arXiv:2412.16352v1 [econ.EM] 20 Dec 2024

Counting Defiers in Health Care with a Design-Based Likelihood for the Joint Distribution of Potential Outcomes*

Neil Christy and Amanda Ellen Kowalski

December 18, 2024

Abstract

We present a design-based model of a randomized experiment in which the observed outcomes are informative about the joint distribution of potential outcomes within the experimental sample. We derive a likelihood function that maintains curvature with respect to the joint distribution of potential outcomes, even when holding the marginal distributions of potential outcomes constant—curvature that is not maintained in a sampling-based likelihood that imposes a large sample assumption. Our proposed decision rule guesses the joint distribution of potential outcomes in the sample as the distribution that maximizes the likelihood. We show that this decision rule is Bayes optimal under a uniform prior. Our optimal decision rule differs from and significantly outperforms a "monotonicity" decision rule that assumes no defiers or no compliers. In sample sizes ranging from 2 to 40, we show that the Bayes expected utility of the optimal rule increases relative to the monotonicity rule as the sample size increases. In two experiments in health care, we show that the joint distribution of potential outcomes that maximizes the likelihood need not include compliers even when the average outcome in the intervention group exceeds the average outcome in the control group, and that the maximizer of the likelihood may include both compliers and defiers, even when the average intervention effect is large and statistically significant.

*Comments very welcome. Previous versions of this paper have been circulated under different working paper numbers and different titles including "Counting Defiers," "A Model of a Randomized Experiment with an Application to the PROWESS Clinical Trial," and "General Finite Sample Inference for Experiments with Examples from Health Care," "Starting Small: Prioritizing Safety over Efficacy in Randomized Experiments Using the Exact Finite Sample Likelihood" (Kowalski, 2019a,b; Christy and Kowalski, 2024). 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 also 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, 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, the Yale Cowles Summer Structural Microeconomics Conference, and the Y-RISE Evidence Aggregation and External Validity Conference for helpful comments. Marian Ewell provided helpful practical information about randomization in clinical trials. 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, Bailey Flanigan, Pauline Mourot, Srajal Nayak, Sukanya Sravasti, and Matthew Tauzer provided excellent research assistance.

1 Introduction

Suppose you have a treatment to improve health care and a nudge to get people to take it. You design a randomized experiment with two people and run it. Now the experiment has ended. The person assigned the nudge intervention has taken the treatment and so has the person assigned control. Why? What would you have seen had the randomization gone differently? What is the joint distribution of potential outcomes in the sample? 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). To answer these questions, we develop a novel decision rule for estimating the joint distribution of potential outcomes within the sample.

In the sample of two people, there are four possible joint distributions of potential outcomes that could explain why you observed one person treated in intervention and another treated in control. Following Angrist et al. (1996), we classify people based on their potential outcomes in intervention and control as always takers, compliers, defiers, and never takers. One possibility is that both people are always takers who would have been treated regardless of their assignment. A second is that only the person assigned intervention was an always taker, and the person assigned control was a defier who was treated in control but would have been untreated intervention. A third is that only the person assigned control was an always taker, and the person assigned intervention was a complier who was treated in intervention but would have been untreated in control. The fourth possibility is that the person assigned intervention was a complier, and the person assigned control was a defier. How can you decide among these possibilities?

Our main innovation is to decide using a design-based model of a randomized experiment with a binary intervention and outcome. The design of the experiment yields a design-based likelihood for the joint distribution of potential outcomes within the sample. We derive the likelihood for an experiment conducted as a series of Bernoulli trials, and we also derive the Copas (1973) likelihood for a completely randomized experiment. The design-based likelihood, which takes potential outcomes as fixed and assignments as random, is different from a sampling-based likelihood that invokes a large sample assumption and takes assignments as fixed and outcomes as random.

The design-based likelihood preserves information about the joint distribution of potential outcomes beyond that contained in the marginal distributions of potential outcomes. A large literature has focused on specifying what we can learn in a sampling-based framework about the joint distribution of potential outcomes from estimates of their marginal distributions through the Boole (1854), Hoeffding (1940), and Fréchet (1957) bounds (see, for example, Balke and Pearl, 1997; Heckman et al., 1997; Manski, 1997a; Tian and Pearl, 2000; Zhang and Rubin, 2003; Fan and Park, 2010; Mullahy, 2018; Ding and Miratrix, 2019; and Tian and Pearl, 2000). We contribute to this literature by demonstrating that the data in our design-based setting can be directly informative about the joint distribution, obviating the need for copula bounds. We provide intuition for this result using simple, novel illustrations and an analogy to the concept of entropy from statistical physics.

We propose a decision rule in the style of Wald (1949) that estimates the joint distribution of potential outcomes in the sample as the maximizer of this likelihood. There are a number of benefits to the statistical decision theory framework in our setting. First, decision theory is easy to apply in our finite sample, 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). Our work contributes to the integration of statistical decision theory into econometrics (Manski, 2004; Dehejia, 2005; Manski, 2007; Hirano, 2008; Stoye, 2012; Kitagawa and Tetenov, 2018; Manski, 2018, 2019; Hirano and Porter, 2020; Manski and Tetenov, 2021), particularly within finite sample settings (Canner, 1970; Manski, 2007; Schlag, 2007; Stoye, 2007, 2009; Tetenov, 2012).

To justify the use of our maximum likelihood decision rule, we demonstrate that it is Bayes optimal under a uniform prior with the appropriate utility function. While one need not be Bayesian to construct our decision rule, we emphasize that Bayes optimality is a desirable property. 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).

For comparison, we also construct a design-based "monotonicity" decision rule inspired by the LATE monotonicity assumption of Imbens and Angrist (1994) and the monotone response assumption of Manski (1997b), commonly invoked in large sample frameworks. To allow for the best possible performance of a monotonicity assumption in our design-based framework, our monotonicity decision rule chooses the constrained maximizer of the likelihood among distributions that contain either no defiers or no compliers, and that match the point estimate of the average intervention effect. Our maximum likelihood decision rule imposes no such restrictions and allows for both compilers and defiers in the same sample.

Using exact computations of the value of the likelihood function over every possible realization of experimental data, we quantify the expected utility gains from our optimal decision rule. We compute the exact expected utility from each decision rule under a uniform prior for all even-numbered sample sizes from 2 to 40. Our maximum likelihood decision rule strictly outperforms the monotonicity decision rule for all sample sizes greater than four, and the Bayes expected utility of the maximum likelihood decision rule relative to the monotonicity decision rule increases with the sample size. In a sample of 40, our maximum likelihood decision rule delivers 1.31 times the Bayes expected utility of the monotonicity decision rule.

