Statistical Decision Theory and Empirical Practice*

Timothy B. Armstrong†
University of Southern California

Toru Kitagawa‡
Brown University

Aleksey Tetenov§
University of Geneva

January 5, 2025

Abstract

This review article discusses statistical decision theory as it relates to empirical practice in economics. We review how hypothesis testing, estimation, preanalysis plans and other widely used aspects of empirical practice can be obtained as special cases of the general decision theoretic setup of Wald. We then review a recent econometrics literature that has debated the role of decision theory in empirical practice. We discuss alternative perspectives from this literature, and how they differ from classical decision theoretic interpretations of empirical practice. To illustrate our arguments, we use a running example in which a researcher performs an experiment, collects data and publishes the results following commonly used guidelines.

1 Introduction

A recent literature has debated the role of decision theory in empirical practice in economics. A common claim in this literature is that a particular aspect of empirical practice, or even current empirical practice as a whole, lacks a satisfactory decision theoretic motivation. Consider this opening statement from Banerjee et al. (2017):

*We thank Chris Hansen, Federico Bugni, Tim Conley, Bruno Ferman, Chishio Furukawa, Max Kasy, Pat Kline, Michal Kolesár, Jonathan Libgober, Chuck Manski, Chen Qiu, Jesse Shapiro and Patrick Vu for helpful comments and discussions.

†email: timothy.armstrong@usc.edu
‡email: toru_kitagawa@brown.edu
§email: aleksey.tetenov@unige.ch

> In the last couple of decades, two of the most successful areas of economic research have been decision theory–and its close cousins, behavioral, and experimental economics–and empirical microeconomics. Despite the fact that both emphasize experimentation as a method of investigation, there is almost no connection between the two literature [*sic*].

Manski (2021) provides another example:

> ...econometrics did not embrace statistical decision theory. Instead, it focused on study of identification, estimation, and statistical inference.

These statements suggest that something has gone very wrong with empirical research in economics. Are economists really ignoring recommendations from decision theory in their empirical practice? Are econometricians and statisticians failing to apply decision theory when developing and prescribing empirical methods?

Addressing these questions requires a working definition of "statistical decision theory." For the purpose of this article, we take "statistical decision theory" to encompass the concepts put forth by Wald in his book (Wald, 1950).[1] We give a brief overview of these concepts here and a more thorough one in Section 3. The main ingredients of statistical decision theory are a *decision rule* that maps data to a decision and a *loss function* that measures the disutility of this decision given a parameter value that describes the true state of the world. The expected loss given an unknown parameter value is called the *risk* of the decision rule. The risk is a function of the unknown parameter value, and statistical decision theory compares different decision procedures by comparing their risk functions. A familiar example is the use of squared error loss $(\hat{\theta} - \theta)^2$ for an estimate $\hat{\theta}$ of a parameter θ, which leads to mean squared error as the risk. As we discuss further in Section 3, statistical decision theory is general enough to treat not only estimation but also other settings such as hypothesis testing and policy recommendations.

Turning back to the quotes at the beginning of this article, it is helpful to distinguish between two critiques of the decision theoretic foundations of a given empirical practice X, such as hypothesis testing, estimation, preanalysis plans, etc.:

> Claim A: *empirical practice X has no known decision theoretic motivation*

and

[1] This definition appears to be consistent with the use of "(statistical) decision theory" by the authors quoted above. Manski (2021) cites Wald (1950). Banerjee et al. (2017) is a "nontechnical discussion" of Banerjee et al. (2020b), which cites Wald (1950) as well as Savage (1954) and more recent papers.

Claim B: *empirical practice X has a well established decision theoretic motivation, but it is not entirely satisfactory for reason Y.*

In the first part of this paper, we argue that Claim A is incorrect when applied to many widely used aspects of empirical practice including hypothesis testing, estimation and preanalysis plans. Indeed, Wald clearly viewed the Neyman-Pearson hypothesis testing framework (introduced in Neyman and Pearson (1933a) and further elaborated by Neyman and others) as well as the problem of statistical estimation as special cases of his theory (see, for example, Sections 1.5 and 1.7 of Wald (1950)). To clarify this point, we review how hypothesis testing and estimation can be derived as special cases of the general decision theoretic setup of Wald (1950), and how other aspects of research design such as preanalysis plans map to this setup. In particular, the size and power of a test correspond to risk with zero-one loss functions corresponding to type I and type II errors. Requirements on size and power typically imposed in hypothesis testing and statistical power analysis are then obtained by applying the minimax criterion to this setting. We refer to this as a "classical decision theoretic interpretation" of the practice of hypothesis testing, since it is consistent with the mapping between the Wald and Neyman-Pearson frameworks used by these authors and their contemporaries.

Fortunately, given our skepticism about Claim A, our view of the recent literature debating the role of decision theory in empirical economics is that one need not make such a strong claim to motivate most of this literature. Rather, many of these papers can be interpreted as making some variation of Claim B, with the papers differing in which aspect of classical decision theoretic interpretations of empirical practice they find unsatisfactory, and in the alternatives to this classical framework that they propose. We review some of the critiques and alternative proposals from this recent econometrics literature and from earlier debates among contemporaries of Wald. In contrast to our strong rebuttal of Claim A, we do not take a strong stance here on the various forms of Claim B that have been made in this literature. Rather, our goal is to provide a discussion organizing the ways in which these proposals differ from classical decision theoretic interpretations of empirical practice.

To map our arguments to empirical practice, we discuss these issues in the context of a running example in which a researcher performs an experiment, collects data and publishes the results following guidelines in Duflo et al. (2007). This provides a concrete setting that allows us to cover much of the recent debate on the role of decision theory in empirical economics. We note, however, that many of these arguments apply to nonexperimental settings and other research designs used in empirical research in economics: the mapping between hypothesis testing and statistical decision theory described in Section 3.2 applies

equally to experimental and observational data.

Our goal is to provide a focused discussion of particular aspects of empirical practice and their decision theoretic interpretation, rather than a comprehensive review of all of the ways that decision theory enters into empirical work. While we review Bayesian decision theory, our main focus is on debates about decision theory as it relates to frequentist procedures such as hypothesis testing. Of course, one form of Claim B for frequentist procedures is an argument that a satisfactory decision theoretic justification should be Bayesian. Sims (2008) argues this point in the context of monetary macroeconomics. While we review Savage's famous arguments in favor of Bayesianism in Section 5.1, we do not provide a comprehensive discussion of more recent debates. We refer to Berger (1985) for further discussion of Bayesian statistics and decision theory. We also focus mostly on hypothesis testing, with only brief discussions of estimation and confidence intervals. It is well known that a confidence interval can equivalently be represented as a family of hypothesis tests about the parameter of interest. Thus, our treatment of hypothesis testing applies to confidence intervals as well. However, debates about confidence intervals and their decision theoretic interpretation often involve considerations that are not captured by this approach.[2]

The remainder of this article is organized as follows. Section 2 introduces our running example. Section 3 reviews decision theory and explains how estimation and hypothesis testing fit into the general decision theoretic framework. Section 4 maps the running example introduced in Section 2 to the decision theoretic environment described in Section 3. Section 5 discusses differing interpretations and critiques of the mapping between decision theory and empirical practiced presented in earlier sections. Section 6 concludes.

2 Running example

We consider the scenario of a researcher conducting a randomized control trial (RCT) that aims to measure the impact of a specific intervention ("treatment") on an outcome of interest. In doing so, the researcher refers to Duflo et al. (2007) or the more recently updated J-PAL online guides for researchers (J-PAL, 2023) as a reference for recommended practice. We now give a stylized description of this process, leaving some details for later.

In designing the trial, the researcher chooses a sample size of individuals to receive the intervention (the treatment group), and a sample size of individuals who do not receive the

[2]One critique of confidence intervals is that they are often used to quantify uncertainty about a parameter, a notion which may not be directly justified by classical decision theoretic interpretations of the corresponding hypothesis tests; see Müller and Norets (2016) for a recent example of such a critique.

intervention (the control group). The treatment and control groups are chosen *at random* from the pool of individuals available for the trial (Duflo et al., 2007, Section 2). Outcomes for each group are then collected after the trial is implemented.

