A Finite Population Likelihood Ratio Test of the Sharp Null Hypothesis for Compliers

Wen Wei Loh
Department of Statistics
University of Washington
wloh@u.washington.edu

Thomas S. Richardson
Department of Statistics
University of Washington
thomasr@u.washington.edu

Abstract

In a randomized experiment with noncompliance, scientific interest is often in testing whether the treatment exposure X has an effect on the final outcome Y. We propose a finite-population significance test of the sharp null hypothesis that X has no effect on Y, within the principal stratum of Compliers, using a generalized likelihood ratio test. We present a new algorithm that solves the corresponding integer programs.

1 INTRODUCTION

Randomized experiments are often employed in order to determine whether a treatment X has a causal effect on an outcome Y. For example, individuals may be randomly assigned to either an active treatment group ($X = 1$) or to the placebo or control group ($X = 0$).

This problem may be formulated in terms of potential outcomes. Denote $Y(x = 1)$ as the outcome that the patient would have if assigned to the treatment arm, while $Y(x = 0)$ is the outcome that would arise under placebo. The absence of an effect of X on Y when the sharp causal null holds is formalized by $Y(x = 1) = Y(x = 0)$, such that every individual in the finite population has the same outcome regardless of the treatment group X to which they were assigned [19].

Randomization of treatment implies that $X \perp\!\!\!\perp \{Y(x = 1), Y(x = 0)\}$. Under the sharp causal null, this then implies $X \perp\!\!\!\perp Y$. Hence testing this latter independence may thus be seen as a test of the sharp causal null. For the case of binary outcomes Y, we may use Fisher's exact test [5], see for example [16, pg. 308].

A key feature of the potential outcome framework is that the set of individuals in the population and the values of their potential outcomes are regarded as fixed. Differences between results over hypothetical replications arise *only* due to different random assignments of this fixed set of individuals to treatment or control.

However, often we are interested in the effect of a treatment X that was not randomized. In this paper 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 1.

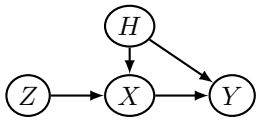

Figure 1: Graphical Representation of the Instrumental Variable Model, where H are unobserved confounding variables.

A common example of this circumstance is a randomized study with 'noncompliance'. In this context Z represents the assigned treatment, while X is the 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, in a randomized psychology experiment, whether or not participants adhere to their assigned treatment depends on their personalities and the type of manipulation (treatment). Patients in a randomized clinical trial may choose not to take their prescribed treatment, possibly due to side-effects. In studies where a randomly selected subset of subjects are offered an incentive to avail themselves of a treatment, or 'encouragement' studies, the inducement may be sufficient for some but not for others.

For each of these randomized experiments, every unit now has a treatment actually received (X) that was *not* randomized, following an *assigned* treatment (Z) that was randomized. We will make the assumption that Z has no (direct) effect on Y except through X, sometimes termed an 'exclusion restriction'.

In such studies with noncompliance with a binary treatment, Angrist et al. [1] and Imbens and Rubin [8] among others, propose to find the effect of treatment on the subset of individuals who would conform with the assigned treatment regardless of the arm to which they are assigned. Sommer and Zeger [18] 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. Balke and Pearl [2] used a symbolic linear program to derive the bounds for counterfactual probabilities, and the average causal effect of X on Y. Rubin [17] uses randomization-based posterior-predictive p-values to test a null treatment effect; Imbens and Rosenbaum [7] use randomization-based inference to obtain valid confidence intervals for the treatment effect under an additive structural model even when the instrument is 'weak'.

In this paper we address the problem of testing the sharp null hypothesis of no effect of X on Y for 'Compliers', the subpopulation or principal stratum where $X(z=0)=0$ and $X(z=1)=1$, where here $X(z)$ indicates the treatment a patient would receive if (counter to fact) assigned to $Z=z$. Under the exclusion restriction, the null hypothesis within this sub-population that X has no effect on Y is equivalent to the null hypothesis that Z has no effect of 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 $Z=1$ arm 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) sub-table, 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 or very low statistical power to reject the null hypothesis, since it is logically possible (though extremely unlikely) that there are no Compliers in a given stratum (in which $Z=X$). 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 patients who consistently do the opposite of their assignment, sometimes called 'Defiers' [3], so for all individuals:

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

It follows from this assumption that all individuals in the $(Z=1,X=0)$ stratum are Never Takers. Under random assignment of treatment (Z), the proportion of Never Takers in the $Z=1$ arm should be approximately the same as in the $Z=0$ arm. This information then reduces the range of probable values (under the randomization distribution) for the number of Compliers in the $(Z=X=0)$ stratum.

Loh and Richardson [10], following [11], use a prespecified 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, the procedure in [10] 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.

In this paper we consider an alternative approach whereby we compare the ratio of the largest probability for the observed data assuming that the sharp null hypothesis holds among Compliers, with the largest probability in the case where we allow a causal effect among Compliers. Such a generalized likelihood ratio (GLR) criterion (see for example [15]) lets us evaluate whether the alternative hypothesis is a significantly better explanation for the observed dataset than the null hypothesis, even when the number of Compliers is unknown.

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.

The remainder of the paper is organized as follows. Section 2 formalizes the potential outcome framework and sets up the motivating examples. The steps to find the maximum likelihood when the null hypothesis holds, and in general, are detailed in Section 3. Section 4 presents the generalized likelihood ratio (GLR) and describes how to find a valid frequentist p-value. The results from applying the procedure to the motivating examples are shown in Section 5. Finally, in Section 6 we briefly describe the extension to include Always Takers.

2 POTENTIAL OUTCOME FRAMEWORK

We now formalize the foregoing development. Recall the following:

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

The potential outcome X_{z_i} is the treatment X a patient *would* be exposed to if assigned $z = i$. Using these potential outcomes we may define four generic compliance 'types' t_X listed in Table 1. We denote the set of such types by $\mathbb{D}_X$.

The potential outcomes are linked to the observed outcomes by the consistency axiom [14], which requires that $Z = z$ implies $X = X_z$.

Table 1: Compliance Types (t_X) based on Potential Outcomes X_z, [8].

X_{z_0}	X_{z_1}	Compliance Type t_X	
0	0	NT	Never Taker
1	0	DE	Defier
0	1	CO	Complier
1	1	AT	Always Taker

As stated above in (1) we will assume that there are no Defiers. We will also focus on the case where there are no Always Takers, so:

$$Z = 0 \;\Rightarrow\; X = 0. \tag{2}$$

This assumption will hold in studies where individuals not assigned to treatment are unable to obtain the active treatment outside of the trial.

