This paper provides the source material for my Haavelmo Lecture at the University of Oslo, December 2019. Earlier drafts were circulated under the title “Statistical Inference for Statistical Decisions.” I am grateful to Olav Bjerkholt, Ivan Canay, Gary Chamberlain, Kei Hirano, Joel Horowitz, Valentyn Litvin, Bruce Spencer, and Alex Tetenov for comments.
1. Introduction: Joining Haavelmo and Wald

Early in the modern development of econometrics, Trygve Haavelmo compared astronomy and planning to differentiate two objectives for econometric modeling: to advance science and to inform decision making. He wrote (Haavelmo, 1943, p. 10):

“The economist may have two different purposes in mind when he constructs a model . . . . First, he may consider himself in the same position as an astronomer, he cannot interfere with the actual course of events. So he sets up the system . . . . as a tentative description of the economy. If he finds that it fits the past, he hopes that it will fit the future. On that basis he wants to make predictions, assuming that no one will interfere with the game. Next, he may consider himself as having the power to change certain aspects of the economy in the future. If then the system . . . . has worked in the past, he may be interested in knowing it as an aid in judging the effect of his intended future planning, because he thinks that certain elements of the old system will remain invariant.”

Jacob Marschak, in a letter supporting Haavelmo’s work, made a related distinction between meteorological and engineering types of econometric inference; see Bjerkholt (2010).

Contrasting astronomy and planning provides a nice metaphor for two branches of econometrics. In 1943, before the advent of space flight, an astronomer might model a solar system or galaxy to advance physical science, but the effort could have no practical impact on decision making. An economist might similarly model a local or national economy to advance social science. However, an economist might also model to inform society about the consequences of contemplated public or private decisions that would change aspects of the economy.

Haavelmo’s seminal doctoral thesis (Haavelmo, 1944) proposed a formal probabilistic structure for econometric modeling, aiming to make econometrics useful for public decision making. To conclude, he wrote (p. 114-115):

“In other quantitative sciences the discovery of “laws,” even in highly specialized fields, has moved from the private study into huge scientific laboratories where scores of experts are engaged, not only in carrying out actual measurements, but also in working out, with painstaking precision, the formulae to be tested and the plans for crucial experiments to be made. Should we expect less in
economic research, if its results are to be the basis for economic policy upon which might depend billions of dollars of national income and the general economic welfare of millions of people?"

Haavelmo's thesis made fundamental contributions that have become thoroughly embedded in subsequent econometric research. Nevertheless, it is unsurprising to find that it did not fully answer all the deep issues that the author raised. Notably, Haavelmo struggled to formalize the implications for decision making of the fact that models can at most seek to approximate actuality. He called attention to the broad issue in his opening chapters on “Abstract Models and Reality” and “The Degree of Permanence of Economic Laws,” but the later chapters did not resolve the matter.

Haavelmo devoted a long chapter to “The Testing of Hypotheses,” expositing the then recent work of Neyman-Pearson and considering its potential use to evaluate the consistency of models with observed sample data. Testing models has subsequently become widespread in economics, both as a topic of study in econometric theory and as a practice in empirical research. However, Neyman-Pearson hypothesis testing does not provide satisfactory guidance for decision making. See Section 2.3 below.

When Haavelmo was writing his thesis, Abraham Wald was initiating his own seminal development of statistical decision theory in Wald (1939, 1945) and elsewhere, which later culminated in his own treatise (Wald, 1950). Wald's work has broad potential application. Indeed, it implicitly provides an appealing formal framework for evaluation of the use of models in decision making. I say that Wald "implicitly" provides this framework because, writing in an abstract mathematical manner, Wald appears not to have explicitly examined decision making with models. Yet it is conceptually straightforward to use statistical decision theory in this way. Explaining this motivates the present paper.

I find it intriguing to join the contributions of Haavelmo and Wald because these pioneering econometrician and statistician interacted to a considerable degree in the United States during the wartime period when both were developing their ideas. Wald came to the U.S. in 1938 as a refugee from Austria. Haavelmo did so in 1939 for what was intended to be a short-term professional visit, but which lasted the entire war when he was unable to return to occupied Norway. Bjerkholt (2007, 2015), in biographical essays on Haavelmo’s period in the United States, describes the many interactions of Haavelmo and Wald, not
only at professional conferences but also in hiking expeditions in Colorado and Maine. Bjerkholt observes that Haavelmo visited Neyman as well, the latter being in Berkeley by then.

Haavelmo’s appreciation of Wald is clear. In the preface of Haavelmo (1944), he wrote (p. v):

“My most sincere thanks are due to Professor Abraham Wald of Columbia University for numerous suggestions and for help on many points in preparing the manuscript. Upon his unique knowledge of modern statistical theory and mathematics in general I have drawn very heavily. Many of the statistical sections in this study have been formulated, and others have been reformulated, after discussions with him.”

The text of the thesis cites several of Wald’s papers. Most relevant is the final chapter on “Problems of Prediction,” where Haavelmo suggests application of the work in Wald (1939) and Mann and Wald (1943) to choose a predictor of a future random outcome. I discuss this further in Section 3.3 below.

Despite Haavelmo’s familiarity with and appreciation of Wald’s work, modern econometrics has applied statistical decision theory only to a limited degree. Instead, it has mainly focused on study of identification and statistical inference. Beginning in Manski (2004) and then in subsequent studies (Manski, 2005, 2007, 2019; Manski and Tetenov, 2007, 2014, 2016, 2019; Dominitz and Manski, 2017, 2019), I have argued that econometrics can and should make productive use of statistical decision theory. Until now, however, I have not specifically written on the potential use of statistical decision theory to evaluate models.

Section 2 reviews the core elements of statistical decision theory in abstraction and uses choice between two actions to illustrate. The basic idea is conceptually simple, although it may be computationally challenging to implement. One specifies a state space, listing all the states of nature that one believes feasible. One considers alternative statistical decision functions (SDFs), which map potentially observed data into decisions. One evaluates an SDF in each state of nature ex ante, by its mean performance across repeated samples. The true state of nature is not known. Hence, one evaluates the performance of an SDF across all the elements of the state space. Three decision criteria have drawn much attention: maximization of subjective expected welfare, the maximin criterion, and the minimax-regret criterion.

Section 3 shows how the Wald framework may be used to evaluate decision making with models. One specifies a model space, which simplifies or approximates the state space in some manner. A model-
based decision uses the model space as if it were the state space. I call attention to the widespread econometric practice of *as-if optimization*: specification of a model, point estimation of its parameters, and use of the point estimate to make a decision that would be optimal if the estimate were accurate. A central theme of statistical decision theory is that one should evaluate as-if optimization or any other model-based decision rule by its performance across the state space, not the model space.

Section 3 uses prediction of a real-valued outcome to illustrate. To illustrate further, Section 4 considers use of the empirical success (ES) rule in treatment choice. This application of as-if optimization is well-grounded when the data are generated by an ideal randomized trial. When the ES rule is used with observational data, it exemplifies a controversial modeling practice, wherein one assumes without good justification that realized treatments are statistically independent of treatment response. Decision-theoretic analysis shows when use of the ES rule with observational data does and does not yield desirable treatment choices.

Although statistical decision theory is conceptually simple, application is computationally challenging in many contexts. Section 5 cites advancement of computation as the primary task to continue building the foundations sketched by Haavelmo and Wald.

### 2. Statistical Decision Theory: Concepts and Practicalities

The Wald development of statistical decision theory directly addresses decision making with sample data. Wald began with the standard decision theoretic problem of a planner (equivalently, decision maker or agent) who must choose an action yielding welfare that depends on an unknown state of nature. The planner specifies a state space listing the states that he considers possible. He must choose an action without knowing the true state.

Wald added to this standard problem by supposing that the planner observes sample data that may be informative about the true state. He studied choice of a *statistical decision function (SDF)*, which maps
each potential data realization into a feasible action. He proposed evaluation of SDFs as procedures, chosen prior to observation of the data, specifying how a planner would use whatever data may be realized. Thus, Wald's theory is frequentist.