In addition, the researcher specifies a statistical hypothesis test for testing the null hypothesis of no effect of the intervention. The choice of sample size is then informed by an *analysis of statistical power* (Duflo et al., 2007, Section 4). In particular, the researcher specifies an effect size that the researcher deems plausible (a "minimum detectable effect"). The researcher then chooses the sample size so that the hypothesis test is guaranteed to have size no greater than a given level (say, 5%) if the null hypothesis holds, and to have some prespecified power (say, 80%) under the alternative hypothesis of the given minimum detectable effect size.

Rather than making these choices while conducting the trial and reporting the results, the researcher commits to all of the steps described so far in a *preanalysis plan*. This preanalysis plan is recorded in a registry of experiments. The researcher may also submit the paper as a *registered report* with discussion of interpretation of possible results to a journal that will make a publication decision based on this report, such as *Journal of Political Economy: Microeconomics* (List, 2023) or *Journal of Development Economics* (Bogdanoski et al., 2020). After the data is collected, the researcher reports the results of the hypothesis test, along with an estimate and confidence interval for the effect size.

3 Decision theory and classical statistics

This section reviews some basic concepts from statistical decision theory (Section 3.1), and shows how the framework of estimation and hypothesis testing from classical statistics fit into statistical decision theory (Section 3.2). We view this as consistent with decision theoretic interpretations of hypothesis testing and estimation in Wald (1950) and other contemporary sources, as we describe in Section 3.3. We therefore refer to this mapping from statistical decision theory to estimation and hypothesis testing as a "classical decision theoretic interpretation" of these practices.

3.1 General setup

We have data Y that follows a distribution P_θ indexed by an unknown parameter θ, taking values in a parameter space Θ. The parameter θ can be finite dimensional as in classical parametric models, or it can be an infinite dimensional object such as a conditional expec-

tation function or the entire unknown distribution of the data. We are faced with an *action space* $\mathcal{A}$ of possible actions. If we choose action $a \in \mathcal{A}$ and the parameter is given by θ, we incur a loss $L(\theta, a)$, called the *loss function*. A *decision rule* is a mapping $\delta(Y)$ that takes the data Y to an action in $\mathcal{A}$. The *risk function* of the decision rule δ is

$$R(\theta, \delta) = E_\theta L(\theta, \delta(Y)) = \int L(\theta, \delta(y))\, dP_\theta(y).$$

We leave randomized decision rules out of the notation in the main text. Appendix A, which covers the randomized sampling scheme in the running example, explicitly incorporates randomized decisions.

The *Bayes risk* of a decision rule δ for a *prior* $\pi(\theta)$ is

$$R_{\text{Bayes}}(\pi, \delta) = \int R(\theta, \delta) d\pi(\theta).$$

The *maximum risk* of a decision rule δ over the parameter space Θ is the worst-case Bayes risk over all priors π supported on Θ or, equivalently, the worst-case risk over $\theta \in \Theta$

$$R_{\max}(\Theta, \delta) = \sup_{\pi \text{ supported on } \Theta} R_{\text{Bayes}}(\pi, \delta) = \sup_{\theta \in \Theta} R(\theta, \delta).$$

The *minimax criterion* evaluates decision rules δ according to their maximum risk $R_{\max}(\Theta, \delta)$. More generally, the Γ-*minimax criterion* takes the maximum of the Bayes risk $R_{\text{Bayes}}(\pi, \delta)$ over π in some restricted set of priors Γ, which may not include all priors supported on any given parameter space Θ. Γ-minimax and other approaches involving multiple priors are considered in the *robust Bayes* literature (see Ríos Insua and Ruggeri, 2000; Giacomini et al., 2021). While many of the discussions of the minimax criterion later in the paper also apply to the Γ-minimax criterion, we focus on the minimax criterion for simplicity.

Finally, rather than considering loss directly, the *minimax regret* criterion considers loss relative to a (typically infeasible) decision that uses knowledge of θ. In particular, minimax regret is defined by first considering the loss function $\tilde{L}(\theta, \delta) = L(\theta, \delta) - \inf_{a \in \mathcal{A}} L(\theta, a)$, and then using $\tilde{L}(\theta, \delta)$ to calculate the maximum risk. For the hypothesis testing and estimation settings considered below, $\inf_{a \in \mathcal{A}} L(\theta, a) = 0$ so that maximum risk and maximum regret are the same. However, the distinction becomes meaningful for welfare loss functions discussed in Section 5.5.

Once one adopts one of these criteria (Bayes risk with a particular prior π, or maximum risk or regret with a particular parameter space Θ), one can use this criterion to compare

decision rules. A *minimax* decision rule minimizes $R_{\max}(\Theta, \delta)$ over δ, whereas a *Bayes* decision rule under prior π minimizes $R_{\text{Bayes}}(\pi, \delta)$ over δ.

3.2 Hypothesis testing and estimation

In this section, we describe how hypothesis testing and estimation can be obtained from the general setup by specializing to particular action spaces $\mathcal{A}$ and loss functions $L(\theta, \delta)$.

For hypothesis testing, we are interested in subsets of the parameter space Θ called the *null hypothesis* H_0 and an *alternative hypothesis* H_1. The action space is $\mathcal{A} = \{0, 1\}$ with 1 denoting rejection of the null and 0 denoting failure to reject. To define the loss function, we consider two types of error. *Type I error* occurs when we incorrectly reject the null ($\theta \in H_0$ and $\delta(Y) = 1$). *Type II error* occurs when we incorrectly fail to reject the null ($\theta \in H_1$ and $\delta(Y) = 0$). Letting w_I and w_{II} denote relative weights on type I and type II errors, this leads to the loss function[3]

$$L(\theta, \delta) = w_I \cdot 1(\theta \in H_0,\, \delta = 1) + w_{II} \cdot 1(\theta \in H_1,\, \delta = 0). \tag{1}$$

Alternatively, we may consider separate loss functions $L_I(\theta, \delta) = 1(\theta \in H_0,\, \delta = 1)$ and $L_{II}(\theta, \delta) = 1(\theta \in H_1,\, \delta = 0)$ for type I and II errors respectively, and evaluate the corresponding risk for each loss function separately.

The parlance of hypothesis testing includes special terminology for the risk function and maximum risk for these loss functions. The *power* of a test at a parameter θ is the probability that the test rejects. Typically, power is used to refer to the rejection probability for θ in the alternative H_1, in which case the power is simply one minus the risk function under the loss $L_{II}(\theta, \delta) = 1(\theta \in H_1,\, \delta = 0)$ given above. The *size* of a test is the maximum rejection probability over θ in the null H_0. It is the maximum risk under the loss function $L_I(\theta, \delta) = 1(\theta \in H_0,\, \delta = 1)$ given above.

While most of our emphasis will be on hypothesis testing, we note that the problem of *estimation* falls into the general setup by taking $\mathcal{A} = \Theta$ and defining a loss function of the form $L(\theta, \delta) = \ell(\delta - \theta)$ for a function ℓ that is decreasing on $(-\infty, 0]$ and increasing on $[0, \infty)$. Common choices are $\ell(t) = t^2$ (squared error loss) and $\ell(t) = |t|$ (absolute error loss). We can also consider estimation of functions of θ by taking $L(\theta, \delta) = \ell(\delta - T(\theta))$ for some function $T(\theta)$. In particular, our running example will involve estimating the difference

[3]Throughout the paper we use $1(\cdot)$ to denote an indicator function equal to one if the condition is satisfied and zero otherwise.

$T(\theta) = \theta_1 - \theta_2$ between two parameters θ_1 and θ_2.

Finally, a $100 \cdot (1-\alpha)\%$ confidence interval (CI) for θ is a set $\mathcal{C}(Y)$ that satisfies $P_\theta(\theta \in \mathcal{C}(Y)) \geq 1-\alpha$ for all $\theta \in \Theta$. One way of placing CIs into the general decision theoretic framework is to view CIs as a way of summarizing a family of tests: the decision rule $\delta_{\theta_0}(Y) = 1(\theta_0 \notin \mathcal{C}(Y))$ is a hypothesis test of the null hypothesis H_{0,θ_0} consisting of the point θ_0, which has size α by construction. The mapping between hypothesis tests and decision theory described above can then be applied to these tests. Alternatively, one may consider the CI as a decision rule directly, with a loss function $L(\theta, \mathcal{C})$ defined in terms of the length of the interval $\mathcal{C}$ and whether it covers the true parameter value θ. Depending on the exact formulation, this may have an equivalent formulation in terms of the corresponding hypothesis tests. Further discussion and results can be found in Pratt (1961), Berger (1985, Section 2.4.3) and Lehmann and Romano (2005, Section 3.5). In the interest of space, we do not a give a separate treatment of CIs and their decision theoretic interpretation.

