General Finite Sample Inference for Experiments with Examples from Health Care*

Amanda E. Kowalski

December 17, 2019

Abstract

I exploit knowledge of the randomization process within an experiment to conduct finite sample inference on quantities that capture heterogeneous intervention effects. The only data that I use are the cross-tabulations of a discrete randomized intervention and a discrete outcome. The inference procedure is general in the sense that it can test hypotheses and construct confidence intervals on various quantities. My main contribution is that I can conduct informative inference on quantities for which previous methods are uninformative, such as the number of participants who respond to the intervention in the opposite direction of the average, sometimes known as the number of defiers. I can also use the same procedure to conduct inference on other quantities, such as the average intervention effect and the fraction affected in either direction, for which previous methods are informative but restricted in the quantities they can consider. I demonstrate the value of general finite sample inference using data from hypothetical drug trials. In one trial, the estimated average intervention effect shows that the lives of 40 out of 100 participants would be saved on average. I reject the null hypothesis that no participants would be killed at the 3% level and infer with 95% confidence that at least 3 participants would be killed.

*I thank Neil Christy, Tory Do, Simon Essig Aberg, Bailey Flanigan, Pauline Mourot, Srajal Nayak, and Matthew Tauzer for excellent research assistance. Don Andrews, Steve Berry, Xiaohong Chen, Peng Ding, Florian Gunsilius, Andreas Hagemann, Guido Imbens, Ang Li, John Pepper, Edward Vytlacil, Christopher Walters, David Wilson, and seminar participants at the Advances with Fields Experiments Conference at the University of Chicago, the Essen Health Conference, Notre Dame, UCLA, UVA, the University of Zurich, and the Yale Cowles Summer Structural Microeconomics Conference provided helpful comments. I also thank Charles Antonelli, Bennett Fauber, and Advanced Research Computing at the University of Michigan, as well as Andrew Sherman and the Yale University Faculty of Arts and Sciences High Performance Computing Center. This paper includes content from earlier working papers circulated as Kowalski (2019a) and Kowalski (2019b).
1 Introduction

A clinical trial shows that an intervention reduces mortality on average. We conclude that it must save the lives of some patients. Can we say with any confidence how many patients in the trial would be killed if they received the intervention? Can we do so with the same finite sample inference procedure that we use to test hypotheses about the average effect?

I propose a general finite sample inference procedure that can be used to test hypotheses about quantities that capture heterogeneous intervention effects. The inference procedure is general in the sense that it can test hypotheses and construct confidence intervals on various quantities. These quantities include but are not limited to the average intervention effect, the number affected by the intervention in the opposite direction of the average intervention effect, and the fraction affected by the intervention in either direction.

For data, the inference procedure relies only on the cross-tabulations of a discrete randomized intervention and a discrete outcome. To fix ideas, consider a randomized experiment with a binary intervention and a binary outcome. If the outcome is mortality, the data consist of the number of participants alive and dead in the intervention and control arms. Using these data, we can estimate the average intervention effect as the fraction of the participants dead in the intervention arm minus the fraction of participants dead in the control arm. The same estimated average intervention effect can be consistent with many different configurations of the data, which consist of four numbers. In an experiment with 100 participants, there are 176,851 possible configurations of the data but only 201 possible values of the estimated average intervention effect. Even though the data that I use are very limited, they contain information beyond the estimated average intervention effect.

For assumptions, the inference procedure relies primarily on the structure of the data generating process. In randomized experiments, the structure of the data generating process is determined by the randomization process, which is a feature of the experimental design. There are many different possible randomization processes. For example, the experimenter could flip a coin for each participant such that the probability that each participant is assigned to the intervention arm is independent and identically distributed. Alternatively, the experimenter could fix the total number of participants to be assigned to each arm and conduct randomization by drawing colored balls out of an urn. Because the experimenter can control the randomization process, the experimenter can also build a compelling rationale for assumptions about the data generating process that follow from it. As long as the randomization process is known, it can be exploited for inference. Even if the randomization process is not known, it might be reasonable to exploit an assumed randomization process for inference, giving deeper meaning to the claim that a natural experiment is “as good as random.”

To build intuition for how the randomization process can inform inference on heterogeneous intervention effects, I draw on the concept of potential outcomes (Neyman, 1923; Rubin, 1974, 1977; Holland, 1986). Suppose that each participant in the experiment has a potential outcome in the intervention arm and another potential outcome in the control arm. In doing so, we assume that the potential outcomes of each participant do not depend on the potential outcomes of any other
participants (this assumption has been referred to as the “stable unit treatment value assumption” (SUTVA) by Rubin (1980), who references an earlier assumption by Cox (1958)). Continuing to fix ideas such that the outcome is mortality, there are four possible types of participants on the basis of the four possible combinations of potential outcomes in intervention and control: those who would live regardless, those who would be saved, those who would be killed, and those who would die regardless.