I describe general decision problems without sample data in Section 2.1 and with data in Section 2.2. Section 2.3 examines the important special case of decisions that choose between two actions. Section 2.4 discusses the practical issues that challenge application of statistical decision theory.

2.1. Decisions Under Uncertainty

Consider a planner who must choose an action yielding welfare that varies with the state of nature. The planner has an objective function and beliefs about the true state. These are considered primitives. He must choose an action without knowing the true state.

Formally, the planner faces choice set $C$ and believes that the true state lies in set $S$, called the state space. The objective function $w(\cdot, \cdot) : C \times S \to \mathbb{R}$ maps actions and states into welfare. The planner ideally would maximize $w(\cdot, s^*)$, where $s^*$ is the true state. However, he only knows that $s^* \in S$.

The choice set is commonly considered to be predetermined. The welfare function and the state space are subjective. The former formalizes what the decision maker wants to achieve and the latter expresses what states of nature he believes could possibly occur.

A close to universally accepted prescription for decision making is that choice should respect dominance. Action $c \in C$ is weakly dominated if there exists a $d \in C$ such that $w(d, s) \geq w(c, s)$ for all $s \in S$ and $w(d, s) > w(c, s)$ for some $s \in S$. Even though the true state $s^*$ is unknown, choice of $d$ is certain to weakly improve on choice of $c$.

There is no clearly best way to choose among undominated actions, but decision theorists have not wanted to abandon the idea of optimization. So they have proposed various ways of using the objective function $w(\cdot, \cdot)$ to form functions of actions alone, which can be optimized. In principle one should only consider undominated actions, but it may be difficult to determine which actions are undominated. Hence,
it is common to optimize over the full set of feasible actions. I define decision criteria accordingly
throughout this paper. I also use max and min notation throughout, without concern for the mathematical
subtleties that sometime make it necessary to suffice with sup and inf operations.

One broad idea is to place a subjective distribution on the state space, average state-dependent
welfare with respect to this distribution and maximize the resulting function. This yields maximization of
subjective average welfare. Let \( \pi \) be the specified probability distribution on \( S \). For each feasible action \( c \),
\( \int w(c, s) d\pi \) is the mean of \( w(c, s) \) with respect to \( \pi \). The criterion solves the problem

\[
\text{(1)} \quad \max_{c \in C} \int w(c, s) d\pi.
\]

Another broad idea is to seek an action that, in some well-defined sense, works uniformly well over
all elements of \( S \). This yields the maximin and minimax-regret (MMR) criteria. The maximin criterion
maximizes the minimum welfare attainable across the elements of \( S \). For each feasible action \( c \), consider
the minimum feasible value of \( w(c, s) \); that is, \( \min_{s \in S} w(c, s) \). A maximin rule chooses an action that solves
the problem

\[
\text{(2)} \quad \max_{c \in C} \min_{s \in S} w(c, s).
\]

The MMR criterion chooses an action that minimizes the maximum loss to welfare that can result
from not knowing the true state. A MMR choice solves the problem

\[
\text{(3)} \quad \min_{c \in C} \max_{s \in S} [\max_{d \in C} w(d, s) - w(c, s)].
\]

Here \( \max_{d \in C} w(d, s) - w(c, s) \) is the \textit{regret} of action \( c \) in state of nature \( s \), that is, the welfare loss associated
with choice of \( c \) relative to an action that maximizes welfare in state \( s \). The true state being unknown, one
evaluates \( c \) by its maximum regret over all states and selects an action that minimizes maximum regret. The
maximum regret of an action measures its maximum distance from optimality across all states. Hence, a MMR choice is uniformly nearest to optimal among all feasible actions.

A planner who asserts a partial subjective distribution on the states of nature could maximize minimum subjective average welfare or minimize maximum average regret. These hybrid criteria combine elements of averaging across states and concern with uniform performance across states. Hybrid criteria may be of interest. However, I will confine discussion to the polar cases in which the planner asserts a complete subjective distribution or none.

2.2. Statistical Decision Problems

Statistical decision problems add to the above structure by supposing that the planner observes data drawn from a sampling distribution. In practice, knowledge of the sampling distribution is generally incomplete. To express this, one extends the concept of the state space $S$ to list the set of feasible sampling distributions, denoted $(Q_s, s \in S)$.

Let $\Psi_s$ denote the sample space in state $s$, that is, $\Psi_s$ is the set of samples that may be drawn under sampling distribution $Q_s$. The literature typically assumes that the sample space does not vary with $s$ and is known. I maintain this assumption and denote the known sample space as $\Psi$, without the $s$ subscript. Then a statistical decision function $c(\cdot): \Psi \rightarrow C$ maps the sample data into a chosen action.

Wald's concept of a statistical decision function embraces all mappings of the form [data $\rightarrow$ action]. An SDF need not perform inference; that is, it need not use data to draw conclusions about the true state of nature. None of the prominent decision criteria that have been studied from Wald's perspective --- maximin, minimax-regret, and maximization of subjective average welfare (minimization of Bayes risk) --- refer to inference. The general absence of inference in statistical decision theory is striking and has been noticed. See Neyman (1962) and Blyth (1970).
Although SDFs need not perform inference, some do so. That is, some have the sequential form [data → inference → action], first performing some form of inference and then using the inference to make a decision. There seems to be no accepted term for such SDFs, so I will call them inference-based.

SDF $c(\cdot)$ is a deterministic function after realization of the sample data, but it is a random function ex ante. Hence, the welfare achieved by $c(\cdot)$ is a random variable ex ante. Wald's central idea was to evaluate the performance of $c(\cdot)$ in state $s$ by $Q_s(w[c(\psi), s])$, the ex-ante distribution of welfare that it yields across realizations $\psi$ of the sampling process.

It remains to ask how a planner might compare the welfare distributions yielded by different SDFs. The planner wants to maximize welfare, so it seems self-evident that he should prefer SDF $d(\cdot)$ to $c(\cdot)$ in state $s$ if $Q_s(w[d(\psi), s])$ stochastically dominates $Q_s(w[c(\psi), s])$. It is less obvious how he should compare rules whose welfare distributions do not stochastically dominate one another.

Wald proposed measurement of the performance of $c(\cdot)$ in state $s$ by its expected welfare across samples, that is, $E_s(w[c(\psi), s]) = \int w[c(\psi), s]dQ_s$. An alternative that has drawn only slight attention is to measure performance by quantile welfare (Manski and Tetenov, 2014). Writing in a context where one wants to minimize loss rather than maximize welfare, Wald used the term risk to denote the mean performance of a SDF across samples and the term inadmissible to denote weak dominance when evaluating performance by risk.

In practice, one does not know the true state. Hence, one evaluates $c(\cdot)$ by the state-dependent expected welfare vector $(E_s(w[c(\psi), s]), s \in S)$. This done, statistical decision theory can use the same decision criteria as does decision theory without sample data. Let $\Gamma$ be a specified set of feasible SDFs that map $\Psi \rightarrow C$. The statistical versions of decision criteria (1), (2), and (3) are

\[
(4) \quad \max_{c(\cdot) \in \Gamma} \int E_s(w[c(\psi), s]) \, d\pi,
\]

\[
(5) \quad \max_{c(\cdot) \in \Gamma} \min_{s \in S} E_s(w[c(\psi), s]),
\]
(6) \[
\min_{c(\cdot) \in \Gamma} \max_{s \in S} \left( \max_{d \in C} w(d, s) - E_r(w[c(\psi), s]) \right).
\]

Criterion (4) is sometimes called a Bayes decision. This terminology may seem odd given the absence of any reference in (4) to Bayes Theorem or Bayesian inference. The criterion simply places a subjective distribution on the state space and optimizes the resulting subjective average welfare. However, when the set of feasible statistical decision functions is unconstrained, this optimization problem yields the same solution as is obtained in conditional Bayes decision making. Conditional Bayes decision making calls on one to first perform Bayesian inference, which transforms the prior distribution on the state space into a posterior distribution without reference to a decision problem. One then chooses an action that maximizes posterior subjective average welfare. It turns out that this procedure, applied to each possible data realization, solves Wald’s problem of minimization of Bayes risk. See, for example, Berger (1985, Section 4.4.1).