2.1 EXCLUSION RESTRICTION

The potential outcome for a given individual $Y_{x_j z_i}$ is the subject's response Y under exposure to treatment $x = j$, and treatment assignment $z = i$. Without further assumptions there are $16 = 2^{2^2}$ possible sets of values for the variables $(Y_{x_0 z_0}, Y_{x_1 z_0}, Y_{x_0 z_1}, Y_{x_1 z_1})$. However, we will assume that there is no (individual-level) direct effect of Z on Y relative X, so that for $j, i, i' \in \{0, 1\}$, we have:

$$Y_{x_j z_i} = Y_{x_j z_{i'}} \equiv Y_{x_j}. \tag{3}$$

Assumption (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 arm. The response type t_Y then simplifies to just four types, with $\mathbb{D}_Y$ as the set of such types, shown in Table 2.

The potential outcomes for Y are again linked to the observed outcomes via the consistency axiom, so that if $X = x$ then $Y = Y_x$.

Table 2: Response Types (t_Y) under Exclusion Restriction (3), [6].

$Y_{x_0 \cdot}$	$Y_{x_1 \cdot}$	Response Type t_Y	
0	0	NR	Never Recover
1	0	HU	Hurt
0	1	HE	Helped
1	1	AR	Always Recover

2.2 RANDOMIZATION ASSUMPTION

We make the following assumption:

$$Z \perp\!\!\!\perp \{X_{z_0}, X_{z_1}, Y_{x_0}, Y_{x_1}\} \tag{4}$$

The assumption states that the distribution of compliance and response types (t_X, t_Y) is the same in both the $z = 1$ and $z = 0$ arms; in other words, that Z is (jointly) independent of the potential outcomes. This will hold whenever treatment assignment Z is physically randomized.

2.3 THE INSTRUMENTAL VARIABLE (IV) MODEL

The model defined by (3) and (4) is known as the Instrumental Variable (IV) model (see for example [1]). A graph corresponding to the IV model given by (3) and (4) is shown in Figure 1. The exclusion restriction (3) corresponds to the absence of a $Z \to Y$ edge while the randomization assumption (4) is indicated by the absence of edges directed into Z.

2.4 AVERAGE CAUSAL EFFECT OF X ON Y

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

$$\text{ACE}(X \to Y) \equiv \text{E}[Y_{x_1} - Y_{x_0}]. \quad (5)$$

The ACE for the sub-population of Compliers is:

$$\text{ACE}_{CO}(X \to Y) \equiv \text{E}[Y_{x_1} - Y_{x_0} \mid t_X = CO]. \quad (6)$$

Since for Compliers $X_z = z$, it follows that $Y_{X=z} \equiv Y_{X_z} = Y_z$ so that

$$\text{ACE}_{CO}(X \to Y) = \text{ITT}_{CO} \equiv \text{E}[Y_{z_1} - Y_{z_0} \mid t_X = CO], \quad (7)$$

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 assumption (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 $\text{ACE}_{t_X}(X \to Y) = 0$ for $t_X \in \{NT, CO, AT\}$ jointly hold. 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.

By definition Never Takers and Always Takers always have the same observed values of $X = 0$ and $X = 1$ respectively (regardless of the Z arm they are assigned to). 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. Thus assuming (1) the only sub-population for which we may observe evidence that $\text{ACE}_{t_X}(X \to Y) \neq 0$ are the Compliers (CO).[1]

Furthermore, with the added assumption that there are no Always Takers, the 'treated' sub-population are simply the Compliers, such that the test of $\text{ACE}_{CO}(X \to Y) = 0$ is the same test for the effect of treatment on the treated, $\text{E}[Y_{x_1} - Y_{x_0} \mid X = 1] = 0$.

2.5 MOTIVATING EXAMPLES

We consider two examples of randomized experiments with noncompliance. The first dataset is from a psychology experiment where individuals were randomly assigned to one of two groups (Table 3). The treatment group was offered a small cup of pop soda ($Z = 1$), while the placebo group was offered a small cup of water ($Z = 0$). Compliance was whether the individual consumed the offered soda ($X = 1$) or not ($X = 0$). Individuals who were not offered soda in the control group ($Z = 0$) had no access to soda, as this was a closed study. There are thus two structural zeros, since $Z = 0$ implies $X = 0$. The response was a binary variable of whether the subject disposed of the cup after the session ($Y = 1$) or left the cup on the table ($Y = 0$). If we were to test the null hypothesis of $Z \perp\!\!\!\perp Y$ with Fisher's Exact Test for the corresponding 2×2 table, we would get a p-value of 0.0085. However, if we disregarded the 30 individuals in the $(Z = 1, X = 0)$ stratum and just tested $Z \perp\!\!\!\perp Y$ among the $(Z = X)$ stratum, we would get a p-value of 0.0546. Finally, Fisher's Exact Test for the null hypothesis that $X \perp\!\!\!\perp Y$ gives a p-value of 0.2157.

Table 3: Psychology Data

z	x	y	count	z	x	y	count
0	0	0	53	1	0	0	13
0	0	1	23	1	0	1	17
0	1	0	*0*	1	1	0	24
0	1	1	*0*	1	1	1	23

The second dataset is from a double-blind placebo-controlled randomized trial of Cholestyramine [4]. Subjects were randomly assigned to one of two arms: subjects in the treatment arm were prescribed Cholestyramine ($Z = 1$), and those in the other arm were given a placebo ($Z = 0$). Compliance was a continuous measure tracking the quantity of prescribed dosage consumed, over several years of treatment during the trial. The response was the average post-treatment cholesterol level, and also a continuous variable. Both continuous measures were dichotomized in [13], and the resulting counts are shown in Table 4. There are also two structural zeros in this dataset, since subjects who are not assigned treatment in the control arm ($Z = 0$) could not obtain the experimental drug Cholestyramine.

Table 4: Cholestyramine/Lipid Data

z	x	y	count	z	x	y	count
0	0	0	158	1	0	0	52
0	0	1	14	1	0	1	12
0	1	0	*0*	1	1	0	23
0	1	1	*0*	1	1	1	78

For both studies in terms of the compliance types, there are no Defiers and no Always Takers, and both (1) and (2) hold. Furthermore, since both studies were double-blind randomized control trials, it may be safely assumed that Z has no effect on Y other than through X, so that the exclusion restriction (3) holds. Thus in this case, there are four response types t_Y, but only two compliance types t_X, which gives us eight combinations for $(t_X, t_Y) \in \{NT, CO\} \times \{HE, HU, AR, NR\}$. We will consider this simpler case during our main development, though the approach extends to the more general case in which there are also Always Takers.