The four potential outcome types reflect heterogeneous intervention effects. By definition, the intervention has no effect on those who would live or die regardless; it decreases mortality for those who would be saved; and it increases mortality for those who would be killed. Therefore, if we knew the number of participants of each type, then we could construct quantities that capture heterogeneous intervention effects such as the average intervention effect, the number affected by the intervention in the opposite direction of the average effect, and the fraction affected in either direction. It is difficult to identify any given participant’s potential outcome type because it is not possible to observe the potential outcome in the intervention arm for a participant assigned to the control arm and vice versa.

I use the randomization process to inform inference on the number of participants of each potential outcome type. I do so by considering all possible ways to divide the participants into the four potential outcome types. In an experiment with 100 participants, there are 176,851 possible ways to divide the participants into the four potential outcome types, just as there are 176,851 possible ways to divide the participants into the four numbers that comprise the data. Therefore, there are 176,851^2—over 31 billion—possible combinations of the data and potential outcome type configurations. However, some potential outcome type configurations cannot be consistent with some data configurations. Furthermore, among the potential outcome type configurations that can be consistent with a given data configuration, I know from the structure of the randomization process that some are more likely than others. I calculate the exact likelihood of all possible combinations of the data and potential outcome type configurations using a supercomputer. I use the calculated values and the observed data configuration to conduct inference on a wide range of quantities that capture heterogeneous intervention effects.

I focus on hypothesis testing and the construction of confidence intervals rather than estimation because there can be more than one potential outcome type configuration that maximizes the likelihood of the observed data configuration. In an experiment with 100 participants, the maximum likelihood estimate of the potential outcome type configuration is multi-valued for 14,940 data configurations, about 8.4% of the possible 176,851. However, the likelihood function still contains information that can be used for inference.

My main contribution is that the general finite sample inference procedure that I propose can conduct informative inference on quantities for which previous methods are uninformative, especially the number of participants who respond to the intervention in the opposite direction of the average. Previous methods are uninformative about number of participants who respond to the intervention in the opposite direction of the average unless they bring in additional data. For example, machine learning brings in additional data on covariates (see Wager and Athey (2018)),
and analysis of “side effects” in medicine brings in additional data on secondary outcomes.

Analysis of heterogeneity in the treatment effect, as opposed to heterogeneity in the intervention effect, brings in data on takeup of a treatment in addition to data on an intervention and outcome. In contrast, I only require data on an intervention and an outcome, which is important because many trials do not collect data on treatment takeup. Furthermore, even for trials that do collect data on treatment takeup, the treatment takeup data are often available before the outcome data. With just the data on the intervention and the treatment, I can examine heterogeneity in the first stage impact of the intervention on the treatment takeup in exactly the same way that I examine heterogeneity in the reduced form impact of the intervention on the outcome; I simply interpret the treatment as an outcome. In doing so, I can conduct informative inference on the number of individuals who respond to the intervention in the opposite direction of the average, also known as the number of “defiers” (Balke and Pearl, 1993; Angrist et al., 1996).

Because I can conduct inference on the number of defiers, I contribute to the literature by offering a test of the LATE monotonicity assumption of Imbens and Angrist (1994), which excludes defiers. The test is effectively a joint test of the LATE monotonicity assumption, SUTVA, and an assumption that randomization was conducted as specified. Previous tests of the LATE monotonicity assumption require data on an intervention, a treatment, and an outcome other than the treatment (Balke and Pearl, 1997; Imbens and Rubin, 1997; Heckman and Vytlacil, 2005; Richardson and Robins, 2010; Huber and Mellace, 2015; Kitagawa, 2015; Mourifié and Wan, 2017; Machado et al., 2019). Furthermore, they are effectively joint tests of the LATE monotonicity assumption, SUTVA, an assumption that randomization was conducted, and the exclusion restriction that the intervention only affects the outcome through the treatment (Angrist et al. (1996), also referred to as the “set-level restriction by Manski (1990)). To my knowledge, previous tests do not provide confidence intervals on the numbers of defiers. The confidence intervals that I provide can be informative even when the test does not reject the null hypothesis of zero defiers because they inform the precision of the assumption claim that there are zero defiers.

Literature on heterogeneity in the treatment effect makes informative inference on the number of defiers and the analogous number of participants who respond to the intervention in the opposite direction of the average seem impossible, especially with more limited data. In the absence of the LATE monotonicity assumption, Balke and Pearl (1997) and Heckman and Vytlacil (2001) show that it is only possible to bound the average treatment effect, and they prove that the bounds that they obtain are the tightest possible. Related Frechet-Hoeffding bounds on the number of participants who respond to the intervention in the opposite direction of the average (Gadbury et al., 2004; Nelsen, 2006) are uninformative in the sense that they always include zero, which makes it seem impossible to say with any confidence that any participants are affected in the opposite direction of the average. Furthermore, Kitagawa (2015) proves that his test is the “strongest possible test for instrument validity,” which makes progress on future tests of the LATE monotonicity assumption seem difficult. I make progress by exploiting the known structure of the data generating process.