2.2.1. Focus on Maximum Regret

In this paper, I mainly measure the performance of SDFs by maximum regret rather than by maximum subjective average welfare or minimum expected welfare. This section explains why.

I have no objection to computation of subjective average welfare when one feels able to place a credible subjective prior distribution on the state space. However, Bayesians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. The controversy suggests that inability to express a credible prior is common in actual decision settings.

When one finds it difficult to assert a credible subjective distribution, Bayesians who believe it essential to use a probability distribution to express uncertainty may suggest use of some default distribution that is variously called a “reference” or “conventional” or “objective” prior, see, for example, Berger (2006). However, there is no consensus on the prior that should play this role. The chosen prior matters for decision making.
In the absence of a credible prior distribution on the state space, practical and conceptual reasons motivate measurement of the performance of SDFs by maximum regret, rather than by minimum expected welfare. From a practical perspective, it has been found that MMR decisions behave more reasonably than do maximin ones in the important context of treatment choice. In common settings of treatment choice with trial data, it has been found that the MMR rule is well approximated by the empirical success rule, which chooses the treatment with the highest observed average outcome in the trial. In contrast, the maximin criterion commonly ignores the trial data, whatever they may be. This was recognized verbally by Savage (1951), who stated that the criterion is “ultrapessimistic” and wrote (p. 63): “it can lead to the absurd conclusion in some cases that no amount of relevant experimentation should deter the actor from behaving as though he were in complete ignorance.” Savage did not flesh out this statement, but it is easy to show that this occurs with trial data. Manski (2004) provides a simple example.

The conceptual appeal of using maximum regret to measure performance is that maximum regret quantifies how lack of knowledge of the true state of nature diminishes the quality of decisions. While the term “maximum regret” has become standard in the literature, this term is a shorthand for the maximum sub-optimality of a decision criterion across the feasible states of nature. An SDF with small maximum regret is uniformly near-optimal across all states. This is a desirable property.

2.3. Binary Choice Problems

SDFs for binary choice problems are relatively simple. They are intellectually interesting because they can always be viewed as hypothesis tests. However, the Wald perspective on testing differs considerably from that of Neyman-Pearson. I explain here.

Let choice set \( C = \{a, b\} \). A SDF \( c(\cdot) \) partitions \( \Psi \) into two regions that separate the data yielding choice of each action. These regions are \( \Psi_{c(a)} = \{\psi \in \Psi: c(\psi) = a\} \) and \( \Psi_{c(b)} = \{\psi \in \Psi: c(\psi) = b\} \).
A hypothesis test motivated by the choice problem partitions state space $S$ into two regions, say $S_a$ and $S_b$, that separate the states in which actions $a$ and $b$ are uniquely optimal. Thus, $S_a$ contains the states $[s \in S: w(a, s) > w(b, s)]$ and $S_b$ contains $[s \in S: w(b, s) > w(a, s)]$. The choice problem does not provide a rationale for allocation of states in which the two actions yield equal welfare. The standard practice in testing is to give one action, say $a$, a privileged status and to place all states yielding equal welfare in $S_a$. Then $S_a = [s \in S: w(a, s) \geq w(b, s)]$ and $S_b = [s \in S: w(b, s) > w(a, s)]$.

In the language of hypothesis testing, SDF $c(\cdot)$ performs a test with acceptance regions $\Psi_{c(\cdot)a}$ and $\Psi_{c(\cdot)b}$. When $\psi \in \Psi_{c(\cdot)a}$, $c(\cdot)$ accepts the hypothesis $\{s \in S_a\}$ by setting $c(\psi) = a$. When $\psi \in \Psi_{c(\cdot)b}$, $c(\cdot)$ accepts the hypothesis $\{s \in S_b\}$ by setting $c(\psi) = b$. I use the word "accepts" rather than the traditional term "does not reject" because choice of $a$ or $b$ is an affirmative action.

Although all SDFs for binary choice are tests, Neyman-Pearson hypothesis testing and statistical decision theory evaluate tests in fundamentally different ways. Sections 2.3.1 and 2.3.2 contrast the two paradigms in general terms. Section 2.3.3 illustrates.

2.3.1. Neyman-Pearson Testing

Let us review the basic practices of classical hypothesis testing, as developed by Neyman and Pearson (1928, 1933). These tests view the two hypotheses $\{s \in S_a\}$ and $\{s \in S_b\}$ asymmetrically, calling the former the null hypothesis and the latter the alternative. The sampling probability of rejecting the null hypothesis when it is correct is the probability of a Type I error. A longstanding convention has been to restrict attention to tests in which the probability of a Type I error is no larger than some predetermined value $\alpha$, usually 0.05, for all $s \in S_a$. In the notation of statistical decision theory, one restricts attentions to SDFs $c(\cdot)$ for which $Q_s[c(\psi) = b] \leq \alpha$ for all $s \in S_a$.

Among tests that satisfy this restriction, Neyman-Pearson testing seeks ones that give small probability of rejecting the alternative hypothesis when it is correct, the probability of a Type II error. However, it generally is not possible to attain small probability of a Type II error for all $s \in S_b$. Letting $S$
be a metric space, the probability of a Type II error typically approaches $1 - \alpha$ as $s \in S_b$ nears the boundary of $S_a$. See, for example, Manski and Tetenov (2016), Figure 1. Given this, the convention has been to restrict attention to states in $S_b$ that lie at least a specified distance from $S_a$.

Let $\rho$ be the metric measuring distance on $S$. Let $\rho_a > 0$ be the specified minimum distance from $S_a$. In the notation of statistical decision theory, Neyman-Pearson testing seeks small values for the maximum value of $Q_s[c(\psi) = a]$ over $s \in S_b$ s. t. $\rho(s, S_a) \geq \rho_a$.

2.3.2. Maximum Regret of Tests

Decision theoretic evaluation of tests does not restrict attention to tests that yield a predetermined upper bound on the probability of a Type I error. Nor does it aim to minimize a constrained maximum value of the probability of a Type II error. Wald's central idea for binary choice as elsewhere is to evaluate the performance of SDF $c(\cdot)$ in state $s$ by the distribution of welfare that it yields across realizations of the sampling process. He first addressed hypothesis testing this way in Wald (1939).

The welfare distribution in state $s$ in a binary choice problem is Bernoulli, with mass points $\max[w(a, s), w(b, s)]$ and $\min[w(a, s), w(b, s)]$. These mass points coincide if $w(a, s) = w(b, s)$. When $s$ is a state where $w(a, s) \neq w(b, s)$, let $R_{c(\cdot)}^s$ denote the probability that $c(\cdot)$ yields an error, choosing the inferior treatment over the superior one. That is,

\begin{align*}
R_{c(\cdot)}^s &= \begin{cases} Q_s[c(\psi) = b] & \text{if } w(a, s) > w(b, s), \\ Q_s[c(\psi) = a] & \text{if } w(b, s) > w(a, s). \end{cases}
\end{align*}

The former and latter are the probabilities of Type I and Type II errors. Whereas Neyman-Pearson testing treats these error probabilities differently, statistical decision theory views them symmetrically.

The probabilities that welfare equals $\max[w(a, s), w(b, s)]$ and $\min[w(a, s), w(b, s)]$ are $1 - R_{c(\cdot)}^s$ and $R_{c(\cdot)}^s$. Wald measured the performance of SDFs by expected welfare. In binary choice problems, expected welfare in state $s$ is
\[ E_s(w[c(\psi), s]) = R_{c(\psi)}(\min \{w(a, s), w(b, s)\}) + [1 - R_{c(\psi)}](\max \{w(a, s), w(b, s)\}) \]

\[ = \max \{w(a, s), w(b, s)\} - R_{c(\psi)}|w(a, s) - w(b, s)|. \]

The expression \( R_{c(\psi)}|w(a, s) - w(b, s)| \) is the regret of \( c(\cdot) \) in state \( s \). Thus, regret is the product of the error probability and the magnitude of the welfare loss when an error occurs.