[1] 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.

3 MAXIMIZING THE LIKELIHOOD UNDER RANDOMIZATION

We first introduce the notation. Let $n_{y_k x_j z_i}$ be the observable count of the number of individuals in the finite population who are assigned to treatment $z = i$, with exposure $x = j$ and outcome $y = k$. We denote marginal tables similarly, for example n_{y_k} and n_{z_i}.

Let $n^{t_X}_{t_Y,z_i}$ be the number of individuals in the finite population of compliance type t_X and response type t_Y, who are assigned to treatment $z = i$. Similarly, let $n^{t_X}_{y_k z_i}$ be the number of individuals of compliance type t_X who are observed to have outcome $y = k$ in the $z = i$ arm, and $n^{t_X}_{y_k} \equiv n^{t_X}_{y_k z_0} + n^{t_X}_{y_k z_1}$ be the total number of individuals in the finite population of compliance type t_X with observed outcome $y = k$. It should be noted that the counts $n^{t_X}_{t_Y,z_i}$, $n^{t_X}_{y_k z_i}$ and $n^{t_X}_{y_k}$ are not all point-identified since they may not be directly observable from the data.

Our interest lies in testing the individual level (or 'sharp') causal null hypothesis that there is no effect of X on Y amongst Compliers:

$$H_0: \quad Y_{x_0} = Y_{x_1}. \tag{8}$$

Under the sharp null hypothesis (8), within the Complier sub-population, each individual would have the same observed outcome Y regardless of whether they took the treatment $(X = Z = 1)$ or did not do so $(X = Z = 0)$. Note that if the individual level causal null hypothesis (8) holds, then there is a zero average causal effect of X on Y for the subpopulation of Compliers (CO) and $\text{ACE}_{CO}(X \to Y) = 0$.

Thus under the null (8), the number of Compliers with observed responses $y = 0$ and $y = 1$ are just the number of Compliers of types Never Recover (NR) and Always Recover (AR) respectively:

$$\begin{aligned} n^{CO}_{y_0} &\underset{H_0}{=} n^{CO}_{NR} \equiv n^{CO}_{NR,z_0} + n^{CO}_{NR,z_1}, \\ n^{CO}_{y_1} &\underset{H_0}{=} n^{CO}_{AR} \equiv n^{CO}_{AR,z_0} + n^{CO}_{AR,z_1}. \end{aligned}$$

If the number of Compliers assigned to $z = 1$ vs. $z = 0$ were pre-specified in advance by the experimental design then, over hypothetical replications, the margins of the 2×2 subtable for Compliers containing the four counts $n^{CO}_{t_Y,z_i}$ for $t_Y \in \{NR, AR\}, i \in \{0, 1\}$ would be fixed. The resulting distribution for one of the cells, for example n^{CO}_{AR,z_1}, would be a hypergeometric distribution under the null hypothesis.

However, since we have no way to ensure a specific number of Compliers are assigned to treatment (or control) this may vary over hypothetical replications, hence none of the four counts $n^{CO}_{t_Y,z_i}$ in the subtable for Compliers will follow a hypergeometric distribution. Further these counts are not directly observable from the data.

3.1 NUISANCE PARAMETERS

Denote ψ^{NT}_k as the total number of Never Takers with observed outcome $y = k$, such that the bivariate parameter $\boldsymbol{\psi}$ is:

$$\boldsymbol{\psi} \equiv \left(\psi^{NT}_0 \equiv n^{NT}_{y_0}, \psi^{NT}_1 \equiv n^{NT}_{y_1}\right).$$

Figure 2 describes the sum relationships between the observed dataset $\{n_{y_k x_j z_i}\}$ and counts $n^{t_X}_{t_Y,z_i}$, $n^{t_X}_{y_k z_i}$ and $n^{t_X}_{y_k}$ under the null (8).

The counts n^{CO}_{NR,z_1} and n^{CO}_{AR,z_1} in the treatment arm $(z = 1)$ are directly observable from the data as $n_{y_0 x_1 z_1}$ and $n_{y_1 x_1 z_1}$ respectively. However, the presence of Never Takers in the finite population prevents us from point-identifying n^{CO}_{NR,z_0} and n^{CO}_{AR,z_0} in the placebo arm $(z = 0)$.

The unknown number of Never Takers $\boldsymbol{\psi} \equiv (\psi^{NT}_0, \psi^{NT}_1)$ may thus be regarded as 'nuisance parameters', since if we knew these quantities, we could simply determine the unobservable counts for the Never Takers in the z_0 arm:

$$\begin{aligned} n^{NT}_{y_0,z_0} &\equiv \psi^{NT}_0 - n^{NT}_{y_0,z_1} = \psi^{NT}_0 - n_{y_0 x_0 z_1}, \\ n^{NT}_{y_1,z_0} &\equiv \psi^{NT}_1 - n^{NT}_{y_1,z_1} = \psi^{NT}_1 - n_{y_1 x_0 z_1}. \end{aligned}$$

This in turn tells us what the exact values of n^{CO}_{NR,z_0} and n^{CO}_{AR,z_0} are, since n^{CO}_{NR,z_0} and $n^{NT}_{y_0,z_0}$ add up to the observable quantity $n_{y_0 x_0 z_0}$, and similarly, n^{CO}_{AR,z_0} and $n^{NT}_{y_1,z_0}$ add up to $n_{y_1 x_0 z_0}$.

Since ψ^{NT}_0 and ψ^{NT}_1 are bounded by the observable quantities in the data $\{n_{y_k x_j z_i}\}$, the space of possible values for the nuisance parameter $\boldsymbol{\psi}$ is the Cartesian product of the respective one-dimensional ranges for ψ_0 and ψ_1:

$$\begin{aligned} \psi^{NT}_0 &\in [n_{y_0 x_0 z_1}, n_{y_0 x_0 z_1} + n_{y_0 x_0 z_0}] = \Psi_0, \\ \psi^{NT}_1 &\in [n_{y_1 x_0 z_1}, n_{y_1 x_0 z_1} + n_{y_1 x_0 z_0}] = \Psi_1, \\ \boldsymbol{\Psi} &= \Psi_0 \times \Psi_1. \end{aligned} \tag{9}$$

3.2 MAXIMIZING THE HYPERGEOMETRIC PROBABILITY IN A 2×2 TABLE

Before analyzing the likelihood in our specific problem, we review the following related result: Suppose an urn contains N balls that are either red or green. $N - k$ balls are drawn from the urn, of which b balls are red. What is the most likely number of red balls x remaining in the urn, or equivalently, the most likely total number of red balls $b + x$ in the urn initially?

