Counting Defiers in Health Care: A Design-Based Model of an Experiment Can Reveal Evidence Against Monotonicity*

Neil Christy and Amanda Ellen Kowalski

December 18, 2024

Abstract

We show that a design-based model of an experiment with a binary intervention and outcome can reveal empirical evidence against a "monotonicity" assumption that the intervention affects all subjects in weakly the same direction. A canonical sampling-based model cannot reveal evidence against monotonicity, but we show that a design-based model and other sampling-based models can. We use statistical decision theory to propose a maximum likelihood decision rule that does not assume monotonicity and provide conditions for its optimality. Under these conditions, the performance of our rule relative to a design-based rule that assumes monotonicity increases with the sample size across all possible experiments with 6 to 40 subjects. In a given experiment, we quantify evidence against monotonicity with a likelihood ratio. Using figures that we develop to visualize potential outcomes, we illustrate evidence against monotonicity in a real experiment that examines the impact of a health care intervention. Even though the experiment shows a large and statistically significant average effect in one direction, our rule reveals positive counts of compilers who respond in that direction and defiers who respond in the other.

*Previous versions of this paper have circulated under different titles (Kowalski, 2019a,b; Christy and Kowalski, 2024a,b). We extend special thanks to Jann Spiess for extensive regular feedback and to Charles Manski, Aleksey Tetenov, Toru Kitagawa, and Donald Rubin for encouraging us to use statistical decision theory and teaching us about it. We thank Guido Imbens for foundational feedback. We also thank Elizabeth Ananat, Don Andrews, Isaiah Andrews, Josh Angrist, Susan Athey, Victoria Baranov, Steve Berry, Stephane Bonhomme, Michael Boskin, Zach Brown, Kate Bundorf, Matias Cattaneo, Xiaohong Chen, Victor Chernozhukov, Janet Currie, Peng Ding, Pascaline Dupas, Brad Efron, Natalia Emanuel, Ivan Fernandez-Val, Michael Gechter, Andrew Gelman, Matthew Gentzkow, Sander Greenland, Florian Gunsilius, Andreas Hagemann, Sukjin Han, Jerry Hausman, Han Hong, Daniel Kessler, Michal Kolesár, Jonathan Kolstad, Ang Li, John List, Bentley MacLeod, Aprajit Mahajan, José Luis Montiel Olea, Derek Neal, Andriy Norets, Matthew Notowidigdo, Elena Pastorino, John Pepper, Demian Pouzo, Tanya Rosenblat, Azeem Shaikh, Elie Tamer, Edward Vytlacil, Stefan Wager, Chris Walker, Christopher Walters, Thomas Wiemann, David Wilson, and seminar participants at the Advances with Fields Experiments Conference at the University of Chicago, the AEA meetings, the Bravo Center/SNSF Workshop on Using Data to Make Decisions, Columbia, the Essen Health Conference, Harvard Medical School, the John List Experimental Seminar, MIT, Notre Dame, NYU, Princeton, the Stanford Hoover Institution, UCLA, UVA, the University of Michigan, the University of Zurich, Yale, and the Y-RISE Evidence Aggregation and External Validity Conference for helpful comments. We thank Charles Antonelli, Bennett Fauber, Corey Powell, and Advanced Research Computing at the University of Michigan, as well as Misha Guy, Andrew Sherman, and the Yale University Faculty of Arts and Sciences High Performance Computing Center. Tory Do, Simon Essig Aberg, Jack Cavanaugh, Bailey Flanigan, Pauline Mourot, Srajal Nayak, Sukanya Sravasti, and Matthew Tauzer provided excellent research assistance.

1 Introduction

Suppose you have a health care intervention that could plausibly affect some patients in one direction and other patients in the opposite direction. You run an experiment with a binary intervention and a binary outcome. In terms of an instrumental variable model, you have the first stage or the reduced form but not both. Using terminology for the first stage, compliers are treated if and only if assigned intervention, defiers are treated if and only if assigned control, always takers are treated regardless, and never takers are untreated regardless (Angrist, Imbens, and Rubin, 1996). Influential work by Imbens and Angrist (1994) proposes a monotonicity assumption that assumes away compilers or defiers in the first stage, and Manski (1997b) proposes an analogous assumption in the reduced form. Monotonicity assumptions are ubiquitous in the analysis of first stages, and they are gaining traction in the analysis of reduced forms (Alsan, Cawley, Doyle, and Skelley, 2025). They are useful for the interpretation of results, but they can be difficult to defend in some contexts, particularly in health care settings where interventions may be helpful for some but harmful for others. In the context of your experiment, you would prefer to decide whether to impose a monotonicity assumption based on evidence.

We show that a design-based model of an experiment can reveal empirical evidence against monotonicity through curvature in the likelihood, which we illustrate with figures that we develop to visualize potential outcomes. As we review, the canonical sampling-based likelihood has flat regions that preclude evidence against monotonicity.[1] However, we show that the likelihood that arises from an alternative sampling-based model in which sampling is done without replacement has curvature that can reveal evidence against monotonicity.

We propose a novel design-based maximum likelihood decision rule in the style of Wald (1949) that does not assume monotonicity. Maximum likelihood estimators are well established and frequently used. There are also straightforward conditions under which our maximum likelihood rule is Bayes optimal, which we review. Alternatively, we can justify our maximum likelihood rule with the principle of maximum entropy, which does not require specification of a Bayesian prior (Jaynes, 1968).

We construct a design-based rule that assumes monotonicity, and we use it to quantify the loss from and evidence against monotonicity. To quantify the loss from monotonicity across all possible experiments with even-numbered sample sizes from 2 to 40, we calculate the exact ratio of the Bayes expected utility from our optimal maximum likelihood rule to that of the suboptimal rule that assumes monotonicity.

[1]For a given pair of marginal distributions, the likelihood function in the canonical sampling-based model is flat over all copulas connecting them—a classic result of Boole (1854), Hoeffding (1940), and Fréchet (1957). A large literature has focused on specifying what we can learn from estimates of these copula bounds in the canonical sampling-based model (see, for example, Balke and Pearl, 1997; Heckman, Smith, and Clements, 1997; Manski, 1997a; Tian and Pearl, 2000; Zhang and Rubin, 2003; Fan and Park, 2010; Mullahy, 2018; Ding and Miratrix, 2019; Li and Pearl, 2019; Bai, Huang, Moon, Shaikh, and Vytlacil, 2024; and Semenova, 2024).

The ratio grows with the sample size, from one in the sample size of 2 to 1.17 in the sample size of 40. To quantify the evidence against monotonicity in a specific experiment, we utilize the likelihood ratio.

We illustrate evidence against monotonicity by applying the maximum likelihood decision rule and the likelihood ratio in a real experiment in health care. The experiment analyzes the effect of assignment to high dose Vitamin C on survival among patients with sepsis (Zabet, Mohammadi, Ramezani, and Khalili, 2016). Sepsis treatments can have serious or even fatal side effects (Warren, Suffredini, Eichacker, and Munford, 2002), so the case for a monotonicity assumption is weak. This trial finds a large and statistically significant effect of the intervention on survival. Both the monotonicity and maximum likelihood decision rules preserve the magnitude of this effect as the difference in the estimated numbers of compliers and defiers; but while the former estimates no defiers by construction, the latter estimates positive numbers of both compliers and defiers with a likelihood 1.19 times larger. This example shows that our proposed decision rule can reveal evidence against monotonicity, even if the estimated effect is large and statistically significant. However, our rule need not always reveal evidence against monotonicity. We show that our rule could have supported monotonicity in an experiment with the same size and estimated average effect if different numbers of subjects had survived in each arm.

Our work reveals evidence against monotonicity by uniting statistical decision theory with a design-based model of an experiment based on causal models of potential outcomes from Neyman (1923), Welch (1937), Kempthorne (1952), Copas (1973), Rubin (1974, 1977), Greenland and Robins (1986), Holland (1986) and others. Copas (1973) derives the design-based likelihood of a completely randomized experiment. We generalize the design-based model to derive the likelihood from a generic randomization process, which we demonstrate with an application to Bernoulli trials, and we extend the model to produce a sampling-based likelihood that can reveal evidence against monotonicity, unlike the canonical sampling-based model. We examine these likelihoods within a statistical decision theory framework to engage with monotonicity assumptions that became ubiquitous decades later. Our work contributes to the integration of statistical decision theory into econometrics (Manski, 2004; Dehejia, 2005; Manski, 2007; Hirano, 2008; Hirano and Porter, 2009; Stoye, 2012; Kitagawa and Tetenov, 2018; Manski, 2018, 2019; Hirano and Porter, 2020; Manski and Tetenov, 2021; Fernández, Olea, Qiu, Stoye, and Tinda, 2024), particularly within finite sample settings (Canner, 1970; Manski and Tetenov, 2007; Schlag, 2007; Stoye, 2007, 2009; Tetenov, 2012).

We illustrate evidence against monotonicity using visualizations of potential outcomes, and we see the development of these visualizations as a secondary contribution. They allow us to depict how the same data can arise from different distributions of potential outcomes, some that satisfy monotonicity and some that do not. Given the design-based randomization process, there are more ways for the data to arise from some distributions of potential outcomes than others, giving rise to different values

of the design-based likelihood. Our visualizations provide intuition for why curvature in the design-based likelihood can provide evidence against monotonicity.

Previous approaches can only reveal evidence against monotonicity with additional data. Machine learning approaches of Wager and Athey (2018) and Semenova (2024) require rich data on covariates. Analysis of side effects in medicine requires data on secondary outcomes. Specification tests of instrumental variable model assumptions can reveal evidence against monotonicity in the first stage (Imbens and Rubin, 1997; Richardson and Robins, 2010; Huber and Mellace, 2013, 2015; Kitagawa, 2015; Mourifié and Wan, 2017; Machado, Shaikh, and Vytlacil, 2019), and marginal treatment effect methods can reveal evidence against monotonicity in the second stage (Björklund and Moffitt, 1987; Heckman and Vytlacil, 1999; Kowalski, 2023a,b), but such approaches require the additional structure of a two-stage model as well as data on both stages.

The remainder of the paper proceeds as follows: Section 2 compares the canonical sampling-based model to the design-based model and an alternative sampling-based model, both of which can reveal evidence against monotonicity. In Section 3, we introduce the statistical decision theory framework and our two proposed decision rules. Section 4 quantifies the losses from and evidence against monotonicity. We show that our proposed decision rule can reveal evidence against monotonicity in a real health care experiment in Section 5. Section 6 concludes.

2 Sampling- and Design-Based Models of an Experiment

2.1 A Canonical Sampling-Based Model with Replacement Cannot Reveal Evidence against Monotonicity

Each individual has a binary potential outcome $Y_I \in \{0,1\}$ in intervention and $Y_C \in \{0,1\}$ in control, where 1 represents "treated" and 0 represents "untreated." Individuals are randomly assigned to intervention ($Z = I$) or control ($Z = C$), and one of their potential outcomes is the observed outcome Y:

$$Y = \mathbf{1}_{\{Z=I\}}(Y_I) + \mathbf{1}_{\{Z=C\}}(Y_C),$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.[2] Assignment to intervention Z is independent of an individual's potential outcomes (Y_I, Y_C).

Individuals are sampled from a superpopulation with replacement, implying that their potential outcomes are independently and identically distributed (I.I.D.). There are four "principal strata," defined by combinations of potential outcomes (Frangakis and Rubin, 2002): let q_{11} represent the share of superpopulation who are always takers ($Y_I = 1$, $Y_C = 1$), q_{10} the share of compliers ($Y_I = 1$, $Y_C = 0$), q_{01} the share of defiers ($Y_I = 0$, $Y_C = 1$), and q_{00} the share never takers ($Y_I = 0$, $Y_C = 0$).

[2] An individual's observed outcome depends only on their own potential outcomes and their inclusion in the intervention or control arm, ruling out network-type effects through a "no interference" (Cox, 1958) or "stable unit treatment value" (Rubin, 1980) assumption.