There is a literature that exploits the known structure of the data generating process for infer-
ence, but to my knowledge, it does not and cannot conduct informative inference on the number of participants who respond to the intervention in the opposite direction of the average, which it considers a nuisance parameter. Fisher (1935) develops a test of the null hypothesis that the fraction affected by the intervention in either direction in either direction is zero. The Fisher (1935) hypothesis is a “sharp” null hypothesis because there is only one potential outcome type configuration that is consistent with the data configuration under the null. In contrast, the Neyman (1923) null hypothesis that the average intervention effect is equal to zero is a “weak” null hypothesis because there are many potential outcome type configurations that are consistent with the data configuration under the null. These configurations can differ in the number of participants who respond to the intervention in the opposite direction of the average, which is thus a nuisance parameter for tests of the null hypothesis that the average intervention effect is zero. The null hypothesis that the number of participants who respond to the intervention in the opposite direction of the average equals zero is also a weak null hypothesis.

My main innovation relative to the literature that exploits the structure of the randomization process to test weak null hypotheses is that I use a test statistic that is informative of the number who respond to the intervention in the opposite direction of the average. Chung et al. (2013) and Wu and Ding (2019) propose asymptotic tests that exploit the structure of the randomization process to test hypotheses on various quantities. However, their tests are limited in the quantities that they can consider because they use the sample analog of the quantity of interest in the test statistic. There is no sample analog for the number who respond to the intervention in the opposite direction of the average, so their methods cannot yield meaningful inference on this quantity. Related tests use the estimated average intervention effect as the test statistic for inference on specific weak null hypotheses either asymptotically (Copas, 1973; Ding and Miratrix, 2019), or by choosing an exact p-value as the maximum p-value from tests of sharp null hypotheses (Chiba, 2015; Rigdon and Hudgens, 2015; Ding et al., 2016; Li and Ding, 2016). I follow the latter approach, but I use a different test statistic, the likelihood ratio: the ratio of the maximum likelihood of the potential outcome type configurations under the null hypothesis to the maximum likelihood of all possible potential outcome type configurations. Data configurations with the same average intervention effect can generate meaningfully different distributions of the likelihood ratio. Therefore, relative to the average intervention effect, the likelihood ratio is more informative of quantities without a sample analog.

Because it uses the likelihood ratio as the test statistic, the general finite sample inference procedure that I propose can conduct inference on any quantity of interest that can be represented as a function of the four counts of the potential outcome types. For example, I demonstrate that I can conduct inference on another quantity for which previous methods are uninformative, the ratio of the number of participants who would be killed to the number of participants who would be saved. The government might want to consider using a threshold in this ratio for approval of new interventions if it is willing to trade off some lives for others, as in the famous Trolley problem (Foot, 1967; Thomson, 1985). I can also use the same general finite sample inference procedure to test hypotheses and construct confidence intervals for which previous methods are
informative but approximate or restricted in the quantities that they can consider, including the average intervention effect and the fraction affected in either direction. For example, a regression can be used for inference on the average intervention effect, but such inference is approximate because it relies on assumptions that hold asymptotically.

Inference using the general finite sample procedure that I propose can change the interpretation of results from experiments. I present results from two hypothetical drug trials with 100 participants that yield the same estimated average intervention effect, which shows that 40 participants would be saved on average. Both estimates are statistically significant at conventional levels. However, each trial has a different data configuration. In one trial, I reject the the null hypothesis that no one would be killed by the intervention at the 3% level, and I infer with 95% confidence that at least 3 participants would be killed. In the other trial, I cannot reject the same null hypothesis at any level.

The paper proceeds as follows. Section 2 introduces the model and derives likelihood functions for two common randomization processes. Section 3 describes the hypothesis testing procedure and the construction of confidence intervals. Section 4 considers hypothetical data from clinical trials for new drugs and demonstrates that I can conduct informative inference on the number of participants who would be killed. Section 5 concludes.

2 Model
2.1 Derivation of Likelihood Functions
Consider a randomized experiment with a binary intervention \( Z \) and a binary outcome \( Y \). To fix ideas, suppose that \( Y \) represents mortality. Each of \( s \) study participants are assigned to either intervention (\( Z = 1 \)) or control (\( Z = 0 \)). At the end of the study period, each participant is observed either dead (\( Y = 1 \)) or alive (\( Y = 0 \)).

There are four possible potential outcome types. The matrix in Figure 1 includes a separate row for each potential outcome type \( i \). The first row includes the \( \theta(1) \) participants who would live regardless of the intervention (\( Y = 0 \) if \( Z = 1 \) and \( Y = 0 \) if \( Z = 0 \)); the second row includes the \( \theta(2) \) participants who would be saved by the intervention (\( Y = 0 \) if \( Z = 1 \) and \( Y = 1 \) if \( Z = 0 \)); the third row includes the \( \theta(3) \) participants who would be killed by the intervention (\( Y = 1 \) if \( Z = 1 \) and \( Y = 0 \) if \( Z = 0 \)); and the fourth row includes the \( \theta(4) = s - \theta(1) - \theta(2) - \theta(3) \) participants who would die regardless of the intervention (\( Y = 1 \) if \( Z = 1 \) and \( Y = 1 \) if \( Z = 0 \)).