Finally, we demonstrate the application of the maximum likelihood decision rule to two real-world experiments in health care. First, we analyze the effect of a nudge intervention intended to increase the uptake of flu vaccination in the experiment of Lehmann et al. (2016). The authors estimate a small, positive effect on vaccination takeup, and the baseline monotonicity decision rule for the joint distribution of potential outcomes in their sample reinforces this conclusion. In contrast, using the maximum likelihood decision rule, we estimate that their sample contained zero defiers and zero compliers—that is, our decision rule suggests that the intervention had no effect in either direction, and that the small observed differences in the average outcomes between the intervention and control groups is due to chance in who was randomized into each group. This example shows that the maximum likelihood decision need not include compliers, even if the average outcome is higher in the intervention group than in the control group.

Second, we analyze the effect of high dose Vitamin C on survival among patients with sepsis in the experiment of Zabet et al. (2016). This small trial finds a large and statistically significant effect of the Vitamin C intervention on survival. Both the baseline monotonicity decision rule and our maximum likelihood decision rule similarly suggest a large effect through the difference in estimated numbers of compliers and defiers; but while the former estimates no defiers by construction, the latter estimates a positive number of both compliers and defiers in the sample. This example highlights that our design-based likelihood can be maximized by a distribution with both compliers and defiers.

The remainder of the paper proceeds as follows: Section 2 exposits the design-based model of a randomized experiment and its implied likelihood function. Section 3 proposes a maximum likelihood decision rule and demonstrates its Bayes optimality. Section 4 quantifies the performance

gains from the maximum likelihood decision rule, and Section 5 applies our decision rule to two randomized experiments in health care. Section 6 concludes.

2 A Design-Based Model of a Randomized Experiment

2.1 Model and Notation

Following the potential outcomes model of Neyman (1923), Rubin (1974, 1977), Holland (1986) and others, we ascribe each individual a binary potential outcome $y_I \in \{0,1\}$ in intervention and $y_C \in \{0,1\}$ in control. Individuals are randomly assigned to intervention ($Z = I$) or control ($Z = C$), and one of their potential treatments is revealed as the observed outcome Y:

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

where $\mathbf{1}_{\{.\}}$ is the indicator function.

An individual's realized 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. Throughout, we define Y=1 as "treated" and Y=0 as "untreated."

Under these assumptions, individuals fall into one of four "principle strata" defined by their combination of potential outcomes (Frangakis and Rubin, 2002). Following Imbens and Angrist (1994) and Angrist et al. (1996), we refer to these four groups as always takers ($y_I = 1$, $y_C = 1$), compliers ($y_I = 1$, $y_C = 0$), defiers ($y_I = 0$, $y_C = 1$), and never takers ($y_I = 0$, $y_C = 0$). Let θ_{y_I,y_C} represent the total number of individuals in the experiment with potential outcomes $(y_I, y_C) \in \{0,1\}^2$. The sum $\theta_{1,1} + \theta_{1,0} + \theta_{0,1} + \theta_{0,0} \equiv n$ is the sample size of the experiment. We represent these four integers compactly as $\boldsymbol{\theta} \equiv \big(\theta_{1,1}, \theta_{1,0}, \theta_{0,1}, \theta_{0,0}\big)$. The value $\boldsymbol{\theta}$ summarizes the joint distribution of potential outcomes within the sample. Following a "design-based" approach, we restrict our attention to the fixed, but unknown, distribution of potential outcomes within the given sample, rather than within some superpopulation.

Let X_{I1} represent the number of treated individuals in the intervention arm ($Z = I, Y = 1$), X_{I0} represent the number of untreated individuals in the intervention arm ($Z = I, Y = 0$), X_{C1} represent the number of treated individuals in the control arm ($Z = C, Y = 1$), and X_{C0} represent the number of untreated individuals in the control arm ($Z = C, Y = 0$). These values constitute the data observed from the so-called "first stage" of an experiment, and we represent the data compactly with $\boldsymbol{X} = \big(X_{I1}, X_{I0}, X_{C1}, X_{C0}\big)$.

2.2 Likelihood Derivation

Let $\boldsymbol{I} \equiv \big(I_{1,1}, I_{1,0}, I_{0,1}, I_{0,0}\big)$ be a random vector whose elements represent the numbers of always takers, compliers, defiers, and never takers randomized into intervention. In an experiment employing simple randomization, each person is assigned to intervention independently with a fixed probability p. Since the assignment of individuals to intervention or control is independent across groups as well as across individuals, we can write the distribution of $\boldsymbol{I}$ as the product of four independent Bernouolli distributions:

$$\mathbb{P}\Big(I_{1,1} = i_{1,1}, I_{1,0} = i_{1,0}, I_{0,1} = i_{0,1}, I_{0,0} = i_{0,0} \mid \boldsymbol{\theta}\Big) = \binom{\theta_{1,1}}{i_{1,1}}\binom{\theta_{1,0}}{i_{1,0}}\binom{\theta_{0,1}}{i_{0,1}}\binom{\theta_{0,0}}{i_{0,0}} \times p^{\sum_{j,k} i_{j,k}}(1-p)^{n-\sum_{j,k} i_{j,k}} \tag{1}$$

Alternatively, in a completely randomized experiment, the experimenter fixes the number of individuals in the intervention group m (often, $m = n/2$) 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 randomization scheme, $\boldsymbol{I}$ follows a multivariate hypergeometric distribution:

$$\mathbb{P}\Big(I_{1,1} = i_{1,1}, I_{1,0} = i_{1,0}, I_{0,1} = i_{0,1}, I_{0,0} = i_{0,0} \mid \boldsymbol{\theta}\Big) = \frac{\binom{\theta_{1,1}}{i_{1,1}}\binom{\theta_{1,0}}{i_{1,0}}\binom{\theta_{0,1}}{i_{0,1}}\binom{\theta_{0,0}}{i_{0,0}}}{\binom{n}{m}} \tag{2}$$

The observable data $\boldsymbol{X}$ can be expressed in terms of the latent $\boldsymbol{I}$ variables and the distribution of potential outcomes $\boldsymbol{\theta}$ by observing that each individual randomized into the intervention group with outcome $Y = 1$ must have either been an always taker or a complier: $X_{I1} = I_{1,1} + I_{1,0}$. Each individual randomized into the intervention group with outcome $Y = 0$ must have either been a never taker or a defier: $X_{I0} = I_{0,0} + I_{0,1}$. In the control group, those observed with outcome $Y = 1$ must be either one of the always takers that were not randomized into intervention, or one of the defiers that were not randomized into intervention: $X_{C1} = \theta_{1,1} - I_{1,1} + \theta_{0,1} - I_{0,1}$. And finally, anyone in the control group with outcome $Y = 0$ must be either one of the never takers that were not randomized into intervention, or one of the compliers that were not randomized into intervention, $X_{C0} = \theta_{0,0} - I_{0,0} + \theta_{1,0} - I_{1,0}$. Thus, we can write the probability of the observed data $\boldsymbol{X}$ conditional on the joint distribution of potential outcomes $\boldsymbol{\theta}$ as:

$$\begin{aligned}
\mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) &= \mathbb{P}\Big(I_{1,1} + I_{1,0} = x_{I1},\ \big(\theta_{1,1} - I_{1,1}\big) + \big(\theta_{0,1} - I_{0,1}\big) = x_{C1}, \\
&\qquad I_{0,0} + I_{0,1} = x_{I0},\ \big(\theta_{0,0} - I_{0,0}\big) + \big(\theta_{1,0} - I_{1,0}\big) = x_{C0} \mid \boldsymbol{\theta}\Big) \\
&= \mathbb{P}\Big(I_{1,1} + I_{1,0} = x_{I1},\ I_{1,1} + I_{0,1} = x_{C1} - \theta_{1,1} - \theta_{0,1}, \\
&\qquad I_{0,0} + I_{0,1} = x_{I0},\ I_{0,0} + I_{1,0} = x_{C0} - \theta_{0,0} - \theta_{1,0} \mid \boldsymbol{\theta}\Big)
\end{aligned}$$