Evaluation of hypothesis tests by expected welfare constitutes a fundamental difference between the perspectives of Wald and of Neyman-Pearson. A planner should care about more than the probabilities of Type I and II error. He should care as well about the magnitudes of the losses to welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between actions is larger. The Neyman-Pearson theory of hypothesis testing does not recognize this.

Computation of regret in a specified state is usually practical. The welfare magnitudes \( w(a, s) \) and \( w(b, s) \) are usually easy to compute. The error probability \( R_{c(\psi)} \) typically does not have a simple explicit form, but it usually can be approximated to any desired precision by Monte Carlo integration. That is, one draws repeated pseudo-realizations of \( \psi \) from the distribution \( Q_s \), computes the fraction of cases in which the resulting \( c(\psi) \) selects the inferior action, and uses this to estimate \( R_{c(\psi)} \).

Whereas computation of regret in one state is not problematic, computation of maximum regret across all feasible states may be burdensome. The state space commonly is uncountable in applications. A pragmatic process for coping with uncountable state spaces is to discretize the space, limiting attention to a finite subset of states that reasonably approximate the full state space.

2.3.3. Illustration: Comparing a Neyman-Pearson Test with an MMR Decision

Manski and Tetenov (2016) compare a Neyman-Pearson test with a decision that minimizes maximum regret, in a simple context where the MMR decision is known. The context is choice between
two treatments, say $t = a$ and $t = b$, when the outcome of interest is binary, with $y(t) = 1$ denoting success and $y(t) = 0$ failure. State $s$ indicates a possible value for the pair of outcome probabilities $(P_s[y(a) = 1], P_s[y(b) = 1])$. The welfare yielded by treatment $t$ in state $s$ is $w(t, s) = P_s[y(t) = 1]$. The sample data are the findings of a balanced randomized trial, assigning the same number $N$ of subjects to each treatment.

In this setting, a widely used classical test is a one-sided two-sample $z$-test, which asymptotically makes the probability of a Type I error equal to 0.05. See Fleiss (1973) for details. Stoye (2009) shows that the MMR decision is the empirical success (ES) rule. This rule, which will be discussed more generally in Section 4, chooses a treatment that maximizes the average sample outcome in the trial.

The regret of any test $c(\cdot)$ in any state $s$ is $R_{c(\cdot), s}[P_s[y(a) = 1] - P_s[y(b) = 1]]$. The quantity $|P_s[y(a) = 1] - P_s[y(b) = 1]|$ is the absolute value in state $s$ of the average treatment effect comparing welfare with treatments $a$ and $b$. Thus, the regret of a test in state $s$ is the product of its error probability and the absolute value of the effect size.

We suppose that the planner has no a priori knowledge of the outcome probabilities. Hence, the state space is the rectangle $[0, 1]^2$. We approximate maximum regret by computing regret over a dense finite grid of states, thus discretizing the state space.

Figure 1 of Manski and Tetenov (2016) shows how the regret incurred by the ES rule and the $z$-test rule varies with the effect size for a sample size of $N = 145$ per treatment arm. Maximum regret is 0.01 for the ES rule and 0.05 for the $z$-test. Maximum regret for each test occurs at an intermediate effect size. Regret is necessarily small for small effect sizes. Regret is also small for large effect sizes, because the probability of error declines with the effect size. The intermediate effect sizes at which regret is maximized differ for the two tests, reflecting the differences in their state-specific error probabilities.

2.4. Practicalities

Statistical decision theory has breathtaking generality. It enables comparison of all SDFs whose risk functions exist. It applies to any sample size, without asymptotic approximations.
The state space may take any form. In Haavelmo’s formalization of econometrics, \( S \) is a space of probability distributions that may possibly describe the economic system under study. The state space may be finite dimensional or larger. The theory is applicable when the true state is a partially identified probability distribution.

Given these features, one might anticipate that statistical decision theory would play a central role in modern statistics and econometrics. Notable contributions emerged in the 1950s and 1960s, as described in the monographs of Ferguson (1967) and Berger (1985). However, the early period of development of statistical decision theory largely closed by the 1970s, except for the conditional Bayes version of Bayesian theory. Conditional Bayes analysis has continued to develop, but as a self-contained field of study disconnected from Wald’s frequentist idea of minimization of Bayes risk.

Why did statistical decision theory lose momentum? One reason may have been diminishing interest in decision making as the motivation for analysis of sample data. Many modern statisticians and econometricians view the objective of empirical research as inference for scientific understanding, rather than use of data in decision making. A different reason may have been the technical difficulty of the subject. Wald’s ideas are easy to describe abstractly, but they can be difficult to apply in practice.

Consider the mathematical problems (4) through (6). These problems are generally well-posed in principle, but they may not be solvable in practice. Each problem requires performance of three nested operations. The most basic inner operation integrates across the sampling distribution of the data to determine expected welfare when a specified SDF \( \alpha(\cdot) \) is used in each feasible state of nature \( s \in S \). The result is evaluation of \( \alpha(\cdot) \) by the state-dependent expected welfare vector \( (E_s\{w[c(\psi), s]\}, s \in S) \). The middle operation integrates or finds an extremum of the result of the inner operation across the state space. The outer operation finds an extremum of the result of the middle operation across all SDFs.

Analytical arguments and numerical computation sometimes yield tractable solutions to these problems. Early analytical work in the conditional Bayes paradigm studied conjugate priors, which pair certain prior distributions on the state space with certain state-dependent sampling distributions for the data to yield simple posterior distributions. An important early analytical solution of an MMR problem was the
study by Hodges and Lehmann (1950) of point prediction of a bounded outcome with data from a random sample. Recently, Dominitz and Manski (2017, 2019) have derived analytical findings on the maximum regret of certain tractable point predictors of bounded outcomes with data from random samples when some outcomes are missing. In another domain, econometricians have proved a sequence of analytical findings on tractable decision rules for treatment choice with data from a randomized experiment; see Manski (2004, 2005, 2007), Hirano and Porter (2009, 2019), Stoye (2009, 2012), Manski and Tetenov (2016, 2019), and Kitagawa and Tetenov (2018).

Numerical computation often was not feasible when statistical decision theory developed in the 1940s, but it has become increasingly possible since then. Modern conditional Bayes analysis has increasingly moved away from use of conjugate priors to numerical computation of posterior distributions. Numerical determination of some maximin and MMR decisions has also become feasible. For example, Manski and Tetenov (2016, 2019) present in tabular form the MMR solutions to certain treatment choice problems with data from a randomized experiment. Computation of state-dependent expected welfare, the inner operation in problems (4) through (6), can now be accomplished numerically by Monte Carlo integration methods. Manski and Tabord-Meehan (2017) do this in the context of point prediction with random-sample data when some outcomes are missing, see Section 3.4 for further discussion.

3. Decision Making with Models

3.1. Basic Ideas

I stated at the outset that standard decision theory begins with a planner who “specifies a state space listing the states that he considers possible.” Thus, the state space should include all states that the planner believes feasible and no others. The terms “considers possible” and “believes feasible” are necessarily subjective and context specific.
The state space may be a large set that is difficult to contemplate in its entirety. Hence, it is common to make decisions using a model. The word “model” is commonly used informally to connote a simplification or approximation of reality. Formally, a model specifies an alternative to the state space. Thus, model \( m \) replaces \( S \) with a model space \( S_m \). A planner using a model acts as if the model space is the state space. Thus, the planner might solve problem (4), (5), or (6) with \( S_m \) replacing \( S \).

The states contained in a model space may or may not be elements of the state space. The statistician George Box famously wrote (Box, 1979): “All models are wrong, but some are useful.” The phrase “all models are wrong” indicates that Box was thinking of models that simplify or approximate reality in a way that one believes could not possibly be accurate, then \( S_m \cap S = \emptyset \). On the other hand, researchers often use models that they believe could possibly be accurate but that are not necessarily so; then \( S_m \subset S \).