We can derive likelihood functions for the superpopulation distribution of potential outcomes $\boldsymbol{q} \equiv (q_{11}, q_{10}, q_{01}, q_{00})$ given the data from the experiment, depending on the randomization procedure. In a sample size of n, if subjects are assigned to intervention or control via a series of Bernoulli trials (also known as "simple randomization"), where p represents the probability of assignment to intervention for each individual, the likelihood takes the form

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \frac{n!}{x_{I1}!x_{I0}!x_{C1}!x_{C0}!} & p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\ & \big(q_{11}+q_{10}\big)^{x_{I1}}\big(1-(q_{11}+q_{10})\big)^{x_{I0}} \\ & \big(q_{11}+q_{01}\big)^{x_{C1}}\big(1-(q_{11}+q_{01})\big)^{x_{C0}}, \end{aligned} \tag{1}$$

where $\boldsymbol{x} = (x_{I1}, x_{I0}, x_{C1}, x_{C0})$ is the observed realization of the experimental outcome $\boldsymbol{X}$, with x_{I1} representing the number of treated individuals in the intervention arm ($Z = I, Y = 1$), x_{I0} the number of untreated individuals in the intervention arm ($Z = I, Y = 0$), x_{C1} the number of treated individuals in the control arm ($Z = C, Y = 1$), and x_{C0} the number of untreated individuals in the control arm ($Z = C, Y = 0$). This likelihood is equivalent to a likelihood appearing in Barnard (1947).

Alternatively, if the experiment is "completely randomized," such that the n individuals are assigned to intervention or control by drawing a fixed number $m \leq n$ of names from a hat, then the likelihood takes the form

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = & \binom{m}{x_{I1}}\big(q_{11}+q_{10}\big)^{x_{I1}}\big(1-(q_{11}+q_{10})\big)^{m-x_{I1}} \\ & \binom{n-m}{x_{C1}}\big(q_{11}+q_{01}\big)^{x_{C1}}\big(1-(q_{11}+q_{01})\big)^{n-m-x_{C1}}. \end{aligned} \tag{2}$$

This likelihood is equivalent to likelihoods from Barnard (1947) and Kline and Walters (2020). We describe the derivation of both likelihoods (1) and (2) in greater detail in Appendix A.

These likelihood functions show that the experiment cannot reveal evidence against monotonicity in the superpopulation. Consider any candidate superpouplation distribution of potential outcomes $\boldsymbol{q}$ with non-zero shares of both compliers and defiers. Without loss of generality, suppose the share of compliers is at least as large as the share of defiers, $q_{10} \geq q_{01}$. Then, there is an alternative distribution $\boldsymbol{q}'$ with no defiers that has the same likelihood as $\boldsymbol{q}$ (specifically, $q'_{11} = q_{11} + q_{01}$, $q'_{10} = q_{10} - q_{01}$, $q'_{01} = 0$, and $q'_{00} = q_{00} + q_{01}$). That is, for every distribution violating the monotonicity assumption, there is an alternative distribution satisfying the monotonicity assumption for which there is equal evidence. In fact, the likelihood functions are flat over all superpopulation joint distributions of potential outcomes with the same values of $q_{11} + q_{01}$, which is the marginal distribution of Y_I, and $q_{11} + q_{01}$, which is the marginal distribution of Y_C, yielding the well-known copula bounds of Boole (1854)-Fréchet

(1957), and Hoeffding (1940).

2.2 A Design-Based Model Can Reveal Evidence against Monotonicity

2.2.1 Derivation of the Design-Based Likelihood

In the design-based framework, we restrict our attention to the fixed, but unknown, joint distribution of potential outcomes *within the sample*, rather than within some superpopulation. We represent this distribution with the sample counts of always takers θ_{11}, compliers θ_{10}, defiers θ_{01}, and never takers θ_{00}. The values $\theta_{11}+\theta_{10}+\theta_{01}+\theta_{00}$ sum to the sample size n. While the sample distribution of potential outcomes could be equivalently represented through shares rather than counts, we focus on counts to make use of known finite-sample probability distributions in the likelihood derivation below.

The design-based likelihood of the joint distribution of potential outcomes in the sample $\boldsymbol{\theta} = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ given the data $\boldsymbol{x}$ is the probability of this realization of the data given $\boldsymbol{\theta}$. To derive the distribution of the data, first let us define the joint distribution of potential outcomes in the intervention arm $\boldsymbol{I} \equiv \big(I_{11}, I_{10}, I_{01}, I_{00}\big)$ as a random vector whose elements represent the numbers of always takers, compliers, defiers, and never takers randomized into intervention. These values are unobserved to the experimenter. However, the randomization process—the design of the experiment—implies a data generating process for the distribution of potential outcomes in the intervention arm $\boldsymbol{I}$ conditional on the distribution of potential outcomes in the sample $\boldsymbol{\theta}$, which we can then use to derive the distribution of the data $\boldsymbol{X}$.

Consider an experiment where the randomization design assigns individuals to the intervention or control arm through a series of Bernoulli trials. Then, each individual's probability of assignment to intervention is p, and assignments are independent across individuals as well as across groups. This independence allows us to write the distribution of $\boldsymbol{I}$ as the product of four independent binomial distributions:

$$\mathbb{P}\Big(I_{11} = I_{11}, I_{10} = I_{10}, I_{01} = I_{01}, I_{00} = I_{00} \mid \boldsymbol{\theta}\Big) = \binom{\theta_{11}}{I_{11}}\binom{\theta_{10}}{I_{10}}\binom{\theta_{01}}{I_{01}}\binom{\theta_{00}}{I_{00}} \times p^{\sum_{j,k} i_{j,k}}(1-p)^{n-\sum_{j,k} i_{j,k}}. \quad (3)$$

Alternatively, in a completely randomized experiment, the randomization design fixes the number of individuals in the intervention group m and selects any of the possible combinations of m individuals in intervention and $n-m$ individuals in control with equal probability, as though drawing names from a hat. Under this design, $\boldsymbol{I}$ follows a multivariate hypergeometric distribution:

$$\mathbb{P}\Big(I_{11} = I_{11}, I_{10} = I_{10}, I_{01} = I_{01}, I_{00} = I_{00} \mid \boldsymbol{\theta}\Big) = \frac{\binom{\theta_{11}}{I_{11}}\binom{\theta_{10}}{I_{10}}\binom{\theta_{01}}{I_{01}}\binom{\theta_{00}}{I_{00}}}{\binom{n}{m}}. \quad (4)$$

In either case, the experimental design implies a data-generating process for the

number of individuals of each type randomized into the intervention arm.

While the distribution of potential outcomes in the intervention arm $\boldsymbol{I}$ is unobservable, we can use its data generating process to derive the distribution of the observable data $\boldsymbol{X}$. Note that each individual randomized into the intervention group with outcome $Y = 1$ must have been either an always taker or a complier: $X_{I1} = I_{11} + I_{10}$. Each individual randomized into the intervention group with outcome $Y = 0$ must have been either a never taker or a defier: $X_{I0} = I_{00} + I_{01}$. In the control group, those observed with outcome $Y = 1$ must be either always takers that were not randomized into intervention, or defiers that were not randomized into intervention: $X_{C1} = \theta_{11} - I_{11} + \theta_{01} - I_{01}$. And finally, those in the control group with outcome $Y = 0$ must be either never takers that were not randomized into intervention, or compliers that were not randomized into intervention: $X_{C0} = \theta_{00} - I_{00} + \theta_{10} - I_{10}$. Thus, we can write the probability of the observed data $\boldsymbol{X}$ conditional on the joint distribution of potential outcomes $\boldsymbol{\theta}$ as:

$$\begin{aligned}
\mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) &= \mathbb{P}\Big(I_{11} + I_{10} = x_{I1},\ \big(\theta_{11} - I_{11}\big) + \big(\theta_{01} - I_{01}\big) = x_{C1}, \\
&\qquad I_{00} + I_{01} = x_{I0},\ \big(\theta_{00} - I_{00}\big) + \big(\theta_{10} - I_{10}\big) = x_{C0} \mid \boldsymbol{\theta}\Big) \\
&= \mathbb{P}\Big(I_{11} + I_{10} = x_{I1},\ I_{11} + I_{01} = x_{C1} - \theta_{11} - \theta_{01}, \\
&\qquad I_{00} + I_{01} = x_{I0},\ I_{00} + I_{10} = x_{C0} - \theta_{00} - \theta_{10} \mid \boldsymbol{\theta}\Big).
\end{aligned}$$

A realization of $\boldsymbol{X}$ may be produced from more than one realization of the distribution of potential outcomes in intervention $\boldsymbol{I}$. To find the probability of a realization of $\boldsymbol{X}$, we sum together the probabilities of each realization of $\boldsymbol{I}$ that could have produced it. We can index these realizations through the realization i of I_{11} and solving the following system of equations for the elements of $\boldsymbol{I}$:

$$\begin{aligned}
I_{11} + I_{10} &= x_{I1}, \\
I_{11} + I_{01} &= \theta_{11} + \theta_{01} - x_{C1}, \\
I_{11} + I_{10} + I_{01} + I_{00} &= x_{I1} + x_{I0}, \\
I_{11} &= i.
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
I_{11} &= i \\
I_{10} &= x_{I1} - i \\
I_{01} &= \theta_{11} + \theta_{01} - x_{C1} - i \\
I_{00} &= x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}.
\end{aligned}$$

The value i is restricted to the set $\mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})$ such that $\boldsymbol{I}$ remains within the support

implied by $\boldsymbol{\theta}$, namely $0 \leq I_{11} \leq \theta_{11}$, $0 \leq I_{10} \leq \theta_{10}$, $0 \leq I_{01} \leq \theta_{01}$, and $0 \leq I_{00} \leq \theta_{00}$. The probability of a realization of $\boldsymbol{X}$ is just the sum of the probability of each of these realizations of $\boldsymbol{I}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} \mathbb{P}\Big(& I_{11} = i, \\ & I_{10} = x_{I1} - i, \\ & I_{01} = \theta_{11} + \theta_{01} - x_{C1} - i, \\ & I_{00} = x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01} \mid \boldsymbol{\theta}\Big). \end{aligned} \tag{5}$$

Equation (5) represents the general design-based likelihood, where the probability of each distribution of potential outcomes in intervention $\boldsymbol{I}$ given the distribution of potential outcomes in the sample $\boldsymbol{\theta}$ is informed by the design of the experiment. When individuals are assigned to intervention by a series of Bernoulli trials, $\boldsymbol{I}$ follows the distribution in (3), yielding:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \\ & \times p^{x_{I1} + x_{I0}} (1-p)^{x_{C1} + x_{C0}}. \end{aligned} \tag{6}$$

Alternatively, in a completely randomized experiment, $\boldsymbol{I}$ follows the distribution in (4), yielding:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}} \Big/ \binom{n}{m} \end{aligned} \tag{7}$$

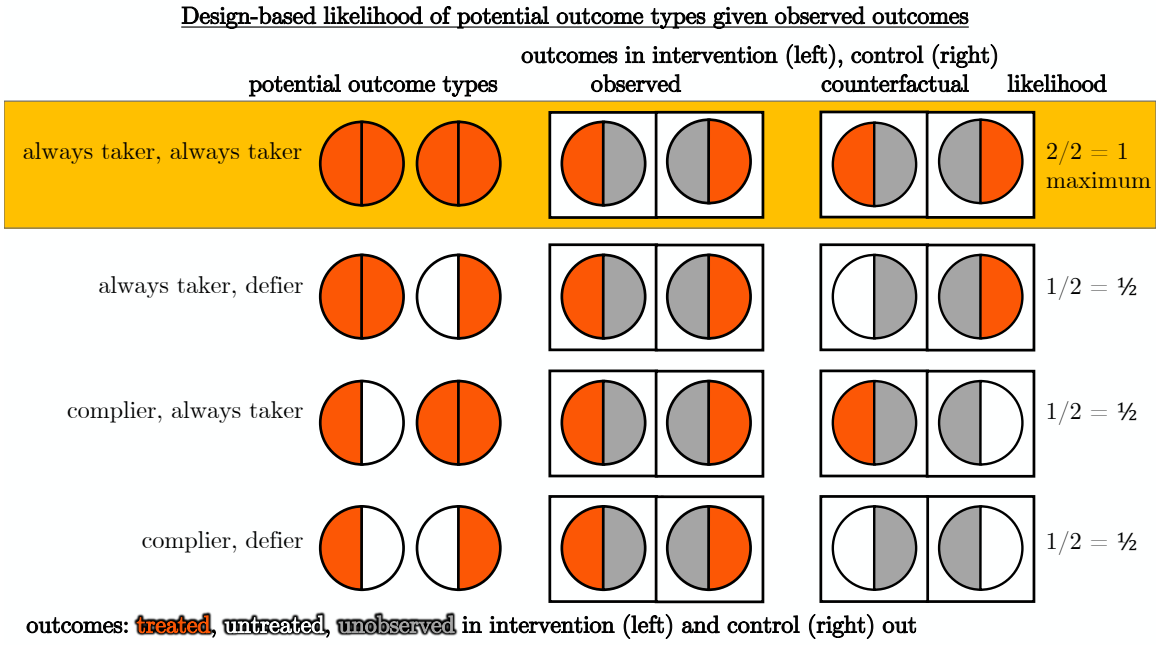
where we have substituted $m = x_{I1} + x_{I0}$. The latter likelihood function is equivalent

to the likelihood in Copas (1973).

