. By maximizing over the additional degree of freedom for the parameter β, one would expect the maximum likelihood under the 'Neyman' null $\hat{h}_\delta(0)$ to be larger than the likelihood under the stricter 'Fisher' null $\hat{h}_\beta(0)$. However, we observe that for datasets I and III respectively, the profile likelihood $\hat{h}_\delta(0)$ in Figure 1.2 has the same value as $\hat{h}_\beta(0)$ in Figure 1.1! This is not a mere coincidence, as we show in the following section.

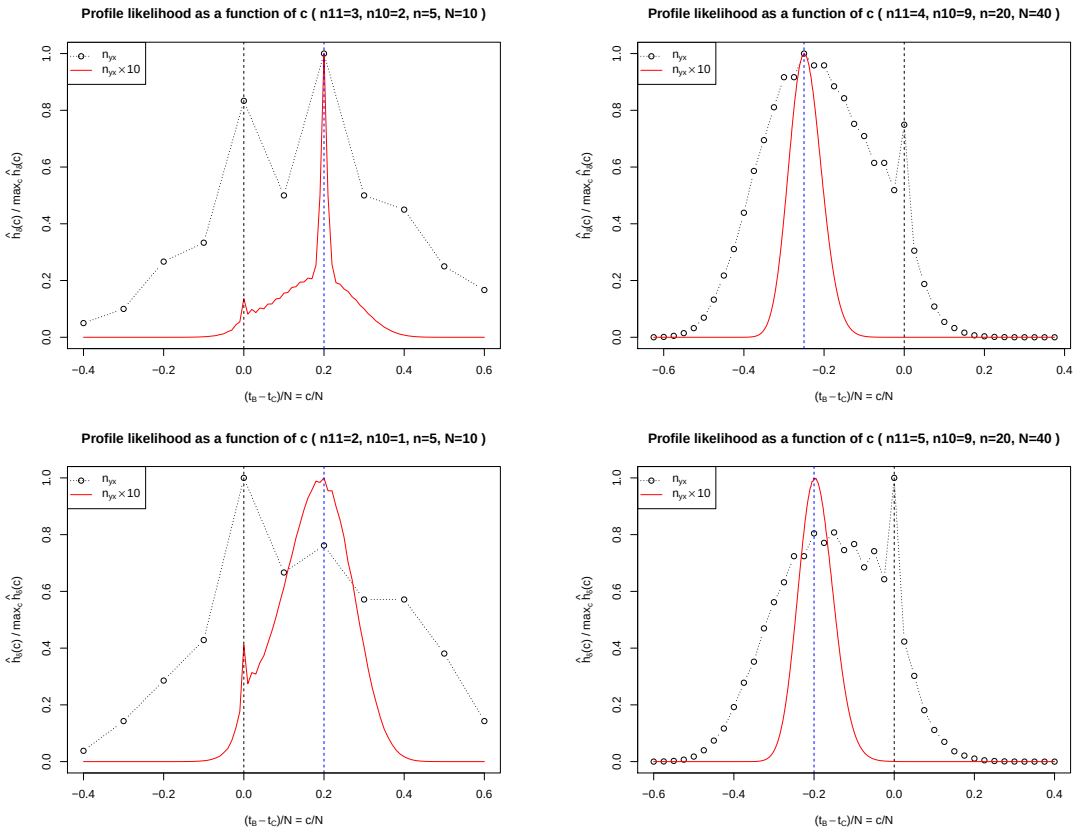

Figure 1.2: Plots of the profile likelihood $\hat{h}_\delta(c)$ against c for datasets I,II,III and IV respectively; the values on the vertical axis have been rescaled as $\hat{h}_\delta(c)/\max_c \hat{h}_\delta(c)$, which is also the GLR for comparing the hypothesis that $\delta = c$ against the hypothesis that $\delta = \widehat{\delta}$; the red solid lines are for larger versions of (ten times) the observed quantities in each dataset; the blue broken line is $\hat{\delta} = \widehat{\text{ACE}} = (n_{11}/n - n_{10}/(N-n))$; the black broken line is $\delta = 0$.

1.7 Maximum Likelihood Under the 'Fisher' and 'Neyman' Null Hypotheses

We first formalize the aforementioned 'Fisher' and 'Neyman' null hypotheses. Under the 'Fisher' causal null, also commonly referred to as the individual or 'sharp' null,

$$H_0(\text{Fisher}) : Y(1) = Y(0). \tag{1.25}$$

When $H_0(\text{Fisher})$ is true, the individual causal effect is zero: each individual in the finite population has the same outcome regardless of the treatment group to which they were assigned. Equivalently, there are no individuals with potential outcomes $Y(1) \neq Y(0)$, so that the only population in $T(\boldsymbol{n})$ satisfying $t_B = t_C = 0$ is $\boldsymbol{t}^0 = (s, 0, 0, N - s)$. $H_0(\text{Fisher})$ is thus a simple null hypothesis, with the likelihood for the observed data being:

$$h(\boldsymbol{t}^0) = \binom{s}{n_{11}, n_{10}} \binom{N-s}{n_{01}, n_{00}}. \tag{1.26}$$

Note $h(\boldsymbol{t}^0)$ is also the kernel of the bivariate hypergeometric probability of the observed dataset in Table 1.1. Assuming that $H_0(\text{Fisher})$ holds hence implies $X \perp\!\!\!\perp Y$ in the observed data. Testing this latter independence using Fisher's exact test [Fisher, 1935] may thus be seen as a test of the null hypothesis that all individual causal effects are exactly zero. This is a possible reason for the adoption of the term 'Fisher' null.

Under the 'Neyman' causal null, variously referred to as the population, 'blunt', or more recently, 'weak' [Chiba, 2015] causal null, the average causal effect - as defined in (1.22) - is zero:

$$H_0(\text{Neyman}) : \text{ACE} = 0 \equiv t_B = t_C. \tag{1.27}$$

Here the total number of individuals with potential outcomes $Y(1) > Y(0)$ is exactly equal to the total number with potential outcomes $Y(1) < Y(0)$. $H_0(\text{Neyman})$ has come to be termed as the 'Neyman' null due to the unbiased estimation of the ACE in [Spława-Neyman et al., 1990].

Denoting ψ as the common value of the column totals $t_B = t_C$, the range of possible values for ψ is:

$$\psi \in \{0, 1, \ldots, \min(n_{11} + n_{00}, n_{01} + n_{10})\} \equiv \Psi(\boldsymbol{n}). \tag{1.28}$$

The 'Neyman' null (1.27) is thus a composite null hypothesis, with ψ being a nuisance parameter for which the ψ-specific 'Neyman' null is:

$$H_0(\psi) : t_B = t_C = \psi. \tag{1.29}$$

The 'Fisher' null (1.25) is equivalent to the 'Neyman' null (1.29) only when $\psi = 0$; in other words, $H_0(\text{Fisher}) \equiv H_0(\psi = 0)$. The set of alternatives against the 'Neyman' null is thus *narrower* than the set of alternatives against the 'Fisher' null [Lang, 2015].

In general, for some given value of $\psi > 0$, there may be multiple populations $\boldsymbol{t} \in T(\boldsymbol{n})$ where $t_B = t_C = \psi$; thus ψ is not the only nuisance parameter determining the 'Neyman' null. The values of t_A compatible with the 'Neyman' null for a given value of ψ are the non-negative integers in $(\{0\} \cup \mathbb{Z}^+)$ that satisfy the following inequalities:

$$0 \le t_A \le n_{11} + n_{10}, \qquad 0 \le \psi \le \min(n_{11} + n_{00}, n_{01} + n_{10}),$$
$$\max(n_{11}, n_{10}) \le t_A + \psi \le \min(N - n_{01}, N - n_{00}), \quad n_{11} + n_{10} \le t_A + 2\psi \le N. \tag{1.30}$$

The inequalities may also be obtained from (1.9) together with the requirement that $t_B = t_C = \psi$.

The maximum likelihood under the ψ-specific 'Neyman' null is then:

$$\widehat{h}_\psi = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) : t_B = t_C = \psi\}} h(\boldsymbol{t}). \tag{1.31}$$

The overall maximum likelihood under the 'Neyman' null (1.27) is then:

$$\widehat{h}(\text{Neyman}) \equiv \max_{\psi \in \Psi(\boldsymbol{n})} \widehat{h}_\psi. \tag{1.32}$$

We will prove in Theorem 2 that the restricted maximum likelihood solution when $t_B = t_C$ under the 'Neyman' null (1.27) is always equal to the likelihood under the 'Fisher' null (1.25) where $t_B = t_C = 0$.

Theorem 2. *Under the 'Neyman' causal null* (1.27) *that the population average causal effect is zero, the maximum likelihood given an observed table* $\boldsymbol{n}$ *is always attained under the stricter 'Fisher' null* (1.25) *for which the individual causal effects are zero. In other words:*

$$\widehat{h}(\textit{Neyman}) = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) : t_B = t_C\}} h(\boldsymbol{t}) = h(\boldsymbol{t}^0) = \binom{s}{n_{11}, n_{10}} \binom{N - s}{n_{01}, n_{00}}. \tag{1.33}$$

Proof. First, relabel the variable X for the potential outcome variables $Y(X = x)$ so that the observed proportion of $Y = 1$ within the $X = 1$ group is less than or equal to the proportion of $Y = 1$ within the $X = 0$ group. Under such a relabelling, individuals with non-zero individual causal effects ($Y(1) \neq Y(0)$) are still either of type B or C, while individuals of type A have potential outcomes $Y(1) = Y(0) = 1$ and those of type D have potential outcomes $Y(1) = Y(0) = 0$. It may thus be assumed that the observed quantities $\boldsymbol{n}$ satisfy the following inequalities without loss of generality:

$$\frac{n_{11}}{n} \leq \frac{n_{10}}{N-n} \iff n_{11}n_{00} \leq n_{01}n_{10} \iff \frac{n_{11}}{s} \leq \frac{n_{01}}{N-s} \iff \frac{n_{00}}{N-s} \leq \frac{n_{10}}{s}. \tag{1.34}$$

For some given value of $\psi \in \Psi(\boldsymbol{n})$ as defined in (1.28), consider a population $\boldsymbol{t} \in T(\boldsymbol{n})$ where $t_B = t_C = \psi$. We will consider the following two possibilities for ψ in turn.

First if $\psi = t_C = t_B \geq s - t_A$, denote $t_{AB} = (t_A + t_B) - s \geq 0$. Using the Chu-Vandermonde convolution, the likelihood for a given set of column totals $\boldsymbol{t}$ satisfies the following inequalities:

$$\begin{aligned}
h(\boldsymbol{t}) &= \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=A,B} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}}}}_{\leq 1} \prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}} \\
&\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}} \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}}}}_{\leq 1} \\
&\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}} \\
&= \binom{s + t_{AB}}{n_{11}, (s - n_{11}) + t_{AB}} \binom{N - s - t_{AB}}{n_{01}, (N - s - n_{01}) - t_{AB}} \\
&\leq \binom{s}{n_{11}, n_{10}} \binom{N - s}{n_{01}, n_{00}} = h(\boldsymbol{t}^0).
\end{aligned} \tag{1.35}$$

The last inequality is obtained with Lemma 3, which may be found in Appendix A.

Otherwise if $\psi = t_C = t_B \leq s - t_A$, then denote $t_{AC} = s - (t_A + t_C) \geq 0$, such that:

$$
\begin{aligned}
h(\boldsymbol{t}) &= \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=A,C} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}}}}_{\leq 1} \prod_{i=B,D} \binom{t_i}{m_{i1}, m_{i0}} \\
&\leq \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}} \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=B,D} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}}}}_{\leq 1} \\
&\leq \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}} \\
&= \binom{N - s + t_{AC}}{n_{00}, n_{01} + t_{AC}} \binom{s - t_{AC}}{n_{10}, n_{11} - t_{AC}} \\
&\leq \binom{N - s}{n_{00}, n_{01}} \binom{s}{n_{10}, n_{11}} = h(\boldsymbol{t}^0). \qquad (1.36)
\end{aligned}
$$

The last inequality is similarly obtained with Lemma 3 in Appendix A.

Since the above is true for all $\boldsymbol{t} \in T(\boldsymbol{n})$ with $t_B = t_C = \psi$, the maximum likelihood for the ψ-specific 'Neyman' null is also bounded above by the likelihood under the 'Fisher' null:

$$
\widehat{h}_\psi = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) \,:\, t_B = t_C = \psi\}} h(\boldsymbol{t}) \leq h(\boldsymbol{t}^0).
$$

Taking the maximum over all possible values of ψ, including $\psi = 0$, the maximum likelihood under the 'Neyman' null is then attained by the population $\boldsymbol{t}^0 = (s, 0, 0, N - s)$ where $\psi = 0$:

$$
\widehat{h}(\text{Neyman}) \equiv \max_{\psi \in \Psi(\boldsymbol{n})} \widehat{h}_\psi = h(\boldsymbol{t}^0).
$$

□

1.7.1 Likelihood ratio for testing the 'Neyman' null

To assess the support in the observed data $\boldsymbol{n}$ for the 'Neyman' null against the unconstrained alternative, we may use the GLR test statistic that compares the maximum likelihood under the composite 'Neyman' null (1.27) against the global maximum likelihood (1.11):

$$
\lambda(\text{Neyman}) \equiv \max_{\psi \in \Psi(\boldsymbol{n})} \widehat{h}_\psi \Big/ \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}) = \widehat{h}(\text{Neyman}) / h(\hat{\boldsymbol{t}}). \qquad (1.37)
$$

From Theorem 2, since $\widehat{h}(\text{Neyman}) = h(\boldsymbol{t}^0)$, the test statistic $\lambda(\text{Neyman})$ is equal to the test statistic $\lambda(\text{Fisher})$ that compares the likelihood under the simple 'Fisher' null against the global maximum likelihood:

$$\lambda(\text{Fisher}) \equiv h(\boldsymbol{t}^0) \Big/ \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}) = h(\boldsymbol{t}^0)/h(\hat{\boldsymbol{t}}). \tag{1.38}$$

In other words, the GLR test statistic used to test the 'Fisher' null hypothesis of zero individual causal effects (1.25) may also be used as a test statistic for testing the 'Neyman' null hypothesis of a zero population average causal effect (1.27):

$$\lambda(\text{Neyman}) = \lambda(\text{Fisher}) \equiv h(\boldsymbol{t}^0)/h(\hat{\boldsymbol{t}}). \tag{1.39}$$

1.8 Alternative Models Where the 'Fisher' Null Does Not Hold

When the 'Fisher' null (1.25) is true, all individuals have potential outcomes $Y(1) = Y(0)$, so that the column totals in the observed table are fixed over hypothetical rerandomizations in the finite population. However, it has been suggested, for example in [Breslow and Day, 1980, Chapter 4.2] and [Rothman et al., 2008, pg. 308], that even when the 'Fisher' null *does not hold*, inference should be performed conditional on fixed column totals in the observed table.

We note the perhaps obvious point that the column totals in the observed table are *not* fixed over hypothetical rerandomizations under the alternative where the population comprises individuals with non-zero individual causal effects. Consider the following counterexample. Suppose that in an observed two-by-two dataset such as Table 1.1, all individuals with observed outcomes $Y = 0$ are of type D, and all but one individual (say 'J') with observed $Y = 1$ are of type A. Individual 'J' is actually of type B, with potential outcomes $Y(x) = x$, and had been assigned to the $X = 1$ group. The total number of individuals with $Y = 1$ and $Y = 0$ are then s and $N - s$, the column totals of the actual observed dataset.

However, if 'J' was assigned to $X = 0$ instead in a hypothetical rerandomization, then 'J' would have an observed outcome of $Y = 0$, not $Y = 1$. All individuals of types A and D would have still the same observed outcomes of $Y = 1$ and $Y = 0$ respectively, regardless of their assigned

treatment group. Then under such assignments, the total number of individuals observed with $Y=1$ and $Y=0$ would be $s-1$ and $N-s+1$ respectively, and not s and $N-s$!

In general, the observed Y margins are not fixed over hypothetical rerandomizations from a finite population. Under a completely randomized design where the row totals n and $N-n$ are fixed, hypothetical resamples $\tilde{\boldsymbol{n}}$ (the tilde $(\sim)$ accent denotes hypothetical resamples) may be described by one entry in each row, for example $(\tilde{n}_{11}, \tilde{n}_{10})$. The sample space of all $\tilde{\boldsymbol{n}}$ given an observed $\boldsymbol{n}$, denoted as $\Omega(\boldsymbol{n})$, is the intersection of the two-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^2$ and the convex polyhedron described with the following inequalities (obtained using `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015]):

$$\max(0, n_{11}+n-N) \leq \tilde{n}_{11} \leq \min(s+n_{00}, n), \qquad \tilde{n}_{01} = n - \tilde{n}_{11} \tag{1.40}$$

$$\max(0, n_{10}-n) \leq \tilde{n}_{10} \leq \min(s+n_{01}, N-n), \qquad \tilde{n}_{00} = N - n - \tilde{n}_{10}. \tag{1.41}$$

Breslow and other authors have proposed the (Fisher) non-central hypergeometric distribution as an alternative model when the 'Fisher' null does not hold, with the unknown parameter being the odds ratio ϕ. The probability of the observed dataset for a given value of ϕ is:

$$\Pr(n_{11}|n,s,\phi) = g(n_{11};n,s,\phi) \Bigg/ \sum_{\tilde{n}_{11}=\tilde{n}_{11}^{(0)}}^{\tilde{n}_{11}^{(1)}} g(\tilde{n}_{11};n,s,\phi),$$

$$\text{where } g(a;n,s,\phi) = \binom{s}{a, s-a}\binom{N-s}{n-a, N-n-s+a}\phi^a; \tag{1.42}$$

the normalizing constant sums over the sample space indexed by $\tilde{n}_{11}$ from $\tilde{n}_{11}^{(0)} = \max(0, s+n-N)$ to $\tilde{n}_{11}^{(1)} = \min(s, n)$.

We now compare the non-central (bivariate) hypergeometric distribution (1.42) with the convolution (central multivariate) hypergeometric probability distribution from the Neyman-Rubin-Copas model (1.8) under the alternative hypothesis where there is a non-zero individual causal effect. Recall the toy dataset II, so that under the true population $\boldsymbol{t} = (3,2,0,5)$, the population causal odds ratio is for example,

$$\phi^* \equiv \frac{t_A + t_B}{t_A + t_C} \times \frac{t_B + t_D}{t_C + t_D} = \frac{7}{3}.$$

Holding the observed column and row totals fixed, $\tilde{n}_{11}$ can take one of four possible values $\{0, 1, 2, 3\}$, with the other three cell counts determined by $\tilde{n}_{11}$ so that

$$\tilde{\boldsymbol{n}} \equiv (\tilde{n}_{11}, \tilde{n}_{10}, \tilde{n}_{01}, \tilde{n}_{00}) = (\tilde{n}_{11}, s-\tilde{n}_{11}, n-\tilde{n}_{11}, N-n-s+\tilde{n}_{11}).$$

Even if we conditioned on the observed margins, such that the conditional distribution based on the Neyman-Rubin-Copas model is $\Pr(\tilde{\boldsymbol{n}}|\boldsymbol{t}, n) \,/\, \sum_{\tilde{n}_{11}=0}^{3} \Pr(\tilde{\boldsymbol{n}}|\boldsymbol{t}, n)$, the non-central hypergeometric distribution $\Pr(n_{11}|n, s, \phi^*)$ differs from the actual (conditional) distribution under $\boldsymbol{t}$.

Conditional Probabilities	$\tilde{n}_{11}$				Row Sum
	0	1	2	3	
Non-central Bivariate Hypergeometric	0.018	0.268	0.536	0.179	1
(Convolution) Central Multivariate Hypergeometric	0.019	0.222	0.518	0.242	1

Table 1.4: Conditional probabilities of each possibly observable dataset in terms of $\tilde{n}_{11}$ under the population $\boldsymbol{t} = (3, 2, 0, 5)$, for the non-central bivariate hypergeometric distribution (Row 1), and the central multivariate hypergeometric distribution under the Neyman-Rubin-Copas convolution likelihood (Row 2); here the observed row and column totals in toy dataset II are fixed.

1.9 Discussion

The goal of understanding the maximum likelihood is to ease finding the GLR test statistic in performing exact significance tests. We conjecture that the global maximum likelihood $\hat{h}$ is attained by at least one population with at most three types, so that finding $\hat{h}$ does not require summing over probabilities for multiple complete tables when evaluating candidate column totals $\boldsymbol{t}$.

Because of the 'fundamental problem of causal inference', causal parameters are generally not identified and cannot be consistently estimated from observed data. A GLR criterion is appropriate for testing hypotheses on the causal parameters since we are affording the observed data the strongest possible support given by values of the unidentified causal parameters. Our interest here is thus not in the maximum likelihood estimates per se, but rather in the *comparison* of the

maximum likelihoods under different causal hypotheses. These results facilitate the construction of exact significance tests using the generalized likelihood ratio in more complex finite population settings. A procedure to construct exact confidence intervals for the ACE using the GLR is also described in Appendix B. In subsequent chapters, the GLR test statistic is extended to carry out exact significance tests in more complicated finite population settings.

Chapter 2

FINITE POPULATION TESTS OF THE SHARP NULL HYPOTHESIS FOR COMPLIERS

2.1 Treatment Noncompliance

We have thus far assumed that the treatment received X was randomly assigned. However, often we are interested in the effect of a treatment X that was not randomized. In this chapter we consider the circumstance where, although X is not randomized, there is another variable Z, called an 'instrument' that is randomized, and influences X, but does not influence Y directly; see Figure 2.1. The assumption that Z has no (direct) effect on Y except through X is sometimes termed an 'exclusion restriction'.

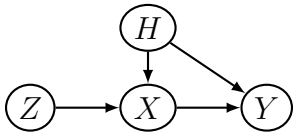

Figure 2.1: Graphical representation of the Instrumental Variable (IV) model, where H are unobserved confounding variables.

A common example of this circumstance is a randomized study with 'noncompliance'. In this context Z represents the randomly assigned treatment, while X is the post-randomization treatment that the patient actually receives; X and Z may differ owing to noncompliance.

Randomized experiments with treatment 'noncompliance' arise in many situations. For example, patients in a randomized clinical trial may choose not to take their prescribed treatment, possibly due to side-effects. In 'encouragement' studies where a randomly selected subset of subjects are offered an incentive to avail themselves of a treatment, the inducement may be sufficient

for some but not for others. In a randomized psychology experiment, whether or not participants adhere to their assigned manipulation (treatment) depends on their individual personalities and the type of manipulation. In each of these situations, every unit now has a treatment actually received (X) that was *not* randomized, even if the *assigned* treatment (Z) was randomized.

In such studies with noncompliance and binary treatment assignment Z, one may use a symbolic linear program to derive the bounds for counterfactual probabilities, and the average causal effect of X on Y [Balke and Pearl, 1994]. Rubin [1998] uses randomization-based posterior-predictive p-values to test a null treatment effect. Imbens and Rosenbaum [2005] use randomization-based inference to obtain valid confidence intervals for the treatment effect under an additive structural model even when the instrument is 'weak'; Keele et al. [2016] propose using attributable effects for Compliers to construct exact confidence intervals for the CACE. A review of different estimation approaches for binary outcomes in the IV model was also conducted in [Clarke and Windmeijer, 2012].

Angrist et al. [1996] and Imbens and Rubin [1997] among others, propose to find the effect of treatment on the subset of individuals who would conform with the assigned treatment regardless of the group to which they are assigned. Sommer and Zeger [1991] describe this subgroup of individuals as 'Compliers': individuals who would take the treatment only if assigned to do so and would not if assigned not to do so.

Another example where the IV model in Figure 2.1 may be assumed for causal inference is in Mendelian randomization analyses for epidemiological research; see for example [Didelez and Sheehan, 2007]. In such applications, interest is in testing the causal effect of a modifiable phenotype or risk exposure X on a disease Y, but inferences based only on the observed (X, Y) data may be biased due to possible confounding between X and Y.

A genetic variant Z that (i) is closely linked to the phenotype or exposure X but has no direct effect on the disease Y, and (ii) is randomly assigned (given the parents' genes) at meiosis such that there is no association between Z and any possible confounding factors with Y, may thus be considered as an instrument. The principal stratum analogous to 'Compliers' are thus individuals whose potential outcome for the phenotype or exposure correspond to the assigned genotype of

interest.

In this chapter we address the problem of testing the 'sharp' null hypothesis of zero individual causal effect of X on Y for 'Compliers', the subpopulation or principal stratum where $X(z) = z$ for $z = 0, 1$; $X(z)$ being the treatment a patient would receive if (counter to fact) assigned to $Z = z$. Under the exclusion restriction assumption, the null hypothesis within this subpopulation that X has no effect on Y is equivalent to the null hypothesis that Z has no effect on Y. Under random assignment for the whole population, each individual in the Complier subpopulation has the same probability of being *assigned* to treatment. Thus we could use the randomization distribution of the outcomes for Compliers under the null hypothesis to carry out a significance test.

However, we face the obvious difficulty that membership in the Complier subpopulation generally cannot be determined from the observed data alone. Although we know that Compliers will have $Z = X$, this condition is necessary but not sufficient. For example, in the $Z = 1$ group, individuals with $X = 1$ may be either Compliers or 'Always Takers', where the latter subgroup are individuals who would always take the active treatment even if (counter to fact) they had been assigned to the placebo group ($Z = 0$). Conversely, an individual with $Z = X = 0$ may be either a Complier or someone who refuses to take treatment regardless of their assigned group, in other words a 'Never Taker'.

If somehow we were told which individuals in the population were Compliers, then we could simply test the 'sharp' null hypothesis by performing a significance test, such as Fisher's exact test, on the (X, Y) subtable, or equivalently the (Z, Y) subtable, for Compliers. One may circumvent the problem of not knowing who the Compliers are by just considering all *logically* possible values for the number of Compliers in any given (Z, X, Y) stratum that may contain them (in which $Z = X$), and then carrying out the significance test for the corresponding (X, Y) subtable for the Compliers. Taking the maximum over all the resulting p-values would then give a valid p-value for the null hypothesis.

There are, however, two concerns with such an approach. The first is that such a procedure will have no statistical power to reject the null hypothesis, since it is always logically possible that there are zero Compliers in any given stratum in which $Z = X$ (such that the maximum p-value for the

Complier subtable will always be 1). The second is that such an approach ignores the information provided by strata that do not contain Compliers, in which $Z \neq X$.

We will assume that there are no individuals who consistently do the opposite of their assignment, sometimes called 'Defiers' [Chickering and Pearl, 1996], so for all individuals:

$$X(z=0) \leq X(z=1). \tag{2.1}$$

It follows from this assumption, also termed 'monotonicity', that all individuals in the $(Z=1, X=0)$ stratum are Never Takers, and all individuals in the $(Z=0, X=1)$ stratum are Always Takers. Under random assignment of treatment Z, the respective proportions of Never Takers and Always Takers in both the $Z=1$ and $Z=0$ groups should be approximately the same. This information then reduces the range of probable values (under the randomization distribution) for the number of Compliers in the $Z = X$ strata. Such an approach of conditioning on the compliance type (Compliers, Never Takers, Always Takers) as if it were a pretreatment covariate is sometimes termed 'principal stratification' [Frangakis and Rubin, 2002]. Forastiere et al. [2015] conduct simulation studies comparing different approaches for imputing the unknown compliance statuses.

In this chapter we discuss two possible approaches that make use of the information in the $Z \neq X$ strata when testing the 'sharp' causal null hypothesis for Compliers. The first approach [Loh and Richardson, 2013], following [Nolen and Hudgens, 2011] and [Berger and Boos, 1994], uses a pre-specified significance level γ to construct a confidence set of values for the number of Compliers in a given (Z, X) stratum. Only values of the number of Compliers that do not indicate large imbalance between the $Z=1$ and $Z=0$ arms, under the randomization distribution, are used to carry out Fisher's exact test in the implied (X, Y) table for Compliers. Taking the maximum over these p-values and adding γ then provides a valid but conservative p-value.