In whatever manner one specifies a model, statistical decision theory provides a clear way to evaluate decision making. What matters is the SDF, say \( c_m(\cdot) \), that one develops using a model. As with any SDF, one measures the performance of \( c_m(\cdot) \) by its vector of state-dependent expected welfares \( (E_s[w(c_m(\psi), s)], s \in S) \). Importantly, the relevant states for evaluation of performance are those in the state space \( S \), not those in the model space \( S_m \).

Thus, statistical decision theory enables one to operationalize Box’s assertion that some models are useful. Useful models are ones whose application to decision making yields acceptably high state-dependent expected welfare across the state space, relative to what is possible in principle. From this perspective, one should not make an abstract assertion that a model is or is not useful. Usefulness depends on the decision context.

3.2. As-If Optimization

A familiar econometric practice specifies a model space, typically called the parameter space. The parameter space is often finite-dimensional in applied research, but this is not essential. Sample data are
used to select a point in the parameter space, called a point estimate of the parameter. The econometric
method used to compute the point estimate typically is motivated by reference to desirable statistical
properties that hold if the model is correct; that is, if the true state of nature lies within the parameter space.

*As-if optimization* chooses an action that optimizes welfare as if the estimate is the true state. Thus, as-if
optimization is a type of inference-based SDF.

Applied economists find as-if optimization appealing because it cleanly separates inference and
decision making. Whereas Wald supposed that a planner both performs research and makes a decision, in
practice there commonly is an institutional separation between research and decision making. Researchers
report inferences and planners use them to make decisions. Thus, planners perform the mapping [inference
← decision] rather than the more basic mapping [data ← decision]. Having researchers report point estimates
and planners use them as if they are accurate exemplifies this process.

Formally, a point estimate is a function $s(\cdot) : \Psi \rightarrow S_m$ that maps data into a point in a model space.

As-if optimization means solution of the problem $\max_{c \in C} w[c, s(\psi)]$. When as-if optimization yields
multiple solutions, one may use some auxiliary rule to select one. The result is an SDF $c[s(\cdot)]$, where

$$\tag{9} c[s(\psi)] \in \arg\max_{c \in C} w[c, s(\psi)], \quad \psi \in \Psi.$$

Solving problem (9) is often simpler than solving problems (4) through (6). Selecting a point
estimate and using it to maximize welfare may be far easier computationally than performing the nested
operations requires to solve problems (4) through (6). However, computational appeal does not suffice to
justify this approach to decision making.

To motivate as-if optimization, econometricians often cite limit theorems of asymptotic theory that
hold if the model is correct. They hypothesize a sequence of sampling processes indexed by sample size $N$
and a corresponding sequence of point estimates $s_N(\cdot) : \Psi_N \rightarrow S_m$. They show that the sequence is consistent
when specified assumptions hold. That is, $s_N(\psi) \rightarrow s^*$ as $N \rightarrow \infty$, in probability or almost surely. They may
prove further results regarding rate of convergence and the limiting distribution of the estimate.
Here and elsewhere, asymptotic arguments sometimes may be suggestive, but they do not prove that as-if optimization provides a well-performing SDF. Statistical decision theory evaluates as-if optimization in state $s$ by the expected welfare $E_s\{w(c[s(\psi)], s)\}$ that it yields across samples of size $N$, not asymptotically. It calls for study of expected welfare across the state space $S$, not the model space $S_m$.

3.2.1. As-If Optimization with Analog Estimates

Econometric research from Haavelmo onward has focused to a considerable degree on a class of problems that connect the state space and the sampling distribution in a simple way. These are problems in which states are probability distributions and the data are a random sample drawn from the true distribution. In such problems, a natural form of as-if optimization is to act as if the empirical distribution of the data is the true population distribution. Thus, one specifies the model space as the set of all possible empirical distributions and uses the observed empirical distribution as the point estimate of the true state.

Goldberger (1968) notably called this the *analogy principle*. He wrote (p. 4): “The *analogy principle* of estimation . . . . proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population.” See also Manski (1988).

The usual justification for using the analogy principle has been asymptotic. The empirical distribution consistently estimates the population distribution and has various further desirable asymptotic properties. This suggests decision making using the empirical distribution as if it were the true population distribution.

3.2.2. As-If Decisions with Set Estimates

As-if optimization uses data to compute a point estimate of the true state of nature and chooses an action that optimizes welfare as if this estimate is accurate. An obvious, but rarely applied, extension is to use data to compute a set-valued estimate of the true state and then act as if the set estimate is accurate. Whereas a point estimate $s(\cdot): \Psi \to S$ maps data into an element of $S$, a set estimate $S(\cdot): \Psi \to 2^S$ maps data into a subset of $S$. For example, $S(\cdot)$ could be a confidence set reported by researchers.
Given data $\psi$, one could act as if the state space is $S(\psi)$ rather than the larger set $S$. Specifically, one could solve these data-dependent versions of problems (1) through (3):

\begin{align*}
(1') \quad \max_{c \in C} \int w(c, s) dm(\psi), \\
(2') \quad \max_{c \in C} \min_{s \in S(\psi)} w(c, s), \\
(3') \quad \min_{c \in C} \max_{s \in S(\psi)} [\max_{d \in C} w(d, s) - w(c, s)].
\end{align*}

In the case of (1'), $\pi(\psi)$ is a subjective distribution on the set $S(\psi)$. These as-if problems are generally easier to solve than are the actual maximin and MMR problems with sample data, stated in (4) through (6). The as-if problems fix $\psi$ and select one action $c$, whereas the actual problems require one to consider all potential samples and choose a decision function $c(\cdot)$. The as-if problems compute welfare values $w(c, s)$, whereas the actual problems must compute more complex expected welfare values $E_s\{w[c(\psi), s]\}$. Section 3.4.1 provides an example.

An alternative type of as-if approach replaces $S$ by $S(\psi)$ in the middle operations of (4) through (6), but it does not replace $E_s\{w[c(\psi), s]\}$ by $w(c, s)$ in the innermost part of each criterion. This approach simplifies (4) through (6) by shrinking the state space over which the middle operations are performed. However, it is more complex than (1') through (3') for two reasons. It requires choice of a decision function $c(\cdot)$ rather than a single action $c$, and it must compute $E_s\{w[c(\psi), s]\}$ rather than $w(c, s)$. Chamberlain (2000) uses asymptotic considerations to suggest this type of as-if decision making and presents an application.

3.3. Prediction with Sample Data

A familiar case of as-if optimization occurs when states are distributions for a real random variable and the decision problem is to predict the value of a realization drawn at random from the true distribution.
When welfare is measured by state-dependent square and absolute loss, the ideal best predictors are well-known to be the population mean and median. When the true distribution is not known but data from a random sample are observed, the analogy principle suggests use of the sample average and median as predictors.

In his final chapter on “Problems of Prediction,” Haavelmo (1944) questions this common application of as-if optimization and instead recommends application of the Wald theory. This final chapter was added by Haavelmo after his distribution, in 1941, of an early version of the thesis. In his section on “General Formulation of the Problem of Prediction,” he writes (p. 109): “We see therefore that the seemingly logical ‘two-step’ procedure of first estimating the unknown distribution of the variables to be predicted and then using this estimate to derive a prediction formula for the variables may not be very efficient.” Citing Wald (1939), he next proposes computation of the state-dependent risk for any proposed predictor function.

Letting $E_2$ denote a predictor function and $(x_1, x_2, \ldots, x_N)$ the sample data, he writes (p. 109): “We have to choose $E_2$ as a function of $x_1, x_2, \ldots, x_N$, and we should, naturally, try to choose $E_2(x_1, x_2, \ldots, x_N)$ in such a way that $r$ (the ‘risk’) becomes as small as possible.” But he immediately recognizes that there generally does not exist a predictor function that minimizes risk across all states of nature. Hence, he goes on to suggest a feasible approach. I quote in full this key passage, which uses the notation $\Omega_1$ to denote the state space (p. 116):

“In general, however, we may expect that no uniformly best prediction function exists. Then we have to introduce some additional principles in order to choose a prediction function. We may then, first, obviously disregard all those prediction functions that are such that there exists another prediction function that makes $r$ smaller for every member of $\Omega_1$. If this is not the case we call the prediction function considered an admissible prediction function. To choose between several admissible prediction functions we might adopt the following principle, introduced by Wald: For every admissible prediction function $E_2$ the ‘risk’ $r$ is a function of the true distribution $p$. Consider that prediction function $E_2$, among the admissible ones, for which the largest value of $r$ is at a minimum (i.e., smaller than or at most equal to the largest value of $r$ for any other admissible
E2). Such a prediction function, if it exists, may be said to be the least risky among the admissible prediction functions.”