The data configuration consists of the total number of participants observed in each of four possible observed outcome groups at the end of the experiment. The matrix in Figure 1 includes a separate column for each observed outcome group \( j \). The first column includes the \( G(1) \) participants observed dead in the intervention arm (\( Z = 1 \) and \( Y = 1 \)); the second column includes the \( G(2) \) participants observed alive in the intervention arm (\( Z = 1 \) and \( Y = 0 \)); the third column includes the \( G(3) \) participants observed dead in the control arm (\( Z = 0 \) and \( Y = 1 \)); and the fourth column includes the \( G(4) = s - G(1) - G(2) - G(3) \) participants observed alive in the control arm (\( Z = 0 \) and \( Y = 0 \)). Note that \( G(j) \) is a random variable, which is why I denote it with a capital letter. The data consist of \( G(1), G(2), G(3), \) and \( G(4) \), which I represent with \( G \), in bold to indicate that
Figure 1: Matrix that Relates Potential Outcome Types and Observed Outcome Groups When Outcome is Mortality

| Potential Outcome Types | Observed Outcome Groups |
|-------------------------|-------------------------|
|                        | Intervention (Z=1)      | Control (Z=0) |
|                        | Dead (Y=1)              | Alive (Y=0)  | Dead (Y=1) | Alive (Y=0) |
| Would Live Regardless  | θ(1)                    | N(1,2)       | N(1,4)     |
| Would Be Saved         | θ(2)                    | N(2,2)       | N(2,3)     |
| Would Be Killed        | θ(3)                    | N(3,1)       | N(3,4)     |
| Would Die Regardless   | s-θ(1) -θ(2)-θ(3)       | N(4,1)       | N(4,3)     |

Notes: Y represents mortality, and Z represents assignment to the intervention arm. N(i,j), the number of participants of potential outcome type i in observed outcome group j, must be equal to zero in all shaded cells. To translate my notation into the potential outcomes notation of Rubin (1974), suppose that each participant has potential outcomes Y(Z), such that Y(1) is the potential outcome in the intervention arm and Y(0) is the potential outcome in the control arm. Then θ(1) = \( \sum \mathbf{1}\{Y(1) = Y(0) = 1\} \), θ(2) = \( \sum \mathbf{1}\{Y(1) = 0, Y(0) = 1\} \), θ(3) = \( \sum \mathbf{1}\{Y(1) = 1, Y(0) = 0\} \), and θ(4) = \( \sum \mathbf{1}\{Y(1) = 1, Y(0) = 0\} \).

\[^\dagger\] Respond to intervention in opposite direction of average if θ(2) < θ(3).

\[^o\] Respond to intervention in opposite direction of average if θ(3) < θ(2).

Note. Y represents mortality, and Z represents assignment to the intervention arm. N(i,j), the number of participants of potential outcome type i in observed outcome group j, must be equal to zero in all shaded cells. To translate my notation into the potential outcomes notation of Rubin (1974), suppose that each participant has potential outcomes Y(Z), such that Y(1) is the potential outcome in the intervention arm and Y(0) is the potential outcome in the control arm. Then θ(1) = \( \sum \mathbf{1}\{Y(1) = Y(0) = 1\} \), θ(2) = \( \sum \mathbf{1}\{Y(1) = 0, Y(0) = 1\} \), θ(3) = \( \sum \mathbf{1}\{Y(1) = 1, Y(0) = 0\} \), and θ(4) = \( \sum \mathbf{1}\{Y(1) = 1, Y(0) = 0\} \).
assignment to the intervention arm will not be observed dead in the control arm \( j = 3 \). The logic for other cells proceeds similarly. I shade all 8 cells that cannot have any participants in them.

Some potential outcome configurations \( \theta \) cannot be consistent with a data configuration \( g \), and those that are consistent may be more or less likely under a given randomization process. By deductive reasoning based on the shaded cells, I can express the distribution of the data as follows:

\[
P(G = g | \theta, s) = P(G(1) = g(1), G(2) = g(2), G(3) = g(3) | \theta, s) \tag{1}
\]

\[
= P(N(3,1) + N(4,1) = g(1),
N(1,2) + N(2,2) = g(2),
N(2,3) + N(4,3) = g(3) | \theta, s) \tag{2}
\]

\[
= P(N(3,1) + N(4,1) = g(1),
N(1,2) + N(2,2) = g(2),
\theta(2) - N(2,2) + s - \theta(1) - \theta(2) - \theta(3) - N(4,1) = g(3) | \theta, s) \tag{3}
\]

I transition from (1) to (2) by expressing each \( G(j) \) as the sum of the nonzero participant counts in column \( j \). The resulting expression is in terms of the participant counts in all 8 cells that are not shaded in Figure 1. I simplify the expression further by recognizing that within any potential outcome type \( i \), the number of participants in the control arm is equal to \( \theta(i) \) minus the number of participants in the intervention arm. Therefore, I can express (3) in terms of only four random variables: \( N(1,2), N(2,2), N(3,1), \) and \( N(4,1) \). These random variables represent the realized number of participants assigned to the intervention arm in each of the four potential outcome types.