3.3 Historical notes

The setup and definitions in Section 3.1 follow Wald (1950), with some minor differences in terminology reflecting more recent usage (e.g. Berger, 1985). Neyman and Pearson (1933a) introduced the problem of trading off type I and type II errors as defined in Section 3.2. The terms "type I error" and "type II error" appear in Neyman and Pearson (1933b). Wald (1950) discusses the Neyman-Pearson framework as a special case of his general setup in Sections 1.5.1, 1.7 and 5.1.2. See, in particular, Wald (1950, Section 5.1.2, p. 131) for the definition of size and power in terms of risk for the loss functions L_I and L_{II} given above. Wald uses "size" for the risk function with loss L_I, whereas we define size to be the maximum of this risk function over the null hypothesis, which appears to be the more common usage (e.g. Lehmann, 1959, p. 61).

The loss function (1) and minimax decision rules under this loss function are considered in Wald (1950, Section 5.1.2), for the special case where w_I and w_{II} are equal. Lehmann (1959, Ch. 8) considers the problem of minimizing $\sup_\theta E_\theta L_{II}(\theta, \delta)$ subject to a bound on $\sup_\theta E_\theta L_I(\theta, \delta)$, which is a constrained optimization problem that is equivalent to the general case where w_I and w_{II} may be different. This form of the problem does not appear explicitly in Wald (1950), although he mentions the possibility of separate and asymmetric consideration of type I and type II errors in his framework (pp. 20-21).

4 Decision theory and the running example

We now return to the running example, and discuss its interpretation through the lens of statistical decision theory. We begin by adding some details that will allow us to map this example to the formal setting in Section 3.

4.1 Details for the running example

The researcher randomly samples a subset of n_0 as the untreated (control) group and a sample of n_1 individuals for treatment. We assume that the population is large, so that we can approximate the distribution as independent and identically distributed (iid) draws $Y_1, \ldots, Y_{n_0}$ from the population distribution of the untreated outcome and draws $Y_{n_0+1}, \ldots, Y_{n_0+n_1}$ from the population distribution of the treated outcome. We provide a formal description of this process using potential outcome notation in Appendix A.

Let μ_0 and μ_1 denote the population means of the distributions of untreated and treated outcomes respectively. The researcher is interested in the average treatment effect (ATE) of the intervention, which is given by $\mu_1 - \mu_0$, and in the null hypothesis $H_0 : \mu_1 - \mu_0 = 0$. The researcher forms the sample means $\bar{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i$, $\bar{Y}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} Y_i$ and the difference-in-means estimate $\bar{Y}_1 - \bar{Y}_0$, which has mean $\mu_1 - \mu_0$ and variance $\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$ where σ_0^2 and σ_1^2 are the population variances of treated and untreated outcomes. We will make the simplifying assumption that σ_0^2 and σ_1^2 are known and that the population distributions of untreated and treated outcomes are normal so that $\bar{Y}_1 - \bar{Y}_0$ is normally distributed:

$$\bar{Y}_1 - \bar{Y}_0 \sim N(\mu_1 - \mu_0, \mathrm{se}^2), \quad \mathrm{se}^2 = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}.$$

Formally, this can be justified in an asymptotic framework using a limit of experiments approach: our statements about the risk function and maximum risk hold asymptotically in a localized parameter space where we do not impose normality or known variance (Hirano and Porter, 2020; van der Vaart, 1998, Ch. 7-9, 15).

Using this sampling framework, the researcher forms a test of H_0 by rejecting when $\bar{Y}_1 - \bar{Y}_0 > \mathrm{se} \cdot z_{1-\alpha}$ where z_q denotes the qth quantile of the $N(0,1)$ distribution, thus guaranteeing that the probability of (falsely) rejecting when $\mu_1 - \mu_0 = 0$ is no greater than α, where α is some prespecified number (say, $\alpha = .05$). As discussed in Section 2, the sample sizes n_0 and n_1 are chosen using an analysis of statistical power in which the researcher specifies a *minimum detectable effect size* τ_{MDE} such that the alternative $H_1 : \mu_1 - \mu_0 \geq \tau_{\mathrm{MDE}}$ is both

plausible and of substantive interest. The researcher then chooses the sample sizes n_0 and n_1 so that the power of the test must be at least β for all μ_0, μ_1 satisfying $H_1 : \mu_1 - \mu_0 \geq \tau_{\text{MDE}}$, where β is some prespecified quantity (say, .8). This power requirement can be written as

$$\tau_{\text{MDE}} \geq (z_{1-\alpha} + z_\beta) \cdot \text{se} = (z_{1-\alpha} + z_\beta)\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}.$$

In addition to this hypothesis test, the researcher reports $\bar{Y}_1 - \bar{Y}_0$ as a point estimate of $\mu_1 - \mu_0$.

4.2 Decision theoretic interpretation

To put this into the general decision theoretic framework of Section 3.1, we take the parameter θ to be (μ_0, μ_1) and the data Y to be $(\bar{Y}_0, \bar{Y}_1)$, with $\bar{Y}_0$ and $\bar{Y}_1$ independent with $\bar{Y}_t \sim N(\mu_t, \sigma_t^2)$ for $t = 0, 1$ (note that σ_0^2 and σ_1^2 do not vary with the parameter θ, reflecting the fact that we are treating them as known).

To map the hypothesis test and statistical power analysis described in Section 4.1 to our general setup, we can use the loss function for hypothesis testing introduced in Section 3.2. In particular, we take $H_0 = \{(\mu_0, \mu_1) : \mu_1 - \mu_0 = 0\}$ and $H_1 = \{(\mu_0, \mu_1) : \mu_1 - \mu_0 \geq \tau_{\text{MDE}}\}$. The choice of size α and power β corresponds to a particular choice of the weights w_I and w_{II} on type I and type II errors in the loss function given in (1) in Section 3.2. In particular, the risk function is

$$R((\mu_0, \mu_1), \delta) = w_I \cdot P_{\mu_0,\mu_1}(\delta(Y) = 1) \cdot 1((\mu_0, \mu_1) \in H_0) + w_{II} \cdot P_{\mu_0,\mu_1}(\delta(Y) = 0) \cdot 1((\mu_0, \mu_1) \in H_1).$$

As discussed above, the researcher has designed the test so that the size is no greater than α, meaning that $P_{\mu_0,\mu_1}(\delta(Y) = 1) \leq \alpha$ for $(\mu_0, \mu_1) \in H_0$, and so that the test has power at least β, meaning that $P_{\mu_0,\mu_1}(\delta(Y) = 0) \leq 1 - \beta$ for all $(\mu_0, \mu_1) \in H_1$. If we set $w_I = 1$ and $w_{II} = \alpha/(1 - \beta)$, then this is equivalent to requiring that $R((\mu_0, \mu_1), \delta) \leq \alpha$.

Empirical practice uses conventional values of α and β without much discussion motivating their choice. In the statistical decision theoretic interpretation of standard practice, these choices imply that researchers, editors, and funding bodies assign costs w_I and w_{II} to type I/II errors. For example, the common benchmark $\alpha = .05$ and $\beta = .8$ correspond to $w_{II} = 1/4$. In Section 5.5 we discuss critiques and alternative specifications of the loss function.

The minimum detectable effect size τ_{MDE} plays an important role in separating the null

hypothesis $H_0 : \mu_1 - \mu_0 \leq 0$ from the alternative $H_1 : \mu_1 - \mu_0 \geq \tau_{\text{MDE}}$. This leads us to assign zero loss for type II error when the treatment effect is less than τ_{MDE}. Separating the null and alternative is a necessary step when using statistical power analysis to choose the sample size or decide whether to proceed with a study: if we instead defined H_1 to be the set where $\mu_1 - \mu_0 > 0$, it would be impossible to guarantee power β while maintaining size α for any $\beta > \alpha$ even with an arbitrarily large sample size. However, separating the null and alternative is not needed for other decision theoretic comparisons of hypothesis tests. For example, the test employed by the researcher in our example is *uniformly most powerful* (van der Vaart, 1998, Proposition 15.2), meaning that it maximizes power among level α tests for *all* values of (μ_0, μ_1) with $\mu_1 - \mu_0 > 0$.