Thus, following Wald, Haavelmo suggests elimination of inadmissible predictors followed by choice of a minimax predictor among those that are admissible.

It may be that econometrics would have progressed early on to make productive use of statistical decision theory if Haavelmo had been able to pursue the above idea further. However, in his next section on “Some Practical Suggestions for the Derivation of Prediction Formulae,” he cautions regarding the practicality of the idea, writing (p. 111): “The apparatus set up in the preceding section, although simple in principle, will in general involve considerable mathematical problems and heavy algebra.”

Aiming for tractability, Haavelmo went on to sketch an example of as-if optimization that chooses an action using a maximum likelihood estimate of a specific finite-dimensional parametric model. He noted that one could study the state-dependent risk of the resulting SDF, but he did not provide any analysis. With this, his chapter on prediction ended. Thus, Haavelmo initiated econometric consideration of statistical decision theory but, stymied by computational intractability, he found himself unable to follow through.

3.4. Prediction under Square Loss

Haavelmo discussed decision theoretic analysis of prediction briefly and abstractly. Subsequent research has focused on the special case of square loss. In this case, the risk of a candidate predictor using sample data is the sum of the population variance of the outcome and the mean square error (MSE) of the predictor as an estimate of the mean outcome. Hence, the regret of a predictor is its MSE as an estimate of the mean. An MMR predictor minimizes maximum mean square error. MMR prediction of the outcome is equivalent to minimax estimation of the population mean.

Among the earliest important practical findings of statistical decision theory was reported by Hodges and Lehman (1950). They derived the MMR predictor under square loss with data from a random sample, when the outcome has known bounded range and all sample data are observed. They assumed no
knowledge of the outcome distribution beyond its bounded support. Normalizing the support to be the interval \([0, 1]\), they proved that the MMR predictor is \(m\sqrt{N + \frac{1}{2}}/(\sqrt{N + 1})\), where \(N\) is sample size and \(m\) is the sample average outcome.

3.4.1. Prediction with Missing Data

Dominitz and Manski (2017, 2019) have recently extended study of prediction of a bounded outcome under square loss to settings in which a random sample is drawn but some realized outcomes are unobserved. It is challenging to determine the MMR predictor when some data are missing. Seeking an approach that is both tractable and reasonable, the paper studies as-if MMR prediction. The analysis assumes knowledge of the fraction of the population with missing data, but it assumes no knowledge of the distribution of missing outcomes beyond its bounded support. It uses the empirical distribution of the observed sample data as if it were the population distribution of observable outcomes.

In the absence of knowledge of the distribution of missing outcomes, the population mean outcome is partially identified when the outcome is bounded. Its identification region is an easy-to-compute interval derived in Manski (1989). If this interval were known, the MMR predictor would be its midpoint. The identification interval is not known with sample data, but one can compute its sample analog and use the midpoint of the sample-analog interval as the predictor.

This midpoint predictor is easy to compute. Its maximum regret is shown to have a simple analytical form. Let \(\delta\) indicate the observability of an outcome. Let \(P(\delta = 1)\) and \(P(\delta = 0)\) denote the fractions of the population whose outcomes are and are not observable. Let \(N\) be the number of observed sample outcomes, which is fixed rather than random under the assumed survey design. The paper proves that the maximum regret of the midpoint predictor is \(\frac{1}{4}[P(\delta = 1)^2/N + P(\delta = 0)^2]\).

The analysis in Dominitz and Manski (2017) presumes a state space that places no restrictions on the distributions of observable and unobservable outcomes. Researchers often assume that data are missing at random. That is, they invoke a model space in which the distributions of observable and unobservable outcomes are the same. They then use the sample average of observed outcomes as the predictor. Dominitz
and Manski caution against this application of modeling when the distributions of observable and unobservable outcomes may differ arbitrarily. They show that the maximum regret of the model-based predictor necessarily exceeds that of the midpoint predictor, in some cases substantially so.

Dominitz and Manski (2019) builds on Dominitz and Manski (2017) to study MMR prediction under square loss of functions of two variables, when some data on one variable are missing. This prediction problem may arise in longitudinal data collection with attrition, if there is a 100-percent response rate in period 1 and some nonresponse in period 2. It may also arise in cross-sectional collection of data on two household members or in surveys with item nonresponse.

Predicting the value of functions of two variables is more complex than functions of one variable, because the form of the function matters. The paper focuses on prediction of linear and indicator functions. It is easy to compute midpoint predictors akin to that posed in the previous study. The paper obtains an analytical expression for maximum regret when predicting the value of an indicator function. It derives an analytical upper bound on maximum regret when predicting the value of linear functions.

3.4.2. Numerical Computation of Maximum Regret in Prediction with Missing Data

The analytical findings on the maximum regret of midpoint predictors described above assume knowledge of the fraction of the population with missing data. Midpoint predictors remain easy to compute when this fraction is not known and instead is estimated by its sample analog. In this case, derivation of an analytical expression for maximum regret does not seem possible, but numerical computation is tractable. I summarize here. This demonstrate how advances in numerical analysis now enable applications of statistical decision theory that were impractical when Haavelmo and Wald made their contributions.

Manski and Tabord-Meehan (2017) describe an algorithm coded in STATA for numerical computation of the maximum regret of the midpoint predictor and other user-specified predictors in the setting of Dominitz and Manski (2017), where the objective is to predict a bounded real outcome. The program, named \texttt{wald_mse}, does not require knowledge of the population fraction of missing data. Instead, P(z = 0) may be estimated by its sample analog.
Letting $y$ denote the outcome of interest, the state space has the form $[P_s(y|\delta = 1), P_s(y|\delta = 0), P_s(\delta = 0)]$, $s \in S$. An important feature of `wald_mse` is that the user can specify the state space flexibly. For example, the user may assume that nonresponse will be no higher than 80% or that the mean value of the outcome for nonresponders will be no lower than 0.5. The user may impose no restrictions connecting the two outcome distributions $P_s(y|\delta = 1)$ and $P_s(y|\delta = 0)$, or he may bound the difference between these distributions.

In any given state $s$, the algorithm uses Monte Carlo integration to approximate the MSE of a user-specified predictor. The quality of the approximation is controlled by user specification of the number of pseudo realizations of $(y, \delta)$ that are drawn. Increasing the number yields a better approximation at the cost of longer computation time.

The algorithm embodies two approaches to maximize MSE across the state space, one applicable when the outcome is binary and the other when the outcome has a continuous distribution. When $y$ is binary, $P_s(y|\delta = 1)$, $P_s(y|\delta = 0)$, and $P_s(\delta = 0)$ are all Bernoulli distributions. The algorithm approximates the state space by a finite grid over the possible Bernoulli parameters for each distribution. It then maximizes MSE over the grid. The user controls the quality of the approximation to the state space by specifying the density of the grid. Increasing the density yields a better approximation at the cost of longer computation time.

When $y$ is continuous, the algorithm presumes that $P_s(y|\delta = 1)$ and $P_s(y|\delta = 0)$ are Beta distributions, while $P_s(\delta = 0)$ is a Bernoulli distribution. Supposing that the two outcome distributions are Beta is a substantive restriction, intended to pose a relatively flexible and tractable state space. The state space is approximated by a finite grid over the possible shape parameters of the Beta distributions and over the possible values of the Bernoulli parameter. As before, the user specifies the density of the grid and the algorithm maximizes mean square error over the grid.
4. Maximum Regret of the Empirical Success Rule for Treatment Choice

4.1. Background

To further flesh out the abstract discussion of Section 3, this section provides decision-theoretic analysis of an important instance of as-if optimization, summarizing findings to date and adding new ones.