2.2.2 An Illustration of Curvature in the Design-Based Likelihood

A simple example with two people can help us visualize curvature in the likelihood function. Suppose one person is treated in intervention and another is treated in control. The rows of Figure 1 show the four joint distributions of potential outcomes that could have produced one person treated in intervention and the other treated in control. We represent each person with a colored circle: the left half of the circle represents a person's potential outcome in intervention, and the right half of the circle represents a person's potential outcome in control (here, orange represents "treated" and white represents "untreated"). In each row, the two circles enter the experiment, represented by a pair of white squares, and one circle falls randomly into each square. The square on the left represents the intervention group and masks the right half of a circle, which we color grey; the square on the right represents the control group and masks the left half of a circle. The first column of pairs of squares represents what the observer would see if the first circle in the respective row were randomized into intervention and the second were randomized into control, while the second column shows the observable data under the counterfactual randomization.

Figure 1: An Illustration of a Randomized Experiment with Two People



The maximizer of the likelihood function indicates that both people are always takers. The intuition is simple. If they are both always takers, then even if the randomization had gone the other way such that the person assigned to intervention

were assigned to control and vice versa, you would have seen the same thing—both the person in intervention and control would still be treated.

Curvature in the likelihood is apparent from the fact that the number of ways that you could have seen what you actually have seen varies across the rows. The "always taker, always taker" row produces the actual observed data in two out of the two possible randomization outcomes. The value of the likelihood here is 1. In the remaining three rows, the actual observed data only occurs under one of the two randomization outcomes, so the value of the likelihood for these rows is 0.5. Paraphrasing the board book "Statistical Physics for Babies," (Ferrie, 2017), physicists refer to the number of ways that you could have seen what you have seen—that is, the numerator of our likelihood—as entropy. Differences in entropy yield curvature in the likelihood.

2.2.3 Curvature in the Design-Based Likelihood can Reveal Evidence Against Monotonicity

Unlike the sampling-based likelihood functions from the previous section, the design-based likelihood functions show that the experiment can reveal evidence against monotonicity. Within the sample, the marginal distribution of Y_I is determined by $\theta_{11} + \theta_{10}$, and the marginal distribution of Y_C by $\theta_{11} + \theta_{01}$. While the sampling-based likelihoods in (1) and (2) vary only with the marginal distributions of the potential outcomes in the superpopulation, both of the design-based likelihoods in (6) and (7) vary with the joint distribution of potential outcomes in the sample $\boldsymbol{\theta}$ even when holding constant the marginal distributions. That is, when both $\theta_{11} + \theta_{10}$ and $\theta_{11} + \theta_{01}$ are fixed, the likelihood functions maintain some curvature, and the experimenter can learn within the Boole-Fréchet-Hoeffding bounds.

We can show this likelihood curvature with a simple example. In the previous example with one person treated in intervention and one person treated in control, the likelihood is maximized when both subjects are always takers, and there is no evidence against monotonicity. It is not until a sample size of six or more that the experiment can provide evidence against monotonicity. Suppose $X_{I1} = 2$ people are treated in intervention, $X_{I0} = 1$ person is untreated in intervention, $X_{C1} = 1$ person is treated in control, and $X_{C0} = 2$ people are untreated in control. We can compare the likelihood values of two joint distributions of potential outcomes with the same marginal distributions: $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00}) = (2, 2, 0, 2)$ satisfies the monotonicity assumption, while $(\theta'_{11}, \theta'_{10}, \theta'_{01}, \theta'_{00}) = (0, 4, 2, 0)$ does not. Under the design of randomization by Bernoulli trials, the likelihood of $\boldsymbol{\theta}$ is exactly 0.125, while the likelihood of $\boldsymbol{\theta}'$ is exactly 0.1875; under the design of a completely randomized experiment, the likelihoods are 0.4 and 0.6. In either case, the experiment provides evidence against the distribution that satisfies the monotonicity assumption.

2.3 Other Sampling-Based Models Can Reveal Evidence Against Monotonicity

The canonical sampling-based model of an experiment from Section 2.1 demonstrates that the experiment cannot provide any evidence against monotonicity in the super-

population. In a seeming paradox, the design-based model from Section 2.2 shows that the experiment *can* provide evidence against monotonicity *in a sample*. How can these statements both be true? Is there any way to exploit the information about monotonicity in the sample to learn about monotonicity in the superpopulation?

To formalize the connection between the distribution of potential outcomes in a sample and the distribution of potential outcomes in a superpopulation, we extend our design-based model with an explicit sampling procedure. Analogously to how the randomization design implies a data generating process for the distribution of potential outcomes in the intervention arm, the sampling design implies a data generating process for the distribution of potential outcomes in the sample, which depends on the distribution in the superpopulation. Formally, we now let $\boldsymbol{\theta}$ be a random variable in $\boldsymbol{\Theta}$ whose distribution is parameterized by some $\boldsymbol{\gamma}$: $\boldsymbol{\theta} \sim \mathbb{P}_{\boldsymbol{\gamma}}$. Following the law of total probability, we can produce a likelihood function for the parameter $\boldsymbol{\gamma}$ by summing the design-based likelihood function in (5) over realizations of $\boldsymbol{\theta}$ weighted by $\mathbb{P}_{\boldsymbol{\gamma}}$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma} \mid \boldsymbol{x}) &= \mathbb{P}(\boldsymbol{x} \mid \boldsymbol{\gamma}) \\
&= \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}(\boldsymbol{x} \mid \boldsymbol{\gamma}, \boldsymbol{\theta}) \mathbb{P}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}(\boldsymbol{x} \mid \boldsymbol{\theta}) \mathbb{P}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \mathbb{P}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \qquad (8)
\end{aligned}$$

In the second to last line, we utilize the fact that $\boldsymbol{X}$ is independent of $\boldsymbol{\gamma}$ conditional on $\boldsymbol{\theta}$.

Equation (8) is the general form of a combined sampling- and design-based likelihood. Formalizing the likelihood this way shows that the information available about the distribution of potential outcomes in the sample $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ is mediated by how probable each sample distribution is in terms of $\mathbb{P}_{\boldsymbol{\gamma}}(\boldsymbol{\theta})$. Whether the curvature in the design-based likelihood persists depends crucially on the details of the sampling process.

A natural and illustrative starting point is the common assumption of I.I.D. sampling, or sampling with replacement. As in Section 2.1, sampling with replacement implies that we may write the probability of each individual's potential outcome realizations as $\boldsymbol{q}$. The joint distribution of potential outcomes in the sample $\boldsymbol{\theta}$ is then the counts of the realizations of independent categorical random variables, which itself is a random variable following the multinomial distribution:

$$\mathbb{P}_{\boldsymbol{\gamma}}\big(\boldsymbol{\theta} = \boldsymbol{\theta}'\big) = \frac{n!}{\theta'_{11}!\theta'_{10}!\theta'_{01}!\theta'_{00}!} q_{11}^{\theta'_{11}} q_{10}^{\theta'_{10}} q_{01}^{\theta'_{01}} q_{00}^{\theta'_{00}}. \qquad (9)$$

In this case, $\boldsymbol{\gamma} = \boldsymbol{q} \equiv (q_{11}, q_{10}, q_{01}, q_{00})$.

As a natural alternative, we also consider the procedure of sampling from a finite superpopulation *without replacement.* Let s_{11} be the number of always takers in the finite superpopulation, s_{10} the number of compliers, s_{01} the number of defiers, and s_{00} the number of never takers. Under sampling without replacement, the potential outcomes of each individual are no longer independently distributed. Instead, the resulting distribution for $\boldsymbol{\theta}$ is the multivariate hypergeometric distribution:

$$\mathbb{P}_{\boldsymbol{\gamma}}(\boldsymbol{\theta} = \boldsymbol{\theta}') = \frac{\binom{s_{11}}{\theta'_{11}}\binom{s_{10}}{\theta'_{10}}\binom{s_{01}}{\theta'_{01}}\binom{s_{00}}{\theta'_{00}}}{\binom{k}{n}}. \tag{10}$$

where $k \equiv s_{11} + s_{10} + s_{01} + s_{00}$ is the finite size of the superpopulation. In this case, $\boldsymbol{\gamma} = \boldsymbol{s} \equiv (s_{11}, s_{10}, s_{01}, s_{00})$.

Random sampling with and without replacement are just two of many sampling models, and different sampling models imply different likelihood functions through their respective distributions $\mathbb{P}_{\boldsymbol{\gamma}}$. Under the specific model of sampling with replacement (I.I.D.), the experiment cannot reveal evidence against monotonicity. In Appendix B, we algebraically derive closed form representations of the likelihood of $\boldsymbol{\gamma} = \boldsymbol{q}$ under sampling with replacement. Under the randomization design of Bernoulli trials, the within-sample likelihood function for $\boldsymbol{\theta}$ takes the form in (6), and the superpopulation likelihood function for $\boldsymbol{\gamma}$ in (8) simplifies exactly to the canonical superpopulation likelihood in (1). Under the design of complete randomization, the within-sample likelihood function for $\boldsymbol{\theta}$ takes the form in (7), and the superpopulation likelihood function for $\boldsymbol{\gamma}$ in (8) simplifies to the canonical superpopulation likelihood function in (2). So, under IID sampling, we are in exactly the cases described in Section 2.1: the likelihood functions are invariant within the Boole-Fréchet-Hoeffding bounds for each pair of marginal distributions, and the experiment can provide no evidence against monotonicity in the superpopulation.

