Remarks on statistical inference for statistical decisions

Charles F. Manski

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP06/19

REMARKS ON STATISTICAL INFERENCE FOR STATISTICAL DECISIONS

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

January 2019

Abstract

The Wald development of statistical decision theory addresses decision making with sample data. Wald's concept of a statistical decision function (SDF) embraces all mappings of the form [data → decision]. An SDF need not perform *statistical inference*; that is, it need not use data to draw conclusions about the true state of nature. *Inference-based* SDFs have the sequential form [data → inference → decision]. This paper offers remarks on the use of statistical inference in statistical decisions. Concern for tractability may provide a practical reason for study of inference-based SDFs. Another practical reason may be necessity. There often is an institutional separation between research and decision making, with researchers reporting inferences to the public. Then planners can perform the mapping [inference → decision], but they cannot perform the more basic mapping [data → decision]. The paper first addresses binary choice problems, where all SDFs may be viewed as hypothesis tests. It next considers as-if optimization, where one uses a point estimate of the true state as if the estimate is accurate. It then extend this idea to as-if decisions using set estimates of the true state, such as confidence sets.

I am grateful to Joel Horowitz and Alex Tetenov for comments.

1. Introduction

The Wald (1950) development of statistical decision theory 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.

Wald's concept of a statistical decision function embraces all mappings of the form [data → decision]. In general, a SDF need not perform *statistical inference*. That is, it need not use sample data to draw conclusions about the true state of nature. The broad word "conclusions" has multiple formal interpretations in the literature on inference. To a Bayesian, it means a posterior distribution on the state space. To a frequentist, it means a point or set-valued estimate of the true state. Set-valued estimates include confidence sets and the results of hypothesis tests.

None of the most prominent decision criteria that have been studied from Wald's perspective—maximin, minimax-regret (MMR), and maximization of subjective average welfare—makes reference 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, they may do so. That is, they may have the sequential form [data → inference → decision], first performing some form of inference and then using the inference to make a decision. There seems to be no accepted term for such sequential procedures, so I shall call them *inference-based* SDFs.

A familiar use of inference in decision making is choice between two actions based on the result of a hypothesis test. This has been particularly common when using data from randomized trials to choose between two medical treatments. A common case views one treatment as the status quo and the other as an innovation. The null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better. A considerable body of medical literature recommends that the status quo treatment be used in clinical practice if the null is not rejected and that the innovation be used if it is rejected. Conventional tests fix the probability of Type I and Type II errors at predetermined values, typically 0.05 and 0.20. Manski and Tetenov (*PNAS*, 2016) give several reasons why such tests generally do not provide a satisfactory basis for treatment choice.

Another familiar practice uses the sample data to compute a point estimate of the true state of nature and then chooses an action that optimizes welfare as if this estimate is accurate. Researchers often motivate use of point estimates in decision making by citing limit theorems of asymptotic theory. They observe that an estimate is consistent, asymptotically normal, and so on. However, a planner does not seek an asymptotically attractive estimate of the true state. He seeks to make a good decision with potentially available sample data. Citation of favorable asymptotic properties of point estimates does not per se provide a firm foundation for their use in decision making.

A famous finding relating decision making to inference is that decisions using Bayesian inference maximize subjective average welfare. Wald studied maximization of subjective average welfare, also known as minimization of Bayes risk. This decision criterion makes no reference to inference. It simply places a subjective distribution on the state space and optimizes the resulting subjective average welfare. However, examination of the optimization problem shows that the solution is an inference-based SDF. One first performs Bayesian inference, which transforms the prior distribution on the state space into a posterior distribution without reference to a decision problem. One then choose an action that maximizes posterior subjective average welfare. See, for example, Berger (1985, Section 4.4.1).

Concern for tractability may provide a practical reason for study of inference-based SDFs. In principle, a planner wanting to apply statistical decision theory just needs to determine the admissible SDFs, specify a decision criterion, and apply it to his choice problem. In practice, it may be intractable to isolate the admissible SDFs and find one that optimizes a specified criterion. Some of the difficulty stems from the generality of Wald's definition of a statistical decision function, which embraces all mappings of the form [data → decision]. This function space encompasses all possibilities in principle, but the breadth of options may be hard to exploit in practice. When one is unable to apply statistical decision theory in generality, one may find it expedient to consider relatively simple rules that are amenable to analysis or computation. Inference-based SDFs may be plausible candidates.

Treatment choice with data from a randomized trial provides an apt example. Exact computation of the MMR treatment rule is generally complex. It is much simpler to compute the *empirical success* rule, which chooses the treatment with the highest observed average outcome in the trial. It has been found that the empirical success rule approximates the MMR rule well in common trial designs. See Manski (2004, 2007), Hirano and Porter (2009), and Stoye (2009).

Another practical reason for study of inference-based SDF may be necessity. Wald supposed that a planner observes data and makes a decision. In practice, there often is an institutional separation between research and decision making. Researchers observe data and report inferences to the public. Planners do not observe the data, only the reported inferences. Then planners can perform the mapping [inference → decision], but they cannot perform the more basic mapping [data → decision]. They can only choose among the inference-based SDFs made available by research reporting practices.

This paper offers some remarks on the use of statistical inference in statistical decisions. My aim is not to prove new theorems, but rather to make some conceptual points that I feel have not been adequately appreciated. Section 2 briefly reviews statistical decision theory and mentions some extensions. I then discuss various forms of inference-based SDFs. Section 3 addresses binary choice problems, where all SDFs may be viewed as hypothesis tests. Section 4 considers the familiar practice of as-if optimization, where one uses a point estimate of the true state as if the estimate is accurate. Section 5 extends this idea to as-if decisions using set estimates of the true state, such as confidence sets. Section 6 concludes.