A realization of $\boldsymbol{X}$ may be produced from multiple realizations of $\boldsymbol{I}$. Thus, 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 $I_{1,1}$ and solving the following system of equations for the elements of $\boldsymbol{I}$:

$$\begin{aligned}
I_{1,1} + I_{1,0} &= x_{I1}, \\
I_{1,1} + I_{0,1} &= \theta_{1,1} + \theta_{0,1} - x_{C1}, \\
I_{1,1} + I_{1,0} + I_{0,1} + I_{0,0} &= x_{I1} + x_{I0}, \\
I_{1,1} &= i
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
I_{1,1} &= i \\
I_{1,0} &= x_{I1} - i \\
I_{0,1} &= \theta_{1,1} + \theta_{0,1} - x_{C1} - i \\
I_{0,0} &= x_{C1} + i - \theta_{1,1} - \theta_{0,1},
\end{aligned}$$

The value i is restricted to the set $\mathcal{I}(\boldsymbol{x}, \boldsymbol{\theta})$ such that $\boldsymbol{I}$ remains within the support of $\boldsymbol{\theta}$, namely $0 \leq I_{1,1} \leq \theta_{1,1}$, $0 \leq I_{1,0} \leq \theta_{1,0}$, $0 \leq I_{0,1} \leq \theta_{0,1}$, and $0 \leq I_{0,0} \leq \theta_{0,0}$. The probability of a realization

of $\boldsymbol{X}$ is just the sum of the probability of each of these realizations of $\boldsymbol{I}$:

$$\begin{aligned}\mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\big) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} \mathbb{P}\Big(& I_{1,1} = i, \\ & I_{1,0} = x_{I1} - i, \\ & I_{0,1} = \theta_{1,1} + \theta_{0,1} - x_{C1} - i, \\ & I_{0,0} = x_{I0} + x_{C1} + i - \theta_{1,1} - \theta_{0,1} \mid \boldsymbol{\theta}\Big).\end{aligned}$$

Substituting either of the distributions for $\boldsymbol{I}$ yields a likelihood expression. Under simple randomization, $\boldsymbol{I}$ follows the distribution in (1), yielding the following likelihood expression:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{X}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{1,1}}{i} \\ & \times \binom{\theta_{1,0}}{x_{I1} - i} \\ & \times \binom{\theta_{0,1}}{\theta_{1,1} + \theta_{0,1} - x_{C1} - i} \\ & \times \binom{\theta_{0,0}}{x_{I0} + x_{C1} + i - \theta_{1,1} - \theta_{0,1}} \\ & \times p^{x_{I1}+x_{I0}}(1-p)^{x_{C1}+x_{C0}}\end{aligned} \tag{3}$$

Alternatively, in a completely randomized experiment, $\boldsymbol{I}$ follows the distribution in (2), yielding:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{X}) = \sum_{i \in \mathcal{I}(\boldsymbol{x},\boldsymbol{\theta})} & \binom{\theta_{1,1}}{i} \\ & \times \binom{\theta_{1,0}}{x_{I1} - i} \\ & \times \binom{\theta_{0,1}}{\theta_{1,1} + \theta_{0,1} - x_{C1} - i} \\ & \times \binom{\theta_{0,0}}{m + x_{C1} + i - \theta_{1,1} - \theta_{0,1} - x_{I1}} \Big/ \binom{n}{m}\end{aligned} \tag{4}$$

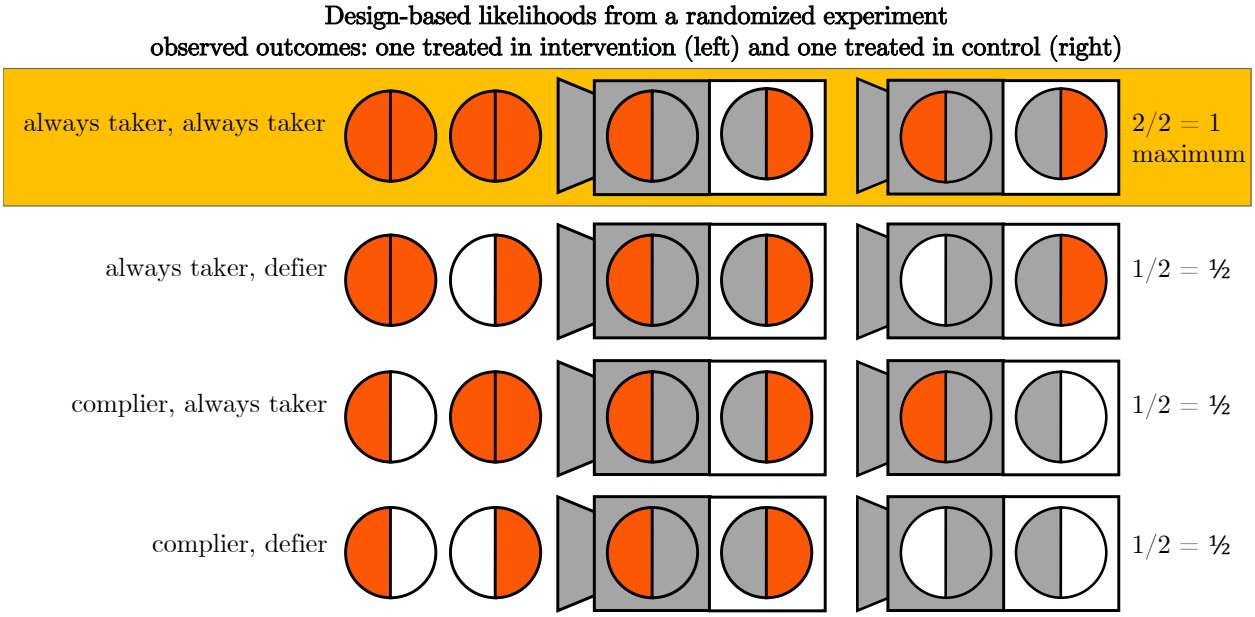
where we have substituted $m = x_{I1} + x_{I0}$. Copas (1973) derives a likelihood function equivalent to (4) to show that large sample tests of the average intervention effect are conservative. We depart from his work by focusing explicitly on the finite sample distribution of potential outcomes and applying insights from statistical decision theory, as detailed below. The likelihood function in (3) is, to the best of our knowledge, novel.