However, such a procedure requires a pre-determined (non-zero) value of γ to eliminate 'unlikely' values for the number of Compliers from consideration when controlling the Type I error rate in a *hypothesis* test. The resulting p-value will hence always be greater than or equal to γ. This is problematic if, as in a *significance* test, we wish to interpret the p-value as measuring the strength of evidence against the null hypothesis.

An alternative approach builds on the generalized likelihood ratio (GLR) test statistic previously presented (in which X was randomly assigned). Under the extended Neyman-Rubin-Copas likelihood for treatment noncompliance, the GLR test statistic compares the ratio of the largest probability for the observed data assuming that the 'sharp' null hypothesis holds among Compliers, with the largest probability allowing a causal effect among Compliers. Such a GLR criterion lets us evaluate whether the alternative hypothesis is a significantly better explanation for the observed dataset than the 'sharp' null, even when the number of Compliers is unknown.

Finding the largest convolution likelihood when the 'sharp' null doesn't hold involves summing over conditional probabilities for multiple complete tables when evaluating candidate values of the column totals. A computationally faster option is to use the maximum conditional mulativariate hypergeometric probability in the complete table instead. Such a maximum probability ratio (MPR) was proposed in [Loh and Richardson, 2015] assuming only Compliers and Never Takers (the latter being individuals who consistently do the opposite of their assigned treatment). The procedure is also available in the `noncompliance` package [Loh and Richardson, 2016] in `R` [R Core Team, 2015]. In this chapter, we will show that a consequence of Conjecture 1, that the maximum likelihood under the binary causal model is achieved by at least one population with at most three types, is that the MPR and the GLR are equal under the binary IV model.

For a given number of Compliers, the relative frequency with which, over hypothetical replications under the null hypothesis, we would obtain a value of the GLR that is as small or smaller than that which we observed, would then be a p-value. Since this relative frequency will depend on the number of Compliers, we maximize the p-value over the number of Compliers. This results in a valid p-value that is suitable to be used in a significance test since it can be arbitrarily close to zero (it does not require specification of some γ). Furthermore, the resulting test has power against some alternatives in which there is a non-zero average causal effect among Compliers.

2.2 Potential Outcomes Framework Under Treatment Noncompliance

We now formalize the foregoing development. Recall the following:

- Z is the randomized treatment *assignment*, where 1 indicates assignment to active treatment;
- X is the treatment *exposure* subsequent to assignment, where 1 indicates treatment received;
- Y is the final response, where 1 indicates a desirable outcome, such as survival.

The potential outcome $X(z)$ is the treatment X a patient *would* be exposed to if assigned $Z = z$. Using these potential outcomes, we may define four generic 'compliance types' c_X listed in Table 2.1. The potential outcomes are linked to the observed outcomes by causal consistency:

$$X = ZX(1) + (1-Z)X(0). \tag{2.2}$$

Compliance Type c_X	$X(1) = 0$	$X(1) = 1$
$X(0) = 0$	Never Taker (NT)	Complier (CO)
$X(0) = 1$	Defier (DE)	Always Taker (AT)

Table 2.1: Compliance Types c_X based on Potential Outcomes $X(z)$, [Imbens and Rubin, 1997].

Under the 'monotonicity' assumption (2.1), there are no 'Defiers' in the finite population, so that individuals in the $Z \neq X$ strata are either Always Takers ($X = 1$) or Never Takers ($X = 0$).

2.2.1 Exclusion Restriction

In general, the potential outcome for a given individual's response Y under exposure to treatment $X = x$, and treatment assignment $Z = z$ is $Y(x, z)$. Denote the set of values for the potential outcomes $\{Y(x, z); x, z = 0, 1\}$ for a single individual as the response type r_Y. Without further assumptions there are $16 = 2^{2^2}$ possible combinations for r_Y. However, under the 'exclusion restriction' assumption, there is no (individual-level) direct effect of Z on Y relative X, so that for $x, z, z' = 0, 1$, we have:

$$Y(x, z) = Y(x, z') \equiv Y(x). \tag{2.3}$$

Assumption (2.3) is guaranteed to hold under double-blind placebo-controlled trials in which the active treatment is without side-effects and unavailable to patients in the control group. The response type r_Y then simplifies to just four types shown in Table 2.2. The potential outcomes $Y(x)$ are again linked to the observed outcomes via the causal consistency axiom, so that if $X = x$ then $Y = Y(x)$.

Response Type r_Y	$Y(x=1)=0$	$Y(x=1)=1$
$Y(x=0)=0$	Never Recover (NR)	Helped (HE)
$Y(x=0)=1$	Hurt (HU)	Always Recover (AR)

Table 2.2: Response Types r_Y under Exclusion Restriction (2.3), [Heckerman and Shachter, 1995].

2.2.2 Randomization Assumption

We also make the following assumption:

$$\Pr\left(Z=1 \mid \{X(0), X(1), Y(0), Y(1)\}\right) = \Pr(Z=1). \tag{2.4}$$

The assumption states that the probability of being assigned to $Z = 1$ is (jointly) independent of an individual's potential outcomes. In other words, the distribution of compliance and response types (c_X, r_Y) is the same in both the $Z = 1$ and $Z = 0$ arms. This will hold whenever treatment assignment Z is physically randomized.

2.2.3 The Instrumental Variable (IV) Model

The model defined by (2.3) and (2.4) is known as the Instrumental Variable (IV) model (see for example [Angrist et al., 1996]). A graph corresponding to the IV model given by (2.3) and (2.4) is shown in Figure 2.1. The exclusion restriction (2.3) corresponds to the absence of a $Z \to Y$ edge while the randomization assumption (2.4) is indicated by the absence of edges directed into Z.

2.2.4 *Intent-To-Treat (*ITT*) Analysis*

One possible approach that gets around the issue of partial compliance is to simply consider the causal effect of treatment *assignment*, rather than treatment *received or exposed to*, also know as the Intent-To-Treat (ITT) effect [Lee et al., 1991]:

$$\text{ITT} \equiv \text{E}[Y(z=1)-Y(z=0)], \tag{2.5}$$

where, in our finite population context, the expectation here is over all possible assignments of individuals to $Z=1$ or $Z=0$.

However, inference for this effect entirely ignores information provided by X. Furthermore, the average difference in final responses Y based on treatment assignment Z does not provide an unbiased or even consistent estimate of the average causal effect of treatment exposure X on final response Y [Angrist et al., 1996].

2.2.5 *Sharp null hypothesis*

Under the 'sharp' causal null hypothesis, the absence of a non-zero individual-level effect of X on Y is formalized by

$$H_0: \; Y(x=1)=Y(x=0); \tag{2.6}$$

each individual in the finite population has the same outcome regardless of the treatment group X to which they were assigned.

2.2.6 *Average Causal Effect of X on Y*

The average causal effect (ACE) of treatment exposure X on outcome Y is defined as:

$$\text{ACE}(X \to Y) \equiv \text{E}[Y(x=1)-Y(x=0)]. \tag{2.7}$$

The ACE for the subpopulation of Compliers, which we also refer to as the Complier Average Causal Effect (CACE), is:

$$\text{ACE}_{CO}(X \to Y) \equiv \text{E}[Y(x=1)-Y(x=0) \mid c_X = CO]. \tag{2.8}$$

Since for Compliers $X(z)=z$, it follows that $Y(X(z)=z)\equiv Y(x=z)=Y(z)$ so that

$$\text{ACE}_{CO}(X\to Y)=\text{E}[Y(z=1)-Y(z=0)\,|\,c_X{=}CO]\equiv \text{ITT}_{CO}, \tag{2.9}$$

or in words, the Average Causal Effect of X on Y for Compliers is equal to the *Intent-to-Treat effect* of Z on Y for Compliers (ITT_{CO}).

Under the 'monotonicity' assumption (2.1) that there are no Defiers, the global null hypothesis $\text{ACE}(X \to Y) = 0$ holds if and only if all the principal stratum-specific null hypotheses $\text{ACE}_{c_X}(X\to Y)=0$ for $c_X \in \{NT, CO, AT\}$ jointly hold. By definition Never Takers and Always Takers always have the same observed values of $X = 0$ and $X = 1$ respectively (regardless of their assigned value of Z). Consequently without further experimentation (to change compliance for these individuals), there is no test for the average causal effect of X on Y in either of these principal strata. In the absence of Defiers, the only subpopulation for which we may observe evidence that $\text{ACE}_{c_X}(X \to Y) \neq 0$ are the Compliers (CO). Evidence against the (narrower) null hypothesis that $\text{ACE}_{CO}(X \to Y) = 0$ hence implies evidence against the global null hypothesis $\text{ACE}(X\to Y)=0$ as well. This is why, even though our procedure is a test of the global null, we describe it as a test of the 'sharp' null for Compliers.

In the absence of Defiers, the 'sharp' causal null hypothesis that X has no effect on Y *within the Complier subpopulation* is equivalent to the null hypothesis that Z has no effect of Y:

$$\text{For Compliers, } \; H_0 : Y(z=1)=Y(X(z=1)=1)=Y(X(z=0)=0)=Y(z=0). \tag{2.10}$$

Similarly, the average effect of treatment on the treated, $\text{E}[Y(x=1)-Y(x=0)\mid X=1]$, is a weighted average of $\text{ACE}_{CO}(X\to Y)$ and $\text{ACE}_{AT}(X\to Y)$ since the 'treated' subpopulation ($X=1$) comprises Compliers only in the treatment group ($Z=1$) but all the Always Takers in the population. Without a test for $\text{ACE}_{AT}(X\to Y)=0$, evidence against the 'sharp' null hypothesis for Compliers would hence imply evidence against a null of zero average effect of treatment on the treated.

2.3 Binary IV model without Defiers

We now introduce the notation for our model, using a double subscript to avoid ambiguity. Let $n_{y_k x_j z_i}$ be the observed number of individuals assigned to treatment $Z=i$, with exposure $X=j$ and outcome $Y=k$. The marginal sums are denoted similarly, for example $n_{y_k x_j} = \sum_{i=0}^{1} n_{y_k x_j z_i}$ and $n_{x_j} = \sum_{k=0}^{1}\sum_{i=0}^{1} n_{y_k x_j z_i}$. We will refer to the observed dataset as $\boldsymbol{n} \equiv \left\{n_{y_k x_j z_i} : i,j,k = 0,1\right\}$, and the size of the finite population as $N = \sum_{k=0}^{1}\sum_{j=0}^{1}\sum_{i=0}^{1} n_{y_k x_j z_i}$. Also denote the total number of individuals in the finite population of compliance type c_X and response type r_Y as $t_{r_Y}^{c_X}$. For example, the total number of Compliers who are Helped and Hurt are respectively t_{HE}^{CO} and t_{HU}^{CO}.

We will assume that there are no Defiers in the population (2.1). Even then, we still need to distinguish the Compliers from the Always Takers and Never Takers to test the 'sharp' null for Compliers. Since our interest here is in the CACE, and there is no direct effect of Z on Y under exclusion restriction (2.3), we will not distinguish between the different response types r_Y for the Never Takers and Always Takers. Instead for $k = 0, 1$, denote the respective total number of Never Takers and Always Takers with observed outcome $Y=k$ in the population as:

$$t_k^{NT} \equiv \sum_{r_Y : Y(0)=k} t_{r_Y}^{NT}, \quad t_k^{AT} \equiv \sum_{r_Y : Y(1)=k} t_{r_Y}^{AT}; \quad k = 0, 1. \tag{2.11}$$

A finite population is therefore uniquely described by the parameter of length eight containing the total number of individuals of each type, which we collectively denote as:

$$\boldsymbol{t} \equiv \left(t_0^{AT}, t_0^{NT}, t_{NR}^{CO}, t_{HE}^{CO}, t_{HU}^{CO}, t_{AR}^{CO}, t_1^{NT}, t_1^{AT}\right). \tag{2.12}$$

2.4 Parameter Space

Let $T(\boldsymbol{n})$ be the parameter space of possible values for $\boldsymbol{t}$ given an observed dataset $\boldsymbol{n}$. Since $t_{AR}^{CO} \equiv N - \left(t_{NR}^{CO} + t_{HE}^{CO} + t_{HU}^{CO} + \sum_{k=0}^{1}\left(t_k^{NT} + t_k^{AT}\right)\right)$ in a finite population of size N, it is sufficient to specify values for the seven column totals $\left(t_0^{AT}, t_0^{NT}, t_{NR}^{CO}, t_{HE}^{CO}, t_{HU}^{CO}, t_1^{NT}, t_1^{AT}\right)$. $T(\boldsymbol{n})$ is then the intersection of the seven-dimensional integer lattice $\left(\{0\} \cup \mathbb{Z}^+\right)^7$, and the convex polyhedron described by the set of inequalities (2.13)$-$(2.25).

$$n_{y_k x_0 z_1} \leq t_k^{NT} \leq n_{y_k x_0}, \quad k = 0, 1 \tag{2.13}$$

$$n_{y_k x_1 z_0} \leq t_k^{AT} \leq n_{y_k x_1}, \quad k = 0, 1 \tag{2.14}$$

$$0 \leq t_{HE}^{CO} \leq \sum_{i=0}^{1} n_{y_i x_i z_i} \tag{2.15}$$

$$0 \leq t_{HU}^{CO} \leq \sum_{i=0}^{1} n_{y_{1-i} x_i z_i} \tag{2.16}$$

$$t_0^{NT} + t_1^{AT} + t_{HE}^{CO} \leq \sum_{i=0}^{1} n_{y_i x_i} \tag{2.17}$$

$$t_1^{NT} + t_0^{AT} + t_{HU}^{CO} \leq \sum_{i=0}^{1} n_{y_{1-i} x_i} \tag{2.18}$$

$$t_{NR}^{CO} \geq 0 \tag{2.19}$$

$$t_0^{NT} + t_{NR}^{CO} + t_{HE}^{CO} \geq n_{y_0 x_0} \tag{2.20}$$

$$t_0^{AT} + t_{NR}^{CO} + t_{HU}^{CO} \geq n_{y_0 x_1} \tag{2.21}$$

$$t_0^{NT} + t_0^{AT} + t_{NR}^{CO} + t_{HE}^{CO} + t_{HU}^{CO} \geq n_{y_0} \tag{2.22}$$

$$\left(\sum_{k=0}^{1} t_k^{AT} \right) + t_0^{NT} + t_{NR}^{CO} + t_{HE}^{CO} \leq n_{y_0 x_0} + n_{x_1} \tag{2.23}$$

$$\left(\sum_{k=0}^{1} t_k^{NT} \right) + t_0^{AT} + t_{NR}^{CO} + t_{HU}^{CO} \leq n_{y_0 x_1} + n_{x_0} \tag{2.24}$$

$$t_0^{NT} + t_0^{AT} + t_{NR}^{CO} \leq n_{y_0}. \tag{2.25}$$

When the 'sharp' null for Compliers (2.10) is true, there are no Compliers of response types Helped or Hurt in the population, and all Compliers are either of type Never Recover or Always Recover. Equivalently the 'sharp' null for Compliers is:

$$H_0 : \quad t_{HE}^{CO} = t_{HU}^{CO} = 0. \tag{2.26}$$

If the observed study comprised only of individuals with $Z = X$, recall from Chapter 1 that $\boldsymbol{t}^0 = (s, 0, 0, N - s)$ is the only population compatible with the observed data and the (global)

'sharp' null, where s is the total number with $Y=1$. However, in the absence of perfect compliance, we no longer know exactly who the Compliers are, even when the (local) 'sharp' null for Compliers holds, without making further assumptions on the Always Taker and Never Takers. If we fixed the number of Always Takers and Never Takers at some value of $\left\{\left(t_k^{NT}, t_k^{AT}\right) : k=0,1\right\}$, then the respective total number of Compliers of types Never Recover and Always Recover, under the 'sharp' null for Compliers, are:

$$t_{NR}^{CO} = n_{y_0} - t_0^{NT} - t_0^{AT}, \quad t_{AR}^{CO} = n_{y_1} - t_1^{NT} - t_1^{AT}. \tag{2.27}$$

The population parameters $\left\{\left(t_k^{NT}, t_k^{AT}\right) : k=0,1\right\}$ are hence nuisance parameters for the 'sharp' null for Compliers. The null parameter space $T^0(\boldsymbol{n})$ under H_0 is then the subspace of $T(\boldsymbol{n})$ where $t_{HE}^{CO} = t_{HU}^{CO} = 0$. Equivalently $T^0(\boldsymbol{n})$ is a Cartesian product of the four (variation-independent) linear spaces corresponding to $\left\{\left(t_k^{NT}, t_k^{AT}\right) : k=0,1\right\}$:

$$T^0(\boldsymbol{n}) = \left\{ \begin{array}{rll} & t_k^{NT} & \in \left\{n_{y_k x_0 z_1}, \ldots, n_{y_k x_0}\right\}, k=0,1; \\ & t_k^{AT} & \in \left\{n_{y_k x_1 z_0}, \ldots, n_{y_k x_1}\right\}, k=0,1; \\ \boldsymbol{t} \in T(\boldsymbol{n}): & t_{NR}^{CO} & = n_{y_0} - t_0^{NT} - t_0^{AT}; \\ & t_{AR}^{CO} & = n_{y_1} - t_1^{NT} - t_1^{AT}; \\ & t_{HE}^{CO} & = t_{HU}^{CO} = 0 \end{array} \right\} \subset T(\boldsymbol{n}). \tag{2.28}$$

2.5 Hypothesis Test of the 'Sharp' Null for Compliers

We now describe how to carry out a hypothesis test of the 'sharp' null for Compliers (2.26). The procedure was previously presented in [Loh and Richardson, 2013] for populations with only Compliers and Never Takers. Here we present the procedure allowing for Always Takers as well.

We first introduce some more notation. Following $\boldsymbol{t}$ in (2.12), in each $Z=i$ group, let $m_{y_k z_i}^{c_X}$ be the number of individuals of compliance type c_X with observed outcome $Y=k$, and m_{r_Y, z_i}^{CO} the number of Compliers of response type r_Y. Denote the vector of length eight containing the number of individuals of each type (in the same order as $\boldsymbol{t}$) in the $Z=i$ group as:

$$\boldsymbol{m}_i \equiv \left(m_{y_0 z_i}^{AT}, m_{y_0 z_i}^{NT}, m_{NR, z_i}^{CO}, m_{HE, z_i}^{CO}, m_{HU, z_i}^{CO}, m_{AR, z_i}^{CO}, m_{y_1 z_i}^{NT}, m_{y_1 z_i}^{AT}\right); \tag{2.29}$$

we will refer to the concatenation of $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$ simply as a *complete* table $\boldsymbol{m}$.

In general, only some of the counts in $\boldsymbol{m}$ may be determined from $\boldsymbol{n}$. By definition, the Always Takers in the control group $m^{AT}_{y_k z_0}$ and Never Takers in the treatment group $m^{NT}_{y_k z_1}$ are directly observable as $n_{y_k x_1 z_0}$ and $n_{y_k x_0 z_1}$ respectively for $k = 0, 1$. For a given population $\boldsymbol{t} \in T(\boldsymbol{n})$, denote the number of Compliers in each $Z = i$ arm with observed outcome $Y = k$ as $n^{CO}_{y_k x_i}(\boldsymbol{t})$, where we use (i) n to emphasize that this is generally not a cell count in the complete table $\boldsymbol{m}$, (ii) the subscript x_i since Compliers have $Z = X$, and (iii) $\boldsymbol{t}$ to indicate the dependence on the fixed population. The four 'observed' counts $n^{CO}_{y_k x_i}(\boldsymbol{t})$, collectively denoted as $\boldsymbol{n}^{CO}(\boldsymbol{t})$, may then be expressed in terms of $\boldsymbol{n}$ and $\boldsymbol{t}$ as:

$$n^{CO}_{y_k x_0}(\boldsymbol{t}) = n_{y_k x_0} - t^{NT}_k, \quad n^{CO}_{y_k x_1}(\boldsymbol{t}) = n_{y_k x_1} - t^{AT}_k; k = 0, 1. \tag{2.30}$$

Under the 'sharp' null for Compliers, there are no Compliers who are Helped or Hurt, so that $m^{CO}_{HE,z_i} = m^{CO}_{HU,z_i} = 0$ for $i = 0, 1$. All compliers with observed outcomes $Y = 0$ must thus be of type Never Recover, while those with $Y = 1$ are of type Always Recover. A two-by-six contingency table such as Table 2.3 may be used to summarize $\boldsymbol{m}$ for an observed dataset $\boldsymbol{n}$ under a given population $\boldsymbol{t} \in T^0(\boldsymbol{n})$.

At this point, one could perform a significance test, such as Fisher's exact test, on the (X, Y) subtable for Compliers $\boldsymbol{n}^{CO}(\boldsymbol{t})$ to test the 'sharp' null. Denoting the *population-specific* p-value as $pv^{CO}(\boldsymbol{t})$, an overall p-value is just the maximum p-value over all *logically* possible values for $\boldsymbol{t}$:

$$pv^{CO} \equiv \max_{\boldsymbol{t} \in T^0(\boldsymbol{n})} pv^{CO}(\boldsymbol{t}). \tag{2.31}$$

Taking the maximum p-value here ensures that rejection of the 'sharp' null $pv^{CO} \leq \alpha$ implies rejection of the 'sharp' null in each of the population-specific tables $pv^{CO}(\boldsymbol{t}) \leq \alpha$, $\forall\, \boldsymbol{t} \in T^0(\boldsymbol{n})$. However, such an approach is undesirable since the test statistic used in calculating the p-value $pv^{CO}(\boldsymbol{t})$ is a function of both $\boldsymbol{n}$ and the nuisance parameter $\boldsymbol{t}$. Furthermore under the population

$$\boldsymbol{t} = (n_{y_0 x_1}, n_{y_0 x_0}, 0, 0, 0, 0, n_{y_1 x_0}, n_{y_1 x_1}),$$

the Complier subtable $\boldsymbol{n}^{CO}(\boldsymbol{t})$ would consist of only zeroes, such that the resulting p-value would

	$Z=0$	$Z=1$	Column
AT, y_0	$n_{y_0x_1z_0}$	$t_0^{AT} - n_{y_0x_1z_0}$	t_0^{AT}
NT, y_0	$t_0^{NT} - n_{y_0x_0z_1}$	$n_{y_0x_0z_1}$	t_0^{NT}
CO, NR	$n_{y_0x_0} - t_0^{NT}$	$n_{y_0x_1} - t_0^{AT}$	$t_{NR}^{CO} = n_{y_0} - t_0^{NT} - t_0^{AT}$
CO, AR	$n_{y_1x_0} - t_1^{NT}$	$n_{y_1x_1} - t_1^{AT}$	$t_{AR}^{CO} = n_{y_1} - t_1^{NT} - t_1^{AT}$
NT, y_1	$t_1^{NT} - n_{y_1x_0z_1}$	$n_{y_1x_0z_1}$	t_1^{NT}
AT, y_1	$n_{y_1x_1z_0}$	$t_1^{AT} - n_{y_1x_1z_0}$	t_1^{AT}
Row	n_{z_0}	n_{z_1}	N

Table 2.3: 'Complete' table $\boldsymbol{m}$ with each entry as a linear function of the observed quantities $\boldsymbol{n}$ and a fixed population total $\boldsymbol{t} \in T^0(\boldsymbol{n})$, assuming that the 'sharp' null for Compliers holds; the two-by-six table has also been transposed to make the cell counts more readable.

just be 1. Such an approach would have no statistical power since the maximum p-value pv^{CO} would then always be 1!

Rather than naïvely considering all possible values of $\boldsymbol{t} \in T^0(\boldsymbol{n})$, we should make use of information from the observed $Z \neq X$ strata. Under randomization (2.3) and exclusion restriction (2.4), the number of Always Takers and Never Takers in the $Z=0$ and $Z=1$ groups of the complete table $\boldsymbol{m}$ should be approximately 'balanced'. For example, if there is no direct effect of Z on Y for the Always Takers, the quantities $m_{y_0z_1}^{AT} = t_0^{AT} - n_{y_0x_1z_0}$ and $m_{y_0z_0}^{AT} = n_{y_0x_1z_0}$ should be 'similar' under random assignment Z. Following [Nolen and Hudgens, 2011], such information is used to limit the range of probable values for $\boldsymbol{t}$ (under the randomization distribution), and subsequently, the set of p-values for the Complier subtables $pv^{CO}(\boldsymbol{t})$.

The procedure from [Loh and Richardson, 2013] is restated as follows. First, for a given significance level α, choose a pre-determined value of $\gamma < \alpha$. For example, in other contexts, Agresti [2013] suggests a 'small' value of $\gamma = 0.001$ if $\alpha = 0.05$. Next, define the 'nuisance' table for a given population $\boldsymbol{t} \in T^0(\boldsymbol{n})$ as $Nuis(\boldsymbol{t})$, obtained by summing the total number of Compliers

within each $Z=i$ group of the complete table $\boldsymbol{m}$ and then disregarding the specific response types. $Nuis(\boldsymbol{t})$ may thus be described by a two-by-five contingency table such as Table 2.4.

	$Z=0$	$Z=1$	Column
AT, y_0	$n_{y_0x_1z_0}$	$t_0^{AT}-n_{y_0x_1z_0}$	t_0^{AT}
NT, y_0	$t_0^{NT}-n_{y_0x_0z_1}$	$n_{y_0x_0z_1}$	t_0^{NT}
CO	$n_{x_0}-t_0^{NT}-t_1^{NT}$	$n_{x_1}-t_0^{AT}-t_1^{AT}$	$N-t_0^{NT}-t_1^{NT}-t_0^{AT}-t_1^{AT}$
NT, y_1	$t_1^{NT}-n_{y_1x_0z_1}$	$n_{y_1x_0z_1}$	t_1^{NT}
AT, y_1	$n_{y_1x_1z_0}$	$t_1^{AT}-n_{y_1x_1z_0}$	t_1^{AT}
Row	n_{z_0}	n_{z_1}	N

Table 2.4: 'Nuisance' table $Nuis(\boldsymbol{t})$ where each cell count is a linear function of the observed quantities $\boldsymbol{n}$ and a fixed population parameter $\boldsymbol{t} \in T^0(\boldsymbol{n})$, assuming that the 'sharp' null for Compliers holds; the two-by-five table has also been transposed to make the cell counts more readable.

Under both randomization and exclusion restriction, treatment assignment Z is independent of the five compliance types in $Nuis(\boldsymbol{t})$. Since the row and column totals are fixed under complete randomization, we carry out Fisher's exact test on $Nuis(\boldsymbol{t})$ as a test of the composite null of randomization and exclusion restriction. The resulting p-value, denoted as $pv^{Nuis}(\boldsymbol{t})$, is then compared against the critical value γ. Populations $\boldsymbol{t}$ where the test of independence is *not* rejected at significance level γ are included in the confidence set:

$$C_\gamma \equiv \left\{ \boldsymbol{t} \in T^0(\boldsymbol{n}) : pv^{Nuis}(\boldsymbol{t}) \geq \gamma \right\}. \tag{2.32}$$

Note that the size of C_γ is hence inversely proportional to γ; setting $\gamma = 0$ corresponds to the entire space $T^0(\boldsymbol{n})$. Taking the maximum Complier subtable p-value $pv^{CO}(\boldsymbol{t})$ among all populations $\boldsymbol{t} \in C_\gamma$ and adding γ then provides a valid but conservative p-value [Berger and Boos, 1994]:

$$pv_\gamma^{CO} \equiv \gamma + \max_{\boldsymbol{t} \in C_\gamma} pv^{CO}(\boldsymbol{t}). \tag{2.33}$$

Such a procedure is relevant if our goal is to control the Type I error rate for some given significance level α. However, such a procedure has two shortcomings. The first is that a pre-determined (non-zero) value of $\gamma < \alpha$ is required to construct C_γ. Whether γ is considered 'small' depends on the significance level α, as well as the strength of evidence against H_0. For example, Loh and Richardson [2013] analyzed a dataset where for $\gamma = 0.01$, the maximum p-value was $\max_{\boldsymbol{t} \in C_\gamma} pv^{CO}(\boldsymbol{t}) = 2 \times 10^{-21}$, so that the overall p-value was just $pv_\gamma^{CO} = 0.01$. This highlights the second drawback that the resulting p-value will always be greater than or equal to γ. This is problematic if, as in a *significance* test, we wish to interpret the p-value as measuring the strength of evidence against the 'sharp' null for Compliers.