Using the approach to deconvolution for discrete random variables, I express (3) as the sum of the probability of each possible realization \( \ell \) of \( N(1,2) \). I then rearrange terms to obtain an expression in terms of the joint probability of \( N(1,2), N(2,2), N(3,1), \) and \( N(4,1) \):

\[
P(G = g | \theta, s) = \sum_{\ell=0}^{\theta(1)} P(N(1,2) = \ell,
N(3,1) + N(4,1) = g(1),
N(1,2) + N(2,2) = g(2),
\]

8
\[
N(2, 2) + N(4, 1) = s - \theta(1) - \theta(3) - g(3) \mid \theta, s
\]

\[
= \sum_{\ell=0}^{\theta(1)} P(N(1, 2) = \ell,
N(2, 2) = g(2) - \ell,
N(3, 1) = \theta(1) + \theta(3) + g(1) + g(2) + g(3) - s - \ell,
N(4, 1) = s + \ell - \theta(1) - \theta(3) - g(2) - g(3) \mid \theta, s).
\] (4)

If the number of participants who would die regardless and are randomized into intervention \(N(1, 2)\) were known, then the remaining parameters would be point identified. Therefore, there is one free variable, normalized to be \(N(1, 2)\), and the likelihood includes a summation over this free variable. Since \(N(1, 2)\) is not known, the observed outcome groups do not point identify the potential outcome types. Under a monotone intervention response assumption in the tradition of the Manski (1997) “monotone treatment response” assumption, the remaining potential outcome types are point identified (Imbens and Rubin, 1997). However, without such an assumption, the observed outcome groups still contain information about all four components of the potential outcome type configuration. The likelihood function captures some of this information because the summation is not flat over the free variable \(N(1, 2)\).

To derive a closed-form expression for the likelihood function, I exploit knowledge of the known randomization process. Suppose the researchers conduct randomization such that \(Z\) is independently and identically distributed. At the start of the experiment, the experimenter has chosen the intended fraction of participants in the intervention arm \(p\). The experimenter flips a weighted coin to determine whether the participant will be randomly assigned to the intervention or control arm. The indicator for whether a participant is assigned to the intervention arm is a random variable that is distributed according to a Bernoulli distribution with probability of success \(p\). The total number of participants of potential outcome type \(i\) who are randomly assigned to the intervention arm is equal to the sum of independent and identically distributed Bernoulli random variables. The sum of independent and identically distributed Bernoulli random variables has a binomial distribution with parameters that represent the number of trials and the probability of success in each trial. Therefore, the total number of participants of potential outcome type \(i\) who are randomly assigned to the intervention arm is distributed according to a binomial distribution with parameters \(\theta(i)\) and \(p\).

The four random variables in (4) correspond to the number of individuals of each type who are randomized into intervention. Therefore, under the I.I.D. model of randomization, these four random variables are independent binomial random variables. Conditional on the intended fraction of participants in the intervention arm \(p\), it is possible to express the distribution of the data as
follows:

$$P(G = g \mid \theta, s, p) = \sum_{\ell=0}^{\theta(1)} \binom{\ell}{\theta(1), p} \times \binom{g(2) - \ell}{\theta(2), p} \times \binom{\theta(1) + \theta(3) + g(1) + g(2) + g(3) - s - \ell}{\theta(3), p} \times \binom{s + \ell - \theta(1) - \theta(3) - g(2) - g(3), s - \theta(1) - \theta(2) - \theta(3), p}$$

where \( \binom{\cdot}{\cdot} \) is the binomial probability mass function,

$$\binom{k}{r, p} = P(K = k \mid r, p) = \{0 \leq k \leq r\} \binom{r}{k} p^k (1 - p)^{r-k}.$$ 

Another prominent randomization process is an urn model, where researchers set a fixed number of participants in intervention and control before randomization occurs. Copas (1973) derives the likelihood function when randomization is conducted via an urn model:

$$P(G = g \mid \theta, s, m) = \sum_{\ell=0}^{\theta(1)} \binom{\theta(1)}{\ell} \times \binom{\theta(2)}{g(2) - \ell} \times \binom{\theta(1) + \theta(3) + g(1) + g(2) + g(3) - s - \ell}{\theta(3)} \times \binom{s + \ell - \theta(1) - \theta(3) - g(2) - g(3), s - \theta(1) - \theta(2) - \theta(3), p}$$

The closed form expressions for the likelihood function when randomization is conducted via the I.I.D. model and the urn model are not meant to be exhaustive. Rather, they illustrate how knowledge of the experimental procedure translates into useful expressions for the likelihood of observed data. The derivation of likelihood functions for other randomization processes follows using similar logic.

### 2.2 Analogy Between the Reduced Form and First Stage

Thus far, I have fixed ideas such that the outcome \( Y \) represents mortality. However, the outcome of interest can also be the takeup of a treatment. Consider a two stage model with binary intervention, binary treatment, and binary outcome. The framework presented here and the corresponding likelihood function can be used to describe the reduced form relationship between the intervention and the outcome as well as the first stage relationship between the intervention and the treatment. Table 1 summarizes the analogy. In the reduced form, the outcome is mortality, and in the first stage, the outcome is a treatment. Using the terminology of Angrist et al. (1996), the \( \theta(1) \) individuals who would live regardless correspond to the always takers, individuals who adopt treatment regardless of assignment to intervention or control. Similarly, \( \theta(2) \) corresponds to compliers, \( \theta(3) \)
to defiers, and \( \theta(4) \) to never takers. Table 1 shows how functions of these parameters yield other quantities of interest including the average intervention affect and the fraction affected in either direction.