This is not, however, the case for all sampling assumptions. Under specifications for $\mathbb{P}_{\boldsymbol{\gamma}}$ that differ from (9), the likelihood function in (8) need not simplify to an expression that depends only on the superpopulation marginal distributions of potential outcomes. In fact, under the distribution implied by sampling without replacement in (10), the likelihood retains some curvature conditional on the marginal distributions, and the experiment can provide evidence against monotonicity in the superpopulation. Under the randomization design of Bernoulli trials, this likelihood takes the

following form:

$$\mathcal{L}(\boldsymbol{s} \mid \boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} \binom{\theta_{11}}{i} \times \binom{\theta_{10}}{x_{I1} - i} \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \times p^{x_{I1} + x_{I0}} (1-p)^{x_{C1} + x_{C0}} \right] \times \frac{\binom{s_{11}}{\theta_{11}} \binom{s_{10}}{\theta_{10}} \binom{s_{01}}{\theta_{01}} \binom{s_{00}}{\theta_{00}}}{\binom{k}{n}}.$$

Under the design of a completely randomized experiment, the likelihood is:

$$\mathcal{L}(\boldsymbol{s} \mid \boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} \binom{\theta_{11}}{i} \times \binom{\theta_{10}}{x_{I1} - i} \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \times \binom{\theta_{00}}{m + x_{C1} + i - \theta_{11} - \theta_{01} - x_{I1}} \Big/ \binom{n}{m} \right] \times \frac{\binom{s_{11}}{\theta_{11}} \binom{s_{10}}{\theta_{10}} \binom{s_{01}}{\theta_{01}} \binom{s_{00}}{\theta_{00}}}{\binom{k}{n}}.$$

We can once again demonstrate curvature in these likelihoods through an example. Using the same data as the example in Section 2.2.3, we have $X_{I1} = 2$, $X_{I0} = 1$, $X_{C1} = 1$ and $X_{C0} = 2$. Suppose that this data represents half of the superpopulation, which was sampled into the experiment without replacement. We can compare the likelihood values of two superpopulation joint distributions with a population size of $k = 12$ and the same marginal distributions: $(s_{11}, s_{10}, s_{01}, s_{10}) = (4, 4, 0, 4)$ satisfies the monotonicity assumption, while $(s'_{11}, s'_{10}, s'_{01}, s'_{00}) = (0, 8, 4, 0)$ does not. Under the design of randomization by Bernoulli trials, the likelihood of $\boldsymbol{s}$ is approximately 0.082, while the likelihood of $\boldsymbol{s}'$ is approximately 0.085; under the design of a completely randomized experiment, the likelihoods are approximately 0.262 and 0.273. In either case, the experiment provides evidence against the distribution that satisfies the monotonicity assumption.

2.4 Choosing Between the Design-Based and Sampling-Based Models

Our results show that the design-based model of an experiment can offer novel evidence against monotonicity in the sample distribution of potential outcomes. In many cases, the sample distribution of potential outcomes is a more appropriate object of inquiry than a superpopulation distribution. Learning about the distribution of potential outcomes in the sample directly addresses the questions of "why" we saw the outcomes we did, and what we would have seen had the randomization gone differently. Counterfactual questions like these have attracted recent interest in the study of causal inference (Gelman and Imbens, 2013; Pearl and Mackenzie, 2018; Imbens, 2020; Dawid and Musio, 2022).

Evidence against monotonicity in the sample distribution of potential outcomes is especially useful when the outcomes of the particular individuals involved are of interest, such as in litigation, or when the sample constitutes the entire superpopulation of interest. Alternatively, some settings imply no sensible superpopulation on which to conduct inference. For example, if a sample consists of geographic regions of a country that are randomly assigned to some treatment, learning about the sample itself could be more useful than learning about a hypothetical superpopulation of geographic regions.

For researchers interested in the superpopulation, our results for sampling-based models suggest that I.I.D. sampling is not an innocuous assumption, as it precludes evidence about monotonicity. In many cases, I.I.D. sampling is invoked for analytical convenience rather than for realism. Many practical sampling procedures are known to differ from sampling with replacement. Clinical trials, for instance, are unlikely to admit the same individual to the trial multiple times. Some sampling processes may be "close enough" to I.I.D. sampling with replacement to justify the assumption, but divergences for small superpopulations can be significant. On the other hand, if the sampling process is truly unknown, focusing on the sample distribution of potential outcomes could be preferable to imposing unfounded sampling assumptions that restrict or eliminate evidence against monotonicity. In the remainder of this paper, we focus our attention the sample distribution of potential outcomes, which is both interesting in its own right, and an important first-step for sampling-based models that can reveal evidence against monotonicity in the superpopulation.

3 Proposed Design-Based Decision Rules with and without Monotonicity

3.1 Preliminaries on Statistical Decision Theory

To utilize the curvature in our design-based likelihood, we propose decision rules in the style of of Wald (1949). This decision theory setting provides two benefits for our framework. First, decision theory is natural to apply in our design-based setting, unlike alternative criteria like consistency that depend on large sample or asymptotic assumptions. Second, statistical decision theory provides straightforward

methods to quantify the gains from exploiting the full curvature in our likelihood over other decision rules. We focus here on the statistical decision problem of choosing the correct distribution of potential outcomes in the sample, rather than on testing hypotheses about the distribution. Classical hypothesis tests that control for test size prioritize a null hypothesis over its alternative, which could limit the amount of information we learn from the likelihood in our setting (Tetenov, 2012).

Suppose a decision maker wishes to guess the joint distribution of potential outcomes in the sample. We write the decision maker's guess as $\widehat{\boldsymbol{\theta}}$, and we define a utility function over a guess $\widehat{\boldsymbol{\theta}}$ and the true distribution $\boldsymbol{\theta}$. The utility function yields one util when the guess is correct and zero utils when the guess is incorrect:

$$u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbf{1}_{\{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}\}}.$$

We may also allow the decision maker to choose a randomized guess, which ascribes a probability distribution over the possible values of $\boldsymbol{\theta}$. We define the decision maker's utility over a randomized guess ρ as the expected utility of guessing according to the probabilities ascribed by ρ:

$$U(\rho, \boldsymbol{\theta}) = \sum_{\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \rho(\widehat{\boldsymbol{\theta}}).$$

The decision maker chooses a decision rule that maps the observable data into (possibly) randomized guesses.[3] We write such a rule as $f : \boldsymbol{\mathcal{X}} \to \Delta(\boldsymbol{\Theta})$, where $\boldsymbol{\mathcal{X}}$ is the space of possible data realizations, $\boldsymbol{\Theta}$ is the space of possible distributions of potential outcomes, and $\Delta(\boldsymbol{\Theta})$ is the space of distributions over $\boldsymbol{\Theta}$. Given a true distribution of potential outcomes $\boldsymbol{\theta}$, the decision maker's expected utility from following a decision rule f is the expected value of $U(f(\boldsymbol{X}), \boldsymbol{\theta})$ with respect to the experimental outcome $\boldsymbol{X}$:

$$\begin{aligned} EU(f, \boldsymbol{\theta}) &= \mathbb{E}\big[U\big(f(\boldsymbol{X}), \boldsymbol{\theta}\big) \mid \boldsymbol{\theta}\big] \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} \sum_{\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} u(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \mathcal{L}(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{x}) f(\boldsymbol{x})(\widehat{\boldsymbol{\theta}}) \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) f(\boldsymbol{x})(\boldsymbol{\theta}). \end{aligned}$$

Under the specified utility function, the decision maker's expected utility is equal to their probability of guessing correctly.

[3] We conflate here the standard definitions of "randomized decision rules" and "behavioral decision rules" (Ferguson, 1967) for expositional clarity. In settings of perfect recall, such as the setting we study here, the space of randomized and behavioral decision rules are equivalent (Kuhn, 1953).

3.2 Proposed Maximum Likelihood Decision Rule

We focus now on what is arguably the simplest and most familiar way to utilize the likelihood: guessing the sample distribution of potential outcomes that maximizes the likelihood. The maximum likelihood decision rule f^* can be defined as follows. Let $\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})$ be the set of $\boldsymbol{\theta}$ values that maximize the likelihood given the observed data $\boldsymbol{x}$:

$$\widehat{\boldsymbol{\Theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

There are finitely many vectors of integers $\boldsymbol{\theta}$ that sum to the actual number of participants in the experiment n, so $\boldsymbol{\Theta}$ is finite and $\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})$ is nonempty. We define the maximum likelihood decision rule f^* as:

$$f^*(\boldsymbol{x})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}(\boldsymbol{x})\}} & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}(\boldsymbol{x}), \\ 0 & \text{o.w.}, \end{cases} \tag{11}$$

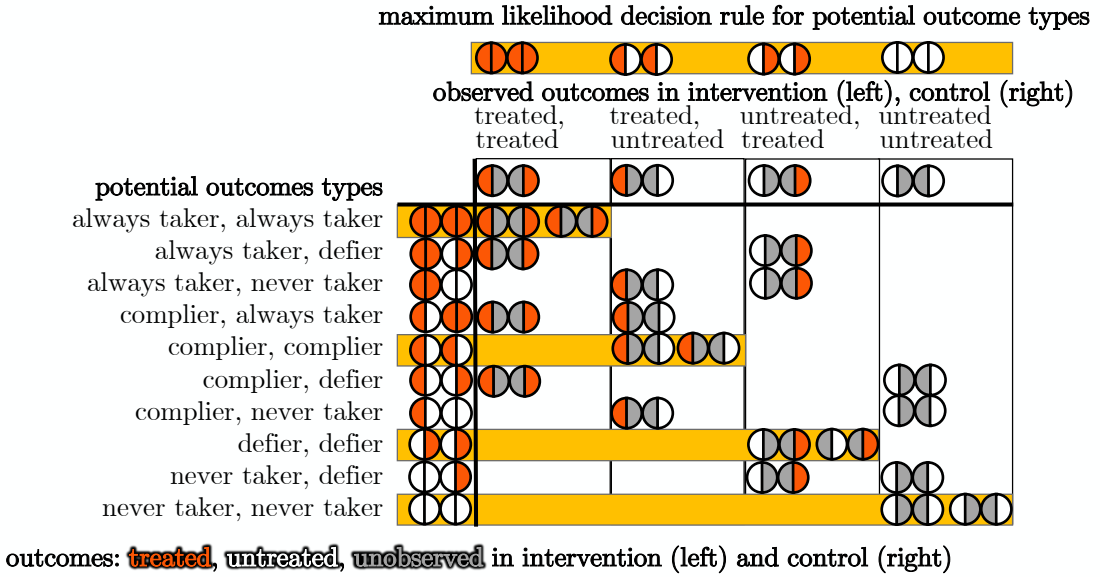
where $\#\{\cdot\}$ is the counting measure. When the likelihood is unimodal, f^* chooses the maximizer with probability one; when the likelihood is multimodal, f^* prescribes an equal probability to each maximizer. The maximum likelihood decision rule does not generically take a convenient analytical form. However, since $\boldsymbol{\Theta}$ is finite, the integer programming problem of maximizing the likelihood can be solved in small samples by an exhaustive grid search.

3.2.1 Illustration of the Maximum Likelihood Decision Rule

Figure 2 illustrates calculation of the maximum likelihood decision rule in an example of an experiment with one person observed treated in intervention and one person treated in control. The rows represent all possible joint distributions of potential outcomes in a sample of two people, while the columns represent all possible data outcomes that could be observed (note that Figure 1 focuses on the first column; here, the decision rule yields a decision for all possible realizations of the data). The cells of the matrix are populated based on the number of randomization outcomes within the given row that would produce the data observed in the given column; when both randomization outcomes would produce the same observation, we place the pairs of circles side by side. The likelihood value is one for every cell with two pairs of circles, 1/2 for every cell with one pair of circles, and zero otherwise. We see that, in the column for each realization of the data, there are four rows whose randomization outcomes could produce that data. Furthermore, each column has one row for which both randomization outcomes produce the relevant data. These rows are the likelihood maximizers, which we highlight in yellow.

Above the columns in Figure 2, we represent the maximum likelihood decision rule, also in yellow. The decision rule maps each column to a row representing a (degenerate) guess for the unobserved joint distribution of potential outcomes. In each row, the unique maximizer of the likelihood indicates that both people have the

Figure 2: Illustration of the Maximum Likelihood Decision Rule in a Sample of Two



same type. People can be the same or different, but in this simple experiment, the maximum likelihood decision rule always indicates that people are the same.[4]

3.2.2 Optimality of the Maximum Likelihood Rule

To evaluate the performance of a decision rule, the decision maker must consider how it performs across the various possible values of the true, unknown joint distribution of potential outcomes $\boldsymbol{\theta}$. One common approach is to measure the decision rule's performance as the average expected utility obtained according to some prior distribution for $\boldsymbol{\theta}$ (Bayes expected utility):

$$EU_{\pi}(f) = \mathbb{E}_{\pi}\big[EU(f, \boldsymbol{\theta})\big] = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} EU(f, \boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where $\pi \in \Delta(\boldsymbol{\Theta})$ is a subjective prior belief about $\boldsymbol{\theta}$. Decision rules that maximize this criterion are said to be "Bayes optimal."