4.3 Additional considerations

Before continuing, let us discuss some aspects of the running example that have not yet received explicit attention in our decision theoretic interpretation. First, recall that individuals are sampled and assigned at random from the pool of individuals available for the trial. As we discuss in more detail in Appendix A, randomized sampling and treatment assignment can be shown to be optimal according to the minimax criterion when the sampling and treatment assignment decisions are incorporated into our framework.

Second, recall that the researcher commits to the research design in a preanalysis plan. This ensures that the decision rule constitutes a binding ex-ante commitment. The bounds on type I and type II error reported by the researcher lead to bounds on the risk function $R((\mu_0, \mu_1), \theta)$ of the decision rule. These objects are computed ex-ante, making ex-ante commitment to a decision rule important for their interpretation. We will return to this point when discussing Savage's group decision-making interpretation of the minimax criterion in Section 5.1 below.

Finally, while our main focus is on hypothesis testing, we can map the point estimate $\bar{Y}_1 - \bar{Y}_0$ to the general decision theoretic setup following Section 3.2: it is a decision rule $\delta(Y) = \bar{Y}_1 - \bar{Y}_0$ under a loss function $L((\mu_0, \mu_1), \delta) = \ell(\delta - (\mu_1 - \mu_0))$ for some function $\ell(\cdot)$. If $\ell(\cdot)$ is symmetric and weakly increasing in the absolute value of its argument, then the estimator $\delta(Y) = \bar{Y}_1 - \bar{Y}_0$ is the minimax decision for this problem (van der Vaart, 1998, Proposition 8.6). Thus, the researcher's reported estimate enjoys a similar decision theoretic justification to the reported hypothesis test: it gives the best possible guarantee on the maximum risk.

5 Interpretations and alternative proposals

Sections 3 and 4 describe a mapping from statistical decision theory to the actual decisions recommended in our running example. As we discussed in Section 3.3, much of this mapping is spelled out in classical sources such as Wald (1950) and Lehmann (1959). We view this as a clear refutation of Claim A in the introduction applied to the aspects of empirical practice described in our running example.

We turn now to Claim B in the introduction, which stated that established decision theoretic motivations for empirical practice are not entirely satisfactory. To examine this claim, we first discuss interpretations of statistical decision theory and its role in empirical practice that have been proposed in the literature (Sections 5.1, 5.2 and 5.3). We then turn to various forms of Claim B, and discuss which aspects of these interpretations they find unsatisfactory.

5.1 Savage's prescriptive interpretation

One possible interpretation of statistical decision theory is that the Bayes, minimax or minimax regret criterion should be used to arrive at a decision that is "best" or "recommended." Perhaps the strongest argument for this point among Wald's contemporaries was given by Savage (1954).[4]

Savage argued for the Bayes criterion by showing that certain behavioral axioms imply the existence of a prior π and a loss function such that preferences are characterized by Bayes risk. Economists may recognize this argument as an axiomatic foundation of expected utility theory. In expected utility theory, $L(\theta, \delta(Y))$ plays the role of the Bernoulli utility function, and Bayes risk $R_{\text{Bayes}}(\pi, \delta)$ plays the role of expected utility (multiplied by -1 so that larger values are preferred) when uncertainty is characterized by the prior π (see, e.g., Mas-Colell et al., 1995, Ch. 6). Thus, one can interpret Savage's argument as stating that researchers should behave like "rational economic agents" when reporting results.

Perhaps less well known among economists today is Savage's justification of the minimax criterion as a solution to a *group decision problem* in the same book (Savage, 1954, Ch. 8-17, see in particular Ch. 10). This group decision interpretation of minimax has been explored in recent work by Banerjee et al. (2020b). The argument involves a thought experiment in which individuals with different priors π agree ex ante to a single decision rule δ. An

[4]In a review of Savage (1954), Anscombe (1956) states: "So far as the reviewer is aware, this is the only intelligent attempt at justifying the minimax principle made by anyone...." See Brown (1994) for further discussion of Savage's arguments and their role in justifying minimax as an "objective" criterion.

individual with prior π evaluates a decision using the Bayes risk $R_{\text{Bayes}}(\pi, \delta)$. If the group includes individuals with all priors π supported on Θ, then the maximum Bayes risk over individuals in the group is equal to the maximum risk $R_{\max}(\Theta, \delta)$. The minimax criterion can then be interpreted as choosing δ to ensure that the maximum Bayes risk (or negative of the minimum expected utility) faced by any member of the group is as small as possible.[5]

When does this interpretation justify using the minimax rule to arrive at a group decision rule $\delta^*_{\text{minimax}} = \arg\min_\delta R_{\max}(\Theta, \delta)$? According to Savage (p. 174), "it cannot be expected that the group minimax rule will, or reasonably should, be accepted by every group faced with every problem." However, "it may happen that, if $[R_{\max}(\Theta, \delta^*_{\text{minimax}})]$ is small, in a rather vague sense, the group will accept the group minimax rule." Since the maximum Bayes risk is small, "no member will feel that the suggestion is a serious mistake." Since the decision rule minimizes the maximum Bayes risk, no one will be able to suggest an alternative that will be preferred to $\delta^*_{\text{minimax}}$ by all members of the group if the minimax rule is unique.

How does this multiperson interpretation of the minimax criterion apply to the process of scientific investigation and publication? The following passage from Savage (1954, p. 175) offers one possibility:

> ... in many situations in which I envisage application of the group minimax principle, the group will in fact be a rather nebulous body of people, for example the group of all specialists in some field. The principle would in such a case be administered by a single member of the group somewhat in the following fashion. In planning an investigation, the results of which he intends to publish, he will endeavor to take account of all opinions, so far as he can know or guess them, that are considered at all reasonable in his field of investigation. And when he publishes his results he will say, in effect, "Whatever reasonable opinions have heretofore been held by members of this specialty, in the light of my investigation and the minimax rule, it is now proper for the members of the specialty, in so far as they are called upon to act in concert, to agree to such and such an action."

To apply this interpretation to our running example as formalized in Section 4, we might imagine a group of specialists who have different priors about the effect size $\mu_1 - \mu_0$. Some strongly believe the null hypothesis $H_0 : \mu_1 - \mu_0 = 0$, while others believe that an effect exists. The researcher conducting the study acts on behalf of this group of specialists to

[5]While we focus on minimax risk, a similar argument can be applied to minimax regret. As noted by Brown (1994), Savage's arguments can be interpreted more generally as a justification for the Γ-minimax criterion $\sup_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta)$, where we interpret Γ as the set of priors held by individuals in the group.

make a decision about whether to state that a nonzero effect exists. While members of the group have different priors about the effect size, they agree on the following utility function:

outcome	utility
incorrectly claiming a positive effect	-1 utile
incorrectly claiming a zero effect when effect is at least τ_{MDE}	$-\alpha/(1-\beta)$ utiles
making either claim when the effect is strictly between 0 and τ_{MDE}	0 utiles

The researcher proposes to make the group decision using the minimax decision rule, which leads to the test that rejects H_0 when $\bar{Y}_1 - \bar{Y}_0 > \text{se} \cdot z_{1-\alpha}$, as described in Section 4. Savage's arguments suggest that this decision rule will be accepted by the group ex ante if the risk is small enough. As described in Section 4.2, the analysis of statistical power used to determine sample size gives bounds on type I and type II error that are equivalent to an upper bound of α on the risk of this decision rule under the loss function described above. By using the test that corresponds to the minimax decision rule, the researcher minimizes the sample size needed to achieve this bound on risk.

While Savage (1954) does not spell out the details of a formal model of strategic interactions that would lead the group of specialists to accept this decision rule, one possibility is to assume that the researcher makes a take-it-or-leave-it offer to each member of the group, and that each member has an outside option worth $-\alpha$ utiles. If the researcher's goal is to ensure unanimous ex ante agreement while minimizing the cost of the study by making the sample size as small as possible, then the researcher will propose the test described above and the group of specialists will unanimously accept.