Table 1: Analogy between Reduced Form and First Stage

| Quantity of Interest | Reduced Form Outcome: Mortality | First Stage Outcome: Treatment |
|----------------------|--------------------------------|--------------------------------|
| \( \theta(1) \)      | Number Who Would Live Regardless | Number of Always Takers        |
| \( \theta(2) \)      | Number Who Would Be Saved\( \dagger \) | Number of Defiers\( \dagger \) |
| \( \theta(3) \)      | Number Who Would Be Killed\( \circ \) | Number of Compliers\( \circ \) |
| \( \theta(4) \)      | Number Who Would Die Regardless | Number of Never Takers         |
| \( \theta(3)/\theta(2) \) | Number Who Would Be Killed / Number Who Would Be Saved | Number of Compliers / Number of Defiers |
| \((\theta(3)−\theta(2))/s\) | Average Intervention Effect     |
| \((\theta(3)+\theta(2))/s\) | Fraction Affected in Either Direction |

\( \dagger \) Respond to intervention in opposite direction of average if \( \theta(2) < \theta(3) \).

\( \circ \) Respond to intervention in opposite direction of average if \( \theta(3) < \theta(2) \).

3 Hypothesis Testing and Confidence Intervals

A closed form expression for the likelihood allows me to conduct inference on quantities of interest. As the test statistic, I use the likelihood ratio, which is the ratio of the maximum likelihood of the potential outcome type configurations under the null hypothesis to the maximum likelihood of all possible potential outcome type configurations. Consider the null hypothesis that the potential outcome configuration \( \theta \) is in the set \( H_0 \) against the alternative hypothesis that the potential outcome configuration \( \theta \) is not in the set \( H_0 \). For observed outcome configuration \( g \), define the likelihood ratio of \( g \) as:

\[
\lambda(g) = \frac{\max_{\theta \in H_0} P(G = g | \theta, s)}{\max_{\theta} P(G = g | \theta, s)}.
\]

The likelihood ratio \( \lambda(g) \) is constrained to \([0, 1]\). A likelihood ratio of one indicates that there exists a potential outcome type configuration \( \theta \) in the null hypothesis that maximizes the likelihood of observed outcome configuration \( g \). On the other hand, a likelihood ratio of zero indicates that observing outcome configuration \( g \) occurs with probability zero if the null hypothesis holds. Smaller values of the likelihood ratio \( \lambda(g) \) indicate that the null hypothesis is relatively less likely than the alternative hypothesis given observed outcome configuration \( g \). Observed outcome configurations with smaller likelihood ratios are more extreme under the null hypothesis.

I conduct a grid search to solve the non-linear integer programming problem in the likelihood
ratio and produce the finite sample distribution of the test statistic. Asymptotic inference with the likelihood ratio relies on Wilks’ Theorem, which states that, under certain regularity conditions, the asymptotic distribution of negative two times the log likelihood ratio is distributed $\chi^2_k$, where $k$ is the reduction in dimensionality of the parameter space when admitting the null. Instead of turning to the asymptotic distribution of the likelihood ratio, I use its exact finite sample distribution to conduct inference for two reasons. First, the regularity conditions of Wilks’ Theorem do not hold because the parameter space is discrete. Second, since I use a grid search, there is relatively low additional computational cost to calculating the finite sample distribution of the test statistic. Exact inference has the advantage of controlling the Type I error, which is the probability of falsely rejecting a true null hypothesis.

For any potential outcome configuration $\theta$ in the null hypothesis, I calculate the exact finite sample conditional distribution of the likelihood ratio $\lambda(G)$ using the known randomization process. In particular, I use the exact finite sample conditional distribution to evaluate the probability of observing a likelihood ratio $\lambda(g)$ less than or equal to a critical value $c$. Given observed outcome configuration $g$, I reject the null hypothesis when

$$\lambda(g) \leq c.$$ 

For a nominal level of the test $\alpha \in (0, 1)$, choose the critical value $c$ such that the Type I error rate is controlled at $\alpha$ for all potential outcome configurations in the null hypothesis. I choose the critical value $c$ to be the smallest number such that

$$\text{for all } \theta \in H_0, P(\lambda(G) \leq c|\theta, s) \leq \alpha.$$ 

By requiring that all potential outcome type configurations $\theta$ in the null hypothesis have controlled Type I error, this rejection rule corresponds to holding a “worst case prior” in a Bayesian framework. The rejection rule is equivalent to assigning each observed outcome configuration $g$ a p-value equal to

$$\max_{\theta \in H_0} P(\lambda(G) \leq \lambda(g)|\theta, s)$$

and rejecting the test exactly when this p-value is at most the nominal level of the test $\alpha$. For computation of this probability, I use the likelihood, which is computationally more efficient than simulating randomization, especially when the null hypothesis is large.