While we need not specify a subjective prior to apply our maximum likelihood

[4] Andrew Gelman and Keith O-Rourke discuss the importance of "sameness" in statistical evidence: "Awareness of commonness can lead to an increase in evidence regarding the target; disregarding commonness wastes evidence; and mistaken acceptance of commonness destroys otherwise available evidence. It is the tension between these last two processes that drives many of the theoretical and practical controversies within statistics" (Gelman and O'Rourke, 2017). In Section 5 we continue our discussion of how the maximum of the likelihood indicates that people are similar, even in a larger experiment where the data would not allow all people to be of the same type.

decision rule, we can establish that it is Bayes optimal under a uniform prior.[5] We review this proof in Appendix C. Bayes optimality implies that our decision rule is admissible (Ferguson, 1967) and that the decision rule cannot be bested in a betting framework (Freedman and Purves, 1969). Additionally, Bayes optimality allows us to quantify the relative losses from using a suboptimal rule that assumes monotonicity, as we do in Section 4.

3.2.3 Maximum Likelihood Rule as Maximum Entropy

We can also motivate our maximum likelihood rule through the principle of maximum entropy (Jaynes, 1957a,b), which unites the theory of information with statistical physics. In Section 2.2.2, we discuss how differences in entropy drive curvature in the likelihood such that the maximum likelihood distribution is also the maximum entropy distribution. Jaynes (1968) advocates for application of the principle of maximum entropy to circumvent the subjective specification of a prior.[6] In Appendix D, we illustrate that in the Monty Hall problem, a celebrated application of Bayesian decision making, it is possible to make the same decision without specifying a Bayesian prior using the maximum likelihood decision rule and the principle of maximum entropy.

3.3 An Alternative Decision Rule with a Monotonicity Assumption

For comparison with our maximum likelihood rule, which does not impose a monotonicity assumption, we now construct a "monotonicity decision rule" that requires either the number of compliers or defiers (or both) to be zero. Our monotonicity decision rule chooses the distribution of potential outcomes that maximizes the design-based likelihood subject to this constraint. The seminal monotonicity assumptions of Imbens and Angrist (1994) or Manski (1997b) are large sample assumptions on an underlying superpopulation. Our monotonicity assumption is a distinct but analagous assumption for the finite sample setting. By constructing our monotonicity rule as a constrained maximum likelihood estimator, we consider the "best" estimate under monotonicity—that is, we do not place restrictions on the relationship between the data and the estimated marginal distributions or average intervention effect, which are often imposed in practice.

We define the restricted "monotonicity" set of distributions as $\boldsymbol{\Theta}^M$, where

$$\Theta^M = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_{10} = 0 \ \vee \ \theta_{01} = 0 \right\}.$$

[5] Thank you to Andriy Norets and Thomas Weimann for bringing these results to our attention.

[6] To make Bayesian decision-making more objective, one option is to choose the least informative prior. Jaynes' alternative to make the process more objective is to choose the least informative updated distribution, which is the distribution that maximizes entropy. In our case, this is the distribution that can generate the data in the greatest number of ways.

Next, we define the set of constrained maximizers of the likelihood:

$$\widehat{\Theta}^M(\boldsymbol{x}) = \underset{\boldsymbol{\theta}\in\Theta^M}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

Finally, we define the monotonicity rule f^M that chooses each of the constrained maximizers with equal probability:

$$f^M(\boldsymbol{x})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\widehat{\Theta}^M(\boldsymbol{x})\}} & \text{if } \boldsymbol{\theta} \in \widehat{\Theta}^M(\boldsymbol{x}), \\ 0 & \text{o.w.} \end{cases}$$

4 Quantifying the Loss From and Evidence Against a Monotonicity Assumption in a Design-Based Model

4.1 Quantifying the Exact Loss From Assuming Monotonicity

Optimality of the maximum likelihood decision rule implies that we can do weakly better (in the sense of Bayes expected utility) than assuming monotonicity. Now we ask: how much better? To answer this question, we compare the exact Bayes expected utility achieved by the maximum likelihood and monotonicity decision rules of Section 3. For a given sample size, we enumerate each of the fintely many feasible sample distributions of potential outcomes $\boldsymbol{\theta}$. For each $\boldsymbol{\theta}$, we compute the expected utility of our two decision rules by evaluating and comparing $\widehat{\boldsymbol{\theta}} = f(\boldsymbol{x})$ with $\boldsymbol{\theta}$ for every possible $\boldsymbol{x}$, and we average the resulting utilities over the values of $\boldsymbol{x}$ according to the likelihood. We then average these expected utilities according to a uniform prior for $\boldsymbol{\theta}$.

Figure 3 shows the ratio of the Bayes expected utilities from the maximum likelihood and monotonicity rules for even sample sizes between 2 and 40. For sample sizes of two and four, the maximum likelihood rule and the monotonicity rule achieve the same Bayes expected utility. As the sample size grows larger, the maximum likelihood rule strictly outperforms the monotonicity rule; in a sample of 40 people, the maximum likelihood rule achieves a Bayes expected utility 1.17 times that of the monotonicity rule.

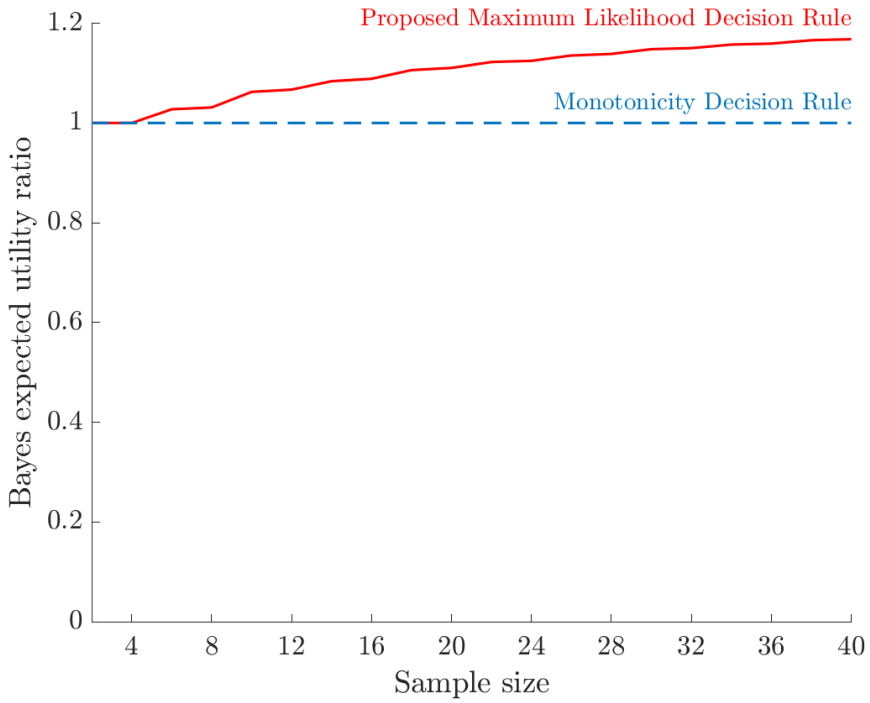
4.2 Quantifying Evidence Against Monotonicity in an Experiment

How strong is the evidence against monotonicity in a given experiment? A likelihood ratio can capture the relative strength of the evidence against monotonicity available in a realization of the data $\boldsymbol{x}$:

$$\mathrm{LR}(\boldsymbol{x}) \equiv \frac{\mathcal{L}\big(f^*(\boldsymbol{x}) \mid \boldsymbol{x}\big)}{\mathcal{L}\big(f^M(\boldsymbol{x}) \mid \boldsymbol{x}\big)}$$

The likelihood ratio is bounded below by 1, as the maximum likelihood rule is an unconstrained maximizer and the monotonicity rule is a constrained maximizer. A

Figure 3: Performance of Decision Rules Relative to Monotonicity Decision Rule



value of exactly 1 implies that the likelihood is maximized at a distribution satisfying montonicity, while values of the likelihood ratio greater than 1 imply that the experiment offers evidence against the monotonicity assumption. The value of the likelihood ratio has an intuitive interpretation—the ratio represents how much higher the probability of the observed data is under the best alternative to monotonicity.

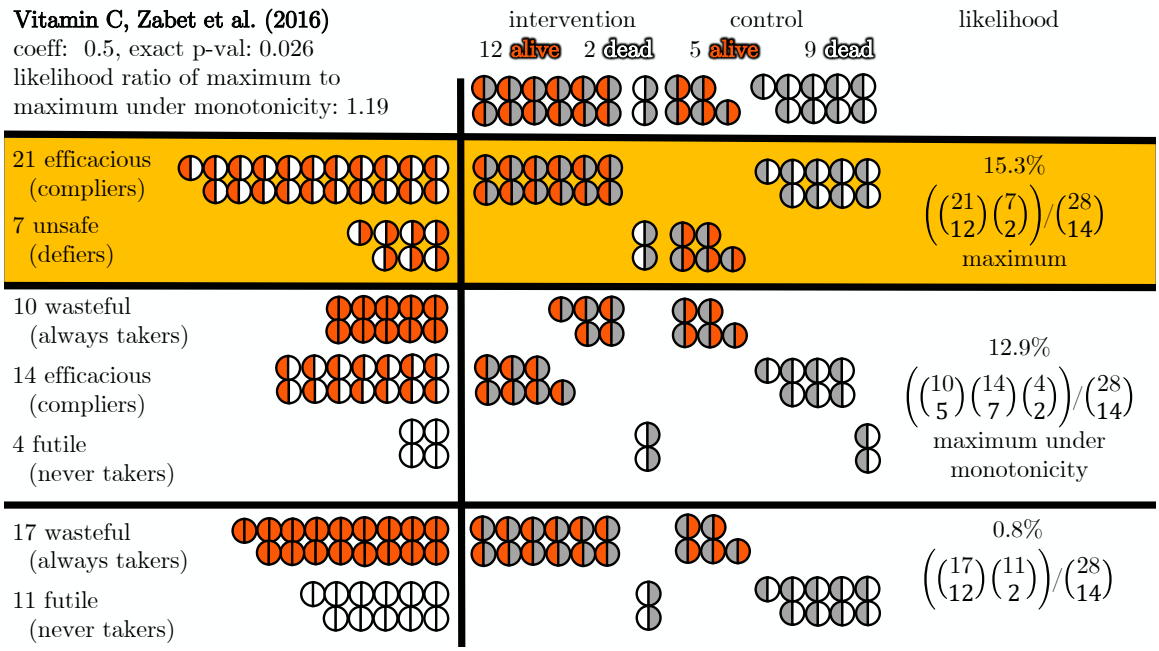
5 A Real Clinical Trial Reveals Evidence Against Monotonicity

We apply our decision rules to a clinical trial of 28 people that examines the impact of high dose Vitamin C on patients with sepsis (Zabet et al., 2016). In terms of an instrumental variable model, this example is a reduced form rather than a first stage, so we introduce additional terminology. The potential outcomes represent "alive" (which maps to treated) and "dead" (which maps to untreated). The four principal strata represent those for whom the intervention is "wasteful" (always takers, who are alive in intervention and alive in control), those for whom the intervention is "efficacious" (compliers, who are alive in intervention and dead in control), those for whom the intervention is "unsafe" (defiers, who are dead in intervention and alive in control), and those for whom the intervention is "futile" (never takers, who are dead in intervention and dead in control).

Figure 4 illustrates the experiment graphically. In intervention, 12 of 14 people

are alive within 28 days, as compared with only 5 of 14 people in control. The Fisher exact test rejects the null hypothesis that the intervention was neither efficacious nor unsafe for anyone at the 2.6% level. The point estimate of the average effect is 0.5, and the p-value of the t-test is 0.004. While we can depict all possible distributions of potential outcomes in an experiment with 2 people in the ten rows of Figure 2, doing so in an experiment with 28 people would require 4,495 rows, so we only focus on three here.