Putting other details aside, a key assumption of Savage's group decision-making interpretation is that the group of specialists makes a *binding ex-ante agreement*. After the experiment has been run, some members of the group will have ex-post regret about accepting the decision rule: an individual with a prior that places high enough prior probability on H_0 (H_1) will wish to reneg when the test rejects (fails to reject) ex-post. Savage (1954, Section 10.4) provides a discussion and an example illustrating this point. Recall that, in our running example, the researcher conducting the trial commits to the trial design in a preanalysis plan and may even submit the plan to a journal as a registered report. These steps can be interpreted as an attempt to strengthen the credibility that a binding commitment has been made, both by the researcher conducting the trial and the journal publishing the results.

5.2 Alternatives to the Savage axioms

Numerous critiques have been made of the Savage axioms and their implication that "rational" single person decisions should be consistent with the formation of prior probabilities. An early example is Ellsberg (1961), who argued through a particular example (now called the "Ellsberg paradox") that a reasonable person may not behave according to the Savage axioms. A large literature has considered alternative behavioral axioms that deal with such critiques. Gilboa and Schmeidler (1989) consider behavioral axioms that are consistent with minimax as a single person decision criterion. Stoye (2012) reviews axiomatizations that lead to related criteria used in statistics including minimax regret and provides further references. Gilboa and Marinacci (2013) provide a comprehensive review and historical discussion. Whereas the group decision interpretation of minimax in Savage (1954) requires an additional step of group consensus to motivate the minimax criterion, axiomatizations from this literature can be used to motivate minimax and other non-Bayesian criteria for a single decision maker.

5.3 Broader uses of statistical decision theory

The prescriptive interpretations in Sections 5.1 and 5.2 suggest using statistical decision theory as a way of arriving at a "recommended decision" after certain primitives have been agreed upon. Brown (2000) suggests a broader interpretation:

> ... the spirit of decision theory is pervasive in contemporary statistical research. Common manifestations include both mathematical and numerical attempts to check the frequentist performance of proposed procedures. This includes comparative investigations of level and power for hypothesis tests or of precision of proposed estimators as, for example, might occur in a Monte Carlo comparison of variances and biases.

Thus, one may report tests along with their size and power or estimates along with their risk, without fully specifying the criterion that should be used to choose between them. Birnbaum (1977) argues that reporting outcomes of statistical hypothesis tests along with their size and power can be a useful way to assess the evidence about the given hypothesis.

As discussed in Section 5.1, a strict application of Savage's group decision interpretation of the minimax criterion requires that decision rules be fully specified and agreed upon in advance. Such interpretations favor inflexible publication norms, requiring explicit preanalysis plans for every published result. Banerjee et al. (2020a) advocate for a more balanced

approach, in which some room is left for exploratory analysis beyond the preanalysis plan. Hypothesis testing and other procedures with ex ante decision theoretic guarantees may still be used in such analysis, with the caveat that an ex ante commitment to a particular decision rule has not actually been made.

5.4 Objections to "behavioral" interpretations of statistics

The decision theoretic interpretations described above treat a hypothesis test as a behavioral rule. In the original article introducing the theory, Neyman and Pearson (1933a) put it this way:

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

Neyman saw this interpretation as applicable "whether [the] study is conducted for purposes of some immediate practical action . . . or for the sake of scientific curiosity" (Neyman, 1957, p. 14). Thus, decision theory can be applied equally to the pursuit of pure science as well as mundane economic tasks. Neyman used the term "inductive behavior" for this interpretation of the process of scientific induction (Neyman, 1957).

R.A. Fisher took issue with this interpretation of scientific practice. How could the lofty ideals of scientific discovery be equated with the mundane business of running a factory or firm? Fisher saw the Neyman-Pearson theory and subsequent developments by Wald as a corruption of the hypothesis testing theory he had pioneered. Fisher's views on this topic represent a particularly strong form of Claim B: according to Fisher, the entire premise of using decision theory to aid in the process of scientific discovery is fundamentally flawed. On the other hand, this critique may not bother policy makers who seek to use the results of an empirical study to make economic decisions, rather than for scientific discovery per se. Further discussion and references for the views of Neyman and Fisher on this topic can be found in Lehmann (2011, Section 4.6).

What alternatives to decision theory does this leave us with? One alternative viewpoint is that researchers should focus on summarizing and communicating their data to readers. This viewpoint is in line with Fisher's statement that "the object of statistical methods is the reduction of data" (Fisher, 1922). Andrews and Shapiro (2021) consider a model in which a researcher reports a function $c(Y)$ of the data Y in the decision theoretic setup

described in Section 3. The departure point from the classical decision theory is that an audience of individuals with different priors or loss functions is free to use this reported information to make different decisions to optimize Bayes risk for their respective decision environments. The authors explore ways in which this leads to prescriptions that differ from those of classical decision theory for reporting a single decision. The question of how to communicate data has also been addressed using data visualization and exploratory data analysis, an approach advocated by Tukey (1977).

While communication and data reduction are worthy goals, they do not speak directly to the question of how to perform scientific induction. Fisher saw "inductive logic" or reasoning "from the particular to the general" as an important part of the theory of statistics (Fisher, 1955). How can one formalize this process without the reference to statistical decisions used by the "inductive behavior" interpretation of Neyman? One possibility is to interpret a Bayesian posterior distribution as encoding inferences that can be made from data. Fisher, however, disliked the subjectivism of Bayesian approaches. A theory of statistical inference that avoids subjective Bayesianism while also avoiding ex ante statements about the performance of behavioral rules remains elusive. Fisher's own attempt at such a theory, the theory of fiducial inference, is now seen as lacking internal consistency: as Efron (1998) states, it is "generally considered to be Fisher's greatest blunder," albeit one that may contain the seeds of interesting ideas. The *objective Bayes* literature has studied how to minimize the subjectivism of the prior specification by defining and constructing objective priors. The approaches to finding objective priors, going back to Jeffreys (1946), are reviewed in Kass and Wasserman (1996).

5.5 Objections to loss functions used in hypothesis testing

The costs and benefits of implementing an economic policy based on an RCT will typically vary with the effect size $\mu_1 - \mu_0$ in ways not captured by the zero-one hypothesis testing loss function (1). A recent literature in econometrics starting with Manski (2004), using a minimax regret approach, and Dehejia (2005), using a Bayesian approach, explicitly models the losses in our running example of an RCT used to determine which treatment is more effective so that it can be implemented.

Assume that treatment $t = 1$ is implemented if $\delta(Y) = 1$ and $t = 0$ is implemented if $\delta(Y) = 0$. If the outcome variable Y_i measures the welfare of individual i net of relevant costs, then the population welfare resulting from implementing decision $\delta(Y)$ is $W((\mu_0, \mu_1), \delta) = \mu_0 + (\mu_1 - \mu_0)\delta(Y)$. The minimax regret criterion, defined in Section 3.1, considers the

difference between the welfare of decision δ and the welfare $\mu_0 + (\mu_1 - \mu_0) \cdot 1(\mu_1 > \mu_0)$ obtained from the oracle decision:

$$L((\mu_0, \mu_1), \delta) = (1 - \delta)(\mu_1 - \mu_0) \cdot 1(\mu_1 > \mu_0) + \delta(\mu_0 - \mu_1) \cdot 1(\mu_0 > \mu_1). \tag{2}$$

A minimax regret decision is then given by a minimax decision for the loss function (2).

Two features distinguish (2) from the hypothesis testing loss (1). First, it is *symmetric*. For a given magnitude of the treatment effect $|\mu_1 - \mu_0|$, a suboptimal treatment decision yields the same loss whether $\mu_1 > \mu_0$ (H_1) or $\mu_0 > \mu_1$ (H_0). Because of this symmetry, the minimax regret decision rule is $\delta^*_{\text{minimax}}(Y) = 1(\bar{Y}_1 > \bar{Y}_0)$.[6]

Second, the loss is proportional to the magnitude of the treatment effect $\mu_1 - \mu_0$. Given the symmetry of (2), proportionality to the treatment effect does not itself affect the resulting decision rule. However, we can also modify (1) to account for the magnitude of the treatment effect without imposing symmetry:

$$L((\mu_0, \mu_1), \delta) = (1 - \delta)(\mu_1 - \mu_0) \cdot 1(\mu_1 > \mu_0) + K\delta(\mu_0 - \mu_1) \cdot 1(\mu_0 > \mu_1). \tag{3}$$

This loss function has been used in Tetenov (2012), Hirano and Porter (2009), and Banerjee et al. (2020b) to capture stronger dislike of making a decision that results in welfare lower than μ_0, which is the average outcome of the status quo treatment. The minimax regret decision takes the form $1(\bar{Y}_1 - \bar{Y}_0 > T(K))$, where $T(K)$ is a threshold characterized by Tetenov (2012).