Note that both likelihood functions vary with the joint distribution of the potential outcomes, even when holding constant the marginal distributions of the potential outcomes. That is, when both $\theta_{1,1}+\theta_{1,0}$ and $\theta_{1,1}+\theta_{0,1}$ are held constant, the likelihood function maintains some curvature. We emphasize that sampling-based models typically refer to the distribution of potential outcomes in a superpopulation from which the sample was drawn, whereas we model the distribution of potential outcomes within a fixed sample, which preserves some information about their joint distribution.

2.3 An Illustration of the Design-Based Likelihood

Our running example of an experiment with two people provides a minimal working example to demonstrate curvature in the likelihood. Suppose one person is treated in intervention and another is treated in control. 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.

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



To illustrate, consider the rows of Figure 1, which 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 ball: the left half of the ball represents a person's potential outcome in intervention, and the right half of the ball represents a person's potential outcome in control (here, orange represents "treated," and white represents "untreated"). In each row, the two balls enter the experiment, represented by a pair of grey and white boxes, and one ball falls randomly into each box. The grey box represents the intervention group and masks the right half of a ball; the white box represents the control group and masks the left half of a ball. The first column of pairs of boxes represents what the observer would see if the first ball in the respective row were randomized into the intervention box and the second were randomized into the control box, while the second column shows the observable data under the alternative randomization outcome.

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 different 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. By the principle of maximum entropy (Jaynes, 1957a,b), the distribution with the greatest entropy (in our case, also the distribution that maximizes the likelihood) is the least informative distribution consistent with the observed data because the observed data could have been generated in the greatest number of ways.[1]

In an experiment with two people, there are four other outcomes that we could have observed, and the unique maximizer of the likelihood function for each indicates that both people have the same type. If the person in intervention is treated and the person in control is untreated, the likelihood is maximized when both people are compliers. If the person in control is treated and the person in intervention is untreated, the likelihood is maximized when both people are defiers. Finally, if the people in intervention and control are both untreated, the likelihood is maximized when both are never takers. In each case, people can be the same or different, but it is most likely that they are the same.[2]

In larger samples, it is not always possible for all the people in the experiment to be of the same type. But, as Fisher (1935) recognized, it is always possible for all the people in the experiment to be of two types—either compliers and defiers or always takers and never takers. However, it need not be the case that the maximizer of the likelihood function includes only two types. Indeed, ascribing each participant to one of two types sometimes implies that assignment to intervention or control within each type is highly imbalanced, while balance between intervention and control within a type is more likely: N choose M is maximized at $M = N/2$ (when N is even). Maximization of the likelihood requires trading off between the higher likelihood of fewer types and the higher likelihood of balance within each type. Just as people are more similar if they belong to fewer types, people of the same type are more similar if they are assigned intervention and control at the same rate.

3 Learning About the Joint Distribution of Potential Outcomes: Insights from Statistical Decision Theory

3.1 Bayes Optimality of the Maximum Likelihood Decision Rule

In the previous section, we presented a design-based model of a random experiment that preserves curvature in the likelihood with respect to the joint distribution of potential outcomes, even when holding constant the marginal distributions. We turn now to the broad setting of statistical decision theory in the style of Wald (1949) to determine the best ways to exploit this novel information. Suppose a decision maker wishes to guess the joint distribution of potential outcomes in the sample. We write the decision maker's guess as $\widehat{\boldsymbol{\theta}}$. The decision maker wishes to guess correctly, so we define a utility function over a guess $\widehat{\boldsymbol{\theta}}$ and the true distribution $\boldsymbol{\theta}$ that yields one util when the guess is

[1] Jaynes' work unites the theory of information with statistical physics. His principle of maximum entropy gives a way to make a decision without the need for a prior. In Bayesian decision-making, the subjective part is to choose a prior. To make it more objective, one option is to choose the least informative prior. However, even the least informative prior can still drive the result in small samples. Jaynes' alternative to make the process more objective is to choose the least informative updated distribution, the distribution that maximizes entropy. Statistical physics considers various functional forms for entropy, but the design of the experiment determines the functional form in our context.

[2] 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).

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 p as the expected utility of guessing according to the probabilities ascribed by p:

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

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.

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 $\boldsymbol{\theta}$. Two common approaches are to measure the decision rule's performance as either the minimum expected utility obtained across all possible values of $\boldsymbol{\theta}$ (minimum expected utility), or the average expected utility obtained according to some prior distribution for $\boldsymbol{\theta}$ (Bayes expected utility). We focus here on the latter criterion. Under our given choice of utility function, the Bayes optimal rule intuitively guesses the mode(s) of the posterior distribution of $\boldsymbol{\theta}$; under a uniform prior, this maximum a-posteriori decision rule simplifies to the maximum likelihood decision rule, which we find desirable not only for its familiarity but also for its sensibility in situations where a strong prior belief is difficult to justify. We emphasize that implementing our maximum likelihood decision rule does not require a subjective prior—only establishing its optimality does. While these results for Bayes optimality under the specified utility function are not novel, we present them here due to their centrality to our discussion.[4]

Consider the candidate decision rule f^*_π which selects the maxima of the posterior distribution of $\boldsymbol{\theta}$ (note that, while each value of $\boldsymbol{\theta}$ itself describes a distribution within the sample, the Bayesian decision maker's subjective belief also induces a distribution over the various values of $\boldsymbol{\theta}$). The decision rule f^*_π can be defined as follows. Let $\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{X})$ be the set of $\boldsymbol{\theta}$ that maximize the posterior

[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).

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

distribution given the observed data $\boldsymbol{X}$, i.e.

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{X}) &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{P}\big(\boldsymbol{\theta} \mid \boldsymbol{X}\big) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{X})\pi(\boldsymbol{\theta}),\end{aligned} \tag{5}$$

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

$$f_\pi^*(\boldsymbol{X})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{x})\}} & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\Theta}}_\pi(\boldsymbol{X}), \\ 0 & \text{o.w.,} \end{cases} \tag{6}$$

where $\#\{\cdot\}$ is the counting measure. Observe that $f_\pi^*(\boldsymbol{X})$ is a well-defined probability distribution over $\boldsymbol{\Theta}$ for all realizations of $\boldsymbol{X}$. When the posterior distribution is unimodal, f_π^* chooses the maximizer with probability one; when the posterior distribution is multimodal, f_π^* prescribes an equal probability to each maximizer.