Figure 4: Illustration of the Zabet et al. (2016) Vitamin C Experiment



As depicted in the highlighted row, our maximum likelihood decision rule shows that the joint distribution of potential outcomes that maximizes the likelihood includes 21 people for whom the intervention was efficacious (compliers), 7 people for whom the intervention was unsafe (defiers), and zero people for whom the intervention was wasteful or futile. The number of compliers net of defiers is 21-7=14, which matches the point estimate of the average effect multiplied by the sample size.

As shown in the second column, in our maximum likelihood configuration, we can deduce how many people of each type were assigned intervention and control. Since the maximizer rules out that any of the 12 people who lived in intervention would have lived regardless, it must have been efficacious for all 12 of them. By similar logic, our decision rule indicates that, by a chance imbalance, more of the people for whom the intervention was efficacious were assigned intervention (12 vs. 9), and fewer of the people for whom the intervention was unsafe were assigned control (2 vs. 5). Using terminology from Pearl (1999), the intervention was "necessary" for the deaths of the

2 people who died in intervention because it was unsafe for them, and they would have lived without it. Similarly, the intervention would have been "sufficient" for the deaths of the 5 people who lived in intervention because it was unsafe for them, so they would have died with it.

The second row of Figure 4 shows the result of the monotonicity decision rule: 10 people for whom the intervention would be wasteful, 14 people for whom the intervention would be efficacious, and 4 people for whom the intervention would be futile. Like the maximum likelihood decision rule, this decision rule matches the point estimate of the average effect. But, the likelihood of this distribution is 12.9%, which is strictly lower than the unconstrained maximum likelihood of 15.3%. The evidence in favor of the maximum likelihood decision, which indicates the presence of both compliers and defiers, is 1.19 stronger than the evidence for the monotonicity decision, as indicated by the likelihood ratio.

The final row shows the Fisher hypothesis distribution in which the intervention has no effect for anyone in the sample. This distribution has a much lower likelihood of 0.8%, and the visualization in the second column provides intuition for this lower value. This configuration implies large imbalances between intervention and control for those for whom the intervention would be wasteful (12 vs. 5) and those for whom the intervention would be futile (2 vs. 9).

While the maximum likelihood and monotonicity decision rules imply different marginal distributions of potential outcomes, our design-based likelihood can also distinguish between distributions with the same marginal distributions.Consider an alternative distribution, not depicted, with 6 people for whom the intervention is wasteful, 18 people for whom the intervention is efficacious, and 4 people for whom the intervention is unsafe. That distribution has the same marginal distributions as the result of the monotonicity decision rule depicted in the second row ($\theta_{11} + \theta_{10} = 24$ and $\theta_{11} + \theta_{01} = 10$), but it has a higher likelihood of 14.5%.

The expressions in the final column help explain why the first row maximizes the likelihood and others do not. In an experiment with 28 people, there are 40,116,600 ways to randomize 14 into intervention and 14 into control (28 choose 14). If the true distribution is the one depicted in the first row, there are 293,930 ways to randomize 12 of the 21 people for whom the intervention is efficacious (compliers) into intervention (21 choose 12) and 21 ways to randomize 2 of the 7 people for whom the intervention is unsafe (defiers) into intervention (7 choose 2). There are then exactly 6,172,530 ways (21 choose 7 times 7 choose 2) for the true distribution to yield the observed data, representing approximately 15.3% of all possible ways. In other rows, there are fewer ways to yield the observed data, so the entropy—-the numerator of the likelihood—-is lower.

It need not be the case that the maximizer of the likelihood includes only two types. As Fisher (1935) recognized, it is always possible for all people in the experiment to be of two types—either compliers and defiers (as shown in the highlighted row) or always takers and never takers (as shown in the last row). However, maxi-

mization of the likelihood requires trading off between the higher likelihood of fewer types and the higher likelihood of balance between intervention and control within each type. In this example, the maximizer of the likelihood has only two types but some imbalance within each type. In the case of the third row, the imbalance within each type is much larger, and the likelihood is much lower. The second row, chosen by the monotonicity rule, is balanced within each type, but it has three types, which decreases its likelihood.[7]

The maximum likelihood decision rule implies that the intervention is unsafe for some patients, and decision makers will need to decide what to do with that information. Those that ascribe to the "do not harm" principle might withhold the intervention (Ben-Michael et al., 2024), and those that only care about the average effect might want to continue using it. However, even those that care only about the average could improve the average if they could better target the intervention away from those for whom it would be unsafe and towards those for whom it would be efficacious. If machine learning methods cannot find heterogeneity in the intervention effect using covariates available by convenience, our maximum likelihood decision can justify collecting a richer set of covariates. It can also motivate the collection of data on a richer set of secondary outcomes. For example, in the Bernard et al. (2001) clinical trial testing the effect of recombinant human activated protein C on patients with sepsis, researchers identified a potential mechanism for how the intervention was unsafe: it led to severe bleeding.

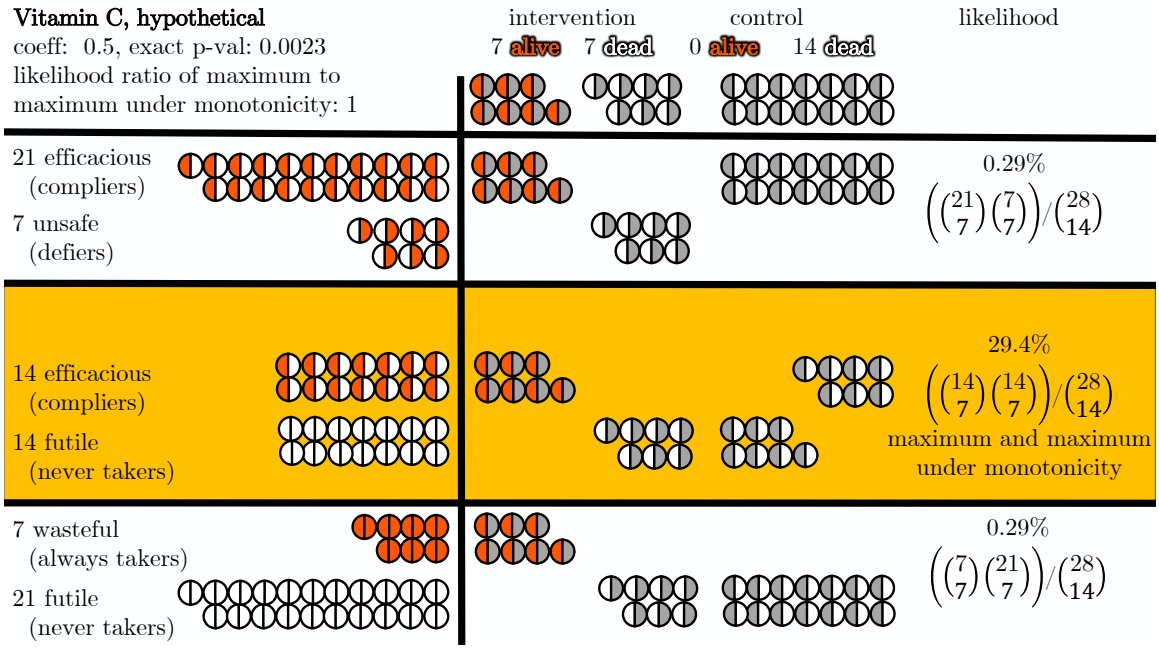
5.1 The Clinical Trial Need Not Reveal Evidence Against Monotonicity

The Zabet et al. (2016) experiment reveals evidence against monotonicity, but the same trial could have supported monotonicity, even with the same estimated average effect, if fewer people had survived in both arms.. Figure 5 shows a hypothetical, alternative outcome of the trial where 7 of the 14 people in the intervention arm are alive and zero of the 14 people in the control arm are alive. As in the actual experiment, the point estimate of the average effect of the intervention on survival is 0.5. But now, the maximum likelihood and monotonicity decision rules both estimate a distribution of potential outcomes consisting of 14 people for whom the intervention would be efficacious (compliers) and 14 people for whom the intervention would be futile (never takers), as shown in Figure 5. The likelihood ratio, then, is exactly 1, and the hypothetical experiment offers no evidence against monotonicity.[8]

[7] Viewing Pascal's triangle as a representation of N choose K can help us to visualize the tradeoff between fewer types (moving to the next row of the triangle, increasing N within each type) and more balance within each type (moving to the middle of the triangle within a row, moving K closer to $N/2$ within each type). We thank Elizabeth Ananat for this point.

[8] As shown in the later columns, the hypothetical experiment does offer evidence against the hypotheses that everyone is affected (the first row) and that no one is affected (the third row). Both require much more unbalanced randomization within each type, and the likelihoods are much smaller.

Figure 5: Illustration of a Hypothetical Vitamin C Experiment



6 Implications

We often take for granted that the average effect estimated from an experiment is sufficient for making a decision. However, considering just this estimate throws away valuable information. We harness information about the randomization process within an experiment to decide whether there is evidence against monotonicity.

Randomized experiments are designed to offer the most credible evidence on causal effects, so their analysis warrants statistical methods tailor-made to harness their design. Athey and Imbens (2017) address this need head on: "we recommend using statistical methods that are directly justified by randomization, in contrast to the more traditional sampling-based approach that is commonly used in econometrics." They quote Freedman (2006), who asserts that "experiments should be analyzed as experiments, not as observational studies." The asymptotic methods used for observational studies were developed, at least in part, due to their analytical convenience—finite sample statistics were sometimes just too hard to compute. In the era of modern computing, large sample approximations may be less useful. The exact design-based model closely follows the actual structure of randomization that produced the data, and as we show, it can produce novel insights over the canonical sampling-based model.

Design-based methods have been around for a long time, and the most practical impediment to their current use is that researchers do not report enough information about their experimental designs (Young, 2019; Bai, 2022). We use large language

models to pull (openAI's GPT-4o-mini) and categorize (openAI's GPT-o3-mini) the randomization processes from 2,080 papers associated with randomized controlled trials from the Abdul Latif Jameel Poverty Action Lab (J-PAL). Only 61% have enough description for the large language model to confidently categorize the randomization method from the paper or associated AEA RCT Registry entry. However, of those, 78% use designs captured by design-based likelihoods that we consider. In the full set of papers, over 60% describe some mechanism through which monotonicity could be violated. Mandated reporting of experimental design by J-PAL and other authorities could facilitate the application of our design-based decision rule, allowing researchers to learn more from their experiments.