Table 5: 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

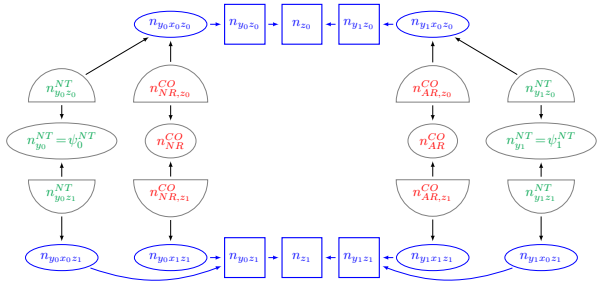

Figure 2: Sum Relationships between the Observed Dataset $\{n_{y_k x_j z_i}\}$ and Possibly Unobserved Counts $n^{t_X}_{t_Y,z_i}, n^{t_X}_{y_k z_i}, n^{t_X}_{y_k}$, Assuming H_0 (8); See Table 6.

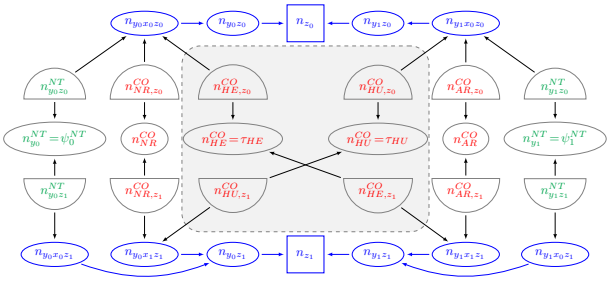

Figure 3: Sum Relationships between the Observed Dataset $\{n_{y_k x_j z_i}\}$ and Possibly Unobserved Counts $n^{t_X}_{t_Y,z_i}$, $n^{t_X}_{y_k z_i}, n^{t_X}_{y_k}$, Without Assuming H_0 (8); See Table 7.

We would thus like to maximize the hypergeometric probability corresponding to Table 5 with respect to x:

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

When $b = 0$, the most likely value of x would just be 0 as well. So if all $N-k$ balls drawn were green, then the most likely number of red balls in the urn is 0.

Theorem 1. *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,

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

The proof for Theorem 1 is given in the supplementary material.

3.3 MAXIMUM LIKELIHOOD UNDER THE NULL

For some given value of the nuisance parameters $\boldsymbol{\psi} = \boldsymbol{u}$, the counts $n^{t_X}_{t_Y,z_i}$, $n^{t_X}_{y_k z_i}$ and $n^{t_X}_{y_k}$ are all point-identified from the observed dataset $\{n_{y_k x_j z_i}\}$. We may hence describe the exact counts in a 2×4 full contingency table such as Table 6.

When we fix the value of $\boldsymbol{\psi}$ at the value $\boldsymbol{u}$, the total number of Never Takers with observed outcomes $y = 0$ and $y = 1$ (u_0^{NT} and u_1^{NT} respectively), as well as the total number of Compliers with observed responses $y = 0$ and $y = 1$ ($n_{y_0} - u_0^{NT}$ and $n_{y_1} - u_1^{NT}$ respectively) are fixed in the population, and would not change under H_0 as we vary over all possible assignments of individuals to $z = 0$ and $z = 1$.

Given the fixed column and row totals in Table 6 over repeated samplings, the randomization distribution under the null hypothesis (8) for the subjects assigned to the $z = 1$ arm is thus the multiple hypergeometric distribution [9, Chapter 39]:

$$\Pr(\{n_{y_k x_j z_i}\} \mid \boldsymbol{\psi} = \boldsymbol{u}, H_0) = \frac{\binom{u_0^{NT}}{n_{y_0 x_0 z_1}}\binom{u_1^{NT}}{n_{y_1 x_0 z_1}}\binom{n_{y_0}-u_0^{NT}}{n_{y_0 x_1 z_1}}\binom{n_{y_1}-u_1^{NT}}{n_{y_1 x_1 z_1}}}{\binom{N}{n_{z_0}}} \tag{10}$$

Note that the sharp null hypothesis for Compliers (8) holding places no restriction on the range of values for the nuisance parameter $\boldsymbol{\psi}$. We shall thus consider the value of the nuisance parameter that lends the strongest support under H_0 to the observed dataset $\{n_{y_k x_j z_i}\}$, by finding the maximum likelihood with respect to $\boldsymbol{\psi}$:

$$q^{H_0}(\{n_{y_k x_j z_i}\}) = \max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} \Pr(\{n_{y_k x_j z_i}\} \mid \boldsymbol{\psi}, H_0). \tag{11}$$

An exhaustive search over the two-dimensional discrete grid of the parameter space $\boldsymbol{\Psi}$ would require calculating $|\boldsymbol{\Psi}| = (n_{y_0 x_0 z_0} + 1) \times (n_{y_1 x_0 z_0} + 1)$ different hypergeometric probabilities. [2]

Instead, we partition Table 6 into two variation-independent 2×2 subtables: one for the Never Takers and Compliers with observed $y = 0$ outcomes (types (NT, y_0) and (CO, NR) respectively), and another for the Compliers and Never Takers with observed $y = 1$ outcomes (types (CO, AR) and (NT, y_1) respectively). The joint probability (10) then factorizes into the corresponding functions of ψ_0^{NT} and ψ_1^{NT} below:

[2] For example in the Lipid data, the search space would be of size $(158 + 1) \times (14 + 1) = 2,385$.

$$\Pr(\{n_{y_k x_j z_i}\} \mid \boldsymbol{\psi}, H_0) = \frac{\binom{n_{y_0}}{n_{y_0 x_0 z_0}}\binom{n_{y_1}}{n_{y_1 x_0 z_0}}}{\binom{N}{n_{z_0}}} \times g_0(\psi_0^{NT} | \{n_{y_k x_j z_i}\}) \times g_1(\psi_1^{NT} | \{n_{y_k x_j z_i}\}); \quad (12)$$

$$g_0(\psi_0^{NT} | \{n_{y_k x_j z_i}\}) = \frac{\binom{\psi_0^{NT}}{n_{y_0 x_0 z_1}}\binom{n_{y_0} - \psi_0^{NT}}{n_{y_0 x_1 z_1}}}{\binom{n_{y_0}}{n_{y_0 x_0 z_0}}} \quad (13)$$

$$g_1(\psi_1^{NT} | \{n_{y_k x_j z_i}\}) = \frac{\binom{\psi_1^{NT}}{n_{y_1 x_0 z_1}}\binom{n_{y_1} - \psi_1^{NT}}{n_{y_1 x_1 z_1}}}{\binom{n_{y_1}}{n_{y_1 x_0 z_0}}}. \quad (14)$$