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:

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

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

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

Thus, since f_π^* achieves the upper bound on the Bayes expected utility of any decision rule, we

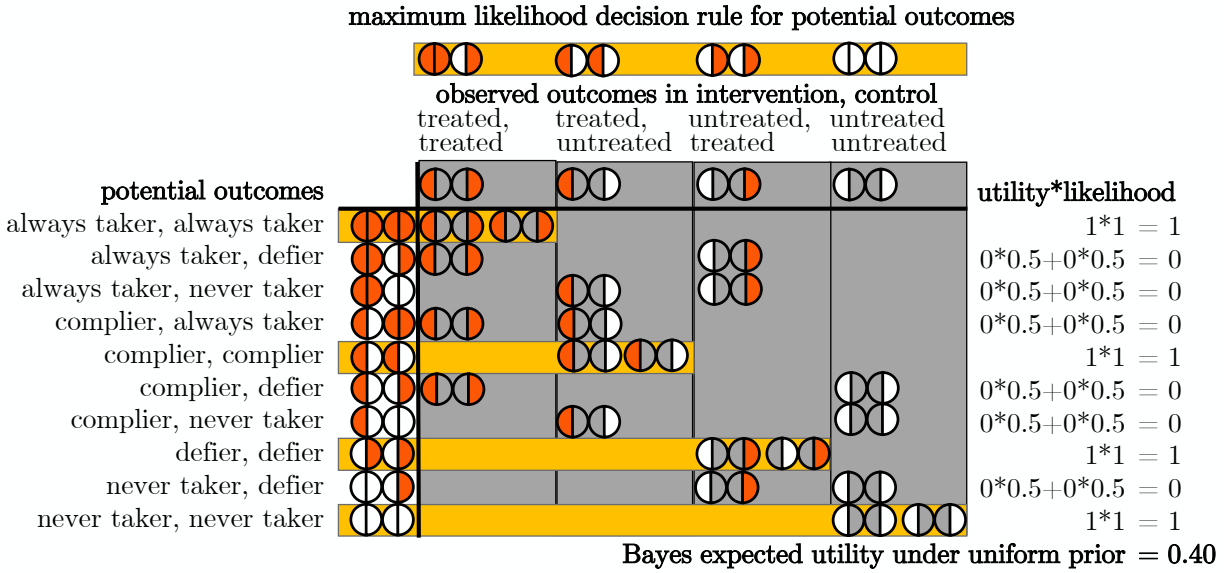
conclude that f_π^* is Bayes optimal.

Finally, observe that when the prior distribution $\pi(\theta)$ is constant, the maximizers of the posterior distribution in (5) are simply the maximizers of the likelihood, and the Bayes rule f_π^* in (6) simplifies to choosing the maximum likelihood estimate for $\boldsymbol{\theta}$ (or randomly choosing across multiple maximizers). While the maximum likelihood decision rule (or, more generally, the maximum aposteriori decision rule) is Bayes optimal, the rule does not generically take a convenient analytical form. However, since $\boldsymbol{\Theta}$ is finite, the integer programming problem of maximizing the likelihood or posterior distribution can be solved in small samples by an exhaustive grid search.

3.2 Illustration of the Maximum Likelihood Decision Rule

Figure 2 illustrates calculation of the Bayes expected utility for the maximum likelihood decision rule in our running example of an experiment with two people. The rows represent all of the possible joint distributions of potential outcomes in a sample of two people, while the columns represent all of the possible data outcomes that could be observed (note that the discussion of Section 2.3 focuses on the first column; here, we extend the discussion to 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 balls side by side. The likelihood value is one for every cell with two pairs of balls, 1/2 for every cell with one pair of balls, 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.

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



Above the columns in Figure 2, we represent the maximum likelihood decision rule, also in a yellow box. The decision rule maps each column to a row representing a (degenerate) guess for the unobserved joint distribution of potential outcomes. In the rightmost column, we calculate the expected utility of following the decision rule in each row as the probability that the rule guesses

correctly. We see that each row produces an expected utility of either one or zero. Finally, we calculate the Bayes expected utility of the maximum likelihood decision rule by averaging over the rows according to a uniform prior. From the preceding discussion, we conclude that 0.40 is the maximum achievable Bayes expected utility under this prior.

4 Performance of the Bayes Optimal Decision Rule

4.1 A Benchmark Monotonicity Decision Rule for Comparison

In the previous section, we established that the maximum likelihood decision rule is Bayes optimal. In this section, we quantify the size of the gain from using the optimal decision rule over suboptimal rules, to demonstrate that the optimal rule offers significant improvement over alternatives. In particular, we compare the performance of the optimal rule to an alternative rule inspired by the "monotonicity" (Imbens and Angrist, 1994) or "monotone response" (Manski, 1997b) assumptions used in sampling-based methods.

We construct the following "monotonicity decision rule," which imposes two restrictions on the estimated joint distribution of potential outcomes. First, the number of compliers or defiers (or both) in the estimated distribution must be zero. Second, the estimated number of compliers or defiers (whichever is nonzero) as a share of the sample must equal the difference in the average outcomes between the intervention and control groups (i.e. the point estimate of the average intervention effect). While this assumption differs fundamentally from the assumptions of Imbens and Angrist (1994) or Manski (1997b), which are large sample assumptions on an underlying superpopulation, we find it a reasonable analogue for the design-based setting.

We formally define the restricted "monotonicity" set of distributions, which is a function of the experimental data, as $\boldsymbol{\Theta}^M(\boldsymbol{X})$, where

$$\Theta^M(\boldsymbol{X}) = \Bigg\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \Big(\theta_{10} = 0 \text{ or } \theta_{01} = 0\Big), \text{ and } \\ \frac{\theta_{10} - \theta_{01}}{\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00}} = \frac{X_{I1}}{X_{I1} + X_{I0}} - \frac{X_{C1}}{X_{C1} + X_{C0}} \Bigg\}$$

Next, we define the set of constrained maximizers of the likelihood (or of the posterior distribution, for nonuniform priors):

$$\widehat{\Theta}^M_\pi(\boldsymbol{X}) = \underset{\boldsymbol{\theta} \in \Theta^M(\boldsymbol{X})}{\arg\max}\ \mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{X}) = \underset{\boldsymbol{\theta} \in \Theta^M(\boldsymbol{X})}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{X})\pi(\boldsymbol{\theta})$$

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

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

We opt for this constrained maximum likelihood approach over a plug-in estimator to ensure a valid, finite sample estimate that lies within $\boldsymbol{\Theta}$. The constrained maximum likelihood approach also guarantees that we compare our proposed decision rule to the "best" monotonicity rule.