The test can generate confidence intervals on quantities of interest. Suppose we want to construct a two-sided confidence interval $[L, U]$ on quantity of interest $q(\theta)$ at a $(1 - \alpha)%$ level. Define the lower end of the confidence interval as the smallest real number $L$ such that the observed outcome configuration $g$ cannot reject the following hypothesis at an $\alpha/2%$ level:

$$H_0 : q(\theta) \leq L,$$

$$H_A : q(\theta) > L.$$
Define the upper end of the confidence interval as the largest real number $U$ such that the observed outcome configuration $g$ cannot reject the following hypothesis at an $\alpha/2\%$ level:

$$
H_0 : q(\theta) \geq U,
$$

$$
H_A : q(\theta) < U.
$$

The construction of one-sided confidence intervals proceeds analogously. It is possible to compute these confidence intervals because the domain of the quantity of interest $q(\theta)$ is finite, so $q(\theta)$ can only admit finitely many values. Therefore, only finitely many hypothesis tests are required to construct the confidence interval.

4 Examples: Clinical Trials for New Drugs

A data analyst at the Food and Drug Administration (FDA) receives the results from randomized clinical trials for two new cancer drugs, Mortem and Vita, where the intervention is access to the new drug and the outcome is mortality. The analyst will use these results to recommend approval or rejection of each drug independently, meaning that the analyst has no reason to compare the trial results. Each trial included 100 participants and used an I.I.D. randomization process. The observed outcome groups for each trial are shown in Table 2.

| Intervention | Control |
|--------------|---------|
| Z=1          | Z=0     |
| Dead         | Alive   |
| Y=1          | Y=0     |
| 15           | 35      |
| Fraction Dead = 0.30 | Fraction Dead = 0.70 |
| Estimated Average Intervention Effect = 0.30 - 0.70 = -0.40 |

The FDA is interested in the average intervention effect. Estimates reveal that both interventions reduce average mortality by 40 percentage points. In both trials, a regression of the outcome on the intervention with robust standard errors indicates that this reduction in mortality is significantly different from zero at conventional levels, as shown in the first line of Table 3. This result is approximate in the sense that it relies on asymptotic assumptions. As shown in the second line of Table 3, the general test, which does not rely on asymptotic assumptions, confirms the finding that the average intervention effect is significantly different from zero. Using evidence on the average intervention effect, the analyst would recommend approval for both drugs at any conventional confidence level.

Sometimes, however, the FDA is only interested in whether a drug would kill any participants, not the average intervention effect. For example, safety is the entire focus of Phase I clinical trials.
The FDA uses additional data on secondary outcomes to infer whether access to the drug would kill any participants, and shuts down trials if side effects seem too large. However, using mortality data alone and the general test, the analyst could directly infer whether access to each drug would kill any participants. As shown in the third line of Table 3, the observed outcomes from the Mortem trial reject the hypothesis that no one would be killed at a 3% level, whereas the observed outcomes from the Vita trial do not reject this hypothesis at any significance level. The corresponding one-sided 95% confidence intervals on the number of participants that would be killed by each drug show that access to Mortem would kill at least 3 of the 100 trial participants, but access to Vita need not kill any participants. If the FDA’s only criterion for rejection is evidence that access to the drug would kill participants, then the analyst would reject Mortem with 97% confidence but would recommend approval for Vita.

Table 3: Inference on Clinical Trial Data

| Quantity of Interest | Test | Inference |
|----------------------|------|-----------|
| Average Intervention Effect | Regression Test | Mortem | Vita |
| \((\theta(3)−\theta(2))/s\) | \(H_0: (\theta(3)−\theta(2))/s = 0; \) \(H_\Lambda: (\theta(3)−\theta(2))/s \neq 0.\) | \([-0.58, -0.22]\) \{0.00003\} | \([-0.54, -0.26]\) \{0.0000001\} |
| Average Intervention Effect | General Test | \(H_0: (\theta(3)−\theta(2))/s = 0; \) \(H_\Lambda: (\theta(3)−\theta(2))/s \neq 0.\) | \([-0.52, -0.24]\) \{0.000006\} | \([-0.50, -0.30]\) \{0.00000007\} |
| Number Who Would Be Killed | General Test | \(H_0: \theta(3) = 0; \) \(H_\Lambda: \theta(3) > 0.\) | \([3, - ]\) \{0.03\} | \([0, - ]\) \{1.00\} |
| Number Who Would Be Killed / Number Who Would Be Saved | General Test | \(H_0: \theta(3)/\theta(2) \leq 0.2; \) \(H_\Lambda: \theta(3)/\theta(2) > 0.2.\) | \([0 , - ]\) \{0.08\} | \([0 , - ]\) \{1.00\} |

* o Respond to intervention in opposite direction of average if \(\theta(3) < \theta(2). \)

Note. p-values for the hypothesis test are in curly braces. 95% confidence intervals corresponding to the hypothesis test are in square brackets. A dash (—) indicates that the confidence interval is one-sided. Regression test conducted using robust standard errors. When \(\theta(2) = 0\) and \(\theta(3) > 0\), the ratio \(\theta(3)/\theta(2)\) is evaluated as infinity. When \(\theta(2) = \theta(3) = 0\), the ratio \(\theta(3)/\theta(2)\) is ill-defined, so potential outcome type configurations with \(\theta(2) = \theta(3) = 0\) are never labeled as elements of the null hypothesis when conducting inference on \(\theta(3)/\theta(2)\).
be killed in the intervention and control arms, which is likely given the randomization process. Despite the fact that the trials produce the same estimated average intervention effect, the data configurations produce different likelihoods for the number who would be killed, and the general finite sample test formalizes examination of those likelihoods.