There is also a subtlety in the procedure. For a given population $\boldsymbol{t} \in T^0(\boldsymbol{n})$, we carried out separate significance tests on the nuisance table $Nuis(\boldsymbol{t})$ and the two-by-two Complier subtable $\boldsymbol{n}^{CO}(\boldsymbol{t})$. The total number of Compliers in each $Z = i$ group differed over hypothetical resamples in $Nuis(\boldsymbol{t})$, but remained fixed (as the row totals) over hypothetical resamples in $\boldsymbol{n}^{CO}(\boldsymbol{t})$. Since only the column and row totals of the complete table $\boldsymbol{m}$ remain fixed over hypothetical rerandomizations from the population $\boldsymbol{t}$, the total number of Compliers in each $Z = i$ group may differ between different resamples. We now describe in the following sections how to carry out a significance test of the 'sharp' null for Compliers using the complete table $\boldsymbol{m}$, without having to consider the nuisance table and Complier subtable separately.

2.6 Likelihood Under a Fixed Population

First we define the likelihood, given the observed data $\boldsymbol{n}$, of a finite population uniquely described by some parameter value $\boldsymbol{t} \in T(\boldsymbol{n})$. In general, without assuming the 'sharp' null for Compliers, each entry in the complete table $\boldsymbol{m}$ may be expressed as a linear function of $\boldsymbol{n}$, $\boldsymbol{t}$, and one of the Complier cell counts, for example m_{HE,z_0}^{CO}. A two-by-eight contingency table such as Table 2.5 may be used to summarize $\boldsymbol{m}$.

Since for any Complier we get to observe exactly one of the potential outcomes $\{Y(1), Y(0)\}$, we generally cannot determine the exact number of Compliers m_{r_Y,z_i}^{CO} even for a fixed value of $\boldsymbol{t}$. Denote the set of complete tables $\boldsymbol{m}$ compatible with both the observed quantities $\boldsymbol{n}$ and population

	$Z=0$	$Z=1$	Column
AT, y_0	$n_{y_0x_1z_0}$	$t_0^{AT}-n_{y_0x_1z_0}$	t_0^{AT}
NT, y_0	$t_0^{NT}-n_{y_0x_0z_1}$	$n_{y_0x_0z_1}$	t_0^{NT}
CO, NR	$(n_{y_0x_0}-t_0^{NT})-b_0$	$t_{NR}^{CO}-(n_{y_0x_0}-t_0^{NT})+b_0$	t_{NR}^{CO}
CO, HE	b_0	$t_{HE}^{CO}-b_0$	t_{HE}^{CO}
CO, HU	$t_{HU}^{CO}+(t_{NR}^{CO}+t_0^{AT}+t_0^{NT})-n_{y_0}+b_0$	$(n_{y_0x_1}-t_0^{AT})-t_{NR}^{CO}+(n_{y_0x_0}-t_0^{NT})-b_0 = n_{y_0}-(t_{NR}^{CO}+t_0^{AT}+t_0^{NT})-b_0$	t_{HU}^{CO}
CO, AR	$t_{AR}^{CO}-(n_{y_1x_1}-t_1^{AT})+t_{HE}^{CO}-b_0 = (t_{AR}^{CO}+t_1^{AT}+t_{HE}^{CO})-n_{y_1x_1}-b_0$	$(n_{y_1x_1}-t_1^{AT})-t_{HE}^{CO}+b_0$	t_{AR}^{CO}
NT, y_1	$t_1^{NT}-n_{y_1x_0z_1}$	$n_{y_1x_0z_1}$	t_1^{NT}
AT, y_1	$n_{y_1x_1z_0}$	$t_1^{AT}-n_{y_1x_1z_0}$	t_1^{AT}
Row	n_{z_0}	n_{z_1}	N

Table 2.5: 'Complete' table $\boldsymbol{m}$ with each entry as a linear function of the observed quantities $\boldsymbol{n}$, the population totals $\boldsymbol{t}$, and one of the Complier cell counts m_{HE,z_0}^{CO} (denoted here as b_0); the two-by-eight table has also been transposed to make the cell counts more readable.

$\boldsymbol{t}$ as $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$. Restricting each entry in $\boldsymbol{m}$ to be a non-negative integer results in the following set of values for m_{HE,z_0}^{CO} that indexes $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$:

$$\left\{b_0^{\min}, b_0^{\min}+1, \ldots, b_0^{\max}\right\}; \tag{2.34}$$

$$b_0^{\min} = \max\left\{(n_{y_0x_0}-t_0^{NT})-t_{NR}^{CO}, n_{y_0}-(t_{HU}^{CO}+t_{NR}^{CO}+t_0^{AT}+t_0^{NT}), t_{HE}^{CO}-(n_{y_1x_1}-t_1^{AT}), 0\right\},$$

$$b_0^{\max} = \min\left\{(n_{y_0x_0}-t_0^{NT}), t_{HE}^{CO}, n_{y_0}-(t_{NR}^{CO}+t_0^{AT}+t_0^{NT}), (t_{AR}^{CO}+t_1^{AT}+t_{HE}^{CO})-n_{y_1x_1}\right\}.$$

Since individuals are sampled without replacement from a finite population, the column totals $\boldsymbol{t}$ remain fixed over hypothetical rerandomizations. Under complete randomization, the row totals of $\boldsymbol{m}$ are also fixed. Note the perhaps obvious point that in general, there is no single entry in $\boldsymbol{m}$ that remains fixed! Hence the randomization distribution for the complete table $\boldsymbol{m}$ is the multiple

hypergeometric distribution:

$$\Pr(\boldsymbol{m} \mid \boldsymbol{t}) = \binom{N}{n_{z_0}, n_{z_1}}^{-1} \prod_{a=1}^{8} \binom{t_{(a)}}{m_{(a),z_1}, m_{(a),z_0}}, \tag{2.35}$$

where the subscript (a) represents the a-th entry in $\boldsymbol{t}$ and $\boldsymbol{m}_i$, and we choose to write the binomial coefficient as:

$$\binom{a+c}{a,c} = \frac{(a+c)!}{a!c!}.$$

It is implicit that the entries $m_{(a),z_i}$ in $\boldsymbol{m}$ are functions of $\boldsymbol{n}$, $\boldsymbol{t}$ and one of the Complier cell counts, for example m^{CO}_{HE,z_0}. The likelihood given the observed data $\boldsymbol{n}$ of a particular population $\boldsymbol{t} \in T(\boldsymbol{n})$ is thus the total probability of of all complete tables $\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$:

$$\begin{aligned} \Pr(\boldsymbol{n} \mid \boldsymbol{t}) &= \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \Pr(\boldsymbol{m} \mid \boldsymbol{t}) = \binom{N}{n_{z_0}, n_{z_1}}^{-1} \text{where } h(\boldsymbol{t}); \\ h(\boldsymbol{t}) &= \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \prod_{a=1}^{8} \binom{t_{(a)}}{m_{(a),z_1}, m_{(a),z_0}}. \end{aligned} \tag{2.36}$$

We shall henceforth refer to the kernel of the likelihood function $h(\boldsymbol{t})$ of an observed dataset $\boldsymbol{n}$ in a given population $\boldsymbol{t}$ as the extended *Neyman-Rubin-Copas*, or *Neyman-Rubin-Copas-Richardson*, likelihood.

2.7 Significance Test of the 'Sharp' Null for Compliers with the Generalized Likelihood Ratio

In a significance test of the 'sharp' null hypothesis for Compliers, our interest is in finding a p-value that measures the strength of evidence in the observed data $\boldsymbol{n}$ against the null, assuming that the null was indeed true. To do so, we will use the generalized likelihood ratio (GLR) test statistic $\lambda(\boldsymbol{n})$ that lets us assess whether allowing a causal effect among Compliers is a *significantly* better explanation for $\boldsymbol{n}$ than assuming no Compliers who are Helped or Hurt. The value of $\lambda(\boldsymbol{n})$ depends only on the observed data $\boldsymbol{n}$, and not on any (unknown) values of the column totals $\boldsymbol{t}$.

We shall first consider the value of $\boldsymbol{t}$ that lends the strongest support to the observed dataset $\boldsymbol{n}$ assuming that there are no Compliers who are Helped or Hurt. The largest probability of $\boldsymbol{n}$ under

the 'sharp' null is:

$$q^{H_0}(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T^0(\boldsymbol{n})} \Pr(\boldsymbol{n} \mid \boldsymbol{t}, H_0) = \binom{N}{n_{z_0}, n_{z_1}}^{-1} \max_{\boldsymbol{t} \in T^0(\boldsymbol{n})} h(\boldsymbol{t}). \tag{2.37}$$

Next, we find the most likely value of $\boldsymbol{t}$ allowing for Compliers who are Helped or Hurt in the population:

$$q(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n})} \Pr(\boldsymbol{n} \mid \boldsymbol{t}) = \binom{N}{n_{z_0}, n_{z_1}}^{-1} \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}). \tag{2.38}$$

The GLR test statistic is then the ratio:

$$\lambda(\boldsymbol{n}) \equiv \frac{q^{H_0}(\boldsymbol{n})}{q(\boldsymbol{n})} = \frac{\max\limits_{\boldsymbol{t} \in T^0(\boldsymbol{n}) \subset T(\boldsymbol{n})} h(\boldsymbol{t})}{\max\limits_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t})}. \tag{2.39}$$

The GLR $\lambda(\boldsymbol{n})$ takes values between 0 and 1, with values closer to 0 indicating stronger evidence *against* the 'sharp' null. However, the 'sharp' null for Compliers is a composite null hypothesis: we must consider a set of sampling distributions for $\lambda(\boldsymbol{n})$, each corresponding to a value of the population parameters $\boldsymbol{t} \in T^0(\boldsymbol{n})$. Similar to the hypothesis testing procedure in Section 2.5, we will evaluate each population $\boldsymbol{t}$ in turn.

For a fixed value of $\boldsymbol{t} \in T^0(\boldsymbol{n})$, the sampling distribution $\Pr(\boldsymbol{n} \mid \boldsymbol{t}, H_0)$ for the observed data $\boldsymbol{n}$ under H_0 induces a null probability distribution for the GLR $\lambda(\boldsymbol{n})$. Using the tilde accent ($\sim$) to indicate random quantities that vary over hypothetical rerandomizations, one may conceptually generate assignments in the complete table $\tilde{\boldsymbol{m}}$, either through complete enumeration using for example [Gail and Mantel, 1977], or via Monte Carlo simulations. The cell counts in $\tilde{\boldsymbol{m}}$ are then aggregated using the linear constraints described in Table 2.3 to obtain the sample space of possibly observable datasets $\tilde{\boldsymbol{n}}$, which we denote $\Omega(\boldsymbol{t})$. For each $\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t})$, we then calculate the value of the test statistic $\lambda(\tilde{\boldsymbol{n}})$. The sampling distribution $\Pr(\tilde{\boldsymbol{n}}|\boldsymbol{t}, H_0)$ induces a randomization distribution for $\lambda(\tilde{\boldsymbol{n}})$. The population-specific p-value is then the total probability of observing a dataset $\tilde{\boldsymbol{n}}$ that is at least as 'far away' from the 'sharp' null as the observed dataset $\boldsymbol{n}$:

$$\begin{aligned} pv^{H_0}(\boldsymbol{n}; \boldsymbol{t}) &= \Pr(\lambda(\tilde{\boldsymbol{n}}) \leq \lambda(\boldsymbol{n}) \mid \boldsymbol{t}, H_0) \\ &= \sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t})} \Pr(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}, H_0) \times \mathbb{1}\{\lambda(\tilde{\boldsymbol{n}}) \leq \lambda(\boldsymbol{n})\}, \end{aligned} \tag{2.40}$$

where $\mathbb{1}\{E\}$ returns the value 1 if condition E is satisfied, and 0 otherwise. The overall p-value is the maximum among all population-specific p-values:

$$pv^{H_0}(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T^0(\boldsymbol{n})} pv^{H_0}(\boldsymbol{n}; \boldsymbol{t}). \tag{2.41}$$

If the 'sharp' null for Compliers is true, then the probability of obtaining a value of the test statistic $\lambda(\tilde{\boldsymbol{n}})$ that is no larger than the observed value $\lambda(\boldsymbol{n})$ is *at most* $pv^{H_0}(\boldsymbol{n})$, regardless of the unknown population $\boldsymbol{t}$. In other words, the overall p-value $pv^{H_0}(\boldsymbol{n})$ is an *upper-bound* on the relative frequency that a dataset at least as 'extreme' as the observed $\boldsymbol{n}$ will arise, over hypothetical rerandomizations from a population with zero Compliers who are Helped or Hurt.

2.8 Comparison of Exact P-values

How does the effort to account for the observed noncompliance using the GLR compare to performing Fisher's exact test for the ITT($Z \to Y$) effect? After all, the ITT p-value is also an exact p-value, and summarizes evidence against the global 'sharp' null, rather than the local 'sharp' null for Compliers.

To compare the resulting p-values from the two procedures, consider a toy observed dataset from a double-blind randomized controlled trial where 13 out of 24 individuals were randomly assigned to treatment $Z = 1$, of which 10 were observed to have a positive outcome $Y = 1$. Of the remaining 11 assigned to control $Z = 0$, only 4 had a positive outcome. The ITT p-value from performing Fisher's exact test on the observed (Z, Y) dataset is then 0.095.

Suppose we then learn there were individuals who did not comply with the assigned treatment, but because patients assigned to the control group $Z = 0$ do not have access to the active treatment outside of the trial, $Z = 0 \Rightarrow X = 0$. Such one-sided compliance implies that there are only Compliers and Never Takers in the population (so that there are no Always Takers). Hence the number of Compliers with observed outcomes $Y = k$ in the treatment group $Z = 1$ is now directly observable from the data as $n_{y_k x_1 z_1}$. The observed counts taking the treatment noncompliance into account are shown in Table 2.6. There are two structural zeros due to the one-sided compliance, and

we shall assume that Z has no effect on Y other than through X, so that the exclusion restriction (2.3) holds.

Number of individuals	$Y=1$		$Y=0$		Row
	$X=1$	$X=0$	$X=1$	$X=0$	
$Z=0$	*0*	4	*0*	7	11
$Z=1$	8	2	2	1	13

Table 2.6: Observed toy dataset $\boldsymbol{n}$ from a randomized controlled trial; note the italicized structural zeroes due to the one-sided noncompliance where $Z=0 \Rightarrow X=0$.

We then carry out our procedure as summarized in Figure 2.2. Given the observed dataset $\boldsymbol{n}$, we enumerate all possible populations $\boldsymbol{t} \in T^0(\boldsymbol{n})$ under the 'sharp' null for Compliers. We denote $\boldsymbol{t}$ in this example simply as (t_0^{NT}, t_1^{NT}), since there are no Always Takers in the population, and the total number of Compliers under the 'sharp' null, given the observed dataset and (t_0^{NT}, t_1^{NT}) is just $t_{NR}^{CO} = n_{y_0} - t_0^{NT}$ and $t_{AR}^{CO} = n_{y_1} - t_1^{NT}$.

For each fixed population $\boldsymbol{t}$, for example $\boldsymbol{t} = (4, 2)$, we generate the sample space $\Omega(\boldsymbol{t})$ and then find the corresponding GLR test statistics $\lambda(\tilde{\boldsymbol{n}})$. The population-specific p-value $pv^{H_0}(\boldsymbol{n}; \boldsymbol{t})$ is then the lower-tail probability $\Pr(\lambda(\tilde{\boldsymbol{n}})|\boldsymbol{t})$ for the critical value $\lambda(\boldsymbol{n})$; for example $pv^{H_0}(\boldsymbol{n}; \boldsymbol{t} = (4, 2)) = 0.018$. The maximum p-value over all populations $\boldsymbol{t} \in T^0(\boldsymbol{n})$ is 0.028, indicating stronger evidence against the 'sharp' null for Compliers (and thus the global 'sharp' null) than the ITT test!

However, whether accounting for the observed noncompliance using the GLR results in a smaller or larger p-value than the ITT test depends on the observed data. In Table 2.6, among the 10 people in the observed $(Z = Y = 1)$ strata, we observed 8 with $X = 1$. How would the overall p-value change if the $(Z = Y = 1)$ margin was fixed, so that the ITT table for the (Z, Y) margins remain the same, but with different observed counts for the $(Z = X = Y = 1)$ strata? We carry out significance tests using the GLR for $n_{y_1 x_1 z_1} = 0, \ldots, 10$ and plot the corresponding GLR p-values in Figure 2.3. For significance levels between 0.02 and 0.08, the GLR test statistic has

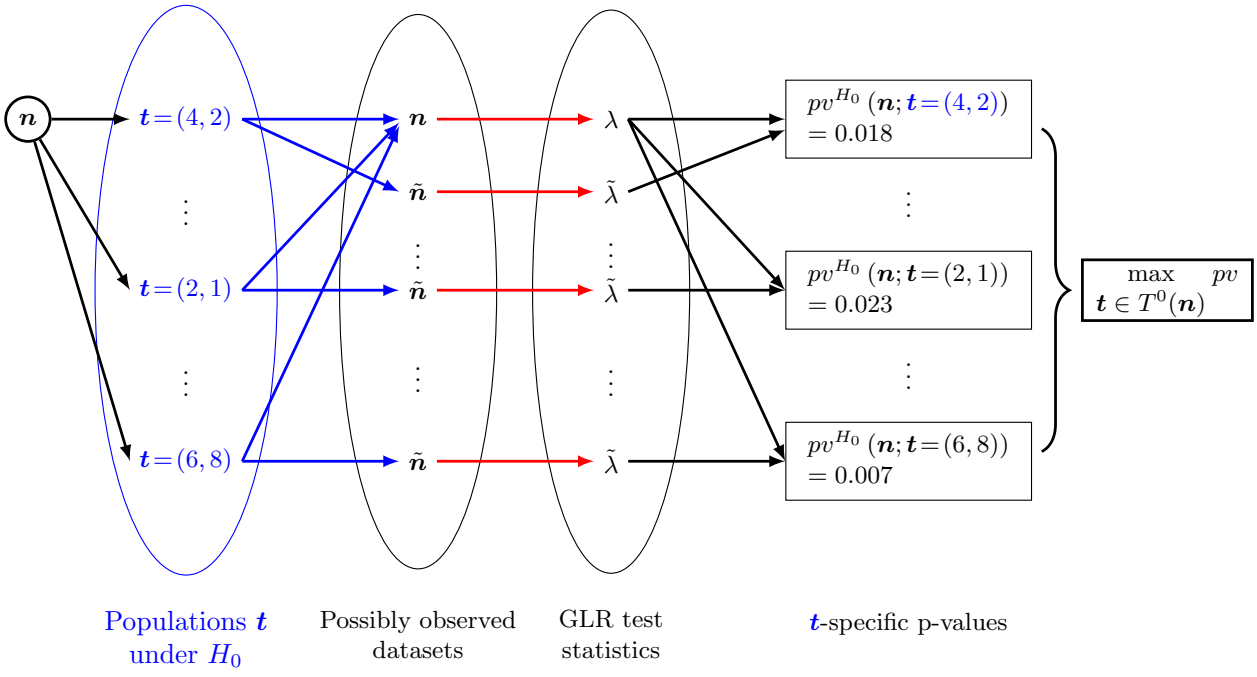

Figure 2.2: Summary of the procedure to carry out a significance test of the 'sharp' null for Compliers H_0, for the observed dataset in Table 2.6; here we assume no Always Takers and no Defiers in the population; the population parameters $\boldsymbol{t}$ are represented here as (t_1^{NT}, t_0^{NT}).

greater statistical power than the other two test statistics.

When $n_{y_1x_1z_1}$ is close to the observed margin of $n_{y_1z_1} = 10$, the resulting p-value for testing the 'sharp' null for Compliers using the GLR (in red) is the same as the ITT p-value (green horizontal line). However, as $n_{y_1x_1z_1}$ decreases, the GLR p-value falls below the ITT p-value before rising past the ITT p-value when $n_{y_1x_1z_1}$ is close to zero. For example, if $n_{y_1x_1z_1} = 1$, the p-value using the GLR in testing the 'sharp' null for Compliers is about 0.5, suggesting a lack of evidence against a zero ACE.

Note that $n_{y_1x_1z_1} = 1$ implies $n_{y_1x_0z_1} = 9$ (since $n_{y_1z_1} = 10$), so that there are nine Never Takers with an observed outcome of $Y=1$ in the $Z=1$ group but at most four in the $Z=0$ group (since $n_{y_1x_0z_0} = 4$)! Such an imbalance suggests a direct effect of Z on Y among the Never Takers.

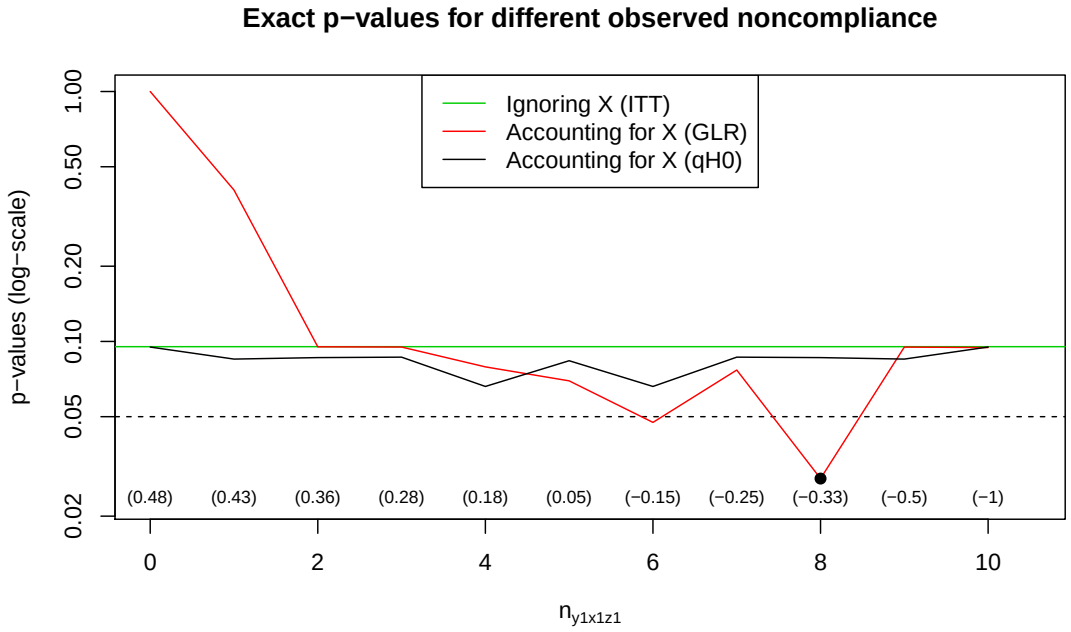

Figure 2.3: Exact p-values using (i) the GLR test statistic (in red) that accounts for the noncompliance X, (ii) Fisher's exact test on the (Z, Y) table (in green), and (iii) the q^{H_0} test statistic that also accounts for X; the significance level of 0.05 is plotted as the broken line; the p-value using the GLR test for the observed dataset in Table 2.6 is indicated by the black solid circle; the values in brackets at the bottom of the plot are the lower bounds of the $ACDE_{NT}(x_0)$ in (2.43).

If we define the Average Controlled Direct Effect for Never Takers as

$$ACDE_{NT}(x_0) = \mathrm{E}[Y(x = 0, z = 1) - Y(x = 0, z = 0) \mid NT], \tag{2.42}$$

then under exclusion restriction where there is no direct effect of Z on Y, $ACDE_{NT}(x_0) = 0$. A lower bound for the $ACDE_{NT}(x_0)$ [Cai et al., 2008] is

$$p(y_1|x_0, z_1) - \min\left\{\frac{p(y_1, x_0|z_0)}{p(x_0|z_1)}, 1\right\}. \tag{2.43}$$

The values of the lower bound (2.43) for $n_{y_1x_1z_1} = 0, \ldots, 10$ are shown at the bottom of Figure 2.3, with positive values equivalent to an IV inequality (under monotonicity) being violated empirically

[Richardson et al., 2011]. For small values of $n_{y_1 x_1 z_1}$, the lower bound is strictly greater than zero, suggesting that empirically, there is a direct effect of Z on Y for Never Takers.

A possible explanation is that under the ITT null, both the 'sharp' null for Compliers *and* exclusion restriction for the Never Takers must hold. Evidence against a zero direct effect of Z on Y for Never Takers then suggests evidence against the ITT null, even if the 'sharp' null for Compliers is true. The positive values of the lower bound for the $ACDE_{NT}(x_0)$ for small values of $n_{y_1 x_1 z_1}$ in this example suggest that there is evidence the overall ITT effect may be due to the direct effect for Never Takers. With larger values of $n_{y_1 x_1 z_1}$, the evidence of the direct effect for Never Takers contributing to the ITT effect is reduced, but the evidence of contribution of the CACE to the ITT effect increases. The evidence against the overall ITT effect then remains the same.

In addition, the (maximum) p-value using the test statistic q^{H_0} (numerator of the GLR) is plotted. Similar to the ITT p-value, it is unable to distinguish between evidence against exclusion restriction and evidence against a non-zero causal effect for the Compliers. Alternatively, if one was to replace the GLR test statistic with the ITT test statistic (the bivariate hypergeometric probability of the induced (Z, Y) margins that ignore X), the resulting (maximum) p-value would just be equal to the p-value from the ITT test. This is because when the 'sharp' null for Compliers is true, the observed Z and Y margins in the complete table (Table 2.3) are the same regardless of the specific population $\boldsymbol{t} \in T^0(\boldsymbol{n})$. By the Chu-Vandermonde convolution, the sampling distribution of the ITT test statistic (ignoring information from X) under some population $\boldsymbol{t}$ (that accounts for X) is just equal to the sampling distribution under the ITT table.

2.8.1 Statistical Power

To examine the statistical power of the GLR significance test, consider a single population that the observed dataset in Table 2.6 could have arisen from, specifically the following population:

$$t_1^{NT} = 4, t_0^{NT} = 1, t_{HE}^{CO} = 15, t_{HU}^{CO} = 4, t_{AR}^{CO} = t_{NR}^{CO} = 0.$$

For each possibly observable dataset in the sample space $\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t})$, we find the maximum p-value in a significance test using each of the three test statistics under comparison in Figure 2.3. The sampling distribution for $\tilde{\boldsymbol{n}}$ then provides an empirical distribution for the respective p-values under this population $\boldsymbol{t}$, which we plot in Figure 2.4 below.

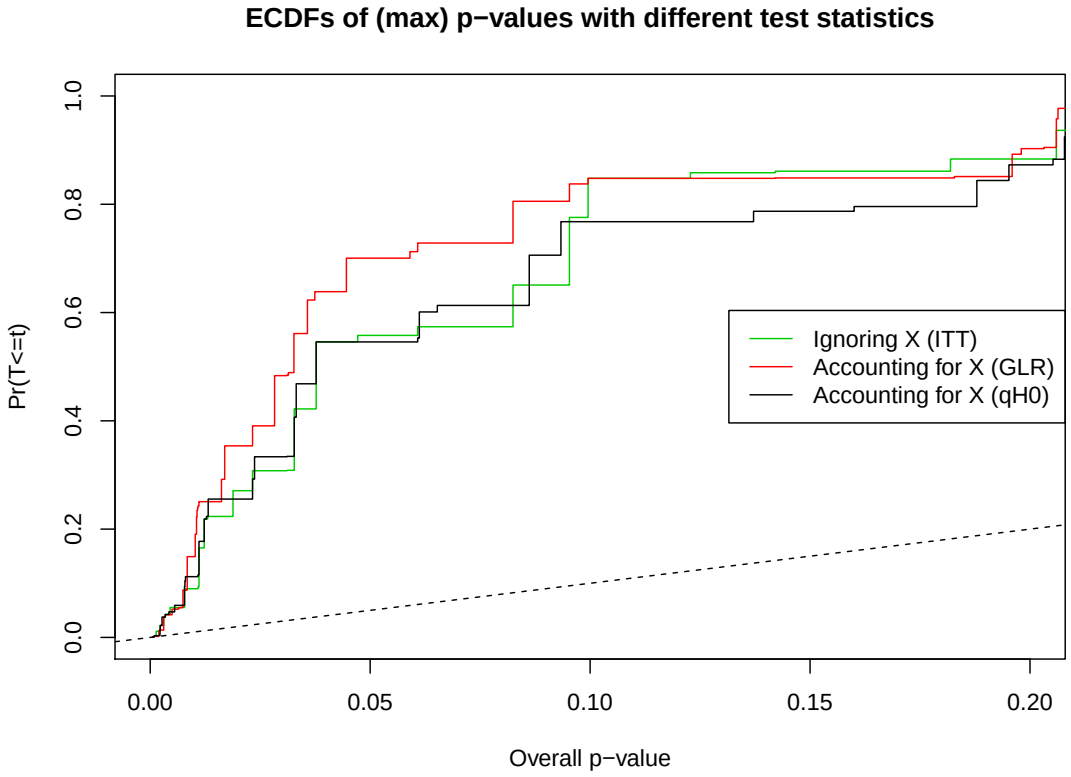

Figure 2.4: Empirical cumulative density functions (ECDFs) of the (maximum) p-values using (i) the GLR test statistic (in red) that accounts for the noncompliance X, (ii) Fisher's exact test on the (Z, Y) table (in green), and (iii) the q^{H_0} test statistic that also accounts for X; all possibly observable datasets were sampled from the population $t_1^{NT} = 4, t_0^{NT} = 1, t_{HE}^{CO} = 15, t_{HU}^{CO} = 4, t_{AR}^{CO} = t_{NR}^{CO} = 0$; the diagonal is plotted as the broken line.