In both subtables, the cell counts in the $z = 1$ arm are fixed, while the row totals for the $z = 0$ arm are $n_{y_0 x_0 z_0}$ and $n_{y_1 x_0 z_0}$ respectively. We may then apply Theorem 1 directly to each subtable to find the following values of ψ_0^{NT} and ψ_1^{NT} that maximise the respective hypergeometric probabilities (13) and (14).

$$\widehat{\psi}_0^{NT} = \begin{cases} \left\lfloor \frac{n_{y_0} n_{y_0 x_0 z_1}}{n_{y_0} - n_{y_0 x_0 z_0}} \right\rfloor & \text{if } \frac{(n_{y_0}+1) n_{y_0 x_0 z_1}}{n_{y_0} - n_{y_0 x_0 z_0}} \leq \left\lceil \frac{n_{y_0} n_{y_0 x_0 z_1}}{n_{y_0} - n_{y_0 x_0 z_0}} \right\rceil, \\ \left\lceil \frac{n_{y_0} n_{y_0 x_0 z_1}}{n_{y_0} - n_{y_0 x_0 z_0}} \right\rceil & \text{otherwise;} \end{cases}$$

$$\widehat{\psi}_1^{NT} = \begin{cases} \left\lfloor \frac{n_{y_1} n_{y_1 x_0 z_1}}{n_{y_1} - n_{y_1 x_0 z_0}} \right\rfloor & \text{if } \frac{(n_{y_1}+1) n_{y_1 x_0 z_1}}{n_{y_1} - n_{y_1 x_0 z_0}} \leq \left\lceil \frac{n_{y_1} n_{y_1 x_0 z_1}}{n_{y_1} - n_{y_1 x_0 z_0}} \right\rceil, \\ \left\lceil \frac{n_{y_1} n_{y_1 x_0 z_1}}{n_{y_1} - n_{y_1 x_0 z_0}} \right\rceil & \text{otherwise.} \end{cases}$$

The largest value of the probability of the observed dataset under the null (11) is then:

$$q^{H_0}(\{n_{y_k x_j z_i}\}) = \Pr(\{n_{y_k x_j z_i}\} \mid (\widehat{\psi}_0^{NT}, \widehat{\psi}_1^{NT}), H_0).$$

3.4 MAXIMUM LIKELIHOOD UNDER THE ALTERNATIVE

When the null hypothesis does not hold, there may be individuals in the Complier sub-population whose treatment exposure $X = j$ has an effect on their observed outcome $Y = k$. Compliers with observed responses $y = 0$ are no longer limited to being only of response type Never Recover (NR): they may also be of types Helped (in the $z = 0$ arm) or Hurt (in the $z = 1$ arm). Similarly, Compliers with observed responses $y = 1$ may also be one of three response types: Always Recover (AR), Helped (in the $z = 1$ arm) or Hurt (in the $z = 0$ arm).

Denote by $\tau_i^{t_Y}(\boldsymbol{\psi})$ the number of Compliers in the finite population with response type t_Y assigned to treatment $z = i$, for some fixed value of $\boldsymbol{\psi}$. For example, $\tau_0^{HE}(\boldsymbol{\psi})$ is the number of Compliers of type Helped in the $z = 0$ arm. The parameter vector $\boldsymbol{\tau}(\boldsymbol{\psi})$ is then:

$$\boldsymbol{\tau}(\boldsymbol{\psi}) \equiv \Big(\tau_0^{HE}(\boldsymbol{\psi}) \equiv n_{HE,z_0}^{CO}(\boldsymbol{\psi}), \quad \tau_0^{HU}(\boldsymbol{\psi}) \equiv n_{HU,z_0}^{CO}(\boldsymbol{\psi}), \quad \tau_1^{HE} \equiv n_{HE,z_1}^{CO}, \quad \tau_1^{HU} \equiv n_{HU,z_1}^{CO} \Big).$$

The sum relationships between the observed dataset $\{n_{y_k x_j z_i}\}$ and counts n_{t_Y,z_i}^{tx}, $n_{y_k z_i}^{tx}$ and $n_{y_k}^{tx}$ may then be described in Figure 3.

The space of possible values for the parameter $\boldsymbol{\tau}(\boldsymbol{\psi})$ depends on the fixed value of $\boldsymbol{\psi}$ and corresponds to a four-dimensional discrete grid:

$$\begin{aligned} \tau_0^{HE}(\boldsymbol{\psi}) &\in [0, n_{y_0 x_0 z_0} - (\psi_0^{NT} - n_{y_0 x_0 z_1})] \equiv \mathrm{T}_0^{HE}(\boldsymbol{\psi}), \\ \tau_0^{HU}(\boldsymbol{\psi}) &\in [0, n_{y_1 x_0 z_0} - (\psi_1^{NT} - n_{y_1 x_0 z_1})] \equiv \mathrm{T}_0^{HU}(\boldsymbol{\psi}), \\ \tau_1^{HE} &\in [0, n_{y_1 x_1 z_1}] \equiv \mathrm{T}_1^{HE}, \\ \tau_1^{HU} &\in [0, n_{y_0 x_1 z_1}] \equiv \mathrm{T}_1^{HU}, \\ \mathrm{T}(\boldsymbol{\psi}) &= \mathrm{T}_0^{HE}(\boldsymbol{\psi}) \times \mathrm{T}_0^{HU}(\boldsymbol{\psi}) \times \mathrm{T}_1^{HE} \times \mathrm{T}_1^{HU}. \end{aligned} \quad (15)$$

For some given value of the nuisance parameters $\boldsymbol{\psi} = \boldsymbol{u} \equiv (u_0^{NT}, u_1^{NT})$, and the primary parameters $\boldsymbol{\tau}(\boldsymbol{u}) = \boldsymbol{t}(\boldsymbol{u}) \equiv (t_0^{HE}(\boldsymbol{u}), t_0^{HU}(\boldsymbol{u}), t_1^{HE}, t_1^{HU})$, the counts n_{t_Y,z_i}^{tx}, $n_{y_k z_i}^{tx}$ and $n_{y_k}^{tx}$ are all point-identified from the observed dataset $\{n_{y_k x_j z_i}\}$. The exact counts may be summarized in a 2×6 full contingency table such as Table 7.