How do these proposals differ from standard practice in our running example? As discussed in Section 4.1, standard practice in our running example corresponds to the size α test $1(\bar{Y}_1 - \bar{Y}_0 > \text{se} \cdot z_{1-\alpha})$. The minimax regret rule $1(\bar{Y}_1 > \bar{Y}_0)$ under the symmetric loss function (2) corresponds to $\alpha = .5$. To obtain the threshold $\text{se} \cdot z_{.95}$ corresponding to a size $\alpha = .05$ test, one must consider the asymmetric loss function (3) with $K = 102.4$ (Tetenov, 2012). The parameter K captures the asymmetry in the costs of type I and type II errors, similar to the ratio $w_I/w_{II} = (1 - \beta)/\alpha$ in Section 4.2. However, the value $K = 102.4$ is much larger than the ratio $w_I/w_{II} = 4$ that corresponds to the benchmark choices $\alpha = .05$ and $\beta = .8$ that motivated the same decision in our running example. Threshold rules derived from classical hypothesis testing loss also differ from decision rules derived from the loss functions (2) and (3) due to how economic costs are incorporated into the outcome Y_i:

[6] See Allen (1953), Hirano and Porter (2009) and Tetenov (2012). Stoye (2009) shows the same is true (up to tie-breaking) for binary outcomes with equal sample sizes $n_0 = n_1$.

in our running example, we did not explicitly incorporate treatment costs into our outcome variable or when defining the null hypothesis.

In our running example, we used the hypothesis testing loss function (1) not only to form a decision rule, but also to provide ex ante risk guarantees on this rule. These risk bounds were obtained using statistical power analysis, and they were used to determine the sample size. Such risk bounds are also subject to criticisms of the hypothesis testing loss function (1). For example, they ignore the costs of making an error when the effect size is nonzero but smaller than the MDE. Manski and Tetenov (2016) propose using risk bounds under the welfare regret loss (2) to choose the sample size, analogous to the role of statistical power analysis for hypothesis testing. If the size of the population to be treated is known, trial size can be chosen optimally to account for the losses both in the trial sample and among those who are treated based on its conclusions (cf. Colton (1963) and references therein).

In addition to this recent econometrics literature, welfare based loss functions such as (2) have been standard in the literature on multi-arm bandits (cf. Berry and Fristedt (1985), Bubeck and Cesa-Bianchi (2012)). Such loss functions also appear in earlier references such as Wald (1950, p. 9) and Allen (1953), albeit without explicit arguments deriving them from welfare or economic profit.

Other authors have critiqued the hypothesis testing loss function used in our running example without explicitly proposing the loss function (2). An early critique of the asymmetry of standard hypothesis tests is given in Simon (1945). Arrow (1960) also argued against the asymmetry of type I/II errors, proposing instead to choose an "economically significant difference" at which the researcher judges "Type II error to be just as costly as a Type I error." Ziliak and McCloskey (2008) criticized the practice of reporting hypothesis tests of an exact zero effect, thereby conflating statistical and economic significance. The American Statistical Association has gone so far as to issue a statement warning, among other things, that "statistical significance is not equivalent to scientific, human, or economic significance" (Wasserstein and Lazar, 2016).

Notwithstanding the above critiques, some recommendations from Duflo et al. (2007) and other sources for standard practice in our running example do tie certain aspects of the hypothesis testing loss function to economic primitives. For example, Duflo et al. (2007, p. 3927) suggest choosing the MDE as "the smallest effect size that is large enough such that the intervention would be cost effective if it were to be scaled up."[7] In practice, however,

[7]Note that this may lead to different decision rules than the loss function (2). For example, if $\alpha = 1 - \beta$ so that type I and type II error are symmetric, then one needs to set the MDE to *twice* the cost of treatment to get the decision rule that is optimal under (2) with cost incorporated into the outcome.

incentives to get a study funded and published play a role in such decisions: the MDE is often reverse engineered to claim adequate power at a sample size determined by budget or subject recruitment constraints (Detsky, 1990; Schulz and Grimes, 2005). We turn to the subject of strategic behavior by researchers and other agents in the next section.

5.6 Lack of explicit modeling of strategic interactions

The group decision interpretation of the minimax rule, proposed in Savage (1954) and discussed in Section 5.1, involves a group of specialists agreeing ex ante to a single decision rule. As discussed in Section 5.1, this interpretation is compatible with a simple model in which specialists with different priors and the same objective function are given a take-it-or-leave-it offer to agree to a decision rule ex ante. However, this model does not explicitly address aspects of the research process which may be important in practice, such as asymmetric information, incentives, and economic interactions between agents.

To address these issues, a recent literature has proposed models that explicitly incorporate these aspects of the research process. A key insight of this literature is that statistical decision rules create incentives for researchers to change how they collect and analyze data. Tetenov (2016) studies regulatory approval decisions based on evidence generated by RCTs by introducing strategic interactions between the regulator and proponents of innovations (e.g., pharmaceutical firms). Given the regulator's commitment to a statistical protocol for approval decisions, the proponent, who is privately informed of the true treatment effect θ of their innovation, decides whether to conduct a trial by maximizing the expected payoff of approval net of the trial costs. If they perform the trial, they send statistical evidence $X \sim P_{X|\theta}$ to the regulator for approval. The regulator, who faces uncertainty about θ, optimizes the statistical approval protocol by taking into account the proponent's optimal response to it. Tetenov (2016) shows that tests which control the probability of type I error by the ratio of proponent's private benefit from approval to sunk trial costs are minimax optimal for the regulator. Bates et al. (2022) consider a similar framework in which the regulator can set the proponent's payoff as a function of the data. Viviano et al. (2022) extend the framework of Tetenov (2016) to settings with multiple treatments or outcomes. McCloskey and Michaillat (2023) propose a model of p-hacking in which researchers, interested in publishing statistically significant findings, optimize their costly hidden data collection in response to the prevailing testing standards. Spiess (2022) studies how a regulator seeking to minimize the mean squared error of the average treatment effect estimator should ex ante constrain estimators available to a researcher with private information and misaligned preferences.

Another strand of literature considers how communication constraints shape statistical reporting of results even when preferences of researchers and policy makers are aligned. Frankel and Kasy (2022) model a journal editor who weighs the benefits of publishing the results of a study against the costs of publication (or of reader's attention). The policy maker then updates their prior about the value of a policy based either on published results or on the signal that no results were published, taking into account the journal's publishing rule. To be published, the results of the study must change the policy maker's beliefs sufficiently to change the policy decision and offset the publication costs. The optimal publication rule for treatment effects estimates takes the form of a one-sided test whose stringency is determined by the prior, the study's standard error, and publication costs. Furukawa (2021) considers a group of researchers who estimate the same treatment effect in independent studies. Each of them can only communicate a restricted signal (positive, negative, or omitted) to the policy maker who then takes a binary action. While the preferences of all researchers and the policy maker are aligned, the equilibrium of this game results in publication bias.

Recent papers such as Kitagawa and Tetenov (2018) study allocation of treatments based on heterogeneous observable covariates of individuals. Such treatment rules can lead to incentives for individuals to strategically modify or misreport their covariates to obtain treatment assignment beneficial to them. Munro (2023) studies a general model of strategic covariate manipulation and shows that optimal treatment allocation can require randomized assignment of individuals reporting identical covariates. Sahoo and Wager (2022) consider strategic interaction between potential treatment recipients who compete (through covariate manipulation) for a desirable treatment when the scoring system is known, but the score cutoff is determined ex post by the distribution of observed covariates to satisfy a capacity constraint.