2.9 Proposed Improvements for Computational Efficiency

2.9.1 Sample space for an observed dataset under H_0

Under the 'sharp' null for Compliers, the same possibly observable dataset $\tilde{\boldsymbol{n}}$ may arise from multiple populations $\boldsymbol{t} \in T^0(\boldsymbol{n})$; in other words, the intersection of the sample spaces for $\boldsymbol{t}$ and $\boldsymbol{t}' \neq \boldsymbol{t}$ is non-empty:

$$\Omega(\boldsymbol{t}) \cap \Omega(\boldsymbol{t}') \neq \emptyset.$$

Naïvely calculating the value of the GLR test statistic $\lambda(\tilde{\boldsymbol{n}})$ every time $\tilde{\boldsymbol{n}}$ arises when generating the sample space $\Omega(\boldsymbol{t})$ under a fixed population $\boldsymbol{t}$ (via either complete enumeration or Monte Carlo simulation) would result in duplicate calculations. Under H_0, $\lambda(\tilde{\boldsymbol{n}})$ is a function only of the observable quantities $\tilde{\boldsymbol{n}}$ *and not of a specific fixed value of* $\boldsymbol{t}$. A computationally more efficient approach would be to first enumerate the sample space of all possibly observed datasets $\tilde{\boldsymbol{n}}$ across all populations $\boldsymbol{t} \in T^0(\boldsymbol{n})$, and then calculate the values of the test statistic $\lambda(\tilde{\boldsymbol{n}})$. The sample space $\Omega^0(\boldsymbol{n})$ is thus defined as:

$$\Omega^0(\boldsymbol{n}) \equiv \bigcup_{\boldsymbol{t} \in T^0(\boldsymbol{n})} \Omega(\boldsymbol{t}). \tag{2.44}$$

The sample space $\Omega^0(\boldsymbol{n})$ is the intersection of the seven-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^7$, and the convex polyhedron described with the inequalities (2.45)-(2.49):

$$0 \leq \tilde{n}_{y_k x_{1-i} z_i} \leq n_{y_k x_{1-i} z_i} + n_{y_k x_{1-i} z_{1-i}}, \qquad i,k = 0,1; \tag{2.45}$$

$$\max\{0, n_{y_i x_i z_{1-i}} - \tilde{n}_{y_i x_i z_{1-i}}\} \leq \tilde{n}_{y_i x_i z_i} \leq n_{y_i} - n_{y_i x_{1-i} z_i} - \tilde{n}_{y_i x_i z_{1-i}}, \qquad i = 0,1; \tag{2.46}$$

$$0 \leq \tilde{n}_{y_{1-i} x_i z_i}, \qquad i = 0,1; \tag{2.47}$$

$$\sum_{j=0}^{1} \sum_{i=0}^{1} \tilde{n}_{y_k x_j z_i} = n_{y_k}, \qquad k = 0,1; \tag{2.48}$$

$$\sum_{k=0}^{1} \sum_{j=0}^{1} \tilde{n}_{y_k x_j z_i} = n_{z_i}, \qquad i = 0,1. \tag{2.49}$$

Detailed steps on how to obtain the inequalities using `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015] are provided in Appendix D. The population-specific probabilities of each $\tilde{\boldsymbol{n}}$ in a

fixed population $\boldsymbol{t}$ may be found using the Neyman-Rubin-Copas-Richardson likelihood (2.36) when evaluating the p-value.

2.9.2 Maximum Complete Probability Ratio (MPR)

One way to find the maximum likelihood solution $q(\boldsymbol{n})$ for a given observed $\boldsymbol{n}$ is to calculate all possible multivariate hypergeometric probabilities $\Pr(\boldsymbol{m} \mid \boldsymbol{t})$ in the two-by-eight complete table $\boldsymbol{m}$ for all possible populations $\boldsymbol{t} \in T(\boldsymbol{n})$, and then sum the probabilities for the (possibly multiple) complete tables in $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ to obtain the likelihood $\Pr(\boldsymbol{n} \mid \boldsymbol{t})$. This is computationally intensive and is $\mathcal{O}(n^4)$, where $n = \max_{i,k} n_{y_k x_i z_i}$.

Instead, consider the maximum conditional probability of the complete table $\boldsymbol{m}$:

$$q^{complete}(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n})} \left\{ \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \Pr(\boldsymbol{m} \mid \boldsymbol{t}) \right\}. \tag{2.50}$$

Following [Loh and Richardson, 2015], we may then maximize over both $\boldsymbol{t}$ and $\boldsymbol{m}$ jointly, using integer programs such as [Oberhofer and Kaufmann, 1987] and [Johnson and Kotz, 1969] to solve the corresponding discrete optimization problems. Finding the global maximum then requires calculating only $\mathcal{O}(n^2)$ hypergeometric probabilities.

Under the 'sharp' null for Compliers, if there are no Compliers who are Helped or Hurt in a given population $\boldsymbol{t}$, then there is only one possible complete table $\boldsymbol{m}$ in $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ since $b_0^{\min} = b_0^{\max} = 0$ (see (2.34)). In other words, for $\boldsymbol{t} \in T^0(\boldsymbol{n})$, the likelihood given the observed data is just the probability of the single complete table:

$$\Pr(\boldsymbol{n} \mid \boldsymbol{t}, H_0) = \Pr(\boldsymbol{m} \mid \boldsymbol{t}). \tag{2.51}$$

The ratio of the maximum complete probabilities (MPR) is then:

$$MPR(\boldsymbol{n}) \equiv \frac{q^{H_0}(\boldsymbol{n})}{q^{complete}(\boldsymbol{n})} \equiv \frac{\max\limits_{\boldsymbol{t} \in T^0(\boldsymbol{n}) \subset T(\boldsymbol{n})} \Pr(\boldsymbol{n} \mid \boldsymbol{t}, H_0)}{\max\limits_{\boldsymbol{t} \in T(\boldsymbol{n})} \left\{ \max\limits_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \Pr(\boldsymbol{m} \mid \boldsymbol{t}) \right\}}. \tag{2.52}$$

The significance test procedure using the MPR is available in the `noncompliance` package [Loh and Richardson, 2016] in `R` [R Core Team, 2015].

2.9.3 Connection to likelihood under perfect compliance

We now examine the relationship between the Neyman-Rubin-Copas-Richardson likelihood function $h(\boldsymbol{t})$ in (2.36) under treatment noncompliance and the Neyman-Rubin-Copas likelihood function under perfect compliance.

Lemma 2. *Under a fixed population $\boldsymbol{t} \in T(\boldsymbol{n})$, the Neyman-Rubin-Copas-Richardson likelihood function $h(\boldsymbol{t})$ given the observed dataset $\boldsymbol{n}$ is a product of the kernel of the multivariate hypergeometric probability for the Always Takers and Never Takers, and the kernel of the Neyman-Rubin-Copas likelihood given the Complier subtable induced by $\boldsymbol{t}$.*

Proof. Denote $\boldsymbol{\psi} \equiv (t_0^{AT}, t_0^{NT}, t_1^{AT}, t_1^{AT})$ as the sub-vector of $\boldsymbol{t}$ containing the total number of Always Takers and Never Takers, and $\boldsymbol{t}^{CO} \equiv (t_{NR}^{CO}, t_{HE}^{CO}, t_{HU}^{CO}, t_{AR}^{CO})$ as the sub-vector of $\boldsymbol{t}$ for the Compliers. Under a fixed population $\boldsymbol{t} \in T(\boldsymbol{n})$, $h(\boldsymbol{t})$ may be decomposed as:

$$
\begin{aligned}
h(\boldsymbol{t}) &= \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \prod_{a=1}^{8} \binom{t_{(a)}}{m_{(a),z_1}, m_{(a),z_0}} \\
&= \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \underbrace{\prod_{k=0}^{1} \binom{t_k^{AT}}{n_{y_k x_1 z_0}, t_k^{AT} - n_{y_k x_1 z_0}} \binom{t_k^{NT}}{t_k^{NT} - n_{y_k x_0 z_1}, n_{y_k x_0 z_1}}}_{= g(\boldsymbol{\psi})} \\
&\quad \times \prod_{r_Y \in \{NR, HE, HU, AR\}} \binom{t_{r_Y}^{CO}}{m_{r_Y,z_0}^{CO}, m_{r_Y,z_1}^{CO}} \\
&= g(\boldsymbol{\psi}) \underbrace{\sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \left[\prod_{r_Y \in \{NR, HE, HU, AR\}} \binom{t_{r_Y}^{CO}}{m_{r_Y,z_0}^{CO}, m_{r_Y,z_1}^{CO}} \right]}_{= f(\boldsymbol{t})}.
\end{aligned}
$$

For $\boldsymbol{\psi} \equiv (t_0^{AT}, t_0^{NT}, t_1^{AT}, t_1^{AT})$, $g(\boldsymbol{\psi})$ is the kernel of the multivariate hypergeometric probability for the Always Takers and Never Takers. We can factor $g(\boldsymbol{\psi})$ out of the sum over $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$ because the terms in $g(\boldsymbol{\psi})$ involving the Always Takers and Never Takers are a function only of $\boldsymbol{\psi}$ and $\boldsymbol{n}$, and not m_{HE,z_0}^{CO} which $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$ is indexed by.

In contrast, $f(\boldsymbol{t})$ is clearly a function of $\boldsymbol{t}, \boldsymbol{n}$ and m^{CO}_{HE,z_0}. We now claim that $f(\boldsymbol{t})$ is actually the Neyman-Rubin-Copas likelihood (1.8) of the population $\boldsymbol{t}^{CO}$ given the two-by-two Complier subtable $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$. $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$ was previously $\boldsymbol{n}^{CO}(\boldsymbol{t})$ in (2.30), but here it is explicit that the entries of $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$ are functions of $\boldsymbol{n}$ and $\boldsymbol{\psi}$ (not the entire parameter $\boldsymbol{t}$):

$$n^{CO}_{y_k x_0}(\boldsymbol{\psi}) = n_{y_k x_0} - t^{NT}_k, \quad n^{CO}_{y_k x_1}(\boldsymbol{\psi}) = n_{y_k x_1} - t^{AT}_k; k = 0, 1. \tag{2.53}$$

The proof then follows the approach under perfect compliance (but with the superscript CO in the notation). For the 'observed' table $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$, express each entry in the *complete two-by-four* table $\boldsymbol{m}^{CO}$ as a linear function of $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$, $\boldsymbol{t}^{CO}$ and the cell count m^{CO}_{HE,z_0}, as per the entries in Table 1.3. The set of complete tables $\boldsymbol{m}^{CO}$ compatible with $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$ and $\boldsymbol{t}^{CO}$ is then $\mathcal{M}(\boldsymbol{t}^{CO}; \boldsymbol{n}^{CO}(\boldsymbol{\psi}))$, and is indexed by the one-dimensional set of possible values for m^{CO}_{HE,z_0} in (1.7) as:

$$\left\{ b_0^{CO,\min}, b_0^{CO,\min} + 1, \ldots, b_0^{CO,\max} \right\}; \tag{2.54}$$

$$b_0^{CO,\min} = \max\left\{ n^{CO}_{y_0}(\boldsymbol{t}) - (t^{CO}_{NR} + t^{CO}_{HU}), n^{CO}_{y_0 x_0}(\boldsymbol{t}) - t^{CO}_{NR}, t^{CO}_{HE} - n^{CO}_{y_1 x_1}(\boldsymbol{t}), 0 \right\},$$

$$b_0^{CO,\max} = \min\left\{ n^{CO}_{y_1 x_0}(\boldsymbol{t}) + n^{CO}_{y_0}(\boldsymbol{t}) - (t^{CO}_{HU} + t^{CO}_{NR}), n^{CO}_{y_0 x_0}(\boldsymbol{t}), t^{CO}_{HE}, n^{CO}_{y_0}(\boldsymbol{t}) - t^{CO}_{NR} \right\}.$$

Substituting the values of $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$ with the corresponding functions of $\boldsymbol{n}$ and $\boldsymbol{\psi}$ from (2.53), we find that the set of values for m^{CO}_{HE,z_0} that indexes $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ in (2.34) is the same as that in (2.54) that indexes $\mathcal{M}(\boldsymbol{t}^{CO}; \boldsymbol{n}^{CO}(\boldsymbol{\psi}))$. In other words,

$$b_0^{\min} = b_0^{CO,\min}, \quad b_0^{\max} = b_0^{CO,\max}. \tag{2.55}$$

This implies that

$$f(\boldsymbol{t}) = \sum_{\boldsymbol{m}^{CO} \in \mathcal{M}(\boldsymbol{t}^{CO}; \boldsymbol{n}^{CO}(\boldsymbol{\psi}))} \left[\prod_{r_Y \in \{NR, HE, HU, AR\}} \binom{t^{CO}_{r_Y}}{m^{CO}_{r_Y, z_0}, m^{CO}_{r_Y, z_1}} \right] \equiv h_{CO}(\boldsymbol{t}^{CO}), \tag{2.56}$$

where $h_{CO}(\boldsymbol{t}^{CO})$ is the Neyman-Rubin-Copas likelihood of the population $\boldsymbol{t}^{CO}$ given the two-by-two Complier subtable $\boldsymbol{n}^{CO}(\boldsymbol{\psi})$. (The subscript CO is used only to distinguish the Neyman-Rubin-Copas likelihood function in (1.8) from the Neyman-Rubin-Copas-Richardson likelihood function also denoted $h(\boldsymbol{t})$ here.) □

Theorem 3. *Given an observed table $\boldsymbol{n}$, the profile likelihood for a fixed value of the nuisance parameter ψ is the product of the kernel of the multivariate hypergeometric probability for the Always Takers and Never Takers under ψ, and the maximum Neyman-Rubin-Copas likelihood given the Complier subtable induced by ψ.*

Proof. For a fixed value of the nuisance parameter ψ, say ψ', let $\boldsymbol{n}^{CO}(\psi')$ be the induced two-by-two Complier subtable, with entries in terms of $\boldsymbol{n}$ and ψ' as stated in (2.53). Denote the total number of Compliers for given ψ' as $N^{CO}(\psi') \equiv \left(N - \sum_{k=0}^{1}\left(t_k'^{NT} + t_k'^{AT}\right)\right) = \left[\sum_{i=0}^{1}\sum_{j=0}^{1} n_{y_j x_i}^{CO}(\psi')\right]$.

Since the parameter space $T(\boldsymbol{n})$ lies in a seven-dimensional convex polyhedron, and we have fixed the values of the four elements in ψ, denote the resulting three-dimensional convex polyhedron for $(t_{NR}^{CO}, t_{HE}^{CO}, t_{HU}^{CO})$ as $T(\boldsymbol{n}; \psi')$. $T(\boldsymbol{n}; \psi')$ is the intersection of $T(\boldsymbol{n})$ and the hyperplanes corresponding to fixed ψ, and is characterized by the following inequalities:

$$
\begin{aligned}
0 \leq t_{NR}^{CO} &\leq n_{y_0x_0}^{CO}(\psi') + n_{y_0x_1}^{CO}(\psi'), && (2.57)\\
0 \leq t_{HE}^{CO} &\leq n_{y_0x_0}^{CO}(\psi') + n_{y_1x_1}^{CO}(\psi'), && (2.58)\\
0 \leq t_{HU}^{CO} &\leq n_{y_1x_0}^{CO}(\psi') + n_{y_0x_1}^{CO}(\psi'), && (2.59)\\
n_{y_0x_0}^{CO}(\psi') \leq t_{NR}^{CO} + t_{HE}^{CO} &\leq N^{CO}(\psi') - n_{y_1x_0}^{CO}(\psi'), && (2.60)\\
n_{y_0x_1}^{CO}(\psi') \leq t_{NR}^{CO} + t_{HU}^{CO} &\leq N^{CO}(\psi') - n_{y_1x_1}^{CO}(\psi'), && (2.61)\\
n_{y_0x_0}^{CO}(\psi') + n_{y_0x_1}^{CO}(\psi') \leq t_{NR}^{CO} + t_{HE}^{CO} + t_{HU}^{CO} &\leq N^{CO}(\psi'). && (2.62)
\end{aligned}
$$

The inequalities (2.57)$-$(2.62) were obtained by rewriting the inequalities (2.13)$-$(2.25) in terms of $(t_{NR}^{CO}, t_{HE}^{CO}, t_{HU}^{CO})$, $\boldsymbol{n}$ and ψ', and then substituting the expressions involving $\boldsymbol{n}$ and ψ' with the entries in $\boldsymbol{n}^{CO}(\psi')$ using (2.53).

The set of inequalities (2.57)$-$(2.62) are then equal to the corresponding set of inequalities describing the parameter space $T_{CO}(\boldsymbol{n}^{CO}(\psi'))$ found with the inequalities in (1.9) for the Complier subtable $\boldsymbol{n}^{CO}(\psi')$. In other words, the parameter space $T_{CO}(\boldsymbol{n}^{CO}(\psi'))$ for the Complier subtable $\boldsymbol{n}^{CO}(\psi')$ induced by ψ', is equivalent to the subspace $T(\boldsymbol{n}; \psi') \subset T(\boldsymbol{n})$ for fixed ψ'. From Lemma 2, for some population $\boldsymbol{t} \in T(\boldsymbol{n})$ where the sub-vector ψ takes the value ψ', the kernel of

the likelihood function given an observed table $\boldsymbol{n}$ may be written as:

$$h(\boldsymbol{t}) = g(\boldsymbol{\psi}')h_{CO}(\boldsymbol{t}^{CO}).$$

The profile likelihood for fixed $\boldsymbol{\psi}'$ is just:

$$\begin{aligned}
\hat{h}(\boldsymbol{\psi}') &\equiv \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) : \boldsymbol{\psi} = \boldsymbol{\psi}'\}} h(\boldsymbol{t}) \\
&= g(\boldsymbol{\psi}') \left[\max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) : \boldsymbol{\psi} = \boldsymbol{\psi}'\}} h_{CO}(\boldsymbol{t}^{CO}) \right] \\
&= g(\boldsymbol{\psi}') \left[\max_{\boldsymbol{t}^{CO} \in T(\boldsymbol{n}; \boldsymbol{\psi}')} h_{CO}(\boldsymbol{t}^{CO}) \right] \\
&= g(\boldsymbol{\psi}') \left[\max_{\boldsymbol{t}^{CO} \in T(\boldsymbol{n}^{CO}(\boldsymbol{\psi}'))} h_{CO}(\boldsymbol{t}^{CO}) \right] \\
&= g(\boldsymbol{\psi}')\hat{h}_{CO}(\boldsymbol{\psi}');
\end{aligned}$$

where $\hat{h}_{CO}(\boldsymbol{\psi}')$ is the global maximum likelihood for the two-by-two Complier subtable $\boldsymbol{n}^{CO}(\boldsymbol{\psi}')$ given $\boldsymbol{n}$ and fixed $\boldsymbol{\psi}'$. □

Conjecture 2 (Extended Never Only Four)**.** *For an observed table $\boldsymbol{n}$, let $\Psi(\boldsymbol{n})$ be the set of unique values for the sub-vector $\boldsymbol{\psi}$ in the parameter space $T(\boldsymbol{n})$. Denote the global maximum likelihood as:*

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}) = \max_{\boldsymbol{\psi} \in \Psi(\boldsymbol{n})} \hat{h}(\boldsymbol{\psi}) = \max_{\boldsymbol{\psi} \in \Psi(\boldsymbol{n})} g(\boldsymbol{\psi})\hat{h}_{CO}(\boldsymbol{\psi}), \tag{2.63}$$

where the last equality is due to Theorem 3. Following Conjecture 1 ('Never Only Four') that the maximum value $\hat{h}_{CO}(\boldsymbol{\psi})$ is never attained by only one population with all four types, it follows that the maximum likelihood $\hat{h}$ is never attained by only one population with all eight types. In other words, there is at least one population $\hat{\mathbf{t}}$ where

$$h(\hat{\boldsymbol{t}}) = \hat{h}, \quad \min_i \hat{\boldsymbol{t}}_i = 0. \tag{2.64}$$

This then implies $\mathcal{M}(\hat{\boldsymbol{t}}; \boldsymbol{n})$ contains just one complete table $\boldsymbol{m}$ (since $b_0^{\min} = b_0^{\max}$), so that the global maximum likelihood $\hat{h}$ given the observed data may be found as the multivariate hypergeo-

metric probability of a single complete table:

$$\Pr(\boldsymbol{n} \mid \hat{\boldsymbol{t}}) = \Pr\left(\boldsymbol{m} \mid \hat{\boldsymbol{t}}\right). \tag{2.65}$$

This implies that the maximum complete probability ratio (MPR) and generalized likelihood ratio (GLR) test statistics are equal:

$$GLR(\boldsymbol{n}) \equiv \frac{q^{H_0}(\boldsymbol{n})}{q(\boldsymbol{n})} = \frac{q^{H_0}(\boldsymbol{n})}{q^{complete}(\boldsymbol{n})} \equiv MPR(\boldsymbol{n}). \tag{2.66}$$

Chapter 3

RANDOMIZATION-BASED SIGNIFICANCE TESTS FOR CAUSAL HYPOTHESES IN PRINCIPAL STRATA

3.1 Introduction

We have presented an exact significance test of the 'sharp' null hypothesis in the Complier subpopulation under the binary instrumental variable (IV) model. The generalized likelihood ratio (GLR) test statistic was used to assess whether allowing for a causal effect among Compliers was a significantly better explanation for the observed data than the 'sharp' null. Here we propose two possible extensions.

First, we consider more general causal hypotheses for the Compliers based on the observed dataset. By testing a set of such hypotheses, a randomization-based confidence interval for the Complier Average Causal Effect (CACE) may be constructed. However, there is a complication that the Complier subpopulation is only partially identified, so that the (implicit) assumption of a fixed population size under perfect compliance does not hold. We propose a parametrization that treats the total number of Compliers as a nuisance parameter, and present the resulting confidence interval for a small toy dataset.

Second, we consider relaxing the exclusion restriction assumption in (2.3) to allow for an (individual-level) direct effect of Z on Y. Specifically we consider a randomized trial with an intermediate post-treatment infection X. The final post-infection outcome Y is only observable for infected individuals ($X = 1$), so that the randomized treatment assignment Z now has a direct effect on the outcome Y, and may no longer be considered an instrument. Under such a circumstance, our interest is in testing the 'sharp' causal null hypothesis of no (individual-level) direct effect of Z on Y for a principal stratum called the 'Always Infected' or 'Doomed':

$$H_0 : \quad Y(x = 1, z = 0) = Y(x = 1, z = 1). \tag{3.1}$$

'Always Infected' individuals have potential outcomes $X(z) = 1$. We will assume that there are no individuals with potential outcomes $\{X(z) = z\}$; in words, no one in the population consistently gets infected under treatment and remains uninfected under control. Infected individuals in the treatment group $Z = 1$ must then be of type 'Always Infected'. However, infected individuals in the control group $Z = 0$ may be either 'Always Infected' or 'Protected' (who would not have been infected had they received treatment; $X(z) = 1 - z$). The third principal stratum in the population are those immune to infection ('Immune', $X(z) = 0$). The observed dataset $\boldsymbol{n}$ and corresponding principal strata or infection types within each observed (Z, X, Y) stratum are described in Table 3.1.

$\boldsymbol{n}$	$Z=0$	$Z=1$
$Y = 0, X = 1$	Always Infected and/or Protected	Always Infected
$Y = 1, X = 1$	Always Infected and/or Protected	Always Infected
$Y = *, X = 0$	Immune	Immune and/or Protected

Table 3.1: Observed dataset for a study with a post-infection outcome and the principal strata or infection types within each observed (Z, X, Y) stratum; an $*$ indicates an outcome that is not observable.

Note that unlike the binary IV model, there is no observed (Z, X, Y) stratum containing only Protected individuals. The resulting Neyman-Rubin-Copas likelihood given an observed dataset of a given population is then a sum over a *two-dimensional (variation-dependent)* space of complete tables. We describe the space of possible complete tables for the observed dataset under a fixed population, and propose a GLR test statistic.

3.2 Randomization-based Confidence Intervals for the Complier Average Causal Effect

3.2.1 Parameters for Population Proportions of Compliers

Our interest here is in the population CACE, which depends on the *proportions* of Compliers of types Helped and Hurt, and in turn, the total number of Compliers in the population. For example, when there is exactly one Complier who is Helped and zero who are Hurt, whether there are 100 Compliers in total in the population or just one Complier would result in quite different values of the CACE.

Denote the proportions of Compliers who are Helped and Hurt in a given population $\boldsymbol{t}$ as:

$$\begin{aligned}\pi(\boldsymbol{t}) &\equiv (\pi_{HE}(\boldsymbol{t}), \pi_{HU}(\boldsymbol{t})) \\ &\equiv \left(\frac{t_{HE}^{CO}}{\max\left(1, N - \sum_{k=0}^{1}\left(t_k^{NT} + t_k^{AT}\right)\right)} \, , \, \frac{t_{HU}^{CO}}{\max\left(1, N - \sum_{k=0}^{1}\left(t_k^{NT} + t_k^{AT}\right)\right)} \right); \end{aligned} \tag{3.2}$$

here we define a population with no Compliers to have zero proportions of Helped and Hurt. We may see that $\pi(\boldsymbol{t})$ is just a deterministic function of the column totals $\boldsymbol{t}$, and is hence a population parameter which will not vary over hypothetical rerandomizations. Since the total number of Compliers in the given population $\boldsymbol{t}$ is just

$$\sum_{r_Y \in \{NR, HE, HU, AR\}} t_{r_Y}^{CO} = N - \sum_{k=0}^{1}\left(t_k^{NT} + t_k^{AT}\right),$$

we consider both $\pi_{HE}(\boldsymbol{t})$ and $\pi_{HU}(\boldsymbol{t})$ as *common fractions* with a nonzero denominator, each taking a *discrete* value between zero and one. The set of all possible unique pairs of values that $\pi(\boldsymbol{t})$ can take, across all possible populations $\boldsymbol{t}$ in the parameter space $T(\boldsymbol{n})$, is denoted as:

$$\Pi(\boldsymbol{n}) = \bigcup_{\boldsymbol{t} \in T(\boldsymbol{n})} \pi(\boldsymbol{t}). \tag{3.3}$$

Note that the set of possible values $\Pi(\boldsymbol{n})$ depends on the observed dataset $\boldsymbol{n}$. For a value of $\pi \in \Pi(\boldsymbol{n})$, the population CACE is the difference between the proportions of Compliers who are Helped and Hurt, and is a function of π:

$$\text{CACE} \equiv \pi_{HE} - \pi_{HU}. \tag{3.4}$$

Since $0 \leq \pi_{HE} + \pi_{HU} \leq 1$ for all values of $\pi \in \Pi(\boldsymbol{n})$ by definition, the CACE will always be between -1 and 1.