Given the fixed column and row totals in Table 7 over repeated samplings, the multiple hypergeometric probability of the subjects assigned to the $z = 1$ arm, when we no longer assume H_0 to hold, is:

$$\begin{aligned} &\Pr(\{n_{y_k x_j z_i}\} \mid (\boldsymbol{\psi}, \boldsymbol{\tau}(\boldsymbol{\psi})) = (\boldsymbol{u}, \boldsymbol{t}(\boldsymbol{u})) \\ &= \frac{\binom{u_0^{NT}}{n_{y_0 x_0 z_1}}\binom{t_0^{HE}(\boldsymbol{u}) + t_1^{HE}}{t_1^{HE}}\binom{t_0^{HU}(\boldsymbol{u}) + t_1^{HU}}{t_1^{HU}}\binom{u_1^{NT}}{n_{y_1 x_0 z_1}}}{\binom{N}{n_{z_0}}} \\ &\times \binom{n_{y_0} - t_0^{HE}(\boldsymbol{u}) - t_1^{HU} - u_0^{NT}}{n_{y_0 x_1 z_1} - t_1^{HU}} \binom{n_{y_1} - t_0^{HU}(\boldsymbol{u}) - t_1^{HE} - u_1^{NT}}{n_{y_1 x_1 z_1} - t_1^{HE}}. \end{aligned} \quad (16)$$

The maximum likelihood for the observed data $\{n_{y_k x_j z_i}\}$, allowing for Compliers who are Helped and Hurt, is then:

$$q^{ML}(\{n_{y_k x_j z_i}\}) = \max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} \max_{\boldsymbol{\tau}(\boldsymbol{\psi}) \in \mathrm{T}(\boldsymbol{\psi})} \Pr\left(\{n_{y_k x_j z_i}\} \mid (\boldsymbol{\psi}, \boldsymbol{\tau}(\boldsymbol{\psi}))\right).$$

Similar to the maximization procedure under the null, we would like to circumvent an exhaustive search of the parameter space $\{\boldsymbol{\Psi} \times \mathrm{T}(\boldsymbol{\Psi})\}$ by decomposing the joint probability (16) into separate objective functions. However, unlike the full contingency table under H_0 in Table 6, we cannot simply partition Table 7 into variation-independent subtables based only on the observed $y = 0$ and $y = 1$ outcomes. This is because if there was an effect of the treatment X on Y, then there is a Complier individual of type Helped or Hurt who would have had a different outcome Y had they been assigned to a different level of Z, and hence received a different exposure level X.

However, when we fix the number of Compliers of types Helped and Hurt in the $z = 1$ arm at some value $(\tau_1^{HE}, \tau_1^{HU}) = (t_1^{HE}, t_1^{HU})$, all six counts in the $z = 1$ arm are now point-identified and fixed. Then Table 7 may be partitioned into two variation-independent

Table 6: Full Contingency Table Under H_0 with Cell Counts that are Point-Identified Given a Value of $\boldsymbol{\psi} = \boldsymbol{u}$.

	$NT, y_0 \equiv$ $NT, (NR/HE)$	CO, NR	CO, AR	$NT, y_1 \equiv$ $NT, (AR/HU)$	Row
z_0	$u_0^{NT} - n_{y_0 x_0 z_1}$	$n_{y_0 x_0 z_0} - [u_0^{NT} - n_{y_0 x_0 z_1}]$	$n_{y_1 x_0 z_0} - [u_1^{NT} - n_{y_1 x_0 z_1}]$	$u_1^{NT} - n_{y_1 x_0 z_1}$	n_{z_0}
z_1	$n_{y_0 x_0 z_1}$	$n_{y_0 x_1 z_1}$	$n_{y_1 x_1 z_1}$	$n_{y_1 x_0 z_1}$	n_{z_1}
Column	u_0^{NT}	$n_{y_0} - u_0^{NT}$	$n_{y_1} - u_1^{NT}$	u_1^{NT}	N

Table 7: Full Contingency Table Allowing for Helped and Hurt with Cell Counts that are Point-Identified Given Values of $\boldsymbol{\psi} = \boldsymbol{u}$ and $\boldsymbol{\tau}(\boldsymbol{u}) = \boldsymbol{t}(\boldsymbol{u}) \equiv \left(t_0^{HE}(\boldsymbol{u}), t_0^{HU}(\boldsymbol{u}), t_1^{HE}, t_1^{HU}\right)$.

	$NT, y_0 \equiv$ $NT, (NR/HE)$	CO, NR	CO, HE	CO, HU	CO, AR	$NT, y_1 \equiv$ $NT, (AR/HU)$	Row
z_0	$u_0^{NT} - n_{y_0 x_0 z_1}$	$n_{y_0 x_0 z_0} - t_0^{HE}(\boldsymbol{u}) - [u_0^{NT} - n_{y_0 x_0 z_1}]$	$t_0^{HE}(\boldsymbol{u})$	$t_0^{HU}(\boldsymbol{u})$	$n_{y_1 x_0 z_0} - t_0^{HU}(\boldsymbol{u}) - [u_1^{NT} - n_{y_1 x_0 z_1}]$	$u_1^{NT} - n_{y_1 x_0 z_1}$	n_{z_0}
z_1	$n_{y_0 x_0 z_1}$	$n_{y_0 x_1 z_1} - t_1^{HU}$	t_1^{HE}	t_1^{HU}	$n_{y_1 x_1 z_1} - t_1^{HE}$	$n_{y_1 x_0 z_1}$	n_{z_1}
	u_0^{NT}	$n_{y_0} - t_0^{HE}(\boldsymbol{u}) - t_1^{HU} - u_0^{NT}$	$t_0^{HE}(\boldsymbol{u}) + t_1^{HE}$	$t_0^{HU}(\boldsymbol{u}) + t_1^{HU}$	$n_{y_1} - t_0^{HU}(\boldsymbol{u}) - t_1^{HE} - u_1^{NT}$	u_1^{NT}	N

2×3 subtables: one for individuals of types (NT, y_0), (CO, NR) and (CO, HE), and another for individuals of types (CO, HU), (CO, AR) and (NT, y_1).