Appendix

Appendix A Likelihood Derivations for the Canonical Sampling-Based Model with Replacement

Here, we provide additional detail for the derivation of the canonical sampling-based likelihoods in Section 2.1. When randomization into intervention or control is conducted as a series of Bernoulli trials, each participant is assigned to intervention independently with a fixed probability p. We can then view the sample as n independent trials with a categorical outcome, defined by the individual's realizations of Y and Z. The data $\boldsymbol{X}$ are the counts of these categorical realizations. The probability that an individual has outcome $Y = 1$ in the intervention group can be expressed in terms of the probability of randomization into the intervention p and the probability of being either an always taker q_{11} or a complier q_{10}:

$$\begin{aligned}\mathbb{P}\big(Y = 1, Z = I\big) &= \mathbb{P}\big(Y_I = 1, Z = I\big) \\ &= \mathbb{P}\big(Y_I = 1\big)\mathbb{P}\big(Z = I\big) && \text{(by independence)} \\ &= (q_{11} + q_{10})p.\end{aligned}$$

The probabilities of the remaining outcomes can be similarly determined:

$$\begin{aligned}\mathbb{P}\big(Y = 0, Z = I\big) &= \big(1 - (q_{11} + q_{10})\big)p \\ \mathbb{P}\big(Y = 1, Z = C\big) &= (q_{11} + q_{01})(1 - p) \\ \mathbb{P}\big(Y = 0, Z = C\big) &= \big(1 - (q_{11} + q_{01})\big)(1 - p).\end{aligned}$$

The probability distribution for categorical counts from repeated, independent Bernoulli trials is the multinomial distribution. Using the above defined probabilities of each outcome, we can write the likelihood for $\boldsymbol{q}$ given $\boldsymbol{x}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) &= \mathbb{P}\big(\boldsymbol{x} \mid \boldsymbol{q}\big) \\ &= \frac{n!}{x_{I1}!x_{I0}!x_{C1}!x_{C0}!} p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\ &\quad \big(q_{11} + q_{10}\big)^{x_{I1}} \big(1 - (q_{11} + q_{10})\big)^{x_{I0}} \\ &\quad \big(q_{11} + q_{01}\big)^{x_{C1}} \big(1 - (q_{11} + q_{01})\big)^{x_{C0}}.\end{aligned}$$

Alternatively, in a completely randomized experiment, a fixed number of participants $m \leq n$ are randomized into the intervention group, and the remaining $n - m$ participants are randomized into the control group (often, $m = n/2$). We can then view the sample as two sets of Bernoulli trials: the first set of m trials represent the individuals sampled into the intervention group, of which there are X_{I1} "successes"; the second set of $n-m$ trials represent the individuals sampled into the control group, of which there are X_{C1} successes. The probability of success in the intervention group

trials is the probability that $Y(1) = 1$, which is $q_{11}+q_{10}$, and the probability of success in the control group trials is the probability that $Y(0) = 1$, which is $q_{11}+q_{10}$. The two sets of trials are independent of each other, allowing us to write the likelihood of a completely randomized experiment as the product of two Binomial random variables, X_{I1} and X_{C1}:

$$\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \binom{m}{x_{I1}} \big(q_{11} + q_{10}\big)^{x_{I1}} \big(1 - (q_{11} + q_{10})\big)^{m - x_{I1}} \binom{n-m}{x_{C1}} \big(q_{11} + q_{01}\big)^{x_{C1}} \big(1 - (q_{11} + q_{01})\big)^{n - m - x_{C1}}.$$

Appendix B Simplification of the Combined Sampling- and Design-Based Likelihoods with I.I.D. Sampling

Under sampling with replacement (I.I.D.), the distribution of $\boldsymbol{\theta}$ takes the form in (9). Here, we show that, under the randomization design of Bernoulli trials, the general combined sampling- and design-based likelihood in (5) simplifies to to the first canonical likelihood from Section 2.1 in (1), and under the design of a completely randomized experiment, it simplifies to the second canonical likelihood in (2).

First, we focus on the design of Bernoulli trials. Substituting (6) and (9) into the general likelihood in (5) yields the following expression:

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Bigg[\sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \\ & \times p^{x_{I1} + x_{I0}} (1-p)^{x_{C1} + x_{C0}} \Bigg] \times \frac{n!}{\theta_{11}! \theta_{10}! \theta_{01}! \theta_{00}!} q_{11}^{\theta_{11}} q_{10}^{\theta_{10}} q_{01}^{\theta_{01}} q_{00}^{\theta_{00}}.\end{aligned} \tag{12}$$

First, we expand the summations:

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \sum_{\theta_{11}=-\infty}^{\infty} \sum_{\theta_{10}=-\infty}^{\infty} \sum_{\theta_{01}=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} & \binom{\theta_{11}}{i} \\ & \times \binom{\theta_{10}}{x_{I1} - i} \\ & \times \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \\ & \times \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \\ & \times p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\ & \times \frac{n!}{\theta_{11}!\theta_{10}!\theta_{01}!\theta_{00}!} q_{11}^{\theta_{11}} q_{10}^{\theta_{10}} q_{01}^{\theta_{01}} q_{00}^{\theta_{00}}.\end{aligned}$$

Note that $\theta_{00} = n - \theta_{11} - \theta_{10} - \theta_{01}$, and the binomial coefficient n-choose-k is zero whenever $n < 0$, $k < 0$, or $n > k$.

Next, we perform a change of variables, eliminating the θ terms and introducing alternative summation indices a, b, and c:

$$a = \theta_{11} - i \qquad b = \theta_{10} + i - x_{I1} \qquad c = \theta_{01} + a - x_{C1},$$

or, equivalently,

$$\begin{aligned}\theta_{11} &= a + i & \theta_{10} &= b + x_{I1} - i \\ \theta_{01} &= c + x_{C1} - a & \theta_{00} &= x_{I0} + x_{C0} - b - c.\end{aligned}$$

Here, we have again used the fact that $\theta_{00} = n - \theta_{11} - \theta_{10} - \theta_{01}$, along with the fact that $n = x_{I1} + x_{I0} + x_{C1} + x_{C0}$. Under this change of variables, the likelihood can be

expressed as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \sum_{a=-\infty}^{\infty} \sum_{b=-\infty}^{\infty} \sum_{c=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} & \binom{a+i}{i} \\ & \times \binom{b + x_{I1} - i}{x_{I1} - i} \\ & \times \binom{c + x_{C1} - a}{c} \\ & \times \binom{x_{I0} + x_{C0} - b - c}{x_{I0} - c} \\ & \times p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\ & \times \frac{n!}{(a+i)!(b+x_{I1}-i)!(c+x_{C1}-a)!(x_{I0}+x_{C0}-b-c)!} \\ & \times q_{11}^{a+i} q_{10}^{b+x_{I1}-i} q_{01}^{c+x_{C1}-a} q_{00}^{x_{I0}+x_{C0}-b-c}.\end{aligned}$$

Expanding the binomial coefficients and simplifying yields:

$$\begin{aligned}\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = n! & p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \\ & \times \left(\sum_{i=-\infty}^{\infty} \frac{q_{11}^{i} q_{10}^{x_{I1}-i}}{i!(x_{I1}-i)!} \right) \left(\sum_{a=-\infty}^{\infty} \frac{q_{11}^{a} q_{01}^{x_{C1}-a}}{a!(x_{C1}-a)!} \right) \\ & \times \left(\sum_{b=-\infty}^{\infty} \frac{q_{10}^{b} q_{00}^{x_{C0}-b}}{b!(x_{C0}-b)!} \right) \left(\sum_{c=-\infty}^{\infty} \frac{q_{01}^{c} q_{00}^{x_{I0}-c}}{c!(x_{I0}-c)!} \right)\end{aligned}$$

Finally, applying the binomial theorem to each sum and recalling $q_{11}+q_{10}+q_{01}+q_{00} = 1$ yields the exact expression for the first canonical sampling-based likelihood in (1).

Next, we turn to the design of a completely randomized experiment and sampling with replacement. We can express the likelihood by substituting (7) and (9) into the

general likelihood expression from (5):

$$\mathcal{L}(\boldsymbol{q} \mid \boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Bigg[\sum_{i \in \mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})} \binom{\theta_{11}}{i} \binom{\theta_{10}}{x_{I1} - i} \binom{\theta_{01}}{\theta_{11} + \theta_{01} - x_{C1} - i} \binom{\theta_{00}}{x_{I0} + x_{C1} + i - \theta_{11} - \theta_{01}} \Big/ \binom{n}{m} \Bigg] \times \frac{n!}{\theta_{11}! \theta_{10}! \theta_{01}! \theta_{00}!} q_{11}^{\theta_{11}} q_{10}^{\theta_{10}} q_{01}^{\theta_{01}} q_{00}^{\theta_{00}}.$$

This expression is identical to the version for Bernoulli trials in (12) divided by $p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}} \big/ \binom{n}{m}$. Therefore, by the previous manipulations, it is also equal to (1) divided by the same, which simplifies to the second canonical sampling-based likelihood in (2).

Appendix C Bayes Optimality of the Maximum Likelihood Decision Rule

In this appendix, we review the result that the maximum *a posteriori* decision rule is Bayes optimal for our chosen loss function, which in turn implies that the maximum likelihood decision rule is optimal under a uniform prior. First, let us define the maximum *a posteriori* decision rule f_π^*, which selects the maxima of the posterior distribution of $\boldsymbol{\theta}$. Formally, let $\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})$ be the set of $\boldsymbol{\theta}$ values that maximize the posterior distribution given the observed data $\boldsymbol{x}$, i.e.

$$\begin{aligned} \widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x}) &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}\big(\boldsymbol{\theta} \mid \boldsymbol{x}\big) \\ &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \pi(\boldsymbol{\theta}), \end{aligned} \tag{13}$$

where $\pi \in \Delta(\boldsymbol{\Theta})$ is the prior belief about $\boldsymbol{\theta}$. The decision rule f_π^* can then be defined as follows:

$$f_\pi^*(\boldsymbol{x})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}} & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x}), \\ 0 & \text{o.w.} \end{cases} \tag{14}$$

Now, let g be an arbitrary decision function. The Bayes expected utility for

decision function g is

$$\begin{aligned}\mathbb{E}\big[EU(g,\boldsymbol{\theta})\big] &= \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} EU(g,\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} \mathcal{L}(\boldsymbol{\theta}\mid\boldsymbol{x})g(\boldsymbol{x})(\boldsymbol{\theta})\right]\pi(\boldsymbol{\theta})\end{aligned}$$

By rearranging terms in the summation, we can bound the Bayes expected utility of g:

$$\begin{aligned}\mathbb{E}\big[EU(g,\boldsymbol{\theta})\big] &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\mathcal{L}(\boldsymbol{\theta}\mid\boldsymbol{x})g(\boldsymbol{x})(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\Big) \\ &\leq \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(g(\boldsymbol{x})(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\Big\{\mathcal{L}(\boldsymbol{\theta}'\mid\boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\}\Big) \\ &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\left[\max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\Big\{\mathcal{L}(\boldsymbol{\theta}'\mid\boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\} \underbrace{\Big(\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{x})(\boldsymbol{\theta})\Big)}_{=1}\right].\end{aligned}$$

This bound is precisely the Bayes expected utility achieved by decision rule f^*_π:

$$\begin{aligned}\mathbb{E}\big[EU(f^*_\pi,\boldsymbol{\theta})\big] &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\Theta} \Big(\mathcal{L}(\boldsymbol{\theta}\mid\boldsymbol{x})f^*_\pi(\boldsymbol{x})(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\Big) \\ &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\sum_{\boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})} \left(\frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}} \max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\Big\{\mathcal{L}(\boldsymbol{\theta}'\mid\boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\}\right) \\ &= \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}\left[\max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\Big\{\mathcal{L}(\boldsymbol{\theta}'\mid\boldsymbol{x})\pi(\boldsymbol{\theta}')\Big\} \underbrace{\Big(\sum_{\boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}}\Big)}_{=1}\right]\end{aligned}$$

Thus, since f^*_π achieves the upper bound on the Bayes expected utility of any decision rule, we conclude that f^*_π is Bayes optimal.

Finally, observe that when the prior distribution $\pi(\theta)$ is constant, the maximizers of the posterior distribution in (13) are simply the maximizers of the likelihood, and the Bayes rule f^*_π in (14) simplifies to f^*.

Given the Bayesian justification offered here, a natural question is what sorts of priors for $\boldsymbol{\theta}$ are sensible? In cases where we think of the sample as being drawn from some superpopulation, the method of sampling can inform the researcher's prior. For example, if it is known that the sample is drawn from a finite superpopulation without replacement, then a reasonable belief distribution may be the distribution for $\boldsymbol{\theta}$ that results from sampling with replacement, i.e. the distribution in (10). In

this case, the superpopulation distribution of potential outcomes serves as a Bayesian hyperparameter indexing priors. The maximum aposterori decision rule that utilizes the sampling-informed prior is optimal in these cases. Other times, the experimenter may not know how the sample was drawn, or there may be no sense in envisioning a sampling process at all. In these cases, the experimenter may be best off choosing an uninformative prior that is uniform across realizations of $\boldsymbol{\theta}$. Indeed it is this case which justifies the maximum likelihood decision rule as Bayes optimal.