A large body of empirical research performed by econometricians, statisticians, and others has studied treatment response in randomized trials and observational settings. The objective of some of this research has been to perform so-called causal inference, without reference to an explicit decision problem. However, much of the research has aimed to inform treatment choice by planners acting in public health or public policy settings. I am concerned with the latter here.

I discuss a standard formalization studied in many sources, for example, in Manski (2004). States of nature are possible distributions of treatment response for the members of a population of observationally identical persons who are subject to treatment. The term “observationally identical” means that these persons share the same observed covariates.

The decision problem is to choose treatments for members of this population. It is assumed that treatment response is individualistic, meaning that each person’s outcome may depend on the treatment he receives but not on the treatments received by others. Welfare is measured by the mean outcome of treatment across the population. It follows that the optimal treatment choice maximizes this mean outcome.

In practice, optimal treatment is infeasible because the true distribution of treatment response is not known. Decision making may use data on the outcomes realized by a random sample of the population. Some research studies settings in which the sample are subjects in a randomized trial, and some studies observational settings in which treatments are determined by another mechanism. Either way, statistical decision theory may be brought to bear to evaluate the performance of SDFs that use the sample data to make treatment decisions.
A simple way to use sample data to make treatment choices is as-if optimization. Applying the analogy principle, one acts as if the empirical distribution of trial outcomes for each treatment equals the population distribution of outcomes for this treatment. Emulating the fact that it is optimal to choose a treatment that maximizes the mean population outcome, one chooses a treatment that maximizes the average sample outcome. This has been called the empirical success (ES) rule in Manski (2004) and elsewhere.

When analyzing data from randomized trials, econometricians and statisticians have long used asymptotic arguments to motivate the ES rule, applying laws of large numbers and central limit theorems. In contrast, a recent econometric literature studies the maximum regret of the ES rule with trial data, in settings where treatment outcomes are bounded. Maximum regret quantifies how lack of knowledge of the true state of nature diminishes the quality of decisions. An SDF with small maximum regret is uniformly near-optimal across all states.

Section 4.2 summarizes findings on the performance of the ES rule with trial data. Section 4.3 provides new analysis of its performance with observational data. Section 4.4 relates the analysis to the early literature in econometrics.

4.2. Maximum Regret of the ES Rule with Trial Data

Researchers often say that treatment selection is random in a trial. In traditional econometric terms, the treatment that each subject receives is exogenous. Whatever language one uses, the formal condition is that the state space only includes states in which treatment response is statistically independent of the treatments that persons in the study population receive.

Study of the regret performance of the ES rule with trial data was initiated by Manski (2004), who used a large-deviations inequality for sample averages of bounded outcomes to derive an upper bound on maximum regret. Subsequently, Stoye (2009) showed that in trials with moderate sample size, the ES rule either exactly or approximately minimizes maximum regret in cases with two treatments and a balanced
design. Hirano and Porter (2009) showed that the ES rule is asymptotically optimal in a formal decision-theoretic sense.

Considering problems with multiple treatments or unbalanced designs, Manski and Tetenov (2016) use large deviations inequalities for sample averages of bounded outcomes to obtain upper bounds on the maximum regret of the ES rule. Their Proposition 1 extends the early finding of Manski (2004) from two to multiple treatments. Proposition 2 derives a new large-deviations bound for multiple treatments.

Let $L$ be the number of treatment arms and let $V$ be the range of the bounded outcome. When the trial has a balanced design, with $n$ subjects per treatment arm, the bounds on maximum regret proved in Propositions 1 and 2 have particularly simple forms, being

\[(10) \quad (2e)^{-\frac{1}{2}}V(L - 1)n^{-\frac{1}{2}},\]

\[(11) \quad V(\ln L)^{\frac{1}{2}}n^{-\frac{1}{2}}.\]

Result (10) provides a tighter bound than (11) for two or three treatments. Result (11) gives a tighter bound for four or more treatments.

4.3 Maximum Regret of the Empirical Success Rule with Observational Data

A planner may use the ES rule with data from an observational study. Then random treatment selection is not a credible assumption constraining the state space. Hence, the bounds on maximum regret given in (10) and (11) do not hold. Maximum regret need not converge to zero as sample size increases. One might nevertheless still apply the ES rule, viewing random treatment selection as a model that may be wrong but that may perhaps still be useful. I pursue this idea here in a setting with two treatments, determining when use of the ES rule does and does not yield a desirable result.
Consider an observational study with two treatments, say (a, b), and a bounded potential outcome taking values in the interval [0, 1]. The planner’s problem is to choose between the two treatments. Each member of the study population has potential outcomes \([y(a), y(b)]\). Binary indicators \([\delta(a), \delta(b)]\) denote whether these outcomes are observable. Realized outcomes are observed, but counterfactual outcomes are unobserved. Hence, the possible values for the observability indicators are \([\delta(a) = 1, \delta(b) = 0]\) and \([\delta(a) = 0, \delta(b) = 1]\). Each element \(s\) of the state space denotes a possible distribution \(P_s[y(a), y(b), \delta(a), \delta(b)]\) of outcomes and observability.

As discussed in Section 2.3, in states of nature where treatment a is better, the regret of any SDF is the product of the probability across repeated samples that the rule commits a Type I error (choosing b) and the magnitude of the loss in mean welfare that occurs when choosing B. Similarly, in states where b is better, regret is the probability of a Type II error (choosing a) times the magnitude of the loss in mean welfare when choosing a. Thus, regret in state \(s\) is \(R_c(\cdot)s \cdot |E_s[y(b)] - E_s[y(a)]|\), where \(R_c(\cdot)\) denotes the error probability. Regret is zero in states where \(E_s[y(b)] = E_s[y(a)]\). Hence, it suffices to consider states where \(E_s[y(b)] \neq E_s[y(a)]\).

With finite sample data, the state-dependent error probabilities for the ES rule may be approximated numerically, but they do not have simple explicit forms. On the other hand, analysis is straightforward in the limiting asymptotic case where all realized outcomes are observed. Given that realized outcomes are observed while counterfactual ones are not, it necessarily holds that \(P[\delta(a) = 1] + P[\delta(b) = 1] = 1\). The data reveal the true values of \(P[y(a)|\delta(a) = 1], P[y(b)|\delta(b) = 1], P[\delta(a)],\) and \(P[\delta(b)],\) provided that \(P[\delta(a) = 1] > 0\) and \(P[\delta(b) = 1] > 0\). This asymptotic setting has long been studied in partial identification analysis of treatment response, beginning with Manski (1990). I proceed likewise henceforth.

The data do not reveal the true counterfactual outcome distributions \(P[y(a)|\delta(a) = 0]\) and \(P[y(b)|\delta(b) = 0]\). In the absence of assumptions restricting these distributions, the state space contains all distributions such that \(0 \leq E[y(a)|\delta(a) = 0] \leq 1\) and \(0 \leq E[y(b)|\delta(b) = 0] \leq 1\). By the Law of Total Probability, the feasible values of \(P[y(a) = 1]\) and \(P[y(b) = 1]\) are
(12a) \[ E[y(a)] \in \left[ E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1], E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1] + P[\delta(a) = 0] \right]. \]

(12b) \[ E[y(b)] \in \left[ E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1], E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1] + P[\delta(b) = 0] \right]. \]

Consider choice of treatment a or b. With choice of a, \( R_c(\cdot) = 0 \) in states where \( E_s[y(b)] < E_s[y(a)] \) and \( R_c(\cdot) = 1 \) in states where \( E_s[y(b)] > E_s[y(a)] \). Maximum regret occurs in the state where \( E_s[y(a)|\delta(a) = 0] = 0 \) and \( E_s[y(b)|\delta(b) = 0] = 1 \). In this state, regret is

(13a) \[ R_c(\cdot) \cdot |E_s[y(b)] - E_s[y(a)]| = E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1] + P[\delta(b) = 0] - E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1]. \]

With choice of b, maximum regret symmetrically is

(13b) \[ R_c(\cdot) \cdot |E_s[y(b)] - E_s[y(a)]| = E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1] + P[\delta(a) = 0] - E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1]. \]

The difference between these maximum-regret expressions is

(14) \[ E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1] + P[\delta(b) = 0] - E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1] - E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1] - P[\delta(a) = 0] + E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1] \]

\[ = 2 \cdot (E[y(b)|\delta(b) = 1] \cdot P[\delta(b) = 1] - E[y(a)|\delta(a) = 1] \cdot P[\delta(a) = 1]) + P[\delta(b) = 0] - P[\delta(a) = 0]. \]

Treatment b (a) uniquely minimizes maximum regret if the value of (14) is positive (negative). The treatments yield the same maximum regret if the value of (14) is zero.
Now consider treatment choice with the ES rule. It is shown below that this rule minimizes maximum regret for some but not all values of the observable quantities. It selects the inferior treatment for other values.