Given a fixed value of (t_1^{HE}, t_1^{HU}), the joint probability (16) then decomposes into a product of functions of $(\psi_0^{NT}, \tau_0^{HE})$ and $(\psi_1^{NT}, \tau_0^{HU})$; see (18) and (19) below. For the given value of (t_1^{HE}, t_1^{HU}), the cell counts in the $z=1$ arm are fixed in each 2×3 variation-independent subtable, while the row totals for the $z = 0$ arms are $n_{y_0 x_0 z_0}$ and $n_{y_1 x_0 z_0}$ respectively.

$$\begin{aligned}
&\Pr\left(\{n_{y_k x_j z_i}\} \,\middle|\, \boldsymbol{\psi}, \tau_0^{HE}(\boldsymbol{\psi}), \tau_0^{HU}(\boldsymbol{\psi}), t_1^{HE}, t_1^{HU}\right) \\
&= \frac{\binom{n_{y_0} + t_1^{HE} - t_1^{HU}}{n_{y_0 x_0 z_0}} \binom{n_{y_1} + t_1^{HU} - t_1^{HE}}{n_{y_1 x_0 z_0}}}{\binom{N}{n_{z_0}}} \\
&\quad \times h_0\left(\psi_0^{NT}, \tau_0^{HE}(\psi_0^{NT}) \,\middle|\, t_1^{HE}, t_1^{HU}, \{n_{y_k x_j z_i}\}\right) \\
&\quad \times h_1\left(\psi_1^{NT}, \tau_0^{HU}(\psi_1^{NT}) \,\middle|\, t_1^{HE}, t_1^{HU}, \{n_{y_k x_j z_i}\}\right);
\end{aligned} \tag{17}$$

$$\begin{aligned}
&h_0\left(\psi_0^{NT}, \tau_0^{HE}(\psi_0^{NT}) \,\middle|\, t_1^{HE}, t_1^{HU}, \{n_{y_k x_j z_i}\}\right) \\
&\equiv \frac{\binom{\psi_0^{NT}}{n_{y_0 x_0 z_1}} \binom{\tau_0^{HE}(\psi_0^{NT}) + t_1^{HE}}{t_1^{HE}} \binom{n_{y_0} - \tau_0^{HE}(\psi_0^{NT}) - t_1^{HU} - \psi_0^{NT}}{n_{y_0 x_1 z_1} - t_1^{HU}}}{\binom{n_{y_0} + t_1^{HE} - t_1^{HU}}{n_{y_0 x_0 z_0}}},
\end{aligned} \tag{18}$$

$$\begin{aligned}
&h_1\left(\psi_1^{NT}, \tau_0^{HU}(\psi_1^{NT}) \,\middle|\, t_1^{HE}, t_1^{HU}, \{n_{y_k x_j z_i}\}\right) \\
&\equiv \frac{\binom{\psi_1^{NT}}{n_{y_1 x_0 z_1}} \binom{\tau_0^{HU}(\psi_1^{NT}) + t_1^{HU}}{t_1^{HU}} \binom{n_{y_1} - \tau_0^{HU}(\psi_1^{NT}) - t_1^{HE} - \psi_1^{NT}}{n_{y_1 x_1 z_1} - t_1^{HE}}}{\binom{n_{y_1} + t_1^{HU} - t_1^{HE}}{n_{y_1 x_0 z_0}}}.
\end{aligned} \tag{19}$$

Since the 2×3 subtables are now variation-independent, we may find the values of:

$$\begin{aligned}
&\left(\widehat{\psi}_0^{NT}(t_1^{HE}, t_1^{HU}), \widehat{\tau}_0^{HE}(\widehat{\psi}_0^{NT}; t_1^{HE}, t_1^{HU})\right), \\
&\left(\widehat{\psi}_1^{NT}(t_1^{HE}, t_1^{HU}), \widehat{\tau}_0^{HU}(\widehat{\psi}_1^{NT}; t_1^{HE}, t_1^{HU})\right)
\end{aligned}$$

that maximise the respective conditional hypergeometric probabilities (18) and (19).

For each fixed value of (t_1^{HE}, t_1^{HU}), a naïve search over the discrete parameter space in each induced 2×3 subtable would involve maximizing over $\binom{n_{y_0 x_0 z_0} + 2}{2}$ and $\binom{n_{y_1 x_0 z_0} + 2}{2}$ hypergeometric probabilities respectively. [3]

Instead, we apply the result from [12], which provides an algorithm to find the most likely values of the cells in the z_0 arms of both subtables, *without calculating any hypergeometric probabilities*. Finally, we need only maximize over the parameter spaces for τ_1^{HE} and τ_1^{HU}, where there are $|\mathrm{T}_1^{HE}| \times |\mathrm{T}_1^{HU}| = (n_{y_1 x_1 z_1} + 1) \times (n_{y_0 x_1 z_1} + 1)$ possible combinations for the point-identified counts in the $Z = 1$ arm. [4]

Allowing Compliers who are Helped and Hurt, the maximum likelihood for the observed dataset $\{n_{y_k x_j z_i}\}$ is then:

$$\begin{aligned}
q^{ML}(\{n_{y_k x_j z_i}\}) \;=\; &\max_{\substack{(\tau_1^{HE}, \tau_1^{HU}) \\ \in \mathrm{T}_1^{HE} \times \mathrm{T}_1^{HU}}} \Pr\left(\{n_{y_k x_j z_i}\} \,\middle|\, \widehat{\boldsymbol{\psi}}, \tau_0^{HE}(\widehat{\boldsymbol{\psi}}), \right. \\
&\left. \tau_0^{HU}(\widehat{\boldsymbol{\psi}}), \tau_1^{HE}, \tau_1^{HU}\right).
\end{aligned} \tag{20}$$

[3] For example in the Lipid data, the search spaces would be of sizes $\binom{158+2}{2} = 12{,}720$ and $\binom{14+2}{2} = 120$ respectively.

[4] In the Lipid data example, we would need to calculate only $(78 + 1) \times (23 + 1) = 1896$ hypergeometric probabilities.

4 GLR AND P-VALUE

A generalized likelihood ratio (GLR) lets us assess the evidence in the observed data both for *and against* the null hypothesis (8) respectively, by comparing the best possible fit of the observed data when H_0 holds, against the best fit without the constraint of H_0. The generalized likelihood ratio for the observed dataset $\{n_{y_k x_j z_i}\}$ is defined as:

$$G(\{n_{y_k x_j z_i}\}) = \frac{q^{H_0}(\{n_{y_k x_j z_i}\})}{q^{ML}(\{n_{y_k x_j z_i}\})}. \tag{21}$$

However, the distribution of the test statistic (21) under H_0 depends on the chosen values of the column totals in the full contingency table (Table 6), which in turn correspond to some value of the nuisance parameter ψ. Under H_0 the value of $\psi = u$ is sufficient to determine the distribution of (21) since the margin totals in the full contingency table (Table 6) are now fixed.