Appendix D The Principle of Maximum Entropy in the Monty Hall Problem

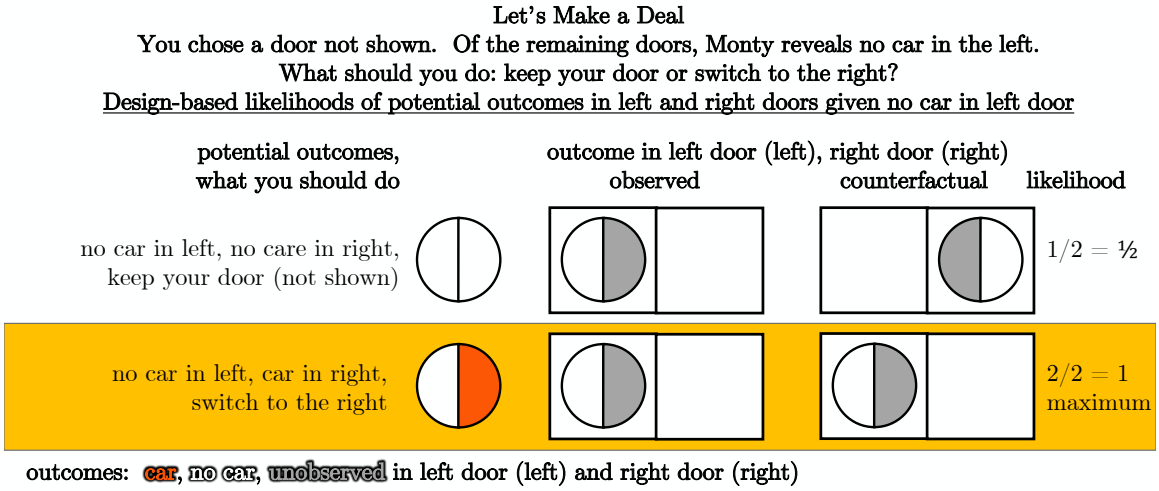
The Monty Hall Problem is a poster child for Bayesian decision making. We illustrate here that the principle of maximum entropy can be used to arrive at the same decision without specifying a Bayesian prior. Wang et al. (2016) discuss the principle of maximum entropy in the same context, but we illustrate it here with our proposed visualization of potential outcomes.

The setup of the Monty Hall problem is as follows. Monty Hall is the host of a game show, and you are a contestant. Monty shows you three doors. He tells you that there is a car behind one of the doors, which you will win if you choose the correct door. You will not lose anything if you choose incorrectly. Monty asks you to guess a door. Of the remaining two doors, Monty opens the one on the left and shows you that it does not contain the car. Assume that he would have only opened a remaining door that did not contain the car. He gives you the option to keep your door or switch to the right remaining door. What should you do?

Figure D.1 illustrates the Monty Hall problem using the approach to visualize potential outcomes that we develop to illustrate the design-based likelihood in Figure 1. In that figure, the binary outcome can take on values of "treated" or "untreated." Here, it can take on values of "car" or "no car." In Figure 1, the binary intervention can take on values of intervention or control. Here, it can take on values of left or right. The observed outcome is no car in left. The potential outcome is unobserved in right. The two rows depict the two possible potential outcomes in left and right given that you observed no car in left. In the first row, there is no car in left or right. Therefore, the car is behind your door, and you should keep your door. In the second row, there is no car in the left remaining door, but the car is behind the right remaining door, so you should switch doors.

To decide using our proposed maximum likelihood rule and the principle of maximum entropy, you need to know the likelihood in each row, depicted in the last column. In the first row, because the car is behind your door, Monty can either open the left or right remaining door. Suppose that Monty has a randomizer that reveals left or right half the time. Given the two possible outcomes of the randomizer, there is only one way that you could have seen him open the left door, so the entropy is 1 and the likelihood is 1/2. In the second row, because the door is behind the right door, Monty can only open the left door. Given the two possible outcomes of the

Figure D.1: An Illustration of the Monty Hall Problem



randomizer, there are two ways that you could have seen him open the left door (he would have opened it either way), so the entropy is 2 and the likelihood is 1. By our proposed maximum likelihood decision rule, the maximum of the likelihood indicates that the car is behind the right remaining door, so you should switch doors. By the principle of maximum entropy, you arrive at the same decision that you would have made under a Bayesian decision rule with a uniform prior.

A further connection between the likelihood in the Monty Hall problem and the designed-based likelihood in the experiment with two people deserves mention. The maximum of the likelihood in both contexts is 1, and the likelihood ratio of the maximum of the likelihood to the other possible value(s) of the likelihood is 2. This likelihood ratio indicates that the strength of the evidence for your decision in an experiment with two people is as strong as the strength of the evidence in the standard Monty Hall problem.

References

Alsan, M., J. Cawley, J. Doyle, Joseph J, and N. Skelley (2025, January). Mean reversion in randomized controlled trials: Implications for program targeting and heterogeneous treatment effects. Working Paper 33369, National Bureau of Economic Research.

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.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review 112*(12), 3911–3940.

Bai, Y., S. Huang, S. Moon, A. M. Shaikh, and E. J. Vytlacil (2024). On the identifying power of monotonicity for average treatment effects. *arXiv preprint arXiv:2405.14104*.

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.

Barnard, G. A. (1947). Significance Tests for 2 x 2 Tables. *Biometrika 34*(1-2), 123–138.

Ben-Michael, E., K. Imai, and Z. Jiang (2024). Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 1–14.

Bernard, G. R., J.-L. Vincent, P.-F. Laterre, S. P. LaRosa, J.-F. Dhainaut, A. Lopez-Rodriguez, J. S. Steingrub, G. E. Garber, J. D. Helterbrand, E. W. Ely, and C. J. Fisher (2001). Efficacy and safety of recombinant human activated protein c for severe sepsis. *New England Journal of Medicine 344*(10), 699–709. PMID: 11236773.

Björklund, A. and R. Moffitt (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.

Boole, G. (1854). Of statistical conditions. In *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*, Chapter 19, pp. 295–319. Walton and Maberly.

Canner, P. L. (1970). Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association 65*(329), 293–306.

Christy, N. and A. E. Kowalski (2024a). Counting defiers in health care with a design-based likelihood for the joint distribution of potential outcomes.

Christy, N. and A. E. Kowalski (2024b). Starting small: Prioritizing safety over efficacy in randomized experiments using the exact finite sample likelihood.

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, NY: Wiley.

Dawid, A. P. and M. Musio (2022). Effects of causes and causes of effects. *Annual Review of Statistics and Its Application 9*(1), 261–287.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics 125*(1-2), 141–173.

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

Fan, Y. and S. S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory 26*(3), 931–951.

Ferguson, T. S. (1967). *Mathematical statistics a decision theoretic approach by Thomas S. Ferguson*. Academic Press.

Fernández, A. A., J. L. M. Olea, C. Qiu, J. Stoye, and S. Tinda (2024). Robust bayes treatment choice with partial identification. *arXiv preprint arXiv:2408.11621*.

Ferrie, C. (2017, December). *Statistical physics for babies*. Baby university. Naperville, IL: Sourcebooks.

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

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

Fréchet, M. (1957). Les tableaux de corrélation et les programmes linéaires. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute 25*(1/3), 23–40.

Freedman, D. (2006). Statistical models for causation: what inferential leverage do they provide? *Eval Rev. 30*(6), 691–713.

Freedman, D. A. and R. A. Purves (1969). Bayes' method for bookies. *The Annals of Mathematical Statistics 40*(4), 1177–1186.

Gelman, A. and G. Imbens (2013). Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.

Gelman, A. and K. O'Rourke (2017). Attitudes toward amalgamating evidence in statistics.

Greenland, S. and J. M. Robins (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology 15*(3), 413–419.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), 487–535.

Heckman, J. J. and E. J. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Hirano, K. (2008). *Decision Theory in Econometrics* (Second ed.)., pp. 1–6. London: Palgrave Macmillan UK.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2020). Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of econometrics*, Volume 7, pp. 283–354. Elsevier.

Hoeffding, W. (1940). Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin 5*(3), 181–233. Translated by Dana Quade in *The Collected Works of Wassily Hoeffding*, ed. Fisher, N. I. and Sen, P. K., pp. 57–107, New York, NY: Springer New York, 1994.

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

Huber, M. and G. Mellace (2013). Relaxing monotonicity in the identification of local average treatment effects. Working paper. https://www.alexandria.unisg.ch/206868/1/LATE%20without%20monotonicity.pdf.

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. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature 58*(4), 1129–1179.

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.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review 106*(4), 620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Physical review 108*(2), 171.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on systems science and cybernetics 4*(3), 227–241.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley.

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

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kline, P. and C. Walters (2020). Reasonable doubt: Experimental detection of job-level employment discrimination. *NBER Working Paper 26861*. https://eml.berkeley.edu/~pkline/papers/Kline_Walters_RD_2020.pdf.

Kowalski, A. E. (2019a). Counting defiers. *NBER Working Paper 25671*. https://www.nber.org/papers/w25671.

Kowalski, A. E. (2019b). A model of a randomized experiment with an application to the PROWESS clinical trial. *NBER Working Paper 25670*. https://www.nber.org/papers/w25670.

Kowalski, A. E. (2023a). Behaviour within a clinical trial and implications for mammography guidelines. *The Review of Economic Studies 90*(1), 432–462.

Kowalski, A. E. (2023b). Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. *The Review of Economics and Statistics 105*(3), 646–664.

Kuhn, H. W. (1953). *11. Extensive Games and the Problem of Information*, pp. 193–216. Princeton: Princeton University Press.

Li, A. and J. Pearl (2019). Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

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. (1997a). The mixing problem in programme evaluation. *The Review of Economic Studies 64*(4), 537–553.

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

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Manski, C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics 139*(1), 105–115.

Manski, C. F. (2018). Reasonable patient care under uncertainty. *Health Economics 27*(10), 1397–1421.

Manski, C. F. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician 73*(sup1), 296–304.

Manski, C. F. and A. Tetenov (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference 137*(6), 1998–2010.

Manski, C. F. and A. Tetenov (2021). Statistical decision properties of imprecise trials assessing coronavirus disease 2019 (covid-19) drugs. *Value in Health 24*(5), 641–647.

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

Mullahy, J. (2018). Individual results may vary: Inequality-probability bounds for some health-outcome treatment effects. *Journal of Health Economics 61*, 151 – 162.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczych 10*, 1–51. Translated by D.M. Dabrowski and T.P. Speed in *Statistical Science 5*(4), pp. 465–472, 1990.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese 121*(1/2), 93–149.

Pearl, J. and D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

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.

Schlag, K. H. (2007). Eleven - designing randomized experiments under minimax regret. *Unpublished manuscript, European University Institute, Florence*.

Semenova, V. (2024). Aggregated intersection bounds and aggregated minimax values.

Stoye, J. (2007). Minimax regret treatment choice with incomplete data and many treatments. *Econometric Theory 23*(1), 190–199.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics 151*(1), 70–81.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics 166*(1), 138–156.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Tian, J. and J. Pearl (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence 28*(1-4), 287–313.

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.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics 20*(2), 165 – 205.

Wang, J. L., T. Tran, and F. Abebe (2016). Maximum entropy and bayesian inference for the monty hall problem. *Journal of Applied Mathematics and Physics 4*(7), 1222–1230.

Warren, H. S., A. F. Suffredini, P. Q. Eichacker, and R. S. Munford (2002). Risks and benefits of activated protein c treatment for severe sepsis. *The New England journal of medicine 347*(13), 1027–1030.

Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika 29*(1/2), 21–52.

Young, A. (2019). Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–598.

Zabet, M. H., M. Mohammadi, M. Ramezani, and H. Khalili (2016). Effect of high-dose ascorbic acid on vasopressor's requirement in septic shock. *Journal of Research in Pharmacy Practice 5*(2), 94–100.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.