Of course, the restrictions imposed by the monotonicity decision rule need not hold in general. Indeed, the sample may be such that there are both compliers and defiers; or, the randomization within the experiment may have occurred in such a way that the share of compliers or defiers is not equal to the point estimate of the average intervention effect (like, for example, if more compliers happen to be randomized into the intervention group than into the control group). In the following

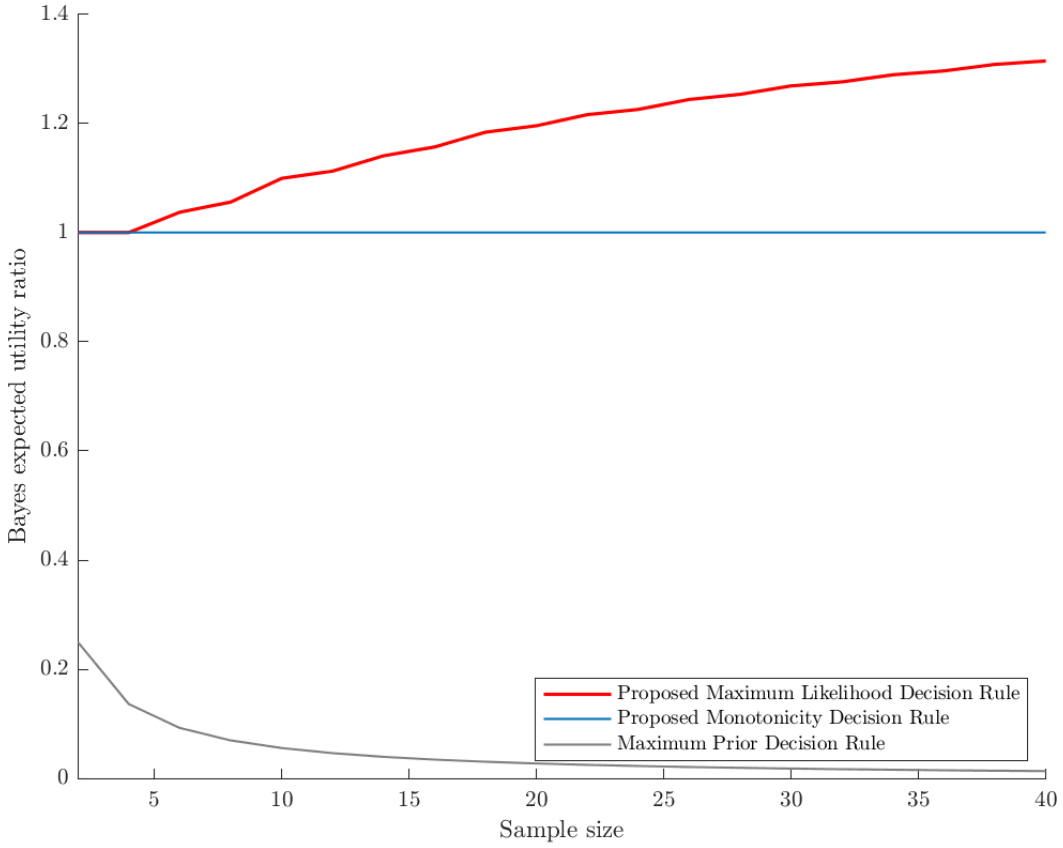
section, we quantify the cost of imposing imposing these assumptions relative to following the optimal rule.

4.2 Relative Performance as a Function of Sample Size

For a given sample size, we evaluate the performance of the maximum likelihood rule by computing the ratio of the Bayes expected utility achieved by this rule to the Bayes expected utility achieved by the monotonicity rule. We impose a uniform prior across $\boldsymbol{\Theta}$, such that our maximum likelihood rule is optimal.

Figure 3 shows this ratio 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 1.31 times that of the monotonicity rule.

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



In addition to the maximum likelihood rule, we also consider the relative performance of a "maximum prior" decision rule that simply chooses the $\boldsymbol{\theta}$ with the highest prior probability, regardless

of the observed data:

$$f_{\pi}^{\text{max prior}}(\boldsymbol{X})(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\#\{\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\}} & \text{if } \boldsymbol{\theta} \in \arg\max_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}'), \\ 0 & \text{o.w.} \end{cases}$$

Note that, in the case of a uniform prior, the maximum prior rule simply selects a value of $\boldsymbol{\theta}$ at random. The maximum prior rule significantly underperforms the monotonicity and maximum likelihood rules, demonstrating that these rules both capture significant learning from the data about the joint distribution of potential outcomes in the sample.

We conclude, then, that the maximum likelihood decision rule performs at least as well as the monotonicity decision rule, and significantly outperforms it as the sample grows. Having motivated its use, we now turn to two example applications in which the optimal rule and the monotonicity rule make meaningfully different decisions in our analysis of the sample.

5 Two Applications to Health Care

5.1 First Stage: A Vaccine Nudge Experiment

We apply our approach to a randomized experiment with a nudge intervention intended to encourage flu vaccination (Lehmann et al., 2016). The researchers used a completely randomized design that assigned 61 of the 122 total workers at a health center to the intervention. Of those assigned intervention, 17 took up the flu vaccine, so we consider them "treated." We consider the remaining 44 who did not "untreated." In control, 10 were treated and 51 were untreated. The point estimate of the average intervention effect of 0.11 (=17/61-10/61) indicates that the intervention increased flu vaccination by 11 percentage points. However, the result is not statistically significant at conventional levels. The p-value from Fisher's exact test is 0.19, and the p-value from a t-test is 0.12. The implied first stage F statistic of 2.4 is below the conventional threshold of 10 for a strong instrument.