We may then enumerate all possible assignments, and consolidate each assignment to obtain the associated *possibly observable dataset* $\{\tilde{n}_{y_k x_j z_i}\}$ based on the sum relationships depicted in Figure 2:

$$\begin{aligned}
&\tilde{n}_{y_0 x_0 z_0} = \tilde{n}^{NT}_{y_0,z_0} + \tilde{n}^{CO}_{NR,z_0}; \tilde{n}_{y_1 x_0 z_0} = \tilde{n}^{NT}_{y_1,z_0} + \tilde{n}^{CO}_{AR,z_0};\\
&\tilde{n}_{y_0 x_0 z_1} = u^{NT}_0 - \tilde{n}^{NT}_{y_0,z_0}; \tilde{n}_{y_0 x_1 z_1} = \left(n_{y_0} - u^{NT}_0\right) - \tilde{n}^{CO}_{NR,z_0};\\
&\tilde{n}_{y_1 x_0 z_1} = u^{NT}_1 - \tilde{n}^{NT}_{y_1,z_0}; \tilde{n}_{y_1 x_1 z_1} = \left(n_{y_1} - u^{NT}_1\right) - \tilde{n}^{CO}_{AR,z_0}.
\end{aligned}$$

For each of these possibly observable datasets $\{\tilde{n}_{y_k x_j z_i}\}$, we then find the corresponding generalized likelihood ratio $G(\{\tilde{n}_{y_k x_j z_i}\}) = q^{H_0}(\{\tilde{n}_{y_k x_j z_i}\})/q^{ML}(\{\tilde{n}_{y_k x_j z_i}\})$. Note that the parameter space, which we denote as $(\tilde{\mathbf{\Psi}}, \tilde{\mathrm{T}}(\tilde{\mathbf{\Psi}}))$, is specific to each dataset $\{\tilde{n}_{y_k x_j z_i}\}$, and differs from the parameter space for the actually observed data $(\mathbf{\Psi}, \mathrm{T}(\mathbf{\Psi}))$.

Given a fixed value of the nuisance parameter ψ, the corresponding ψ-specific p-value is then the total probability under H_0 of all datasets $\{\tilde{n}_{y_k x_j z_i}\}$ with generalized likelihood ratios $G(\{\tilde{n}_{y_k x_j z_i}\})$ that are at least as extreme as that for the observed data $\{n_{y_k x_j z_i}\}$:

$$p^{H_0}(\{n_{y_k x_j z_i}\}; \psi) = \sum_{\substack{\{\tilde{n}_{y_k x_j z_i}\}:\\ G(\{\tilde{n}_{y_k x_j z_i}\}) \leq G(\{n_{y_k x_j z_i}\})}} \Pr(\{\tilde{n}_{y_k x_j z_i}\} \mid \psi, H_0).$$

Since each fixed value of ψ corresponds to a different number of Compliers and hence a different instance of the null hypothesis, we may report the maximum among all the ψ-specific p-values as the p-value from our significance test:

$$p^{H_0}(\{n_{y_k x_j z_i}\}) \equiv \max_{\psi \in \mathbf{\Psi}} p^{H_0}(\{n_{y_k x_j z_i}\}; \psi). \tag{22}$$

The probability of obtaining a value of the test statistic as extreme as the one observed (21) will never be larger than $p^{H_0}(\{n_{y_k x_j z_i}\})$, irrespective of the number of Compliers in the population, and is thus a valid frequentist p-value.

5 APPLICATIONS

5.1 PSYCHOLOGY DATA EXAMPLE

For the observed dataset $\{n_{y_k x_j z_i}\}$ in the motivating example from Table 3, the largest probability under H_0 is $q^{H_0}(\{n_{y_k x_j z_i}\}) = 1\times10^{-4}$; the maximum likelihood without the constraint of no Compliers who were Helped or Hurt in the population is $q^{ML}(\{n_{y_k x_j z_i}\}) = 2.3\times10^{-3}$. The generalized likelihood ratio (21) for this dataset is: $G(\{n_{y_k x_j z_i}\}) = \frac{q^{H_0}(\{n_{y_k x_j z_i}\})}{q^{ML}(\{n_{y_k x_j z_i}\})} = 0.052$.

There were $1296 = (23+1)\times(53+1)$ possible values of the nuisance parameter ψ, and the maximum among all the ψ-specific p-values is:

$$p^{H_0}(\{n_{y_k x_j z_i}\}) \equiv \max_{\psi \in \mathbf{\Psi}} \left\{p^{H_0}(\{n_{y_k x_j z_i}\}; \psi)\right\} = 0.0137.$$

5.2 LIPID DATA EXAMPLE

For the observed dataset $\{n_{y_k x_j z_i}\}$ in the motivating example from Table 4, the largest probability under H_0 is $q^{H_0}(\{n_{y_k x_j z_i}\}) = 6\times10^{-23}$; the maximum likelihood without the constraint of no Compliers who were Helped or Hurt in the population is $q^{ML}(\{n_{y_k x_j z_i}\}) = 0.0019$. The generalized likelihood ratio (21) for this dataset is then: $G(\{n_{y_k x_j z_i}\}) = \frac{q^{H_0}(\{n_{y_k x_j z_i}\})}{q^{ML}(\{n_{y_k x_j z_i}\})} = 3\times10^{-20}$.

There were $2385 = (14+1)\times(158+1)$ possible values of the nuisance parameter ψ, and the maximum among all the ψ-specific p-values is:

$$p^{H_0}(\{n_{y_k x_j z_i}\}) \equiv \max_{\psi \in \mathbf{\Psi}} \left\{p^{H_0}(\{n_{y_k x_j z_i}\}; \psi)\right\} = 2\times10^{-21}.$$

In comparison, using a pre-specified value of $\gamma = 0.01$ gives a p-value of $0.01 + (1\times10^{-21}) \approx 0.01$ [10].

6 CONCLUSIONS

We have proposed a finite population significance test of the sharp null hypothesis for Compliers using the generalized likelihood ratio. The resulting p-value may be arbitrarily close to zero and summarizes the strength of evidence against the sharp null hypothesis for Compliers (8).

While our development has assumed that there are no Always Takers, the approach extends to the more general case in which there are also Always Takers. However, this would increase the dimension of the nuisance parameter and hence the size of the nuisance parameter space, such that finding the generalized likelihood ratio test statistic would be computationally more intensive. For example, even when the sharp null hypothesis for Compliers (8) holds, the number of Compliers in the $z = 1$ arm $\left(n^{CO}_{NR,z_1}\right.$ and $\left.n^{CO}_{AR,z_1}\right)$ would no longer be point-identified from the observed counts $n_{y_0 x_1 z_1}$ and $n_{y_1 x_1 z_1}$ respectively.

References

[1] 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.

[2] 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.

[3] 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.

[4] B Efron and D Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

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

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

[7] 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.

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

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

[10] 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.

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

[12] 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.

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

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

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

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

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

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

[19] J Spława-Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.*, 5(4):465–472, 1990. Translated from the Polish and edited by D M Dąbrowska and T P Speed.