The ES rule chooses treatment a if \( E[y(a)|\delta(a) = 1] > E[y(b)|\delta(b) = 1] \) and chooses b if \( E[y(a)|\delta(a) = 1] < E[y(b)|\delta(b) = 1] \). The ES rule does not prescribe a treatment if \( E[y(a)|\delta(a) = 1] = E[y(b)|\delta(b) = 1] \). Then some auxiliary criterion must be used. I henceforth consider the cases where the ES rule yields a determinate choice.

The ES rule yields the MMR choice when the ordering of \( E[y(a)|\delta(a) = 1] \) and \( E[y(b)|\delta(b) = 1] \) reverses the ordering of maximum regret across the two treatments. This necessarily occurs when \( P[\delta(a) = 1] = P[\delta(b) = 1] = \frac{1}{2} \). Then (14) reduces to \( E[y(b)|\delta(b) = 1] - E[y(a)|\delta(a) = 1] \).

When \( P[\delta(a) = 1] \neq P[\delta(b) = 1] \), the ES rule may minimize or maximize maximum regret, depending on the values of the observable quantities \( E[y(a)|\delta(a) = 1] \), \( E[y(b)|\delta(b) = 1] \), \( P[\delta(a)] \), and \( P[\delta(b)] \). I consider here cases in which treatment b has greater empirical success than a, so the ES rule chooses b. Analysis of the contrary case is symmetric.

To simplify notation, let \( E[y(b)|\delta(b) = 1] = m \) and \( E[y(a)|\delta(a) = 1] = m - \varepsilon \), where \( 0 < \varepsilon \leq m \leq 1 \). The findings vary depending on the relative sizes of \( P[\delta(b) = 1] \) and \( P[\delta(a) = 1] \). First consider cases where \( P[\delta(b) = 1] = \frac{1}{2} + k \) and \( P[\delta(a) = 1] = \frac{1}{2} - k \), where \( 0 < k < \frac{1}{2} \). Then the value of (14) is

\[
2[m\left(\frac{1}{2} + k\right) - (m - \varepsilon)\left(\frac{1}{2} - k\right) - k] = 2[mk + \varepsilon\left(\frac{1}{2} - k\right) - k] = 2[\varepsilon\left(\frac{1}{2} - k\right) - k(1 - 2m)].
\]

Hence, treatment b uniquely minimizes maximum regret if \( \varepsilon > k(1 - 2m)/(\frac{1}{2} - k) \) and treatment a does so if \( \varepsilon < k(1 - 2m)/(\frac{1}{2} - k) \). A sufficient condition for the former result is that \( m \geq \frac{1}{2} \). The latter result occurs if \( m < \frac{1}{2} \) and \( \varepsilon \) is sufficiently small.

Now consider cases where \( P[\delta(b) = 1] = \frac{1}{2} - k \) and \( P[\delta(a) = 1] = \frac{1}{2} + k \), where \( 0 < k < \frac{1}{2} \). Then the value of (14) is
\[2[m(\frac{1}{2} - k) - (m - \varepsilon)(\frac{1}{2} + k) + k)] = 2[-2mk + \varepsilon(\frac{1}{2} + k) + k] = 2[\varepsilon(\frac{1}{2} + k) - k(2m - 1)].\]

Hence, treatment b uniquely minimizes maximum regret if \(\varepsilon > k(2m - 1)/(\frac{1}{2} + k)\) and treatment a does so if \(\varepsilon < k(2m - 1)/(\frac{1}{2} + k)\). A sufficient condition for the former result is that \(m \leq \frac{1}{2}\). The latter result occurs if \(m > \frac{1}{2}\) and \(\varepsilon\) is sufficiently small.

Recall Box's statement that a model may be wrong but useful. The above derivation shows that using the model of random treatment selection to motivate the ES rule minimizes maximum regret in some contexts. However, it maximizes maximum regret in other contexts. Thus, the model is useful for treatment choice in some contexts but harmful in others.

4.4. Analysis of Treatment Response and Study of Systems of Jointly Determined Variables

The terminology “analysis of treatment response” has become widespread in empirical microeconomics since the 1990s, but it does not appear in early writing on econometrics. When Haavelmo wrote his thesis, a central focus of econometricians was identification and estimation of models of systems of jointly determined variables. An important objective was to use estimated models to predict the impacts of contemplated public policies. Adding further notation to denote realized treatments and outcomes, it is easy to see that analysis of treatment response lies within this longstanding concern of econometrics.

Let \(z\) denote the treatment that a person in the study population receives; thus, \(z = 1\) if \(\delta(b) = 1\) and \(z = 0\) if \(\delta(b) = 0\). Let \(y\) denote the outcome that a person realizes; thus, \(y = y(b)\) if \(z = 1\) and \(y = y(a)\) if \(z = 0\). Realizations of \((y, z)\) are observable. Random sampling of the study population asymptotically reveals the distribution \(P(y, z)\).

Early econometricians analyzed data on realized treatments and outcomes in settings where treatments are chosen purposefully rather than randomly. Haavelmo (1944) put it this way (p. 7):
“the economist is usually a rather passive observer with respect to important economic phenomena; he usually does not control the actual collection of economic statistics. He is not in a position to enforce the prescriptions of his own designs of ideal experiments.”

Analyzing settings with passive observation of treatments and outcomes, econometricians found it sensible to model \( y \) and \( z \) as jointly determined variables. A simple example of joint determination occurs if treatments are determined by outcome optimization with perfect foresight, as considered by Roy (1951). This model assumes that each person in the study population know the outcomes he would experience under each treatment and chooses the treatment that yields the better outcome.

When treatment selection is random, \( E(y|z = 1) = E[y(b)] \) and \( E(y|z = 0) = E[y(a)] \). Among the central contributions of early econometrics was to show that these equalities do not generally hold when \( y \) and \( y \) are jointly determined. Haavelmo (1943), Section 1 shows this in a simple context.

Viewing random treatment selection as implausible in settings with passive observation, econometricians have long studied other models that point-identify \( E[y(b)] \) and \( E[y(a)] \), given empirical knowledge of \( P(y, z) \). The early literature mainly studied linear models and distributional assumptions involving instrumental variables. Recent research weakens the assumption that relations are linear but continues to use instrumental variables.

Throughout all this long period, econometricians have struggled to measure model performance in decision making. As discussed in this paper, statistical decision theory provides a coherent framework for model evaluation.

5. Conclusion

To reiterate the central theme of this paper, use of statistical decision theory to evaluate econometric models is conceptually coherent and simple. A planner specifies a state space listing all the states of nature deemed feasible. One evaluates the performance of any contemplated decision rule by the state-dependent
vector of expected welfare that it yields. Decisions made using models are evaluated in this manner. Importantly, statistical decision theory evaluates model-based decision rules by their performance across the state space, not across the model space.

The primary challenge to use of statistical decision theory in practice is computational. Recall that, in his discussion sketching application of statistical decision theory to prediction, Haavelmo (1944) remarked that such application (p. 111): “although simple in principle, will in general involve considerable mathematical problems and heavy algebra.”

Many mathematical operations that were infeasible in 1944 are tractable now, as a result of important advances in analytical methods and revolutionary ones in numerical computation. Hence, it has increasingly become possible to use statistical decision theory when performing econometric research that aims to inform decision making. Future advances in analysis and numerical computation should continue to expand the scope of applications.
References