3.2.2 Significance Tests of Causal Hypotheses for Compliers

For a given value of $\pi' \in \Pi(\boldsymbol{n})$, define the causal hypothesis that the proportions of Compliers who are Helped and Hurt are equal to π' as:

$$H_0(\pi') : \pi(\boldsymbol{t}) = \pi'. \tag{3.5}$$

We use π' to refer to some value of the proportions in $\Pi(\boldsymbol{n})$, and $\pi(\boldsymbol{t})$ as the deterministic function of the population parameters $\boldsymbol{t}$ as defined in (3.2). To perform a significance test of $H_0(\pi')$, we require a test statistic that is a function only of the observed data $\boldsymbol{n}$ and the fixed value of π', but not $\boldsymbol{t}$. We propose the generalized likelihood ratio (GLR) test statistic to assess whether allowing an unrestricted value of the CACE is a significantly better explanation for the observed dataset $\boldsymbol{n}$ than restricting the value of the CACE at π'.

In general, there may be multiple populations $\boldsymbol{t} \in T(\boldsymbol{n})$ where the population proportions $\pi(\boldsymbol{t}) = \pi'$. For example, $\pi' = (1, 0)$ is compatible with any population $\boldsymbol{t}$ where all the Compliers in that population are of type Helped, in other words, $t_{HE}^{CO} = N - \sum_{k=0}^{1} \left(t_k^{NT} + t_k^{AT}\right)$. For a particular value of $\pi' \in \Pi(\boldsymbol{n})$, denote the subspace of $T(\boldsymbol{n})$ containing populations that are compatible with π' as:

$$T(\pi'; \boldsymbol{n}) \equiv \{\boldsymbol{t} \in T(\boldsymbol{n}) : \pi(\boldsymbol{t}) = \pi'\}. \tag{3.6}$$

The largest probability given $\boldsymbol{n}$ under the given hypothesis $H_0(\pi')$ is obtained by finding the value(s) of $\boldsymbol{t} \in T(\pi'; \boldsymbol{n})$ that lend the strongest support to $\boldsymbol{n}$ under the fixed value of π':

$$q(\pi'; \boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T(\pi'; \boldsymbol{n})} \Pr(\boldsymbol{n} \mid \boldsymbol{t}), \tag{3.7}$$

where $\Pr(\boldsymbol{n} \mid \boldsymbol{t})$ is the probability given $\boldsymbol{n}$ in a given population $\boldsymbol{t}$. Note that we are explicitly maximizing over the size of the Complier subpopulation, since only the population proportions $\pi(\boldsymbol{t}) = \pi'$, and not the actual values of the population totals in $\boldsymbol{t}$, are specified under $H_0(\pi')$. The

GLR test statistic for $H_0(\pi')$ is then:

$$\lambda(\pi';\boldsymbol{n}) \equiv \frac{q(\pi';\boldsymbol{n})}{q(\boldsymbol{n})} \equiv \frac{\max\limits_{\boldsymbol{t}\,\in\,T(\pi';\boldsymbol{n})} \Pr(\boldsymbol{n}\mid\boldsymbol{t})}{\max\limits_{\boldsymbol{t}\,\in\,T(\boldsymbol{n})} \Pr(\boldsymbol{n}\mid\boldsymbol{t})} = \frac{\max\limits_{\boldsymbol{t}\,\in\,T(\pi';\boldsymbol{n})} h(\boldsymbol{t})}{\max\limits_{\boldsymbol{t}\,\in\,T(\boldsymbol{n})} h(\boldsymbol{t})}, \tag{3.8}$$

where $h(\boldsymbol{t})$ is the Neyman-Rubin-Copas-Richardson likelihood function defined in (2.36), and $q(\boldsymbol{n})$ is the global maximum likelihood given the observed dataset. There will hence be a value of the GLR $\lambda(\pi';\boldsymbol{n})$ for the observed data $\boldsymbol{n}$ corresponding to each unique value of $\pi' \in \Pi(\boldsymbol{n})$, such that there are a total of $|\Pi(\boldsymbol{n})|$ values of the GLR test statistic for the observed dataset.

Since $H_0(\pi')$ may be a composite 'null' hypothesis, we need to consider the sampling distribution for $\boldsymbol{n}$ for each possible value of the population totals $\boldsymbol{t}$. For a fixed value of $\boldsymbol{t} \in T(\pi';\boldsymbol{n})$, the sample space of possibly observable datasets $\tilde{\boldsymbol{n}}$, denoted as $\Omega(\boldsymbol{t})$, is obtained by first conceptually generating assignments (via either complete enumeration or Monte Carlo sampling) in the two-by-eight complete table $\tilde{\boldsymbol{m}}$, and then summing the cell counts in $\tilde{\boldsymbol{m}}$ based on the linear constraints in Table 2.5 to obtain $\tilde{\boldsymbol{n}}$. If $\Omega(\boldsymbol{t})$ was obtained by complete enumeration, the probability $\Pr(\tilde{\boldsymbol{n}}|\boldsymbol{t})$ is calculated using the convolution likelihood (2.36).

We then find the value of the test statistic $\lambda(\pi';\tilde{\boldsymbol{n}})$ for each possibly observable dataset $\tilde{\boldsymbol{n}}$ to obtain a population-specific randomization distribution for $\lambda(\pi';\boldsymbol{n})$. The population-specific p-value under $H_0(\pi')$ is the total probability of observing a value of the test statistic *for the given* π' at least as extreme as the observed quantity:

$$\begin{aligned} pv^{\pi'}(\boldsymbol{n};\boldsymbol{t}) &= \Pr\left(\lambda(\pi';\tilde{\boldsymbol{n}}) \le \lambda(\pi';\boldsymbol{n}) \mid \boldsymbol{t}\right) \\ &= \sum_{\tilde{\boldsymbol{n}}\,\in\,\Omega(\boldsymbol{t})} \Pr\left(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}\right) \times \mathbb{1}\{\lambda(\pi';\tilde{\boldsymbol{n}}) \le \lambda(\pi';\boldsymbol{n})\}. \end{aligned} \tag{3.9}$$

The overall p-value for a significance test of $H_0(\pi')$ is then the maximum among all population-specific p-values with $\pi(\boldsymbol{t}) = \pi'$:

$$pv^{\pi'}(\boldsymbol{n}) \equiv \max_{\boldsymbol{t}\,\in\,T(\pi';\boldsymbol{n})} pv^{\pi'}(\boldsymbol{n};\boldsymbol{t}). \tag{3.10}$$

By definition, the 'sharp' null for Compliers where there are no Compliers who are Helped or Hurt ($t^{CO}_{HE} = t^{CO}_{HU} = 0$) is equivalent to $H_0(\pi' = (0,0))$. There is also no (further) restriction

placed on the null parameter space $T^0(\boldsymbol{n})$, so that $T((0,0);\boldsymbol{n}) = T^0(\boldsymbol{n})$. Since the test statistics $\lambda((0,0);\boldsymbol{n})$ and $\lambda(\boldsymbol{n})$ are also equal, the p-value $pv^{(0,0)}(\boldsymbol{n})$ is equal to the p-value under the 'sharp' null $pv^{H_0}(\boldsymbol{n})$.

3.2.3 Confidence interval for the Complier Average Causal Effect

A $100(1-\alpha)\%$ joint confidence region for the proportions of Compliers who are Helped and Hurt given the observed dataset $\boldsymbol{n}$ is then:

$$\Pi_\alpha(\boldsymbol{n}) \equiv \left\{\pi' \in \Pi(\boldsymbol{n}) : pv^{\pi'}(\boldsymbol{n}) \geq \alpha\right\}. \tag{3.11}$$

Denote the set of all unique values for the CACE for a given observed dataset as:

$$\Delta(\boldsymbol{n}) \equiv \bigcup_{\pi' \in \Pi(\boldsymbol{n})} \pi'_{HE} - \pi'_{HU}. \tag{3.12}$$

To construct a $100(1-\alpha)\%$ confidence interval for the CACE, we first find the maximum p-value over the subset of $\Pi(\boldsymbol{n})$ that maps onto each unique discrete value of CACE $= \delta$ as follows:

$$pv^{\delta}(\boldsymbol{n}) \equiv \max_{\{\pi' \in \Pi(\boldsymbol{n}) \,:\, \pi'_{HE} - \pi'_{HU} = \delta\}} pv^{\pi'}(\boldsymbol{n}). \tag{3.13}$$

The $100(1-\alpha)\%$ confidence interval for the CACE is just:

$$\Delta_\alpha(\boldsymbol{n}) \equiv \left\{\delta \in \Delta(\boldsymbol{n}) : pv^{\delta}(\boldsymbol{n}) \geq \alpha\right\}. \tag{3.14}$$

By inverting a set of randomization-based significance tests, $\Delta_\alpha(\boldsymbol{n})$ contains the values of the CACE that would not be 'surprising' at the α significance level.

3.2.4 Toy dataset

We now apply our procedure to a toy dataset in Table 3.2. All possible unique values for $\pi' \in \Pi(\boldsymbol{n})$ for the observed dataset are plotted in Figure 3.1. Values of π' in the joint confidence region $\Pi_\alpha(\boldsymbol{n})$ for $\alpha = 0.05$ are marked in green. The resulting 95% confidence interval for the CACE is $\{0.18, \ldots, 1\}$.

Number of individuals	$Y=1$		$Y=0$		Row
	$X=1$	$X=0$	$X=1$	$X=0$	
$Z=0$	0	3	2	8	13
$Z=1$	8	2	2	1	13

Table 3.2: Observed toy dataset $\boldsymbol{n}$ from a randomized controlled trial.

We may also compare our result to a Bayesian credible interval. First, we check whether the Instrumental Variable (IV) inequalities assuming monotonicity have been satisfied empirically. The inequalities (Equation (8.24) in [Pearl, 2009] and Equations (12)-(13) in [Richardson et al., 2011]) are:

$$\begin{aligned} &\Pr(Y = y, X = 0|Z = 0) \geq \Pr(Y = y, X = 0|Z = 1) \\ &\Pr(Y = y, X = 1|Z = 1) \geq \Pr(Y = y, X = 1|Z = 0), \end{aligned} \tag{3.15}$$

for $y = 0, 1$. For the observed dataset, the inequalities satisfy the IV inequalities empirically:

$$\begin{aligned} &\Pr(Y = 0, X = 0|Z = 0) - \Pr(Y = 0, X = 0|Z = 1) = 7/13 \\ &\Pr(Y = 1, X = 0|Z = 0) - \Pr(Y = 1, X = 0|Z = 1) = 1/13 \\ &\Pr(Y = 0, X = 1|Z = 1) - \Pr(Y = 0, X = 1|Z = 0) = 0 \\ &\Pr(Y = 1, X = 1|Z = 1) - \Pr(Y = 1, X = 1|Z = 0) = 8/13 \end{aligned}$$

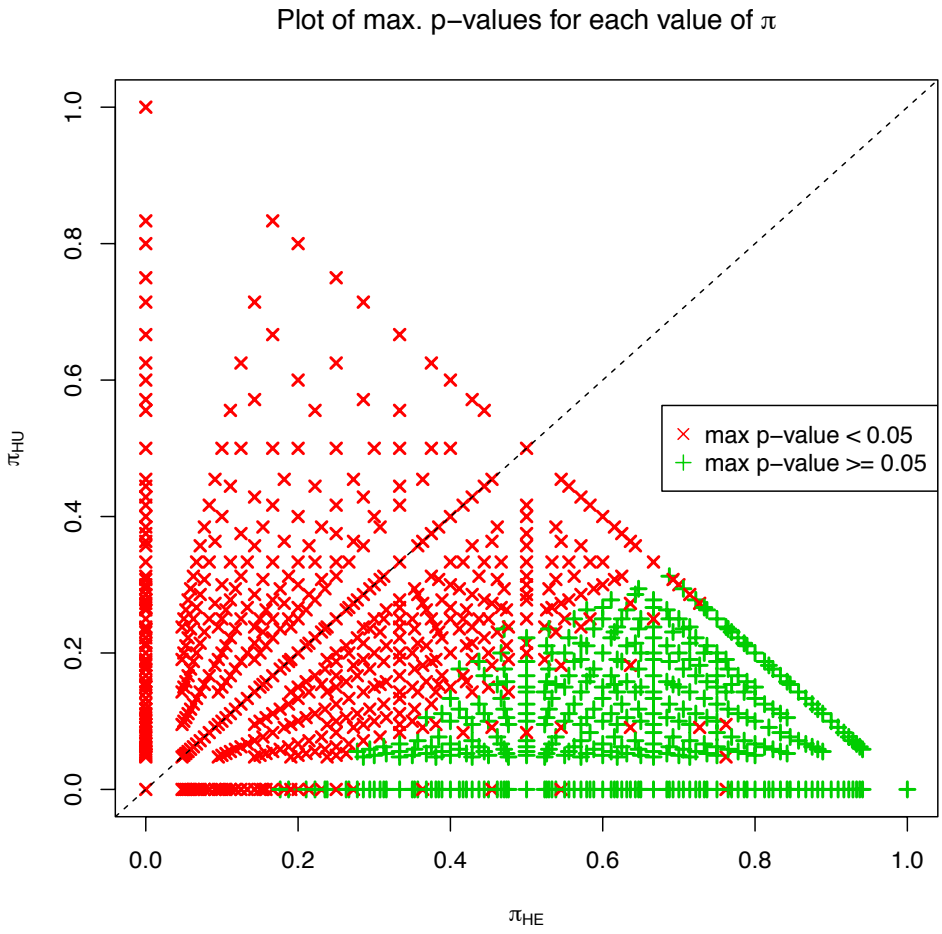

Figure 3.1: Plot of possible unique values for $\pi \equiv (\pi_{HE}, \pi_{HU})$ based on the observed dataset in Table 3.2; values of π' where $pv^{\pi'}(\boldsymbol{n}) \geq 0.05$ are plotted in green.

Note that under exclusion restriction and monotonicity, the Complier Average Causal Effect (CACE) is point-identified [Angrist et al., 1996]:

$$\text{CACE} = \frac{p(y_1|z_1) - p(y_1|z_0)}{p(x_1|z_1) - p(x_1|z_0)}.$$

Following [Richardson et al., 2011], we use a Dirichlet prior for the population proportions $p(y, x|z)$ for example, $\text{Dir}(0.25, 0.25, 0.25, 0.25) \times \text{Dir}(0.25, 0.25, 0.25, 0.25)$ under a saturated model, and a multinomial likelihood for $\boldsymbol{n}$. We then sample $\tilde{p}(y, x|z)$ from the truncated con-

jugate Dirichlet posterior distribution via Monte Carlo simulation, where draws that violate the IV inequalities in (3.15) were discarded (about 37%). The posterior distribution for the CACE is then obtained from the remaining $\tilde{p}(y, x|z)$ and plotted in Figure 3.2. The symmetric two-sided 95% credible interval (with tail probabilities of 2.5% on either side) is $[0.11, 0.93]$; the lower bound of a one-sided 95% credible interval is 0.19.

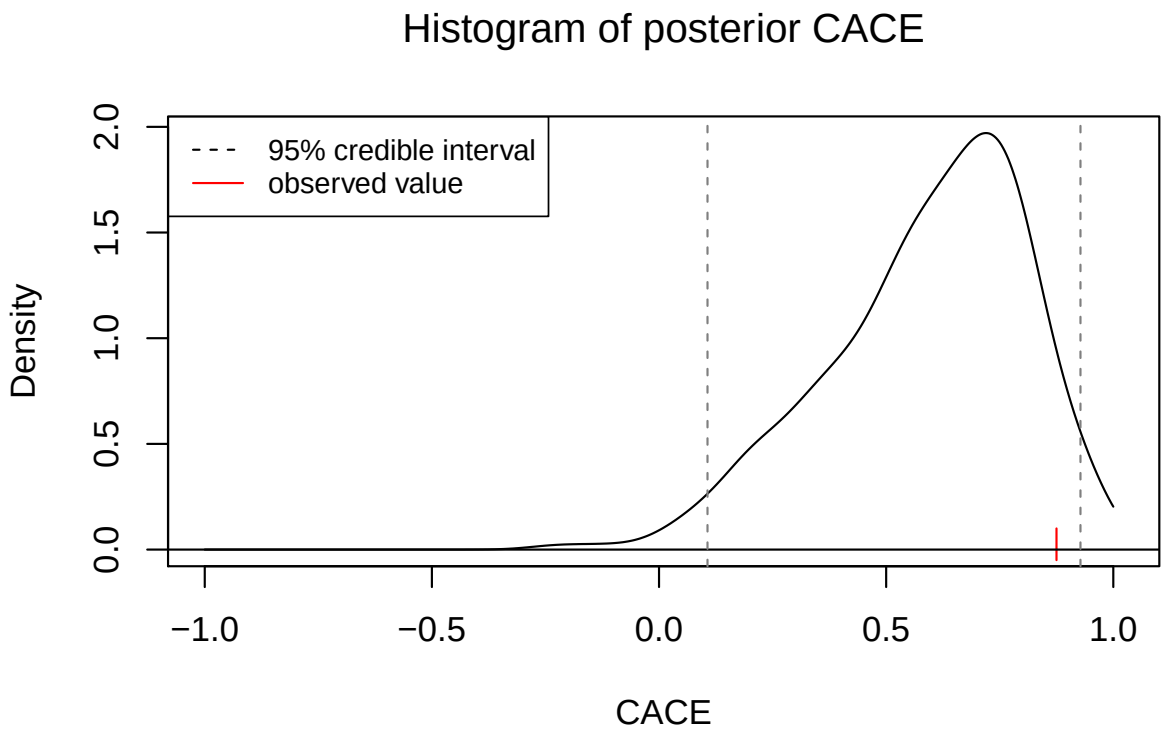

Figure 3.2: Posterior distribution for the CACE based on samples from a truncated Dirichlet posterior distribution with prior $\text{Dir}(0.25, 0.25, 0.25, 0.25) \times \text{Dir}(0.25, 0.25, 0.25, 0.25)$; the two-sided 95% credible interval is indicated by the broken lines; and the observed value of 0.875 is marked by the red line.

3.3 GLR for Testing the Sharp Null in the Always Infected Principal Stratum

3.3.1 Population Parameters

We now consider the scenario of a randomized trial with an intermediate post-treatment infection X, where the final post-infection outcome Y is only observable for infected individuals ($X = 1$). Here we allow for a direct effect of treatment Z on Y.

We first introduce the notation. The potential outcome for an individual's final response Y under intermediate infection $X = x$, and treatment assignment $Z = z$ is $Y(x, z)$, such that for binary treatment Z, an individual's joint potential outcomes for Y are $\{Y(X(0), 0), Y(X(1), 1)\}$. The potential outcomes are linked to the observed outcomes via causal consistency:

$$X = ZX(1) + (1-Z)X(0); \quad Y = ZY(X(1), 1) + (1-Z)Y(X(0), 0). \tag{3.16}$$

In general, principal strata [Frangakis and Rubin, 2002] are just mutually-exclusive groups in the population as defined by the joint potential outcomes $\{X(z)\}$, where $X(z)$ denotes the potential post-treatment outcome had the individual been assigned to treatment $Z = z$. We label individuals with potential infection outcomes $X(z) \equiv 1-z$ as 'Protected'; their potential outcomes $\{Y(1, 0), Y(0, 1)\}$ are either $\{0, *\}$ or $\{1, *\}$, where $*$ indicates an outcome that is not observable or possibly undefined. For $k = 0, 1$, denote t_k^{PD} as the total number of Protected (PD) individuals in the population with potential outcomes $\{k, *\}$. The total number of 'Immune' individuals, with potential outcomes $X(z) \equiv 0$ so that $Y(0, 0) = Y(0, 1) \equiv *$, is just t^{Im}.

Since 'Always Infected' individuals will be infected regardless of their treatment assignment, their joint potential outcomes $\{Y(1, 0), Y(1, 1)\}$ can then take one of four possible combinations for binary response Y. Assuming that $Y = 1$ denotes a desirable outcome such as recovery, we label each one of the four combinations as a response type r_Y, and describe them in Table 3.3.

Denote $t_{r_Y}^{AI}$ as the total number of Always Infected (AI) individuals of response type r_Y in the population. Under the 'monotonicity' assumption, there are no individuals in the population who consistently get infected under treatment and remain uninfected under control, so that

$$X(z=0) \geq X(z=1). \tag{3.17}$$

Response Type r_Y	$Y(x=1,z=1)=0$	$Y(x=1,z=1)=1$
$Y(x=1,z=0)=0$	Never Recover (NR)	Helped (HE)
$Y(x=1,z=0)=1$	Hurt (HU)	Always Recover (AR)

Table 3.3: Response Types r_Y for the Always Infected (AI) principal stratum.

A finite population may then be uniquely described by the population parameters:

$$\boldsymbol{t} \equiv (t_0^{PD}, t_1^{PD}, t_{NR}^{AI}, t_{HE}^{AI}, t_{HU}^{AI}, t_{AR}^{AI}, t^{IM}). \tag{3.18}$$

Let $T(\boldsymbol{n})$ be the parameter space of possible values for $\boldsymbol{t}$ given an observed dataset $\boldsymbol{n}$. It is sufficient to specify values for $(t_0^{PD}, t_1^{PD}, t_{NR}^{AI}, t_{HE}^{AI}, t_{HU}^{AI}, t_{AR}^{AI})$ since $t^{IM} = N - (t_0^{PD} + t_1^{PD} + t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} + t_{AR}^{AI})$. $T(\boldsymbol{n})$ is then the intersection of the six-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^6$, and the convex polyhedron described with the inequalities in Figure 3.3.

3.3.2 Likelihood under a given population

Following the notation for $\boldsymbol{t}$, let $\boldsymbol{m}_i$ be the number of individuals of each type represented in $\boldsymbol{t}$ within each $Z=i$ group; a *complete* table $\boldsymbol{m}$ is then the concatenation of $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$. For a given set of column totals $\boldsymbol{t} \in T(\boldsymbol{n})$, each entry in $\boldsymbol{m}$ may be expressed as a linear function of $\boldsymbol{n}$, $\boldsymbol{t}$, and *two of the cell counts*, for example m_{HE,z_0}^{AI} and $m_{y_0 z_0}^{PD}$. A two-by-seven contingency table such as Table 3.4 may be used to summarize $\boldsymbol{m}$.

Denote the set of complete tables $\boldsymbol{m}$ compatible with the observed dataset $\boldsymbol{n}$ and the given column totals $\boldsymbol{t}$ as $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$. $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$ is then a two-dimensional space indexed by a variation-dependent set of possible values for m_{HE,z_0}^{AI} and $m_{y_0 z_0}^{PD}$. The inequalities in terms of m_{HE,z_0}^{AI} and $m_{y_0 z_0}^{PD}$ that characterize $\mathcal{M}(\boldsymbol{t};\boldsymbol{n})$ are shown in Figure 3.4. The likelihood given the observed data $\boldsymbol{n}$ of a particular population $\boldsymbol{t} \in T(\boldsymbol{n})$ is thus the total probability of of all complete tables $\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})$:

	$Z=0$	$Z=1$	Total
Protected (PD), $Y(1,0)=0$	c_0	$t_0^{PD} - c_0$	t_0^{PD}
Always Infected (AI), NR	$n_{y_0x_1z_0} - b_0 - c_0$	$t_{NR}^{AI} - (n_{y_0x_1z_0} - b_0 - c_0)$	t_{NR}^{AI}
Always Infected (AI), HE	b_0	$t_{HE}^{AI} - b_0$	t_{HE}^{AI}
Always Infected (AI), HU	$t_{HU}^{AI} - n_{y_0x_1z_1} + t_{NR}^{AI} - (n_{y_0x_1z_0} - b_0 - c_0)$	$n_{y_0x_1z_1} - t_{NR}^{AI} + (n_{y_0x_1z_0} - b_0 - c_0)$	t_{HU}^{AI}
Always Infected (AI), AR	$t_{AR}^{AI} - n_{y_1x_1z_1} + (t_{HE}^{AI} - b_0)$	$n_{y_1x_1z_1} - (t_{HE}^{AI} - b_0)$	t_{AR}^{AI}
Protected (PD), $Y(1,0)=1$	$n_{x_1} - \sum t_{r_Y}^{AI} - c_0$	$t_1^{PD} - n_{x_1} + \sum t_{r_Y}^{AI} + c_0$	t_1^{PD}
Immune (IM)	$n_{x_0z_0}$	$t^{IM} - n_{x_0z_0}$	t^{IM}
Total	n_{z_0}	n_{z_1}	N

Table 3.4: Complete table $\boldsymbol{m}$ with each entry as a linear function of the observed quantities $\boldsymbol{n}$, the population totals $\boldsymbol{t}$, and two of the cell counts m_{HE,z_0}^{AI} and $m_{y_0z_0}^{PD}$ (denoted here as b_0 and c_0 respectively); we write $\sum_{r_Y \in \{NR,HE,HU,AR\}} t_{r_Y}^{AI}$ simply as $\sum t_{r_Y}^{AI}$; the two-by-seven table has also been transposed to make the cell counts more readable.

$$\Pr(\boldsymbol{n} \mid \boldsymbol{t}) = \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \Pr(\boldsymbol{m} \mid \boldsymbol{t}) = \binom{N}{n_{z_0}, n_{z_1}}^{-1} h(\boldsymbol{t});$$
$$h(\boldsymbol{t}) = \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t};\boldsymbol{n})} \prod_{a=1}^{7} \binom{t_{(a)}}{m_{(a),z_1}, m_{(a),z_0}}, \tag{3.39}$$

where the subscript (a) represents the a-th element in $\boldsymbol{t}$ and $\boldsymbol{m}_i$, and the entries $m_{(a),z_i}$ for $i = 0, 1$ are functions of $\boldsymbol{n}, \boldsymbol{t}, m_{HE,z_0}^{AI}$ and $m_{y_0z_0}^{PD}$.

Under the 'sharp' null for the 'Always Infected' (AI) principal stratum (3.1), there are no individuals in the AI subpopulation who are Helped or Hurt:

$$H_0 : Y(x=1, z=0) = Y(x=1, z=1) \quad \Longleftrightarrow \quad t_{HE}^{AI} = t_{HU}^{AI} = 0. \tag{3.43}$$

For each population $\boldsymbol{t}$ that satisfies (3.43), $t^{AI}_{HE} = 0 \Rightarrow m^{AI}_{HE,z_0} = 0$, so that the set of complete tables $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$ then reduces to a linear space indexed by $m^{PD}_{y_0 z_0}$:

$$\begin{aligned} &\left\{c_0^{\min}, c_0^{\min}+1, \ldots, c_0^{\max}\right\} \\ c_0^{\min} &= \max\left(0, t^{IM} - n_{x_0}, n_{y_0 x_1} - t^{AI}_{NR}, n_{y_0 x_1 z_0} - t^{AI}_{NR}\right) \\ c_0^{\max} &= \min\left(t^{PD}_0, n_{x_1} - t^{AI}_{NR} - t^{AI}_{AR}, n_{y_0 x_1} - t^{AI}_{NR}, n_{y_0 x_1 z_0}\right). \end{aligned} \tag{3.44}$$

So even under the 'sharp' null for the Always Infected, there may be multiple complete tables in a given population, so that the likelihood requires summing over multiple hypergeometric probabilities.