6 Conclusion

The general framework of statistical decision theory encompasses hypothesis testing, estimation, statistical power analysis and other aspects of empirical practice. Indeed, a mapping between these procedures and the general framework of statistical decision theory, which we reviewed in Section 3, was largely spelled out by Wald in his book introducing the theory (Wald, 1950), and in other contemporary sources. What is less clear is whether this mapping provides a satisfactory description of the role of decision theory in empirical practice or, more broadly, what this role should be. Rather than attempting to answer these questions

ourselves, our goal has been to provide an organizing discussion of a subset of the literature in statistics, economics and econometrics that has debated this topic. We hope this article has provided a useful introduction to these debates.

References

Allen, Jr., S. G. (1953): "A Class of Minimax Tests for One-sided Composite Hypotheses," *The Annals of Mathematical Statistics*, 24, 295–298.

Andrews, I. and J. M. Shapiro (2021): "A Model of Scientific Communication," *Econometrica*, 89, 2117–2142.

Anscombe, F. J. (1956): "Review of The Foundations of Statistics." *Journal of the American Statistical Association*, 51, 657–659, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Arrow, K. J. (1960): "Decision theory and the choice of a level of significance for the t-test," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. by I. Olkin, Stanford University Press, Stanford studies in mathematics and statistics, 70–78.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020a): "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics," .

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020b): "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*, 110, 1206–1230.

Banerjee, A. V., S. Chassang, and E. Snowberg (2017): "Chapter 4 - Decision Theoretic Approaches to Experiment Design and External Validity," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Field Experiments*, 141–174.

Bates, S., M. I. Jordan, M. Sklar, and J. A. Soloff (2022): "Principal-Agent Hypothesis Testing," *arXiv*.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, New York: Springer, 2nd ed. 1985. corr. 3rd printing 1993 edition ed.

Berry, D. and B. Fristedt (1985): *Bandit Problems: Sequential Allocation of Experiments*, Monographs on Statistics and Applied Probability, Springer Netherlands.

Birnbaum, A. (1977): "The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-savage argument for Bayesian theory," *Synthese*, 36, 19–49.

Bogdanoski, A., A. Foster, D. Karlan, and E. Miguel (2020): "Pre-results Review at the Journal of Development Economics: Lessons learned," MetaArXiv 5yacr, Center for Open Science.

Brown, L. D. (1994): "Minimaxity, More or Less," in *Statistical Decision Theory and Related Topics V*, ed. by S. S. Gupta and J. O. Berger, Springer New York, 1–18.

——— (2000): "An Essay on Statistical Decision Theory," *Journal of the American Statistical Association*, 95, 1277–1281.

Bubeck, S. and N. Cesa-Bianchi (2012): *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*, Foundations and Trends® in Machine Learning Series, Now Publishers.

Colton, T. (1963): "A Model for Selecting One of Two Medical Treatments," *Journal of the American Statistical Association*, 58, 388–400.

Dehejia (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

Detsky, A. S. (1990): "Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials," *Statistics in Medicine*, 9, 173–184.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Chapter 61 Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, ed. by T. P. Schultz and J. A. Strauss, Elsevier, vol. 4, 3895–3962.

Efron, B. (1998): "R. A. Fisher in the 21st Century," *Statistical Science*, 13, 95–114.

Ellsberg, D. (1961): "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 75, 643–669.

Fisher, R. (1955): "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society. Series B (Methodological)*, 17, 69–78, publisher: [Royal Statistical Society, Wiley].

Fisher, R. A. (1922): "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.

Frankel, A. and M. Kasy (2022): "Which Findings Should be Published," *American Economic Journal: Microeconomics*, 14, 1–38.

Furukawa, C. (2021): "Publication Bias Reexamined: A NewCommunication Model and Correction Method," *unpublished*.

Giacomini, R., T. Kitagawa, and M. Read (2021): "Robust Bayesian Analysis for Econometrics," *CEPR Discussion paper*.

Gilboa, I. and M. Marinacci (2013): "Ambiguity and the Bayesian Paradigm," in *Advances in Economics and Econometrics: Volume 1, Economic Theory: Tenth World Congress*, Cambridge University Press, vol. 49, 179.

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

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

——— (2020): "Chapter 4 - Asymptotic analysis of statistical decision rules in econometrics," in *Handbook of Econometrics*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics, Volume 7A*, 283–354.

J-PAL (2023): "Introduction to randomized evaluations," .

Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186, 453–461, publisher: Royal Society.

Kass, R. E. and L. Wasserman (1996): "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Kitagawa, T. and A. Tetenov (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

Lehmann, E. L. (1959): *Testing Statistical Hypotheses*, New York: John Wiley & Sons, Inc.

——— (2011): *Fisher, Neyman, and the Creation of Classical Statistics*, New York, NY: Springer.

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses*, Springer.

List, J. A. (2023): "Editor's Introduction to JPE Micro," *Journal of Political Economy Microeconomics*, 1, 1–6.

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

——— (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17985.

Manski, C. F. and A. Tetenov (2016): "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 113, 10518–10523.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, New York: Oxford University Press, 1 edition ed.

McCloskey, A. and P. Michaillat (2023): "Critical Values Robust to P-hacking," *arXiv*.

Munro, E. (2023): "Treatment Allocation with Strategic Agents," *Management Science*.

Müller, U. K. and A. Norets (2016): "Credibility of Confidence Sets in Nonstandard Econometric Problems," *Econometrica*, 84, 2183–2213.

Neyman, J. (1957): ""Inductive Behavior" as a Basic Concept of Philosophy of Science," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 25, 7–22.

Neyman, J. and E. S. Pearson (1933a): "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231, 289–337.

——— (1933b): "The testing of statistical hypotheses in relation to probabilities a priori," *Mathematical Proceedings of the Cambridge Philosophical Society*, 29, 492–510.

Pratt, J. W. (1961): "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56, 549–567.

Ríos Insua, D. and F. Ruggeri, eds. (2000): *Robust Bayesian Analysis*, New York, NY: Springer.

Sahoo, R. and S. Wager (2022): "Policy Learning with Competing Agents," *arXiv*.

Savage, L. J. (1954): *The Foundations of Statistics*, John Wiley & Sons.

Schulz, K. F. and D. A. Grimes (2005): "Sample size calculations in randomised trials: mandatory and mystical," *The Lancet*, 365, 1348–1353.

Simon, H. A. (1945): "Statistical Tests as a Basis for "Yes-No" Choices," *Journal of the American Statistical Association*, 40, 80–84.

Sims, C. (2008): "Improving Monetary Policy Models," *Journal of Economic Dynamics and Control*, 32, 2460–2475.

Spiess, J. (2022): "Optimal Estimation when Researcherand Social Preferences are Misaligned," *unpublished*.

Stoye, J. (2009): "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70–81.

——— (2012): "New Perspectives on Statistical Decisions Under Ambiguity," *Annual Review of Economics*, 4, 257–282, _eprint: https://doi.org/10.1146/annurev-economics-080511-110959.

Tetenov, A. (2012): "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," *Journal of Econometrics*, 166, 157–165.

——— (2016): "An Economic Theory of Statistical Testing," *cemmap working paper*.

Tukey, J. (1977): *Exploratory Data Analysis*, Reading, Mass: Pearson, 1st edition ed.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge, UK ; New York, NY, USA: Cambridge University Press.

Viviano, D., K. Wüthrich, and P. Niehaus (2022): "(When) Should you Adjust Inferences for Multiple Hypothesis Testing?" *arXiv*.

Wald, A. (1950): *Statistical decision functions.*, Wiley publications in statistics, New York: Wiley.

Wasserstein, R. L. and N. A. Lazar (2016): "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133.

Ziliak, S. and D. McCloskey (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Economics, Cognition, And Society, University of Michigan Press.

A Details of randomized sampling

This appendix provides a formal discussion of the random sampling scheme and statements regarding minimaxity of randomized sampling made in Section 4.2.