In other scenarios, however, the FDA is willing to approve drugs that would kill sufficiently few patients for each patient that would be saved. This may be the case for a novel treatment of a rare terminal disease. For instance, the FDA may desire that no more than 1 participant would be killed by access to a drug for each 5 participants who would be saved by access to a drug. To evaluate this criterion, the analyst could use the general test on the hypothesis that the ratio of the number of people that the intervention would kill to the number of people it would save is less than $1/5 = 0.2$. Table 3 shows that the observed outcome groups of the Mortem trial reject the hypothesis at an 8% confidence level. In contrast, the observed outcome groups of the Vita trial cannot reject this hypothesis at any level. If the FDA’s only criterion for rejection is evidence that access to the drug would kill more than one participant for each five that access to the drug would save, then the analyst would reject Mortem with 92% confidence but would recommend approval for Vita.

5 Conclusion

Using knowledge of the randomization process within an experiment, I develop a general finite sample test to conduct inference on quantities that capture heterogeneous intervention effects. I demonstrate that I can conduct informative inference on quantities for which previous methods are uninformative, such as the number of participants who respond to the intervention in the opposite direction of the average. I also use the same approach to conduct inference on other quantities, such as the average intervention effect, for which previous methods are informative but approximate or restricted in the quantities that they can consider. Inference on quantities that capture heterogeneity can play an important role in decision making in health economics, as I demonstrate in hypothetical examples relating to drug approval.

The presence of participants that respond to an intervention in the opposite direction of the average can pose an ethical dilemma: is it permissible to scale up an experimental intervention that kills some in order to save others? The general finite sample test can help to determine the presence of such a dilemma. In my ongoing research, I am extending the general finite sample test with the goal of resolving such a dilemma. By incorporating additional data on covariates, secondary outcomes, and treatment takeup, I can perform richer inference. The main goal of such inference is to inform targeting of interventions toward individuals likely to be saved and away from individuals likely to be killed. Such inference can also provide novel tests of previous econometric models. Notably, in Kowalski (2019c) I consider a model that incorporates data on treatment takeup in which I can extend general finite sample inference to test the exclusion restriction separately from the LATE monotonicity assumption.

An important area for future work is to collect data on the structure of the randomization processes for existing experiments and to analyze them using general finite sample inference. Al-
though the computations are intensive, they only need to be executed once for an experiment of a given size with a given randomization process. To make application of the general finite sample test feasible for researchers with limited computational resources, I can provide reference tables for common sample sizes and randomization processes.

References

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

Balke, A. and J. Pearl (1993). Nonparametric bounds on causal effects from partial compliance data. Working paper.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439), 1171–1176.

Chiba, Y. (2015). Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *Journal of Biometrics & Biostatistics* 6(244).

Chung, E., J. P. Romano, et al. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* 41(2), 484–507.

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

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Ding, P., A. Feller, and L. Miratrix (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(3), 655–671.

Ding, P. and L. W. Miratrix (2019). Model-free causal inference of binary experimental data. *Scandinavian Journal of Statistics* 46(1), 200–214.

Fisher, R. (1935). *Design of Experiments* (1st ed.). Oliver and Boyd.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review* 5, 5–15.

Gadbury, G. L., H. K. Iyer, and J. M. Albert (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference* 121(2), 163–174.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2001). Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric Evaluation of Labour Market Policies*, pp. 1–15. Springer.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81(396), 945–960.

Huber, M. and G. Mellace (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics* 97(2), 398–411.
Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica* 62(2), 467–475.

Imbens, G. W. and D. B. Rubin (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies* 64(4), 555–574.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica* 83(5), 2043–2063.

Kowalski, A. E. (2019a). Counting defiers. *NBER Working Paper* 25671.

Kowalski, A. E. (2019b). A model of a randomized experiment with an application to the prowess clinical trial. *NBER Working Paper* 25670.

Kowalski, A. E. (2019c). Recovering heterogeneity from an instrumental variable model without exclusion or monotonicity. Working paper.

Li, X. and P. Ding (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine* 35(6), 957–960.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics* 212, 522–555.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* 80(5), 319–323.

Manski, C. F. (1997). Monotone treatment response. *Econometrica* 65(6), 1311–1334.

Mourifié, I. and Y. Wan (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics* 99(2), 305–313.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Berlin, Heidelberg: Springer Science & Business Media.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10, 1–51.

Richardson, T. S. and J. M. Robins (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 415–444.

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

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics* 2(1), 1–26.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher Randomization Test comment. *Journal of the American Statistical Association* 75(371), 591–593.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal* 94(1395), 1395–1415.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523), 1228–1242.

Wu, J. and P. Ding (2019). Randomization tests for weak null hypotheses. *arXiv preprint arXiv:1809.07419*. 

17