3.3.3 Finding the Generalized Likelihood Ratio test statistic

The maximum likelihood given the observed dataset assuming that H_0 holds is just:

$$q^{H_0}(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T^0(\boldsymbol{n})} \Pr\left(\boldsymbol{n} \mid \boldsymbol{t}\right), \tag{3.45}$$

while the global maximum likelihood is:

$$q(\boldsymbol{n}) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n})} \Pr\left(\boldsymbol{n} \mid \boldsymbol{t}\right). \tag{3.46}$$

The GLR is then:

$$q^{H_0}(\boldsymbol{n}) / q(\boldsymbol{n}) \,. \tag{3.47}$$

It remains to be shown if the likelihood $\Pr\left(\boldsymbol{n} \mid \boldsymbol{t}\right)$ may also be decomposed so that results from the perfect compliance setting may be applied to the Always Infected principal stratum. It would also be useful to examine whether there still remain multiple complete tables $\boldsymbol{m}$ under the maximum likelihood solution $\hat{\boldsymbol{t}}$.

$$0 \leq t_{NR}^{AI} \leq n_{y_0 x_1} \tag{3.19}$$

$$0 \leq t_{HE}^{AI} \leq \sum_{i=0}^{1} n_{y_i x_1 z_i} \tag{3.20}$$

$$0 \leq t_{HU}^{AI} \leq \sum_{i=0}^{1} n_{y_{1-i} x_1 z_i} \tag{3.21}$$

$$0 \leq t_{AR}^{AI} \leq n_{y_1 x_1} \tag{3.22}$$

$$t_{NR}^{AI} + t_{HU}^{AI} \geq n_{y_0 x_1 z_1} \tag{3.23}$$

$$t_{AR}^{AI} + t_{HE}^{AI} \geq n_{y_1 x_1 z_1} \tag{3.24}$$

$$t_{NR}^{AI} + t_{HE}^{AI} \leq n_{x_1} - n_{y_1 x_1 z_0} \tag{3.25}$$

$$t_{AR}^{AI} + t_{HU}^{HU} \leq n_{x_1} - n_{y_0 x_1 z_0} \tag{3.26}$$

$$t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} + t_{AR}^{AI} \leq n_{x_1} \tag{3.27}$$

$$t_0^{PD} \geq 0 \tag{3.28}$$

$$t_1^{PD} \geq 0 \tag{3.29}$$

$$t_0^{PD} + t_{NR}^{AI} + t_{HE}^{AI} \geq n_{y_0 x_1 z_0} \tag{3.30}$$

$$t_1^{PD} + t_{AR}^{AI} + t_{HU}^{AI} \geq n_{y_1 x_1 z_0} \tag{3.31}$$

$$t_0^{PD} + t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} \geq n_{y_0 x_1} \tag{3.32}$$

$$t_1^{PD} + t_{AR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} \geq n_{y_1 x_1} \tag{3.33}$$

$$t_0^{PD} + t_{NR}^{AI} + t_{HE}^{AI} + t_{AR}^{AI} \geq \sum_{i=0}^{1} n_{y_i x_1 z_i} \tag{3.34}$$

$$t_1^{PD} + t_{NR}^{AI} + t_{HU}^{AI} + t_{AR}^{AI} \geq \sum_{i=0}^{1} n_{y_{1-i} x_1 z_i} \tag{3.35}$$

$$t_0^{PD} + (t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} + t_{AR}^{AI}) \geq n_{x_0} - n_{y_1 x_1 z_0} \tag{3.36}$$

$$t_1^{PD} + (t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} + t_{AR}^{AI}) \geq n_{x_0} - n_{y_0 x_1 z_0} \tag{3.37}$$

$$n_{x_1} \leq t_0^{PD} + t_1^{PD} + (t_{NR}^{AI} + t_{HE}^{AI} + t_{HU}^{AI} + t_{AR}^{AI}) \leq n_{x_1} + n_{x_0 z_1}. \tag{3.38}$$

Figure 3.3: Inequalities characterizing the convex polyhedron for the parameter space $T(\boldsymbol{n})$ given an observed dataset $\boldsymbol{n}$; these were obtained using `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015].

$$\begin{aligned}
\max\left(0, t^{AI}_{HE} - n_{y_1x_1z_1}\right) &\leq m^{AI}_{HE,z_0} \\
&\leq \min\left(t^{AI}_{HE}, t^{AI}_{AR} + t^{AI}_{HE} - n_{y_1x_1z_1}\right) \qquad (3.40) \\
\max\left(0, t^{IM} - n_{x_0}\right) &\leq m^{PD}_{y_0z_0} \\
&\leq \min\left(t^{PD}_0, n_{x_1} - \sum_{r_Y \in \{NR,HE,HU,AR\}} t^{AI}_{r_Y}\right) \qquad (3.41) \\
\max\left(n_{y_0x_1} - t^{AI}_{NR} - t^{AI}_{HU}, n_{y_0x_1z_0} - t^{AI}_{NR}\right) &\leq m^{AI}_{HE,z_0} + m^{PD}_{y_0z_0} \\
&\leq \min\left(n_{y_0x_1} - t^{AI}_{NR}, n_{y_0x_1z_0}\right) \qquad (3.42)
\end{aligned}$$

Figure 3.4: Inequalities for m^{AI}_{HE,z_0} and $m^{PD}_{y_0z_0}$ indexing $\mathcal{M}(\boldsymbol{t}; \boldsymbol{n})$, the two-dimensional variation-dependent set of complete tables $\boldsymbol{m}$ compatible with the observed dataset $\boldsymbol{n}$ and some fixed value of the column totals $\boldsymbol{t}$; these were obtained using `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015].

BIBLIOGRAPHY

A Agresti. *Categorical data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2013.

J D Angrist, G W Imbens, and D B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

A Balke and J Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc., San Francisco, 1994.

R L Berger and D D Boos. P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

N E Breslow and N E Day. Statistical methods in cancer research. volume I - the analysis of case-control studies. (32):5–338, 1980.

Z Cai, M Kuroki, J Pearl, and J Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701 2008.

Y Chiba. Exact tests for the weak causal null hypothesis on a binary out come in randomized trials. *J Biom Biostat*, 06(03), 2015. doi: 10.4172/2155-6180.1000244.

D M Chickering and J Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, pages 1269–1276. AAAI Press, 1996.

P S Clarke and F Windmeijer. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652, 2012.

J B Copas. Randomization models for the matched and unmatched 2×2 tables. *Biometrika*, 60 (3):467–476, 1973. ISSN 00063444.

V Didelez and N A Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.*, 16(4):309–330, 2007. ISSN 0962-2802.

R A Fisher. *The design of experiments.* Oliver & Boyd, 1935.

L Forastiere, F Mealli, and L Miratrix. Posterior predictive P-values with fisher randomization tests in noncompliance settings: Test statistics vs discrepancy variables. *arXiv preprint arXiv:1511.00521*, 2015.

C E Frangakis and D B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

M Gail and N Mantel. Counting the number of $r \times c$ contingency tables with fixed margins. *Journal of the American Statistical Association*, 72(360a):859–862

C J Geyer, G D Meeden, and incorporates code from cddlib written by Komei Fukuda. *rcdd: Computational Geometry*, 2015. URL `https://CRAN.R-project.org/package=rcdd`. R package version 1.1-9.

D Heckerman and R Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

G W Imbens and P R Rosenbaum. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126, 2005.

G W Imbens and D B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.*, 25(1):305–327, 1997.

N L Johnson and S Kotz. *Discrete distributions*. Houghton Mifflin, Boston, 1969.

L J Keele, D S Small, and R Grieve. Randomization based instrumental variables methods for binary outcomes with an application to the improve trial. *Journal of the Royal Statistical Society, Series A*, Revise and Resubmit, 2016.

J B Lang. A closer look at testing the “no-treatment-effect" hypothesis in a comparative experiment. *Statistical Science*, 30(3):352–371, 2015.

Y J Lee, J H Ellenberg, D G Hirtz, and K B Nelson. Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine*, 10(10):1595–1605, 1991.

E L Lehmann. *Nonparametrics: statistical methods based on ranks*. Springer New York, 2006.

W W Loh and T S Richardson. A finite population test of the sharp null hypothesis for compliers. *UAI Workshop on Approaches to Causal Structure Learning, 15 July, Bellevue, Washington*, 2013.

W W Loh and T S Richardson. A finite population likelihood ratio test of the sharp null hypothesis for compliers. *Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

W W Loh and T S Richardson. *noncompliance: Causal Inference in the Presence of Treatment Noncompliance Under the Binary Instrumental Variable Model*, 2016. URL `https://CRAN.R-project.org/package=noncompliance`. R package version 0.2.2.

T L Nolen and M G Hudgens. Randomization-based inference within principal strata. *Journal of the American Statistical Association*, 106(494):581–593, 2011.

W Oberhofer and H Kaufmann. Maximum likelihood estimation of a multivariate hypergeometric distribution. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002)*, 49(2):188–191, 1987.

J Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge, second edition, 2009.

J Pearl. On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872, 2010.

M D Perlman and L Wu. The emperor's new tests. *Statistical Science*, 14(4):355–369, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

T S Richardson, R J Evans, and J M Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

J Rigdon. *RI2by2: Randomization inference for treatment effects on a binary outcome*, 2014. URL `https://CRAN.R-project.org/package=RI2by2`. R package version 1.2.

J Rigdon and M G Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924–935, 2015.

K J Rothman, S Greenland, and T L Lash. *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 2008.

D B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.

D B Rubin. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17(3):371–385, 1998.

A Sommer and S L Zeger. On estimating efficacy from clinical trials. *Statistics in Medicine*, 10 (1):45–52, 1991.

J Spława-Neyman, D M Dabrowska, and T P Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

H Zhang. A note about maximum likelihood estimator in hypergeometric distribution. *Comunicaciones En Estadística*, 2(2):169–174, 2009.

Appendix A

MAXIMUM LIKELIHOOD FOR PERFECTLY BALANCED DESIGNS

We first state the following lemmas used to order products of binomial coefficients in proving Theorem 1.

Lemma 3. *For the non-negative integers a, b, c, d, k in the integer lattice:*

$$\left\{ a, b, c \in \left(\{0\} \cup \mathbb{Z}^+\right)^3 ; d, k \in \mathbb{Z}^+ : a + c, a + b \geq 1, d \geq k, \frac{a}{a+c} \leq \frac{b}{b+d} \right\},$$

the following inequality holds:

$$\binom{a+c+k}{a, c+k} \binom{b+d-k}{b, d-k} < \binom{a+c}{a, c} \binom{b+d}{b, d}.$$

Proof. The respective binomial coefficients are proportional to the bivariate hypergeometric probability mass functions for the following two-by-two tables:

	Col. I	Col. II
Row I	a	b
Row II	$c+k$	$d-k$
Total	$a+c+k$	$b+d-k$

	Col. I	Col. II
Row I	a	b
Row II	c	d
Total	$a+c$	$b+d$

First,

$$\frac{a}{a+c} \leq \frac{b}{b+d} \iff \frac{a(d+b)}{b(c+a)} = \frac{ad+ab}{bc+ab} \leq 1 \iff \frac{ad}{bc} \leq 1 \iff \frac{a}{c} \leq \frac{b}{d},$$

such that for $1 \leq i \leq k$,

$$\frac{a}{c+k} \leq \frac{a}{c+i} < \frac{a}{c} \leq \frac{b}{d} \leq \frac{b}{d-(i-1)} \leq \frac{b}{d-(k-1)} < \frac{b}{d-k}. \tag{A.1}$$

The binomial coefficients are then:

$$\begin{aligned}
\binom{a+c+k}{a,c+k}\binom{b+d-k}{b,d-k} &= \binom{a+c}{a,c}\frac{(a+c+k)!c!}{(a+c)!(c+k)!} \times \binom{b+d}{b,d}\frac{(b+d-k)!d!}{(b+d)!(d-k)!} \\
&= \binom{a+c}{a,c}\prod_{i=1}^{k}\frac{a+c+i}{c+i} \times \binom{b+d}{b,d}\prod_{i=1}^{k}\frac{d-(i-1)}{b+d-(i-1)} \\
&= \binom{a+c}{a,c}\binom{b+d}{b,d} \times \prod_{i=1}^{k}\underbrace{\left(1+\frac{a}{c+i}\right)\left(1+\frac{b}{d-(i-1)}\right)^{-1}}_{<1 \text{ from (A.1)}} \\
&< \binom{a+c}{a,c}\binom{b+d}{b,d}.
\end{aligned}$$

□

Lemma 4. *For the integers* $a, b, c \geq 1$*,*

$$\binom{2(a+c)}{a+c,a+c}\binom{2(a+b)}{a+b,a+b} < \binom{2a}{a,a}\binom{2(a+b+c)}{a+b+c,a+b+c}. \tag{A.2}$$

Proof. Denote the ratio of the following binomial coefficients for the integer $x \geq 1$ as:

$$r(x) = \binom{2(x+1)}{x+1,x+1}\Big/\binom{2x}{x,x} = \frac{2(2x+1)}{x+1} = 2\left(2-\frac{1}{1+x}\right).$$

We may see that $r(x) < r(x') \iff x < x'$. Without loss of generality, assume that $c \leq b$. Then:

$$\begin{aligned}
\frac{\binom{2(a+c)}{a+c,a+c}\binom{2(a+b)}{a+b,a+b}}{\binom{2a}{a,a}\binom{2(a+b+c)}{a+b+c,a+b+c}} &= \prod_{i=0}^{c-1}\frac{\binom{2(a+c-i)}{a+c-i,a+c-i}\binom{2(a+b+i)}{a+b+i,a+b+i}}{\binom{2(a+c-i-1)}{a+c-i-1,a+c-i-1}\binom{2(a+b+i+1)}{a+b+i+1,a+b+i+1}} \\
&= \prod_{i=0}^{c-1}\frac{r(a+c-i)}{r(a+b+i+1)} \\
&\leq \prod_{i=0}^{c-1}\frac{r(a+b-i)}{r(a+b+i+1)} \\
&< 1.
\end{aligned}$$

□

We now restate Theorem 1 as follows, and provide the proof below.

Theorem 1. *For an observed two-by-two dataset $\boldsymbol{n}$ in the form of Table 1.1, denote the global maximum likelihood as:*

$$\hat{h} = \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}).$$

If $\boldsymbol{n}$ is perfectly balanced so that $n = N - n$, then there exists a population $\hat{\boldsymbol{t}}$ where $h(\hat{\boldsymbol{t}}) = \hat{h}$ and

$$\min_i \hat{t}_i = 0.$$

Proof. First, relabel the potential outcome variables $Y(1)$ and $Y(0)$ so that so that (i) the total number of individuals with $Y = 1$ is fewer than or equal to the total number with $Y = 0$, (ii) the proportion of $Y = 1$ within the $X = 1$ group is less than or equal to the proportion of $Y = 1$ within the $X = 0$ group, and (iii) the proportion of $Y = 1$ in the $X = 0$ group is less than or equal to 1/2. The linear constraints in Table 1.3 for a given population $\boldsymbol{t}$ are isomorphic under such a relabelling; only the definition of the response types are changed. It may thus be assumed that the observed quantities $\boldsymbol{n}$ satisfy the following inequalities without loss of generality:

$$\text{(i)}\, s \equiv n_{11}+n_{10} \leq n_{01}+n_{00} \equiv N-s, \quad \text{(ii)}\, \frac{n_{11}}{n} \leq \frac{n_{10}}{N-n} \iff n_{11}n_{00} \leq n_{01}n_{10}, \quad \text{(iii)}\, n_{10} \leq n_{00}. \tag{A.3}$$

Now consider the following partitions of $T(\boldsymbol{n})$, as defined in Equation (1.9), in turn.

1. $t_A + t_B \geq s$.

   Denote $t_{AB} = (t_A + t_B) - s \geq 0$. Using the Chu-Vandermonde convolution, the likelihood for a given set of column totals $\boldsymbol{t}$ satisfies the following inequalities:

$$
\begin{aligned}
h(\boldsymbol{t}) &= \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=A,B} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}}}}_{\leq 1} \prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}} \\
&\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}} \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}}}}_{\leq 1} \\
&\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \binom{N - (t_A + t_B)}{n_{01}, N - (t_A + t_B) - n_{01}} \\
&= \binom{s + t_{AB}}{n_{11}, (s - n_{11}) + t_{AB}} \binom{N - s - t_{AB}}{n_{01}, (N - s - n_{01}) - t_{AB}} \\
&\leq \binom{s}{n_{11}, n_{10}} \binom{N - s}{n_{01}, n_{00}} = h(\boldsymbol{t}^0). \qquad \text{(A.4)}
\end{aligned}
$$

   The last inequality is obtained with Lemma 3, and $\boldsymbol{t}^0 = (s, 0, 0, N - s)$.

2. $t_A + t_C \leq s$.

   Denote $t_{AC} = s - (t_A + t_C) \geq 0$, such that:

$$
\begin{aligned}
h(\boldsymbol{t}) &= \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=A,C} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}}}}_{\leq 1} \prod_{i=B,D} \binom{t_i}{m_{i1}, m_{i0}} \\
&\leq \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}} \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=B,D} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}}}}_{\leq 1} \\
&\leq \binom{t_A + t_C}{(t_A + t_C) - n_{10}, n_{10}} \binom{N - (t_A + t_C)}{N - (t_A + t_C) - n_{00}, n_{00}} \\
&= \binom{N - s + t_{AC}}{n_{00}, n_{01} + t_{AC}} \binom{s - t_{AC}}{n_{10}, n_{11} - t_{AC}} \\
&\leq \binom{N - s}{n_{00}, n_{01}} \binom{s}{n_{10}, n_{11}} = h(\boldsymbol{t}^0). \qquad \text{(A.5)}
\end{aligned}
$$

   The last inequality is similarly obtained with Lemma 3.

3. $t_A + t_B < s < t_A + t_C$.

The maximum likelihood $h(\hat{\boldsymbol{t}})$ must be at least as large as $h(\boldsymbol{t}^0)$. Furthermore, if there is a set of column totals $\boldsymbol{t}$ with a value of $h(\boldsymbol{t})$ that is strictly larger than $h(\boldsymbol{t}^0)$, then $\boldsymbol{t}$ must lie in the partition of $T(\boldsymbol{n})$ where $t_A + t_B < s < t_A + t_C$. For each set of column totals in $T(\boldsymbol{n})$ that satisfies $t_A + t_B < s < t_A + t_C$, an upperbound for the likelihood is then:

$$\begin{aligned} h(\boldsymbol{t}) &= \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \underbrace{\frac{\prod_{i=A,B} \binom{t_i}{m_{i1}, m_{i0}}}{\binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}}}}_{\leq 1} \prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}} \\ &\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \max_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}} \\ &\equiv u(\boldsymbol{t}). \end{aligned} \tag{A.6}$$

Denoting

$$\widehat{\boldsymbol{m}} = \operatorname*{arg\,max}_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{t}; \boldsymbol{n})} \prod_{i=C,D} \binom{t_i}{m_{i1}, m_{i0}}, \tag{A.7}$$

in addition to the linear constraint $\widehat{m}_{C1} + \widehat{m}_{D1} = n_{01}$, the values of $(\widehat{m}_{C1}, \widehat{m}_{D1})$ are restricted in terms of the column totals such that $\widehat{m}_{C1} + \widehat{m}_{C0} = t_C, \widehat{m}_{D1} + \widehat{m}_{D0} = N - (t_A + t_B + t_C)$. A further upperbound for $u(\boldsymbol{t})$ may thus be obtained by relaxing the restriction of the column totals. To do so, we find the most likely values of (m_{C1}, m_{D1}) for the given values of $(\widehat{m}_{C0}, \widehat{m}_{D0})$ and $\boldsymbol{n}$, subject to just the linear constraint $m_{C1} + m_{D1} = n_{01}$. From [Johnson and Kotz, 1969, page 146], the most likely values are then deterministic functions of $(\widehat{m}_{C0}, \widehat{m}_{D0})$ and $\boldsymbol{n}$:

$$m^*_{C1} = \left\lceil \frac{\widehat{m}_{C0}}{\widehat{m}_{C0} + \widehat{m}_{D0}} (n_{01} + 1) \right\rceil - 1, \tag{A.8}$$

$$m^*_{D1} = n_{01} - m^*_{C1}. \tag{A.9}$$

Then

$$\begin{aligned} u(\boldsymbol{t}) &= \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \prod_{i=C,D} \binom{t_i}{\widehat{m}_{i1}, \widehat{m}_{i0}} \\ &\leq \binom{t_A + t_B}{n_{11}, (t_A + t_B) - n_{11}} \prod_{i=C,D} \binom{m^*_{i1} + \widehat{m}_{i0}}{m^*_{i1}, \widehat{m}_{i0}} \\ &\equiv u^*(\widehat{m}_{C0}, \widehat{m}_{D0}). \end{aligned}$$

In general, $u^*(\widehat{m}_{C0}, \widehat{m}_{D0})$ is not necessarily a valid likelihood for $\boldsymbol{n}$: neither $(t_A + t_B, 0, m^*_{C1} + \widehat{m}_{C0}, m^*_{D1} + \widehat{m}_{D0})$ nor $(0, t_A + t_B, m^*_{C1} + \widehat{m}_{C0}, m^*_{D1} + \widehat{m}_{D0})$ may be in the parameter space $T(\boldsymbol{n})$. The exception is if either $\widehat{m}_{C0}$ or $\widehat{m}_{D0}$ is equal to one of the observed values n_{10} or n_{00}, then $u^*(\widehat{m}_{C0}, \widehat{m}_{D0})$ corresponds to a valid likelihood for the observed dataset $\boldsymbol{n}$.

First, to simplify the notation, we replace $\widehat{m}_{C0}$ as b_0, and $\widehat{m}_{D0}$ as c_0, so that the function $u^*(b_0, c_0)$ takes the following form:

$$u^*(b_0, c_0) = \binom{n_{11} + n_{10} - c_0 + b_0}{n_{11}, n_{10} - c_0 + b_0} \binom{c^*_1 + c_0}{c^*_1, c_0} \binom{n_{01} - c^*_1 + n_{00} - b_0}{n_{01} - c^*_1, n_{00} - b_0}; \tag{A.10}$$

$$c_0 = 1, \ldots, n_{10} - 1; b_0 = 0, \ldots, \min(c_0 - 1, n_{00}); c^*_1 = \left\lceil \frac{c_0}{n_{00} + (c_0 - b_0)} (n_{01} + 1) \right\rceil - 1.$$

The function $u^*(b_0, c_0)$, in terms of the observed quantities in $\boldsymbol{n}$ is just the kernel of the multiple hypergeometric probability for the following two-by-three table:

	A	C	D	Row
$X=1$	n_{11}	c^*_1	$n_{01} - c^*_1$	n
$X=0$	$n_{10} - (c_0 - b_0)$	c_0	$n_{00} - b_0$	n

From hereon, we assume a perfectly balanced design where $n = N - n$. The assumptions in (A.3) imply that the observed quantities $\boldsymbol{n}$ satisfy the following ordering:

$$n_{11} \leq n_{10} \leq n_{00} \leq n_{01}. \tag{A.11}$$

The observed quantities in $\boldsymbol{n}$ may be equivalently expressed as:

$$n_{11}, n_{10} \equiv n_{11} + k, n_{00} \equiv n_{11} + k + k_{00}, n_{01} \equiv n_{11} + 2k + k_{00}, \tag{A.12}$$

where $k, k_{00} \geq 0$ are observed quantities from $\boldsymbol{n}$. The cell count $c^*_1 \equiv c^*_1(b_0, c_0)$ is a deterministic function of (b_0, c_0) and $\boldsymbol{n}$:

$$\begin{aligned} c^*_1 &= \left\lceil \frac{c_0}{(n_{11} + k + k_{00}) + (c_0 - b_0)} (n_{11} + 2k + k_{00} + 1) \right\rceil - 1 \\ &= \left\lceil c_0 \left(1 + \frac{(k+1) - (c_0 - b_0)}{(n_{11} + k + k_{00}) + (c_0 - b_0)} \right) \right\rceil - 1 \\ &\begin{cases} < c_0 & , c_0 - b_0 \geq k + 1, \\ \geq c_0 & , c_0 - b_0 \leq k. \end{cases} \end{aligned} \tag{A.13}$$

The function $u^*(b_0, c_0)$, in terms of (n_{11}, k, k_{00}) is just the kernel of the multiple hypergeometric probability for the following two-by-three table:

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11}+k+k_{00})+(k-c_1^*)$	n
$X=0$	$(n_{11}+k)-(c_0-b_0)$	c_0	$(n_{11}+k+k_{00})-b_0$	n

We now consider each of the following possible scenarios corresponding to (A.13) in turn. For each scenario, we will represent the kernel of the multiple hypergeometric probability as a contingency table for clarity. An ordering between two contingency tables implies an ordering between the corresponding product of binomial coefficients. Our goal is to show that, in each case, the probability of the initial table $u^*(b_0, c_0)$ has an upper bound corresponding to the likelihood of another table $h(\boldsymbol{t})$ that leads to the same observed data $\boldsymbol{n}$ (via the linear constraints in Equation (1.2)), and has at most three non-zero response types.

Scenario	Ordering of cell counts	
1.	$c_0 - b_0 = k$	
2i.	$0 < c_0 - b_0 < k \Rightarrow c_0 \leq c_1^*$	$0 < c_0 \leq c_1^* \leq k$
2ii.		$0 < c_0 \leq k < c_1^*$
3.		$k < c_0 \leq \min(n_{11}+k, c_1^*)$
4.	$k < c_0 - b_0 \leq n_{11}+k \Rightarrow c_1^* < c_0$	$\max(k, c_1^*) < c_0 \leq n_{11}+k$

1. $c_0 - b_0 = k$. This implies that $b_0 = c_0 - k \leq n_{11}$.

	A	C	D	Row
$X=1$	n_{11}	$k + b_0$	$(n_{11} + k + k_{00}) - b_0$	n
$X=0$	n_{11}	$k + b_0$	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

2i. $0 < c_0 \leq c_1^* \leq k$.

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) + (k - c_1^*)$	n
$X=0$	$n_{11} + (k - c_0) + b_0$	c_0	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{n_{11}}{n_{11}+(k-c_0)} \leq 1 \leq \frac{(n_{11}+k+k_{00})+(k-c_1^*)}{n_{11}+k+k_{00}}$

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) + (k - c_1^*)$	n
$X=0$	$n_{11} + (k - c_0)$	c_0	$(n_{11} + k + k_{00})$	n

Note that the above table is compatible with the observed data $\boldsymbol{n}$ via the linear constraints in Equation (1.2).

2ii. $0 < c_0 \leq k < c_1^*$. This implies that $k - c_0 \geq 0$. Note that by definition of c_1^* in (A.13), $c_1^* < 2c_0 \leq 2k \Rightarrow c_1^* - k < k$.

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + b_0 + (k - c_0)$	$c_0 = k - (k - c_0)$	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{n_{11}}{n_{11}+b_0} \leq 1 \leq \frac{c_1^*}{k}$

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

First if $b_0 > (c_1^* - k) > 0$, then:

	A	C	D	Row
$X=1$	$n_{11} = n_{11} + (c_1^* - k) - (c_1^* - k)$	$c_1^* = k + (c_1^* - k)$	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{k}{k} = 1 \leq \frac{n_{11}+b_0}{n_{11}+(c_1^*-k)}$

	A	C	D	Row
$X=1$	$n_{11} + (c_1^* - k)$	k	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

$=$

	A	C	D	Row
$X=1$	$n_{11} + (c_1^* - k)$	k	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + (c_1^* - k) + b_0 - (c_1^* - k)$	k	$(n_{11} + k + k_{00}) - (c_1^* - k) - b_0 + (c_1^* - k)$	n

$\leq$ by Lemma 3 since $b_0 - (c_1^* - k) > 0$

	A	C	D	Row
$X=1$	$n_{11} + (c_1^* - k)$	k	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} + (c_1^* - k)$	k	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

Otherwise if $b_0 \leq c_1^* - k \leq k$, then:

	A	C	D	Row
$X=1$	n_{11}	$c_1^* = (k + b_0) + (c_1^* - k) - b_0$	$(n_{11} + k + k_{00}) - b_0 - (c_1^* - k) + b_0$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{k}{k+b_0} \leq 1 = \frac{(n_{11}+k+k_{00})-b_0}{(n_{11}+k+k_{00})-b_0}$

	A	C	D	Row
$X=1$	$n_{11} = n_{11} + b_0 - b_0$	$k + b_0$	$(n_{11} + k + k_{00}) - b_0$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n
$X=0$	$n_{11} + b_0$	k	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

3. $0 < c_0 - b_0 < k, k < c_0 \leq n_{11} + k$. This implies that $0 < b_0 - (c_0 - k) = k - (c_0 - b_0) < k$. For this setting, we will use the following notation instead:

$$\begin{aligned} \delta_0 &= (n_{11} + k) - c_0, & 0 \leq \delta_0 < n_{11}; \\ \delta_b &= b_0 - (c_0 - k), & 0 < \delta_b < k; \\ \delta_1 &= c_1^* - c_0, & 0 \leq \delta_1. \end{aligned}$$

The intital table may then be written as:

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k) - \delta_0 + \delta_1$	$(k+k_{00}) + (\delta_0 - \delta_1)$	n
$X=0$	$n_{11} + \delta_b$	$(n_{11}+k) - \delta_0$	$(k+k_{00}) + (\delta_0 - \delta_b)$	n

In addition, note the relationship between $(c_1^* - c_0)$ and b_0 under this setting:

$$\begin{aligned} c_1^* - c_0 &= \left\lceil c_0 \frac{(k+1) - (c_0 - b_0)}{(n_{11} + k + k_{00}) + (c_0 - b_0)} \right\rceil - 1 \\ &= \left\lceil \underbrace{(k - c_0 + b_0)}_{>0} \underbrace{\frac{c_0}{(n_{11} + k + k_{00} - b_0) + c_0}}_{<1} + \underbrace{\frac{c_0}{(n_{11} + k + k_{00} - b_0) + c_0} - 1}_{<0} \right\rceil \\ &\leq k - c_0 + b_0; \\ \Rightarrow \delta_1 \leq \delta_b &\equiv \delta_b - \delta_1 \geq 0. \end{aligned}$$

There are three possible orderings for δ_0 relative to $\delta_1 \leq \delta_b$, and we shall consider each in turn.