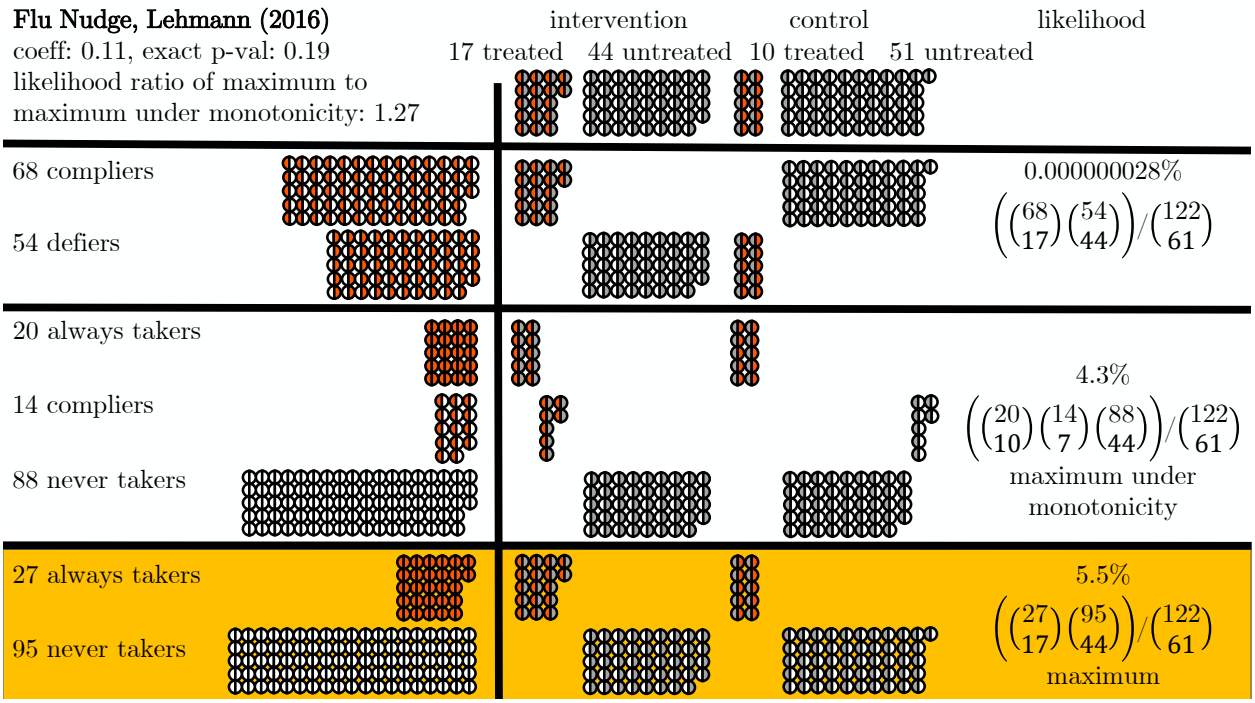
Suppose that the researchers want a data-driven approach to determine if they should make a LATE monotonicity assumption. Using downstream data on flu cases for the same people, they want to produce an instrumental variable estimate, which they would like to interpret as a local average treatment effect on compliers. However, they are concerned that there could have been defiers. They recognize that the intervention could have been off-putting for some people because it made a flu vaccination appointment that some people had to cancel or reschedule. Maybe there were so many defiers that they diluted the point estimate of the average intervention effect, thereby reducing its magnitude and statistical significance. In that case, the instrumental variable estimate would give a weighted average of the treatment effect on compliers and the opposite of the treatment effect on defiers. On the other hand, perhaps the intervention did not have any effect at all, and the observed average intervention effect just occurred by chance, in which case the instrumental variable estimate would be undefined.

Our decision rule shows that the joint distribution of potential outcomes that maximizes the likelihood includes 27 always takers and 95 never takers. This distribution is consistent with the Fisher null hypothesis that the intervention did not have an impact on anyone—there were no compilers and no defiers. The researchers decide not to proceed with an instrumental variable estimate because they are concerned about first stage relevance.

What is the strength of the evidence behind their decision, and is there any intuition behind it? In Figure 4, we report a graphical illustration of the experiment. In the experiment with only 2 people shown in Figure 2, there are 10 rows and 4 columns. However, in an experiment with 122 people, there are 317,750 rows and 3,844 columns, so we focus on the single column that represents the observed outcomes in intervention and control and three rows of interest. The last row is the

row with the highest likelihood, so we highlight it. This row, which indicates that there are 27 always takers and 95 never takers, has a likelihood of 5.5%.

Figure 4: Illustration of the Lehmann et al. (2016) Vaccine Nudge Experiment



For comparison to the distribution that maximizes the likelihood, the other two rows report distributions that preserve the average intervention effect. The average intervention effect implies that the nudge increased flu vaccination by 7 people among the 61 in intervention, consistent with 14 additional vaccinations in the full sample of 122 people. Therefore, the average intervention effect implies 14 net compliers, 14 more compliers than defiers. The potential outcome distributions in the first two rows both have 14 more compliers than defiers, but they have very different likelihoods. The first row depicts the distribution consistent with the sharp null hypothesis that everyone was affected by the intervention in one direction or the other such that the experiment includes 68 compliers and 54 defiers. The likelihood is 0.000000028%. The middle row depicts the result of our monotonicity decision rule. The likelihood is 4.3%.

To interpret the strength of the evidence behind the decision, we report the ratio of the maximum likelihood to the maximum likelihood under late monotonicity: 1.27. This likelihood ratio can be interpreted as a Bayes factor comparing the hypothesis that the joint distribution of potential outcomes is the final row of Figure 4 versus the hypothesis that the distribution is the middle row of Figure 4 (note that a prior is not needed to compute the Bayes factor between two sharp hypotheses, which is why we prefer the likelihood ratio terminology). While this value is not particularly large relative to conventional levels for Bayes factors, it is remarkable that the ratio of the maximum likelihood relative to the likelihood of the distribution shown in the first row is over 196 million, providing very strong evidence for the maximum likelihood decision over the alternative decision that everyone was affected.

The cells of the figure provide some intuition for the variation in the likelihoods. They depict the implied numbers of each of the principle strata randomized into intervention and control. In the first row, for a truth of 68 compliers and 54 defiers to be consistent with the observed outcomes,

the randomization would have assigned way more compliers to intervention than control and at the same time assigned way more defiers to control vs. intervention. Thus, the likelihood is small.

In the next two rows, randomization is balanced between intervention and control within always takers, compliers, and never takers, yielding much higher likelihoods. These likelihoods differ, though, so randomization imbalance cannot explain all variation across likelihoods. The last column shows the derivation of the likelihoods. There are $3.83e35$ ways to randomize 122 people into two groups with 61 each (122 chose 61). If the true distribution includes 20 always takers, 14, compliers and 88 never takers, 4.3% of those ways will yield the observed data. The entropy is $1.66e34 = (20 \text{ choose } 10) \times (14 \text{ choose } 7) \times (88 \text{ choose } 44)$. In contrast, if the true distribution has 27 always takers and 95 never takers, the entropy is much higher.[5]