A.1 Setup

We consider a finite population potential outcomes model, in which a pool of individuals $j = 1, \ldots, N$ with potential outcomes $Y_j(0)$ and $Y_j(1)$ is available for the trial. These potential outcomes are modeled as nonrandom. To form our sample, we draw $n_0 + n_1$ random numbers $j(1), \ldots, j(n_0+n_1)$ independently from the uniform distribution on the set $\{1, \ldots, N\}$, and we set $Y_i = Y_{j(i)}(0)$ (i.e. we sample unit $j(i)$, assign it to non-treatment, and record the outcome as Y_i) for $i = 1, \ldots, n_0$ and $Y_i = Y_{j(i)}(1)$ (i.e. we sample unit $j(i)$, assign it to treatment, and record the outcome as Y_i) for $i = n_0 + 1, \ldots, n_0 + n_1$. This leads to $Y_1, \ldots, Y_{n_0}$ and $Y_{n_1+1}, \ldots, Y_{n_0+n_1}$ being sampled iid with expectations μ_0 and μ_1 respectively, where $\mu_1 - \mu_0$ is the average treatment effect $\frac{1}{N}\sum_{j=1}^{N}[Y_j(1) - Y_j(0)]$ in the potential outcome model. This is the model used in the main text, without the normality assumption. In practice, we draw $j(1), \ldots, j(n_1 + n_0)$ without replacement: we sample each $j(i)$ only from the remaining indices not equal to $j(i')$ for any $i' < i$. This incorporates the constraint that the researcher cannot sample both $Y_j(0)$ and $Y_j(1)$ for the same individual j. The iid sampling scheme is an approximation that is accurate when N is large relative to n_0 and n_1.[8]

To formally describe results on minimaxity of randomization and necessity of randomization for forming nontrivial tests, we extend our framework to include the sampling decision, and to allow for randomization. A decision function $\delta()$ maps the potential outcomes $Y = \{Y_j(0), Y_j(1)\}_{j=1}^{N}$ to a decision to reject ($\delta(Y) = 1$) or fail to reject (coded as $\delta(Y) = 0$). The researcher is constrained to sample only n_0 units for non-treatment and n_1 units for treatment. Formally, this means that a nonrandomized decision takes the form $\delta(Y) = \tilde{\delta}(\{Y_j(0)\}_{j\in\mathcal{J}_0}, \{Y_j(1)\}_{j\in\mathcal{J}_1}; \mathcal{J}_0, \mathcal{J}_1)$, where $\mathcal{J}_0$ and $\mathcal{J}_1$ are nonoverlapping sets with cardinality n_0 and n_1 respectively. A randomized decision $\delta(Y, U)$ then randomizes over such de-

[8]Note that randomization plays two roles here: the overall sample $\mathcal{J} = \{j(1), \ldots, j(n_0+n_1)\}$ is a random sample of the population $\{1, \ldots, N\}$ and the sample of treated individuals $\mathcal{J}_1 = \{j(n_0+1), \ldots, j(n_0+n_1)\}$ is a random subset of the sample $\mathcal{J}$. If individuals who show up for the experiment cannot be randomly sampled from a relevant population, one may take the n_0+n_1 individuals to be the entire sample. In this case only the latter source of randomization is present. This leads to some differences in the analysis, although basic points made here such as minimaxity of randomization and necessity of randomization for nontrivial tests still hold.

cisions, thereby taking the form $\delta(Y,U) = \tilde{\delta}(\{Y_j(0)\}_{j\in\mathcal{J}_0(U)}, \{Y_j(1)\}_{j\in\mathcal{J}_1(U)}; \mathcal{J}_0(U), \mathcal{J}_1(U); U)$. The entire set of potential outcomes $Y = \{Y_j(0), Y_j(1)\}_{j=1}^N$ plays the role of the unknown parameter θ. Letting $\text{ATE}(Y) = \frac{1}{N}\sum_{j=1}^N [Y_i(1) - Y_i(0)]$ denote the average treatment effect, the null hypothesis is $H_0 : \text{ATE}(Y) = 0$, and we consider the power under the alternative $H_1 : \text{ATE}(Y) \geq \tau_{\text{MDE}}$. The weighted loss function $L(Y, \delta(Y,U))$ that corresponds to the minimax testing problem is equal to w_I if $\text{ATE}(Y) = 0$ and $\delta(Y,U) = 1$, w_{II} if $\text{ATE}(Y) \geq \tau_{\text{MDE}}$ and $\delta(Y,U) = 0$ and 0 otherwise. The risk function is given by

$$R(Y,\delta) = \int L(Y, \delta(Y,u))\, dP_U(u)$$

where P_U is the (known) distribution of the randomization device U. The minimax criterion optimizes the worst-case risk

$$R_{\max}(\mathcal{Y},\delta) = \sup_{Y\in\mathcal{Y}} \int L(Y, \delta(Y,u))\, dP_U(u)$$

where $\mathcal{Y}$ is the set of configurations of $Y = \{Y_j(0), Y_j(1)\}_{j=1}^N$ that are deemed plausible. Here $\mathcal{Y}$ plays the role of the parameter space Θ in the main text.

A.2 Necessity of randomized sampling for a nontrivial test

We now describe a result that shows that randomized sampling is necessary to form a test that does better than a random guess that does not use the data. The result adapts the discussion in Lehmann and Romano (2005, Section 5.10, pp. 181-183) to the potential outcome setting with purely design-based randomization considered here (the result appears in Section 5.9 in the original edition Lehmann 1959). Since tests with nontrivial minimax power are possible in this setting with randomized sampling and treatment assignment, this implies that the minimax procedure (test and sampling scheme) uses randomized sampling. Earlier results on minimaxity of randomized sampling in a purely design based setting with sampling from a single population can be found in Savage (1954, Ch. 14, Section 8).

For simplicity, suppose that $Y_j(0)$ and $Y_j(1)$ are known to be in some set $\mathcal{Z}$. Let $\mathcal{Y} = \mathcal{Z}^N$ be the set of all possible configurations of the potential outcomes. Let $\delta(Y,U) = \tilde{\delta}(\{Y_j(0)\}_{j\in\mathcal{J}_0(U)}, \{Y_j(1)\}_{j\in\mathcal{J}_1(U)}; \mathcal{J}_0(U), \mathcal{J}_1(U); U)$ be a procedure that does not involve randomized sampling. Formally, this means that $\mathcal{J}_0(U)$ does not depend on U and is equal to some fixed set $\mathcal{J}_0$, and similarly for $\mathcal{J}_1$. It can be seen that no such decision rule can have power greater than its size at *any* $Y^{\text{alt}} = \left\{Y^{\text{alt}}(0), Y^{\text{alt}}(1)\right\}_{j=1}^N$.

Indeed, consider any population $Y^{\text{alt}} = \left\{Y_j^{\text{alt}}(0), Y_j^{\text{alt}}(1)\right\}_{j=1}^{N}$ in the alternative H_1. Since $\mathcal{J}_0$ and $\mathcal{J}_1$ are nonoverlapping, there exists another population Y^{null} satisfying the null H_0 such that $\tilde{\delta}(Y^{\text{null}}, \mathcal{J}_0, \mathcal{J}_1, U) = \tilde{\delta}(Y^{\text{alt}}, \mathcal{J}_0, \mathcal{J}_1, U)$, so that the decision under this distribution in the null H_0 is identical to the one made under the distribution Y^{alt} in the alternative H_1. In particular, we can construct Y^{null} by starting with Y^{alt} and changing the $Y_j(1)$ observations in $\mathcal{J}_0$ to be identical to $Y_j(0)$, and similarly for the set $\mathcal{J}_1$ (for $j \notin \mathcal{J}_0 \cup \mathcal{J}_1$, we can set $Y_j(0) = Y_j(1)$ to be any value in $\mathcal{Z}$ we choose).

Thus, without a sampling scheme that involves randomization of some form, it is impossible to form a test that controls size while achieving nontrivial power at any alternative. In contrast, using the randomization scheme described above in which $\mathcal{J}_0$ and $\mathcal{J}_1$ are chosen completely at random from all nonoverlapping sets, one can form a test with size arbitrarily close to zero and power arbitrarily close to one for all Y satisfying $H_1 : \text{ATE}(Y) \geq \tau_{\text{MDE}}$ for any fixed τ_{MDE} once n_0 and n_1 are large enough, so long as the set $\mathcal{Z}$ that contains the potential outcomes is bounded. For example, the test that rejects when $\bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} Y_i - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i \geq \tau_{\text{mde}/2}$ achieves this goal. It follows that any sampling scheme that does not involve randomization will be strictly suboptimal from a minimax perspective.