(a) $\delta_1 \leq \delta_b < \delta_0$.

	A	C	D	Row
$X=1$	$n_{11} = n_{11} + (\delta_b - \delta_1) - (\delta_b - \delta_1)$	$(n_{11} + k) - \delta_0 + \delta_1$	$(k + k_{00}) + (\delta_0 - \delta_b) + (\delta_b - \delta_1)$	n
$X=0$	$n_{11} + \delta_b$	$(n_{11} + k) - \delta_0$	$(k + k_{00}) + (\delta_0 - \delta_b)$	n

$\leq$ by Lemma 3 since $\frac{(k+k_{00})+(\delta_0-\delta_b)}{(k+k_{00})+(\delta_0-\delta_b)} = 1 \leq \frac{n_{11}+\delta_b}{n_{11}+(\delta_b-\delta_1)}$

	A	C	D	Row
$X=1$	$n_{11} + \delta_b - \delta_1$	$(n_{11} + k) - \delta_0 + \delta_1$	$(k + k_{00}) + (\delta_0 - \delta_b)$	n
$X=0$	$n_{11} + \delta_b$	$(n_{11} + k) - \delta_0$	$(k + k_{00}) + (\delta_0 - \delta_b)$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	$n_{11} + \delta_b$	$k + (n_{11} - \delta_0)$	$(k + k_{00}) + (\delta_0 - \delta_b)$	n
$X=0$	$n_{11} + \delta_b$	$k + (n_{11} - \delta_0)$	$(k + k_{00}) + (\delta_0 - \delta_b)$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	$n_{11} + \delta_b$	k	$(n_{11} + k + k_{00}) - \delta_b$	n
$X=0$	$n_{11} + \delta_b$	k	$(n_{11} + k + k_{00}) - \delta_b$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

(b) $\delta_1 < \delta_0 \leq \delta_b$.

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k)-\delta_0+\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n
$X=0$	$n_{11}+\delta_1+(\delta_b-\delta_1)$	$(n_{11}+k)-\delta_0$	$(k+k_{00})+(\delta_0-\delta_1)-(\delta_b-\delta_1)$	n

$\leq$ by Lemma 3 since $\frac{n_{11}}{n_{11}+\delta_1} \leq 1 = \frac{(k+k_{00})+(\delta_0-\delta_1)}{(k+k_{00})+(\delta_0-\delta_1)}$

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k)-\delta_0+\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n
$X=0$	$n_{11}+\delta_1$	$(n_{11}+k)-\delta_0+\delta_1-\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	n_{11}	$k+(n_{11}-\delta_0)+\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n
$X=0$	n_{11}	$k+(n_{11}-\delta_0)+\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11}+k+k_{00})$	n
$X=0$	n_{11}	k	$(n_{11}+k+k_{00})$	n

(c) $\delta_0 \leq \delta_1 \leq \delta_b$. This implies that $\delta_b - \delta_0 \geq \delta_1 - \delta_0 \geq 0$.

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k)-\delta_0+\delta_1$	$(k+k_{00})+(\delta_0-\delta_1)$	n
$X=0$	$n_{11}+(\delta_b-\delta_0)+\delta_0$	$(n_{11}+k)-\delta_0$	$(k+k_{00})+(\delta_0-\delta_b)$	n

$\leq$ by Lemma 3 since $\frac{n_{11}}{n_{11}+(\delta_b-\delta_0)} \leq 1 \leq \frac{(n_{11}+k)+(\delta_1-\delta_0)}{n_{11}+k}$

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k)+(\delta_1-\delta_0)$	$(k+k_{00})+(\delta_0-\delta_1)$	n
$X=0$	$n_{11}+(\delta_b-\delta_0)$	$(n_{11}+k)$	$(k+k_{00})+(\delta_0-\delta_b)$	n

$=$

	A	C	D	Row
$X=1$	n_{11}	$(n_{11}+k)+(\delta_1-\delta_0)$	$(k+k_{00})-(\delta_1-\delta_0)$	n
$X=0$	$n_{11}+(\delta_b-\delta_1+\delta_1-\delta_0)$	$(n_{11}+k)$	$(k+k_{00})-(\delta_b-\delta_1+\delta_1-\delta_0)$	n

$\leq$ by Lemma 3 since $\frac{n_{11}}{n_{11}+(\delta_1-\delta_0)} \leq 1 = \frac{(k+k_{00})-(\delta_1-\delta_0)}{(k+k_{00})-(\delta_1-\delta_0)}$

	A	C	D	Row
$X=1$	$n_{11}+(\delta_1-\delta_0)-(\delta_1-\delta_0)$	$(n_{11}+k)+(\delta_1-\delta_0)$	$(k+k_{00})-(\delta_1-\delta_0)$	n
$X=0$	$n_{11}+(\delta_1-\delta_0)$	$n_{11}+k$	$(k+k_{00})-(\delta_1-\delta_0)$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	$n_{11}+(\delta_1-\delta_0)$	$n_{11}+k$	$(k+k_{00})-(\delta_1-\delta_0)$	n
$X=0$	$n_{11}+(\delta_1-\delta_0)$	$n_{11}+k$	$(k+k_{00})-(\delta_1-\delta_0)$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	$n_{11}+k$	$k+k_{00}$	n
$X=0$	n_{11}	$n_{11}+k$	$k+k_{00}$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11}+k+k_{00})$	n
$X=0$	n_{11}	k	$(n_{11}+k+k_{00})$	n

4. $k < c_0 - b_0 \leq n_{11} + k$. This implies that $k \leq k + b_0 < c_0$.

   Note that since $k < c_0 - b_0 \Rightarrow c_1^* < c_0 \leq n_{11} + k$, this implies that $c_1^* - k < n_{11} \leq n_{11} + k_{00}$.

   (i) First if $c_1^* \leq k < c_0 - b_0$, then $k - c_1^* \geq 0$ and:

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) + (k - c_1^*)$	n
$X=0$	$n_{11} + (k - c_0) + b_0$	$(c_0 - b_0) + b_0$	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{c_1^*}{c_0 - b_0} \leq 1 \leq \frac{(n_{11}+k+k_{00})+(k-c_1^*)}{n_{11}+k+k_{00}}$

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) + (k - c_1^*)$	n
$X=0$	$n_{11} - [(c_0 - b_0) - k]$	$k + [(c_0 - b_0) - k]$	$(n_{11} + k + k_{00})$	n

$\leq$ by Lemma 3 since $\frac{c_1^*}{k} \leq 1 = \frac{n_{11}}{n_{11}}$

	A	C	D	Row
$X=1$	n_{11}	$c_1^* = k - (k - c_1^*)$	$(n_{11} + k + k_{00}) + (k - c_1^*)$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

   (ii) Otherwise if $k < c_1^* \equiv c_1^* - k > 0$, then there are two possible orderings:

   (a) $k \leq c_1^* - b_0 < c_0 - b_0$ and

   (b) $c_1^* - b_0 < k < c_0 - b_0$.

   We consider each ordering in turn.

(a) $k \leq c_1^* - b_0 < c_0 - b_0$. This implies that $(c_1^* - b_0) - k \geq 0$ and $c_0 - c_1^* > 0$.

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11}+k+k_{00})-(c_1^*-k)$	n
$X=0$	$n_{11}-(c_1^*-b_0)+k-(c_0-c_1^*)$	$c_0 = c_1^*+(c_0-c_1^*)$	$(n_{11}+k+k_{00})-b_0$	n

$\leq$ by Lemma 3 since $\frac{c_1^*}{c_1^*} = 1 \leq \frac{n_{11}}{n_{11}-(c_1^*-b_0)+k}$

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11}+k+k_{00})-(c_1^*-k)$	n
$X=0$	$n_{11}-(c_1^*-b_0)+k$	c_1^*	$(n_{11}+k+k_{00})-(c_1^*-k)+(c_1^*-b_0)-k$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	n_{11}	$c_1^* = k+(c_1^*-k)$	$(n_{11}+k+k_{00})-(c_1^*-k)$	n
$X=0$	n_{11}	$c_1^* = k+(c_1^*-k)$	$(n_{11}+k+k_{00})-(c_1^*-k)$	n

$\leq$ by Lemma 4

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11}+k+k_{00})$	n
$X=0$	n_{11}	k	$(n_{11}+k+k_{00})$	n

(b) $c_1^* - b_0 < k < c_0 - b_0$. This implies that $(c_0 - b_0) - k > 0, c_1^* < k + b_0$ and $k - (c_1^* - b_0) > 0$.

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	$n_{11} - (c_0 - b_0) + k$	$c_0 = k + b_0 + (c_0 - b_0) - k$	$(n_{11} + k + k_{00}) - b_0$	n

$\leq$ by Lemma 3 since $\frac{c_1^*}{k+b_0} \leq 1 = \frac{n_{11}}{n_{11}}$

	A	C	D	Row
$X=1$	n_{11}	c_1^*	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	n_{11}	$k + b_0 = c_1^* + k - (c_1^* - b_0)$	$(n_{11} + k + k_{00}) - (c_1^* - k) - k + (c_1^* - b_0)$	n

$\leq$ by Lemma 3

	A	C	D	Row
$X=1$	n_{11}	$c_1^* = k + (c_1^* - k)$	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n
$X=0$	n_{11}	$c_1^* = k + (c_1^* - k)$	$(n_{11} + k + k_{00}) - (c_1^* - k)$	n

$\leq$ by Lemma 4 since $c_1^* - k > 0$

	A	C	D	Row
$X=1$	n_{11}	k	$(n_{11} + k + k_{00})$	n
$X=0$	n_{11}	k	$(n_{11} + k + k_{00})$	n

In all possible scenarios, we have shown that $u^*(b_0, c_0)$ is bounded above by the likelihood of another table $h(\boldsymbol{t})$ that leads to the same observed data $\boldsymbol{n}$ (via the linear constraints in Equation (1.2)), and has at most three non-zero response types. For a perfectly balanced dataset where $n = N - n$, the maximum likelihood solution is never achieved by just populations that have all four response types.

□

Appendix B

RANDOMIZATION-BASED SIGNIFICANCE TESTS WITH THE GENERALIZED LIKELIHOOD RATIO UNDER THE NEYMAN-RUBIN BINARY CAUSAL MODEL

In this Appendix we describe how to carry out significance tests for causal hypotheses under the Neyman-Rubin binary causal model using the generalized likelihood ratio (GLR) test statistic.

B.1 Significance Tests of Causal Hypotheses

Broadly the three components in a significance test of a causal hypothesis are:

1. Causal hypothesis H_0 that states the assumptions or restrictions on the causal parameters in the finite population $\boldsymbol{t}$ that remain fixed over hypothetical rerandomizations.
2. Observed data $\boldsymbol{n}$ and test statistic TS, where TS is a function of both $\boldsymbol{n}$ and H_0. Here we propose the GLR λ as the test statistic, with *smaller* values of λ being further deviations from H_0. One may choose a different test statistic, for example the (maximum) likelihood under H_0.
3. Sampling distribution $\Pr(\boldsymbol{n} \mid H_0)$, which in turn induces a null probability distribution for TS. Depending on H_0, there may be more than one sampling distribution for $\boldsymbol{n}$, where each sampling distribution corresponds to a specified value of the population parameters $\boldsymbol{t}$. We will use the tilde accent $(\sim)$ to indicate random quantities that vary over hypothetical rerandomizations, so that possibly observed datasets and the corresponding test statistics are $\tilde{\boldsymbol{n}}$ and $\tilde{\text{TS}}$ respectively. The sample space of $\tilde{\boldsymbol{n}}$ that may arise from a fixed population $\boldsymbol{t}$ is denoted $\Omega(\boldsymbol{t})$, so that

$$\sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t})} \Pr(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}) = 1,$$

where $\Pr(\tilde{\boldsymbol{n}} \mid \boldsymbol{t})$ is the probability of observing $\tilde{\boldsymbol{n}}$ given $\boldsymbol{t}$ under the *Neyman-Rubin-Copas* likelihood in (1.8).

We will consider two classes of causal hypotheses here. The first is the 'Fisher' null $H_0(\text{Fisher})$, which is a simple null hypothesis since $\boldsymbol{t}^0 = (s, 0, 0, N-s)$ is the only population in the parameter space $T(\boldsymbol{n})$ that satisfies $H_0(\text{Fisher}) : t_B = t_C = 0$. The sampling distribution $\Pr(\boldsymbol{n} \mid H_0(\text{Fisher}))$ is thus well-defined. The p-value for the 'Fisher' null is just the total probability of observing a dataset $\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t}^0)$ with test statistic $\tilde{\text{TS}}$ at least as extreme as the observed value:

$$pv(\text{Fisher}) = \sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t}^0)} \Pr(\tilde{\boldsymbol{n}} \mid H_0(\text{Fisher})) \times \mathbb{1}\{\tilde{\text{TS}} \leq \text{TS}\}. \tag{B.1}$$

Here $\mathbb{1}\{E\}$ returns a value of 1 if the condition E is met, and 0 otherwise. Choosing TS to be the likelihood under the 'Fisher' null $\Pr(\boldsymbol{n} \mid H_0(\text{Fisher}))$ thus returns the two-sided p-value from Fisher's exact test [Fisher, 1935].

Next we define $H_0(\delta = c)$ as a causal hypothesis where for some given value c, the population average causal effect (ACE) is fixed at $\text{ACE} \equiv \delta/N \equiv (t_B - t_C)/N = c/N$. By definition, the 'Neyman' null corresponds to a value of $c = 0$, so that $H_0(\delta = 0) \equiv H_0(\text{Neyman})$. In general, $H_0(\delta = c)$ is a composite causal hypothesis: there may be multiple populations $\boldsymbol{t} \in T(\boldsymbol{n})$ where $t_B - t_C = c$. Denote the subspace of $T(\boldsymbol{n})$ containing populations $\boldsymbol{t} \in T(\boldsymbol{n})$ that satisfy $H_0(\delta = c)$ as:

$$T(\boldsymbol{n}; \delta = c) \equiv \{\boldsymbol{t} \in T(\boldsymbol{n}) : t_B - t_C = c\}. \tag{B.2}$$

B.2 Randomization-based Confidence Intervals for the Average Causal Effect

Rigdon and Hudgens [2015], henceforth RH, have described how, given an observed dataset $\boldsymbol{n}$, to invert a sequence of permutation tests for the complete tables $\boldsymbol{m} \in \mathcal{N}(\boldsymbol{n})$ to construct a two-sided confidence interval for the $\text{ACE} \equiv \delta/N$.

We restate the procedure here in terms of the causal hypotheses $H_0(\delta = c)$. First the set of all possible values that δ can take given an observed dataset is just:

$$\Delta(\boldsymbol{n}) \equiv \bigcup_{\boldsymbol{t} \in T(\boldsymbol{n})} t_B - t_C; \tag{B.3}$$

the total number of unique causal hypotheses $H_0(\delta = c)$ is then $|\Delta(\boldsymbol{n})|$. As RH point out, the set of values for $\delta \in \Delta(\boldsymbol{n})$ are bounded between $-N$ and N, so that the $-1 \leq \delta/N \leq 1$.

For a given causal hypothesis $H_0(\delta = c)$, define the GLR for testing $H_0(\delta = c)$ as:

$$\lambda(\delta = c) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n}; \delta = c) \subset T(\boldsymbol{n})} \Pr(\boldsymbol{n}|\boldsymbol{t}) \Big/ \max_{\boldsymbol{t} \in T(\boldsymbol{n})} \Pr(\boldsymbol{n}|\boldsymbol{t}). \tag{B.4}$$

RH propose the absolute difference between the estimate of the ACE and c as the test statistic:

$$|\widehat{\text{ACE}} - c| = \left| \left(\frac{n_{11}}{n} - \frac{n_{10}}{N-n} \right) - c \right|. \tag{B.5}$$

Note that both $\lambda(\delta = c)$ and $|\widehat{\text{ACE}} - c|$ are functions only of $\boldsymbol{n}$ and the fixed constant c, and *not* the specific population $\boldsymbol{t}$: the exact population totals $\boldsymbol{t}$ are treated as a nuisance parameter under $H_0(\delta = c)$.

For each population $\boldsymbol{t}$ in $T(\boldsymbol{n}; \delta = c)$, the population-specific p-value is then:

$$pv(\delta = c; \boldsymbol{t}) = \sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t})} \Pr\left(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}\right) \times \mathbb{1}\Big\{ \tilde{\lambda}(\delta = c) \leq \lambda(\delta = c) \Big\}. \tag{B.6}$$

The overall p-value for $H_0(\delta = c)$ is then the maximum over all population-specific p-values:

$$pv(\delta = c) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n}; \delta = c)} pv(\delta = c; \boldsymbol{t}). \tag{B.7}$$

$pv(\delta = c)$ is hence a valid but conservative p-value for $H_0(\delta = c)$, since under $H_0(\delta = c)$, the probability of observing a value of $\tilde{\lambda}(\delta = c)$ that is at least as extreme as the observed value $\lambda(\delta = c)$ is *at most* $pv(\delta = c)$. As RH point out, for a pre-specified value of α, there may not be a need to find the exact value of $pv(\delta = c)$ if there is some population $\boldsymbol{t}'$ where $pv(\delta = c; \boldsymbol{t}') \geq \alpha$, since

$$\alpha \leq pv(\delta = c; \boldsymbol{t}') \leq \max_{\boldsymbol{t} \in T(\boldsymbol{n}; \delta = c)} pv(\delta = c; \boldsymbol{t}) \equiv pv(\delta = c).$$

An exact randomization-based $100(1-\alpha)\%$ confidence interval for δ is then the set of values for δ where $pv(\delta) \geq \alpha$:

$$\Delta_\alpha(\boldsymbol{n}) \equiv \{\delta \in \Delta(\boldsymbol{n}) : pv(\delta) \geq \alpha\}. \tag{B.8}$$

B.3 Statistical Inference for the 'Fisher' and 'Neyman' Null Hypotheses

In this section we examine the 'Neyman' null corresponding to $\delta = 0$, so that $H_0(\delta = 0) \equiv H_0(\text{Neyman})$. We show that using the GLR λ, the resulting p-value from a significance test of the stricter 'Fisher' null can never exceed the p-value from a significance test of the 'Neyman' null.

Theorem 4. *For a given observed table $\boldsymbol{n}$, the 'Neyman' null assumes a zero population average causal effect while the stricter 'Fisher' null assumes all individual causal effects are zero. Under the Neyman-Rubin-Copas likelihood, the overall maximum p-value for testing the composite 'Neyman' null using the GLR test statistic is at least as large as the p-value for testing the simple 'Fisher' null.*

Proof. The 'Neyman' null $H_0(\text{Neyman}) \equiv H_0(\delta = 0)$ is a composite null hypothesis. Denote the subspace of $T(\boldsymbol{n})$ containing populations $\boldsymbol{t} \in T(\boldsymbol{n})$ that satisfy $H_0(\text{Neyman})$ as:

$$T(\boldsymbol{n}; \delta = 0) \equiv \{\boldsymbol{t} \in T(\boldsymbol{n}) : t_B = t_C\}. \tag{B.9}$$

One such population that satisfies 'Neyman' null is $\boldsymbol{t}^0 = (s, 0, 0, N - s)$, which is also the only population in $T(\boldsymbol{n})$ that satisfies the 'Fisher' null $H_0(\text{Fisher})$. The GLR $\lambda(\text{Neyman})$ is:

$$\lambda(\text{Neyman}) \equiv \max_{\boldsymbol{t} \in T(\boldsymbol{n}; \delta = 0) \subset T(\boldsymbol{n})} h(\boldsymbol{t}) \Big/ \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}); \tag{B.10}$$

whereas the GLR $\lambda(\text{Fisher})$ is:

$$\lambda(\text{Fisher}) \equiv h(\boldsymbol{t}^0) \Big/ \max_{\boldsymbol{t} \in T(\boldsymbol{n})} h(\boldsymbol{t}) = h(\boldsymbol{t}^0)/h(\hat{\boldsymbol{t}}).$$

From Theorem 2, both test statistics are equal: $\lambda(\text{Neyman}) = \lambda(\text{Fisher})$.

Under the fixed population $\boldsymbol{t}^0$, denote the sample space for $\tilde{\boldsymbol{n}}$ as $\Omega(\boldsymbol{t}^0)$. The p-value for testing the 'Fisher' null using $\lambda(\text{Fisher})$ is then equal to the p-value for testing the 'Neyman' null $\lambda(\text{Neyman})$:

$$
\begin{aligned}
pv(\text{Fisher}) &\equiv \sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t}^0)} \Pr\left(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}^0\right) \times \mathbb{1}\left\{\tilde{\lambda}(\text{Fisher}) \leq \lambda(\text{Fisher})\right\} \\
&= \sum_{\tilde{\boldsymbol{n}} \in \Omega(\boldsymbol{t}^0)} \Pr\left(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}^0\right) \times \mathbb{1}\left\{\tilde{\lambda}(\text{Neyman}) \leq \lambda(\text{Neyman})\right\} \\
&\equiv pv(\text{Neyman}; \boldsymbol{t}^0).
\end{aligned}
$$

The p-value for testing $H_0(\text{Fisher})$ is thus a lower bound for the overall p-value for testing $H_0(\text{Neyman})$, since

$$
pv(\text{Fisher}) = pv(\text{Neyman}; \boldsymbol{t}^0) \leq \max_{\boldsymbol{t} \in T(\boldsymbol{n}; \delta = 0)} pv(\text{Neyman}; \boldsymbol{t}) \equiv pv(\text{Neyman}). \tag{B.11}
$$

This implies the following inferential relationships between the 'Neyman' and 'Fisher' null hypotheses:

$$
pv(\text{Fisher}) \geq \alpha \Rightarrow pv(\text{Neyman}) \geq \alpha; \quad pv(\text{Neyman}) \leq \alpha \Rightarrow pv(\text{Fisher}) \leq \alpha.
$$

□

B.4 Efficient Complete Enumeration of the Sample Space $\Omega(\boldsymbol{n})$

In general, to generate the sample space of possibly observable datasets $\Omega(\boldsymbol{t})$ for a given population $\boldsymbol{t}$, one must first generate assignments in the complete table $\tilde{\boldsymbol{m}}$, either through complete enumeration using for example [Gail and Mantel, 1977], or via Monte Carlo simulation. The entries in $\tilde{\boldsymbol{m}}$ are then aggregated using the linear constraints in (1.2) to obtain $\tilde{\boldsymbol{n}}$.

However, this is computationally inefficient since different complete tables $\boldsymbol{m}$ and $\boldsymbol{m}' \neq \boldsymbol{m}$ (either from the same population or from different populations) may result in the same dataset $\tilde{\boldsymbol{n}}$. Denote the sample space of all possibly observable datasets $\tilde{\boldsymbol{n}}$ given an observed dataset $\boldsymbol{n}$ as:

$$
\Omega(\boldsymbol{n}) = \bigcup_{\boldsymbol{t} \in T(\boldsymbol{n})} \Omega(\boldsymbol{t}). \tag{B.12}
$$

Under a completely randomized design, where the sizes of the treatment groups remain fixed over hypothetical rerandomizations, it is sufficient to specify values for one cell count in each row, for example $(\tilde{n}_{11}, \tilde{n}_{10})$. $\Omega(\boldsymbol{n})$ is then the intersection of the two-dimensional integer lattice $(\{0\} \cup \mathbb{Z}^+)^2$ and the convex polyhedron described with the inequalities (1.40)−(1.41), which we restate here as:

$$\begin{aligned} \max(0, n_{11} + n - N) \leq \tilde{n}_{11} \leq \min(s + n_{00}, n), \qquad & \tilde{n}_{01} = n - \tilde{n}_{11} \\ \max(0, n_{10} - n) \leq \tilde{n}_{10} \leq \min(s + n_{01}, N - n), \qquad & \tilde{n}_{00} = N - n - \tilde{n}_{10}. \end{aligned}$$

In general, a test statistic $\tilde{\text{TS}}$ is a function of the possibly observed dataset $\tilde{\boldsymbol{n}}$, and not the complete tables $\tilde{\boldsymbol{m}}$ nor the population totals $\boldsymbol{t}$. By enumerating $\Omega(\boldsymbol{n})$ once, we may calculate $\tilde{\text{TS}}$ for each $\tilde{\boldsymbol{n}}$ once (for each H_0), and avoid any redundant calculations of $\tilde{\text{TS}}$ for $\tilde{\boldsymbol{n}}$ that may arise under multiple populations. Furthermore, such an approach reduces the need for Monte Carlo simulation (especially for small datasets), and does away with the intermediate step of generating complete tables $\tilde{\boldsymbol{m}}$ for each population $\boldsymbol{t}$ in turn. Datasets $\tilde{\boldsymbol{n}}$ that could not have arisen under some given population $\boldsymbol{t}$ then have probability zero; in other words, $\boldsymbol{t} \notin T(\tilde{\boldsymbol{n}}) \iff \Pr(\tilde{\boldsymbol{n}} \mid \boldsymbol{t}) = 0$.

B.5 Simulation Results

We now compare the confidence intervals constructed using (i) the procedure described in RH with Monte Carlo simulations (specifically using the `RI2by2::Perm.CI` function [Rigdon, 2014] in `R` [R Core Team, 2015]), (ii) $\Omega(\boldsymbol{n})$ with the $|\widehat{\text{ACE}} - \delta|$ test statistic, and (iii) $\Omega(\boldsymbol{n})$ with the GLR λ test statistic. Intervals constructed using the last two procedures are exact in the sense that there is no Monte Carlo error.

The simulations under Scenario 1 in RH for sample size 20 are then replicated 5000 times for each true value of the ACE between 0.2 and 0.95. The average widths of the confidence intervals are plotted in Figure B.1 below. The intervals constructed with Monte Carlo (MC) simulation are the widest due to the approximation of the randomization distribution, while the widths of the exact intervals using either test statistic are quite similar. Under the setting where $\Pr(X = 1) = 0.5$, the

confidence intervals using the GLR test statistic appear to be slightly narrower on average than those using $|\widehat{\text{ACE}} - \delta|$.