5.2 Reduced Form: A Clinical Trial for Sepsis Treatment

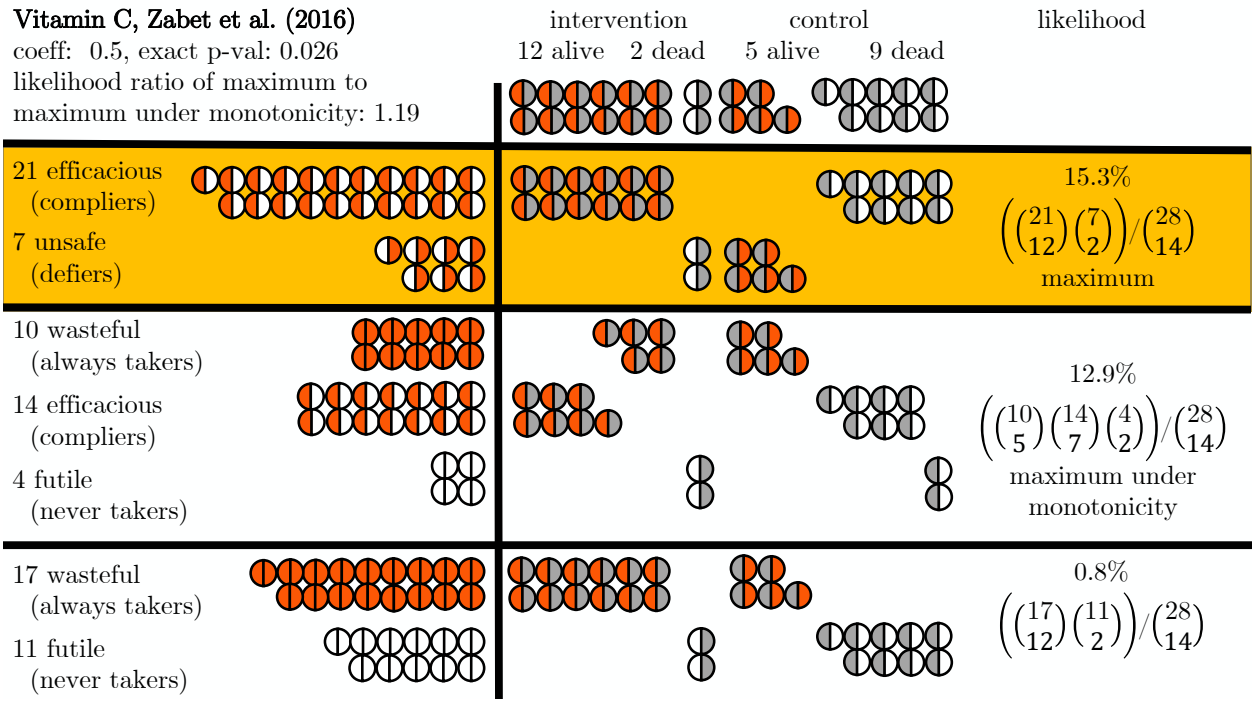
We next apply our decision rule to a clinical trial of 28 people that examined the impact of high dose Vitamin C on patients with sepsis (Zabet et al., 2016). With this example, we consider a "reduced form" setting, in which the potential outcomes represent "survival" (which maps to "treated") and "death" (which maps to "untreated"). Under the outcome of survival, we can interpret the principle strata in this experiment as those for whom the intervention is wasteful (alive in intervention and alive in control, i.e. always takers), those for whom the intervention is efficacious (alive in intervention and dead in control, i.e. compliers), those for whom the intervention is unsafe (dead in intervention and alive in control, i.e. defiers), and those for whom the intervention is futile (dead in intervention and dead in control, i.e. never takers). In control, 9 of 14 people died within 28 days, as compared with only 2 of 14 people in intervention. The Fisher exact test rejects the null hypothesis that the intervention was neither efficacious nor safe for anyone at the 2.6% level. The point estimate of the average intervention effect is 0.5, and the p-value of the t-test is 0.004.

Suppose the researchers feel confident that the intervention significantly reduced mortality on average, but they fear that high dose Vitamin C may have also had side effects that ultimately killed some patients. Absent data on alternative outcomes, how could the researchers assess whether the intervention was unsafe for any patients (i.e. whether there are both compliers and defiers)? Our decision rule shows that the joint distribution of potential outcomes that maximizes the likelihood is one with 21 people for whom the intervention was efficacious, 7 people for whom the intervention was unsafe, and zero people for whom the intervention was wasteful or futile. The number of net compliers is estimated to be 21-7=14, which matches the point estimate of the average intervention effect multiplied by the sample size. Following our decision rule, the researchers should conclude that high dose Vitamin C had adverse side effects. With access to additional covariates or outcomes, researchers could potentially identify a mechanism that could be unsafe for some patients. For example, in the Bernard et al. (2001) clinical trial testing the effect of recominbant human activated protein C on patients with sepsis, researchers identified a potential mechanism for harm by observing an additional outcome among some patients: severe bleeding. Our decision rule, which does not assume away the presence of those for whom the intervention was unsafe, could be informative about when such mechanisms are likely to be present.

Figure 5 represents the experiment graphically. The first row shows the joint distribution of potential outcomes that maximizes the likelihood at 15.3%. We can also deduce how many people

[5] Pascal's triangle provides some intuition. Within a row of Pascal's triangle, N choose k grows as k gets closer to $n/2$ (randomization gets closer to balanced); moving down the triangle, we also see that N choose k grows as N increases (the sample size increases). However, moving down a row typically increases N choose k by more than moving across a row. Therefore, even though the last two rows both have balanced randomization, the last row has larger numbers of two types instead of smaller numbers of three types, yielding a larger value for entropy. We are grateful to Liz Ananat for sharing this point.

Figure 5: Illustration of the Zabet et al. (2016) Vitamin C Experiment



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 effective for all 12 of them. By similar logic for the people who died in intervention and lived and died in control, our decision rule suggests that it just so happened via the randomization process that more of the people for whom the intervention was effective 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), our best guess is that 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 5 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. By construction, this decision rule matches the point estimate of the average intervention effect. The likelihood of this distribution is 12.9%, which is strictly lower than the unconstrained maximum. The ratio of the maximum likelihood to the maximum under monotonicity is 1.19, suggesting that the evidence in favor of our maximum likelihood decision rule is 1.19 times stronger than the evidence in favor of the monotonicity decision rule. 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%, which is intuitive given the implied amount of imbalance between intervention and control among those for whom the intervention would be wasteful and those for whom the intervention would be futile.

6 Implications

In many experiments, we take for granted that the point estimate of the average intervention effect is sufficient for making a decision. However, considering just this point estimate throws away

valuable information: what was the randomization process? How many people are observed treated and untreated in intervention and control? With this paper, we try to exploit more information about the experiment and its outcomes.

A randomized experiment is widely consiered to offer the most credible evidence on causal effects. The analysis of randomized experiments, then, warrants statistical methods tailor-made to the tool. 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." Going further, 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, these restrictions are less limiting, and large sample approximations may be less useful. Exact design-based methods closely follow the actual structure of randomization that produced the data, and as we have seen here, they can produce novel insights over large sample methods.

Sometimes, a decision maker really is just interested in their finite sample. In Lehmann et al. (2016), researchers sampled the entire population of interest—the 122 employees at a particular health care provider. Other times, decision makers wish to use an experiment to draw conclusions about a separate population. An important goal for the design-based decision rules developed here and elsewhere, then, is understanding how to extend what we learn in a finite sample to groups outside the sample. Our work provides an important motivating example. Applying experimental data to learn directly about the joint distribution of potential outcomes in a superpopulation faces well-known limitations; but if a given sample is drawn from a superpopulation, and the sample contains both compliers and defiers, then the superpopulation must as well.

References

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

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

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.

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.

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 (2024). 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.

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.

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.

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

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.

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.

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.

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

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.

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

Lehmann, B. A., G. B. Chapman, F. M. Franssen, G. Kok, and R. A. Ruiter (2016). Changing the default to promote influenza vaccination among health care workers. *Vaccine 34*(11), 1389–1392.

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

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.

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*.

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.

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

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.