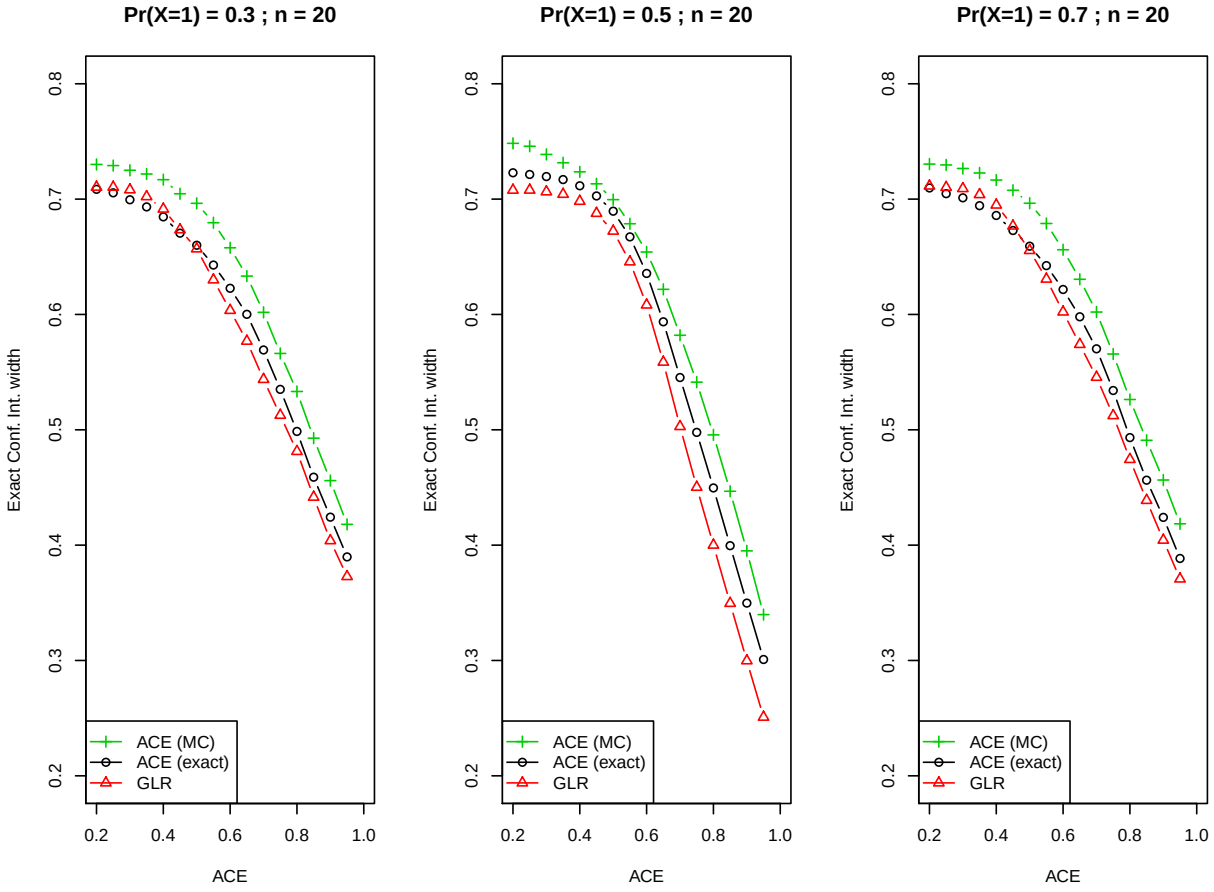

Figure B.1: Comparison of randomization-based confidence intervals in simulations under Scenario 1 in [Rigdon and Hudgens, 2015] for sample size 20; 5000 replicates were made at each true value of the ACE, located at equally spaced intervals of 0.05 between 0.2 and 0.95 (indicated by the points).

Appendix C

MAXIMUM LIKELIHOOD OF A 2×2 OBSERVED DATASET

Here we present the results needed to find the maximum likelihood of an observed table $\hat{h}$ as defined in Equation (1.11) of Chapter 1.

C.1 Maximizing the Hypergeometric Probability in a 2×2 Table

We review some existing results from [Johnson and Kotz, 1969] and [Zhang, 2009] below. Recall the following 2×2 table where the row totals $(k, N-k)$ and the counts in one row $(b, (N-k)-b)$ are fixed.

Table C.1: 2×2 Table With Unknown Column Totals

	Red	Green	Row
Not drawn	x	$k-x$	k
Drawn	b	$(N-k)-b$	$N-k$
Column	$b+x$	$N-(b+x)$	N

Our interest is in maximizing the hypergeometric probability with respect to x:

$$\Pr(x \,|\, (k, b, N)) = \binom{b+x}{x}\binom{N-(b+x)}{k-x}\Big/\binom{N}{k}.$$

Theorem 5. *In a 2×2 table where the row totals $(k, N-k)$ and the counts in one row $(b, (N-k)-b)$ are fixed, the most likely value of $x \in [0, k]$ under the randomization assumption is:*

$$\hat{x} = \arg\max_{x\in[0,k]} \left\{ x < (k+1)\frac{b}{N-k} \right\} = \left\lfloor \left\lfloor (k+1)\frac{b}{N-k} \right\rfloor \right\rfloor,$$

where the 'basement' function $\lfloor\lfloor a \rfloor\rfloor$ is defined as:

$$\lfloor\lfloor a \rfloor\rfloor = \max\{0, \lceil a \rceil - 1\}.$$

Equivalently,

$$b+\hat{x} = \begin{cases} \lfloor b\frac{N}{N-k} \rfloor & \textit{if } b\frac{N+1}{N-k} \leq \lceil b\frac{N}{N-k} \rceil, \\ \lceil b\frac{N}{N-k} \rceil & \textit{otherwise.} \end{cases}$$

Proof. We are interested in the following $k+1$ binomial coefficient products to find the most likely value of x:

$$\left\{ \binom{b+x}{x} \binom{N-(b+x)}{k-x}, \quad x \in [0,k] \right\}.$$

The first terms in each product are an increasing sequence in x, the second terms are a decreasing sequence in x for all values of $x \in \{0,\ldots,k-1\}$.

$$\begin{aligned} \binom{b+x}{x} &= \binom{b+x+1}{x+1} - \binom{b+x}{x+1} \\ &< \binom{b+x+1}{x+1}; \\ \binom{N-(b+x)}{k-x} &= \binom{N-(b+x+1)}{k-x} + \binom{N-(b+x+1)}{k-(x+1)} \\ &> \binom{N-(b+x+1)}{k-(x+1)}. \end{aligned}$$

The relationship between consecutive products due to a unit increase in x is thus:

$$\begin{aligned} &\binom{b+x+1}{x+1}\binom{N-(b+x+1)}{k-(x+1)} \\ &\quad = \binom{b+x}{x}\binom{N-(b+x)}{k-x} \times \frac{b+x+1}{x+1} \times \frac{k-x}{N-(b+x)} \\ &\quad > \binom{b+x}{x}\binom{N-(b+x)}{k-x}; \end{aligned}$$

$$\begin{aligned} &\iff && \frac{b+x+1}{x+1} > \frac{N-(b+x)}{k-x} \\ &\iff && \frac{k-x}{x+1} > \frac{N-(b+x)}{b+x+1} \\ &\iff && \frac{k-x+(x+1)}{x+1} > \frac{N-(b+x)+(b+x+1)}{b+x+1} \\ &\iff && \frac{b+x+1}{x+1} > \frac{N+1}{k+1} \\ &\iff && x+1 < (k+1)\frac{b}{N-k}. \end{aligned} \tag{C.1}$$

Since (k, b, N) are fixed, the sequence of binomial coefficient products will increase with x until it exceeds the critical value $(k+1)\frac{b}{N-k}$ and the inequality no longer holds. The most likely value of x is then:

$$\begin{aligned}\hat{x} &= \arg\max_{x\in[0,k]}\left\{x < (k+1)\frac{b}{N-k}\right\} \\ &= \begin{cases} \left\lfloor (k+1)\frac{b}{N-k}\right\rfloor & \text{if } (k+1)\frac{b}{N-k} > \left\lfloor (k+1)\frac{b}{N-k}\right\rfloor, \\ (k+1)\frac{b}{N-k} - 1 & \text{if } (k+1)\frac{b}{N-k} = \left\lfloor (k+1)\frac{b}{N-k}\right\rfloor. \end{cases} \\ &= \left\lfloor \left\lfloor (k+1)\frac{b}{N-k}\right\rfloor \right\rfloor,\end{aligned}$$

where the 'basement' function is defined as $\lfloor\lfloor x\rfloor\rfloor = \max\{0, \lceil x\rceil - 1\}$. This is the same result previously shown in [Zhang, 2009]. Johnson and Kotz [1969, page 146] give an equivalent result in terms of the column total $(b + x)$.

We may also rewrite the critical value $(k+1)\frac{b}{N-k}$ as follows:

$$\begin{aligned}(k+1)\frac{b}{N-k} &= k\frac{b}{N-k} + \frac{b}{N-k} \\ &= \left\lceil k\frac{b}{N-k}\right\rceil + \frac{b}{N-k} - \left(\left\lceil k\frac{b}{N-k}\right\rceil - k\frac{b}{N-k}\right).\end{aligned}$$

Since $0 < \frac{b}{N-k} < 1$ and $0 \le \lceil k\frac{b}{N-k}\rceil - k\frac{b}{N-k} < 1$, we see that:

$$-1 < (k+1)\frac{b}{N-k} - \left\lceil k\frac{b}{N-k}\right\rceil < 1.$$

There are now two cases to consider:

1. If $1 > (k+1)\frac{b}{N-k} - \lceil k\frac{b}{N-k}\rceil > 0$, then

$$\begin{aligned}&(k+1)\frac{b}{N-k} > \left\lceil k\frac{b}{N-k}\right\rceil, \\ &\Rightarrow \left\lceil k\frac{b}{N-k}\right\rceil = \arg\max_{x\in[0,k]}\left\{x < (k+1)\frac{b}{N-k}\right\}.\end{aligned}$$

When $k\frac{b}{N-k}$ is a positive integer, such as in a balanced table where $k = N - k$, then $k\frac{b}{N-k} = \lceil k\frac{b}{N-k}\rceil$. Since $\frac{b}{N-k} > 0$,

$$\Rightarrow (k+1)\frac{b}{N-k} > k\frac{b}{N-k} = \left\lceil k\frac{b}{N-k}\right\rceil.$$

2. If $-1 < (k+1)\frac{b}{N-k} - \left\lceil k\frac{b}{N-k} \right\rceil \leq 0$, then

$$\begin{aligned}
&\left\lceil k\frac{b}{N-k} \right\rceil - 1 < (k+1)\frac{b}{N-k} \leq \left\lceil k\frac{b}{N-k} \right\rceil \\
\Rightarrow &\arg\max_{x\in[0,k]} \left\{ x < (k+1)\frac{b}{N-k} \right\} \\
&= \left\lceil k\frac{b}{N-k} \right\rceil - 1 \\
&= \left\lfloor k\frac{b}{N-k} \right\rfloor .
\end{aligned}$$

This gives us the following result:

$$\begin{aligned}
\hat{x} &= \arg\max_{x\in[0,k]} \left\{ x < (k+1)\frac{b}{N-k} \right\} \\
&= \begin{cases} \left\lfloor k\frac{b}{N-k} \right\rfloor & \text{if } (k+1)\frac{b}{N-k} \leq \left\lceil k\frac{b}{N-k} \right\rceil, \\ \left\lceil k\frac{b}{N-k} \right\rceil & \text{otherwise.} \end{cases}
\end{aligned}$$

□

C.2 Maximizing the hypergeometric probability in a 2×3 *table*

Consider the 2×3 table below where we only get to observe the exact cell counts n_1, n_2 and the sums $a > 0$ and $b > 0$. By symmetry of the rows, we shall assume that $a \leq b$ without loss of generality.

A	B	C	Row total
n_1	x	$a-x$	$a+n_1$
y	n_2	$b-y$	$b+n_2$

The pair of missing values (x, y) takes values in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{0, \ldots, a\}, \mathcal{Y} = \{0, \ldots, b\}$. Our interest is in finding the global maximum hypergeometric probability $\hat{h}$, where:

$$\begin{aligned} h(x,y) &= \binom{n_1+y}{n_1}\binom{n_2+x}{n_2}\binom{(a-x)+(b-y)}{(a-x)} \\ &= \frac{(n_1+y)!}{n_1!y!}\frac{(n_2+x)!}{n_2!x!}\frac{[(a-x)+(b-y)]!}{(a-x)!(b-y)!} \\ &\propto \frac{(n_1+y)!}{y!}\frac{(n_2+x)!}{x!}\frac{[(a-x)+(b-y)]!}{(a-x)!(b-y)!}; \\ \hat{h} &\equiv \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} h(x,y). \end{aligned}$$

First if $n_1 = 0$, then $\hat{h}$ is attained by $y = 0$ and $\hat{x} = \left\lfloor\left\lfloor(a+1)\frac{n_2}{b+n_2}\right\rfloor\right\rfloor$, where $\hat{x}$ is from applying Theorem 5 to the 2×2 subtable with columns B and C. Similarly if $n_2 = 0$, $\hat{h}$ is attained by $x = 0$ and $\hat{y} = \left\lfloor\left\lfloor(b+1)\frac{n_1}{a+n_2}\right\rfloor\right\rfloor$.

For tables where $\min(n_1, n_2) > 0$, $\hat{h}$ may be found by evaluating $h(x, y)$ for all values in $\mathcal{X}\times\mathcal{Y}$. Instead, first condition on $x \in \mathcal{X}$ (since $|\mathcal{X}| \leq |\mathcal{Y}|$), then find the corresponding (conditional) maximum $h(x, \hat{y}(x))$ by applying Theorem 5 to the 2×2 subtable for columns A and C, so that

$$\hat{y}(x) = \left\lfloor\left\lfloor(b+1)\frac{n_1}{n_1+a-x}\right\rfloor\right\rfloor. \tag{C.2}$$

The search space for $\hat{h}$ may then be reduced to a linear space (of size a) since

$$\hat{h} = \max_{x\in\mathcal{X}} h(x, \hat{y}(x)).$$

However, by the symmetry of the rows, if we conditioned on $y \in \mathcal{Y}$ first and found the (conditional) maximum $h(\hat{x}(y), y)$, where

$$\hat{x}(y) = \left\lfloor\left\lfloor(a+1)\frac{n_2}{n_2+b-y}\right\rfloor\right\rfloor, \tag{C.3}$$

then the global maximum would be:

$$\hat{h} = \max_{y\in\mathcal{Y}} h(\hat{x}(y), y).$$

This implies that the equalities (C.2) and (C.3) are *necessary* conditions for $\hat{h}$. There is also at least one pair of values for (x, y) that satisfies both (C.2) and (C.3) since if we fix $x = a$, then

$$\widehat{y}(a) = \left\lfloor \left\lfloor (b+1)\frac{n_1}{n_1 + a - a} \right\rfloor \right\rfloor = (b+1) - 1 = b, \quad \widehat{x}(b) = \left\lfloor \left\lfloor (a+1)\frac{n_2}{n_2 + b - b} \right\rfloor \right\rfloor = a.$$

We may then reduce the search space to a sub-linear space simply by checking for each $x \in \mathcal{X}$ whether

$$x = \hat{x}(\hat{y}(x)), \quad \hat{x}(\hat{y}(x)) = \left\lfloor \left\lfloor (a+1)\frac{n_2}{n_2 + b - \hat{y}(x)} \right\rfloor \right\rfloor. \tag{C.4}$$

Then the global maximum likelihood is:

$$\hat{h} = \max_{x \in \mathcal{X} : x = \hat{x}(\hat{y}(x))} h(x, \hat{y}(x)). \tag{C.5}$$

C.3 Maximum Likelihood under the Never Only Four Conjecture

It follows from the Never Only Four Conjecture (Conjecture 1) that the maximum likelihood $\hat{h}$ for a given observed table $\boldsymbol{n}$ may be found by comparing only the populations $\boldsymbol{t} \in T(\boldsymbol{n})$ with three or fewer types. In other words, we need only evaluate values of $\boldsymbol{t}$ along the edges and (both) vertices to find $\hat{h}$, where the likelihood $h(\boldsymbol{t})$ is the probability of a single complete table $\boldsymbol{m}$. For each possible edge where $t_i = 0$, the complete table $\boldsymbol{m}$ reduces to a 2×3 table in which we may apply the result in Section C.2 to find the conditional maximum likelihood, say

$$\hat{h}_i = \max_{\{\boldsymbol{t} \in T(\boldsymbol{n}) \,:\, t_i = 0\}} h(\boldsymbol{t}).$$

There are generally only two vertices in $T(\boldsymbol{n})$ where $\boldsymbol{t}$ has exactly two types: $\boldsymbol{t}^0 \equiv (s, 0, 0, N - s)$ and $\boldsymbol{t}^1 \equiv (0, n_{11} + n_{00}, n_{10} + n_{01}, 0)$. Then $\hat{h}$ is just

$$\hat{h} = \max \left\{ h(\boldsymbol{t}^0), h(\boldsymbol{t}^1), \max_i \hat{h}_i \right\}.$$

Appendix D

PARAMETER AND SAMPLE SPACE FOR AN OBSERVED DATASET UNDER THE BINARY IV MODEL ASSUMING MONOTONICITY

In this Appendix, we explain how to obtain the parameter space $T(\boldsymbol{n})$ and sample space $\Omega(\boldsymbol{n})$ for an observed dataset $\boldsymbol{n}$ under the binary IV model without Defiers. $\boldsymbol{n}$, $T(\boldsymbol{n})$ and $\Omega(\boldsymbol{n})$ are jointly described by a convex polyhedron (a set of solutions to a finite system of linear inequalities). Each observed quantity in $\boldsymbol{n}$ is a single dimension in the polyhedron, but given the constraint that $\boldsymbol{n}$ sums to N, we choose to write $n_{y_1x_0z_1}$ as the difference between N and the other seven quantities $\boldsymbol{n}$. We do the same for $\tilde{\boldsymbol{n}}$. However, we will keep the eight parameters in $\boldsymbol{t}$ intact. There are hence a total of $(8-1)+(8-1)+8=22$ dimensions in the polyhedron.

First, we list the vertices of the polyhedron using a matrix, say `V0`, with 22 columns (one for each dimension) and each row representing an extreme point. `V0` is shown in Figure D.1. We use the following examples to explain how to derive `V0`.

Consider a population containing a single Always Taker named 'R' with an observed outcome $Y=0$. If R is assigned to the $Z=0$ group, R would have observed values $(Y=0, X=1, Z=0)$, so that $n_{y_0x_1z_0}=1$. If R is assigned to the $Z=1$ group instead, the observed values would be $(Y=0, X=1, Z=1)$, so that $n_{y_0x_1z_1}=1$. There are then 2×2 possible combinations of the observed values $\boldsymbol{n}$ and the possibly observable values $\tilde{\boldsymbol{n}}$ (over hypothetical rerandomizations) for R. Each of these combinations is described in the first four rows of Figure D.1.

Now consider another population with a single Complier named 'H' who is of type Helped. If H is assigned to the $Z=0$ group, we would observe $n_{y_0x_0z_0}=1$. Whereas if H is assigned to the $Z=1$ group, we would observe $n_{y_1x_1z_1}=1$, (which is equivalent to all the other seven values of $\boldsymbol{n}$ being zero). There are again 2×2 possible combinations of $\boldsymbol{n}$ and $\tilde{\boldsymbol{n}}$ for H, as described in rows 13 to 16 of Figure D.1. The remaining rows of `V0` are then obtained by considering the other

population parameters similarly, so that there are a total of $8 \times 4 = 32$ rows in `V0`.

We then use `rcdd` [Geyer et al., 2015] in `R` [R Core Team, 2015] to convert `V0` to a H-representation (via facet enumeration). The linear constraint that $n_{z_0} = \tilde{n}_{z_0}$ is then added to the H-representation since under complete randomization, the sizes of the treatment groups are fixed over hypothetical rerandomizations. Finally, we convert back to a V-representation (via vertex enumeration) of the convex polyhedron again. The `R` commands are:

1. `qux.ppt <- makeV(V0)`

   This appends the two columns `l` and `b` to `V0` to obtain a V-representation of the convex polyhedron. Here `l` and `b` are just indicators with values 0 and 1 respectively, and are used to flag which vertices in `V0` the points in the convex polyhedron are linear combinations of.

2. `qux <- z2q(qux.ppt, rep(1, length(qux.ppt)))`

   This formats the integer inputs as rational arithmetic representations.

3. ```
   out <- scdd(qux, representation = "V")
   H0 <- out$output
   colnames(H0) <- c("type", "b", colnames(V0))
   ```

   This converts the V-representation to a H-representation, which characterizes the polyhedron as the intersection of a finite collection of closed half spaces. The H-representation is then saved as a matrix `H0`.

4. ```
   H1 <- rbind(H0, 0)
   H1[nrow(H1), c("type", "n000", "n100", "n010", "n110",
                  "nt000", "nt100", "nt010", "nt110")] <-
     c(1, rep(1,4), rep(-1,4))
   ```

   The constraint that $n_{z_0} = \tilde{n}_{z_0}$ is represented as $\sum_{j,k=0}^{1} n_{y_k x_j z_0} - \sum_{j,k=0}^{1} \tilde{n}_{y_k x_j z_0} = 0$. `type`=1 is used to indicate that this is a linear equality.

5. ```
   out <- scdd(H1, representation = "H")
   V1 <- out$output
   ```

   This converts the H-representation that includes the linear constraint under complete randomization back to a V-representation.
   ```

To obtain the parameter space $T(\boldsymbol{n})$ for the observed dataset $\boldsymbol{n}$, we have to find the projection of the 22-dimensional convex polyhedron `V1` onto a 15-dimensional convex polyhedron for $T(\boldsymbol{n})$ and $\boldsymbol{n}$. We do this by dropping the dimensions corresponding to $\tilde{\boldsymbol{n}}$ from `V1`, and then converting to a H-representation to obtain the set of linear (in)equalities for $\boldsymbol{n}$ and $\boldsymbol{t}$.

```
Vpsi <- V1[, -c(18:ncol(V1))]
out <- scdd(Vpsi, representation = "V")
Hpsi <- out$output
```

`Hpsi` is shown in Figure D.2. Denote v as the vector of dimensions (corresponding to the column names of `Hpsi`), b_i as the i-th element of the column vector `b`, and a_i as the i-th row of `Hpsi` but without the first two elements for `type` and `b`. If `type`=0, then that row in the H-representation is read as

$$b_i + a_i v \geq 0;$$

otherwise if `type`=1, then that row is:

$$b_i + a_i v = 0.$$

To obtain the sample space $\Omega^0(\boldsymbol{n})$ under the 'sharp' null for the observed dataset $\boldsymbol{n}$, we have to first add the following constraints for zero Compliers who are Helped or Hurt to the H-representation `H1`:

```
H2 <- rbind(H1, 0, 0)
H2[nrow(H2)-1, c("type","CoHE")] <- c(1, 1)
H2[nrow(H2), c("type","CoHU")] <- c(1, 1)
```

We then convert back to a V-representation `V2`, drop the dimensions corresponding to $\boldsymbol{t}$ from `V2`, and then convert to a H-representation again to obtain the set of linear inequalities for $\boldsymbol{n}$ and $\tilde{\boldsymbol{n}}$.

```
out <- scdd(H2, representation = "H")
V2 <- out$output
```

```
Vppt <- V2[, -c(10:17)]
out <- scdd(Vppt, representation = "V")
Hppt <- out$output
```

`Hppt` is shown in Figure D.3, and interpreted in the same manner as `Hpsi`.

	$n_{y_0x_0z_0}$	$n_{y_1x_0z_0}$	$n_{y_0x_1z_0}$	$n_{y_1x_1z_0}$	$n_{y_0x_0z_1}$	$n_{y_0x_1z_1}$	$n_{y_1x_1z_1}$	t_0^{AT}	t_0^{NT}	t_{NR}^{CO}	t_{HE}^{CO}	t_{HU}^{CO}	t_{AR}^{CO}	t_1^{NT}	t_1^{AT}	$\tilde{n}_{y_0x_0z_0}$	$\tilde{n}_{y_1x_0z_0}$	$\tilde{n}_{y_0x_1z_0}$	$\tilde{n}_{y_1x_1z_0}$	$\tilde{n}_{y_0x_0z_1}$	$\tilde{n}_{y_0x_1z_1}$	$\tilde{n}_{y_1x_1z_1}$
1			1					1										1			
2			1					1													1
3						1		1													1
4						1		1										1			
5	1								1							1					
6	1								1											1	
7					1				1											1	
8					1				1							1					
9	1									1						1					
10	1									1											1
11						1				1											1
12						1				1						1					
13	1										1					1					
14	1										1										1
15							1				1										1
16							1				1					1					
17		1										1					1				
18		1										1									1
19						1						1									1
20						1						1					1				
21		1													1		1				
22		1													1						1
23							1								1						1
24							1								1		1				
25		1											1				1				
26		1											1								
27													1								
28													1				1				
29				1										1					1		
30				1										1							1
31							1							1							1
32							1							1					1		

Figure D.1: V-representation of the convex polyhedron for the observed dataset $\boldsymbol{n}$, parameter space $T(\boldsymbol{n})$ and the sample space $\Omega(\boldsymbol{n})$; missing values here are zeroes.

	type	b	$n_{y_0x_0z_0}$	$n_{y_1x_0z_0}$	$n_{y_0x_1z_0}$	$n_{y_1x_1z_0}$	$n_{y_0x_0z_1}$	$n_{y_0x_1z_1}$	$n_{y_1x_1z_1}$	t_0^{AT}	t_0^{NT}	t_{NR}^{CO}	t_{HE}^{CO}	t_{HU}^{CO}	t_1^{NT}	t_1^{AT}	t_{AR}^{CO}
1							-1				1							
2			1			1	1		1		-1		-1				-1	
3					-1			-1		1		1		1				
4												1						
5			1				1				-1							
6		1				-1			-1	-1	-1	-1		-1	-1			
7													1					
8			1		1		1	1		-1	-1	-1						
9							1											
10					-1					1								
11			-1				-1				1	1	1					
12		1	-1			-1	-1		-1	-1				-1	-1			
13					1			1		-1								
14														1				
15			1		1	1	1	1	1	-1	-1	-1	-1				-1	
16			-1		-1		-1	-1		1	1	1	1	1				
17						1			1								-1	
18						-1											1	
19		1								-1	-1	-1	-1	-1	-1	-1	-1	
20		1	-1	-1	-1	-1	-1	-1	-1									
21		1	-1		-1	-1	-1	-1	-1						-1			
22		-1	1	1	1	1	1	1	1						1			
23					1													
24						1												
25	1	-1								1	1	1	1	1	1	1	1	1

Figure D.2: H-representation of the convex polyhedron for the observed dataset $\boldsymbol{n}$, and the parameter space $T(\boldsymbol{n})$; missing values here are zeroes.

	type	b	$n_{y_0x_0z_0}$	$n_{y_1x_0z_0}$	$n_{y_0x_1z_0}$	$n_{y_1x_1z_0}$	$n_{y_0x_0z_1}$	$n_{y_0x_1z_1}$	$n_{y_1x_1z_1}$	$\tilde{n}_{y_0x_0z_0}$	$\tilde{n}_{y_1x_0z_0}$	$\tilde{n}_{y_0x_1z_0}$	$\tilde{n}_{y_1x_1z_0}$	$\tilde{n}_{y_0x_0z_1}$	$\tilde{n}_{y_0x_1z_1}$	$\tilde{n}_{y_1x_1z_1}$
1							-1			1				1		
2										1						
3			1													
4			1				1							-1		
5							1									
6			1		1		1	1		-1		-1		-1		
7			1				1	1		-1				-1		
8					1			1				-1				
9								1								
10														1		
11												1				
12									1							
13											1					
14			-1		-1				1	1	1	1				-1
15			-1	-1	-1				1	1	1	1				
16			1	1	1					-1	-1	-1				1
17			1	1	1	1				-1	-1	-1				
18			1	1	1				-1	-1		-1				1
19		1	-1	-1	-1	-1	-1	-1	-1							
20				1												
21																1
22		1	-2	-1	-2	-1	-1	-1		1		1				-1
23					1											
24						1										
25	1		-1	-1	-1	-1				1	1	1	1			
26	1		-1		-1		-1	-1		1		1		1	1	

Figure D.3: H-representation of the convex polyhedron for the observed dataset $\boldsymbol{n}$, and the sample space $\Omega(\boldsymbol{n})$; missing values here are zeroes.

VITA

Wen Wei was a graduate student in the Statistics department at the University of Washington in Seattle from 2011 to 2016. Prior to returning to graduate school, he was an analyst with Singapore Airlines. He has an M.A. in Statistics from Harvard University and a B.Sc. in Mathematics and Statistical Science from University College London. He was a former national Under-19 rugby player for Singapore, and has played for the Harvard Business School rugby team. In his spare time, he enjoys visiting Sonic Boom Records in Ballard, and listening to artists on labels such as Raster-Noton and Tri▼Angle.