This paper does not address motivations for statistical inference other than decision making. Researchers sometimes remark that they perform inference in the service of science or knowledge. I do not attempt to interpret this motivation.

2. Concepts of Statistical Decision Theory

Section 2.1 reviews standard concepts of decision theory for choice without sample data. Section 2.2 reviews standard concepts of statistical decision theory. Section 2.3 calls attention to some alternatives to the manner in which Wald measured the performance of SDFs.

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 knows (or believes) that the true state lies in set S, called the state space. The objective function $w(\cdot, \cdot)$: $C \times S \rightarrow R^1$ 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 $s^* \in S$.

A close to universally accepted prescription for decision making is that the planner should not choose a dominated action. 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, a planner who wants to maximize $w(\cdot, s^*)$ knows that choice of d weakly improves on choice of c.

Let D denote the undominated subset 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 consider only the undominated actions D, but it often is difficult to determine which actions are undominated. Hence, it is common to maximize over the full set C of feasible actions.

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

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

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 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

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

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

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

Here $\max_{d \in C} w(d, s) - w(c, s)$ is the *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.

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 and have received some attention; see, for example, Berger (1985). However, I will confine discussion to the polar cases in which the planner asserts a complete subjective distribution or none at all.

2.2. Statistical Decision Problems

Statistical decision problems add to the above structure by supposing that the planner observes data drawn from a sampling process that is informative about the true state of nature. Let Q denote the sampling distribution and let Ψ denote the sample space; that is, Ψ is the set of samples that may be drawn under Q. A statistical decision function $c(\cdot)$: $\Psi \rightarrow C$ maps the sample data into a chosen action. Let Γ denote the space of all SDFs mapping $\Psi \rightarrow C$.

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 $Q\{w[c(\psi), s]\}$, the ex ante distribution of welfare that it yields across realizations ψ 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 $c(\cdot)$ if $Q\{w[d(\psi), s]\}$ stochastically dominates $Q\{w[c(\psi), s]\}$. 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(·) by its expected welfare across samples; that is, $E\{w[c(\psi), s]\} \equiv \int w[c(\psi), s]dQ$. 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.

In practice, knowledge of the sampling distribution is incomplete. To express this, one extends the concept of the state space S to list the set of feasible sampling processes, namely $(Q_s, s \in S)$, and one evaluates c(·) by the state-dependent expected welfare vector $(E_s\{w[c(\psi), s]\}, s \in S)$. This done, the Wald theory uses the same ideas as does decision theory without sample data.

One first eliminates dominated SDFs. SDF c(·) is weakly dominated (inadmissible in Wald's terminology) if there exists another SDF d(·) such that $E_s\{w[d(\psi), s]\} \geq E_s\{w[c(\psi), s]\}$ for all $s \in S$ and $E_s\{w[d(\psi), s]\} > E_s\{w[c(\psi), s]\}$ for some $s \in S$. The subjective average welfare, maximin, and MMR criteria solve the problems

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

$$(5) \qquad \max_{c(\cdot) \in \Gamma} \ \min_{s \in S} \ E_s\{w[c(\psi), s]\},$$

$$(6) \qquad \min_{c(\cdot) \in \Gamma} \ \max_{s \in S} \ (\max_{d(\cdot) \in \Gamma} E_s\{w[d(\psi), s]\} - E_s\{w[c(\psi), s]\}).$$

The state-dependent optimization problem $\max_{d(\cdot) \in \Gamma} E_s\{w[d(\psi), s]\}$ within the MMR problem (6) is solved by choosing an action that maximizes state-dependent welfare w(·, s), whatever value the realized data may take. Hence, (6) reduces to

$$(6') \qquad \min_{c(\cdot) \in \Gamma} \ \max_{s \in S} \ (\max_{d \in C} w(d, s) - E_s\{w[c(\psi), s]\}).$$

Inspection of criteria (4) through (6′) shows that none of them makes explicit reference to inference.

Instead, each directly chooses an SDF that solves the specified problem. As mentioned earlier, the SDF that maximizes subjective average welfare turns out to perform Bayesian inference and then maximize posterior subjective expected welfare. However, the maximin and MMR criteria need not yield inference-based SDFs.

2.3. Alternatives to Expected Welfare for Evaluation of Performance

I wrote above that Wald's central idea was to evaluate a SDF by the distribution of welfare that it yields across realizations of the sampling process. His specific proposal to measure mean performance across realizations of the sampling process is reasonable. The mean of a probability distribution respects stochastic dominance. A considerable part of theoretical and applied study of probability has focused on the means of distributions. Nevertheless, measurement of mean performance across samples is not the only reasonable way to evaluate SDFs. One is to measure quantile welfare.

Considering decision making with probabilistic uncertainty but without sample data, Manski (1988) proposed measurement of the performance of a decision criterion by a quantile of the welfare distribution that it yields. I observed that maximization of expected and quantile welfare differ in important respects. Whereas the ranking of actions by expected welfare is invariant only to cardinal transformations of the objective function, the ranking by quantile welfare is invariant to ordinal transformations. Whereas expected welfare conveys risk preferences through the shape of the objective function, quantile welfare does so through the specified quantile, with higher values conveying more risk preference. Whereas expected welfare is not well-defined when the distribution of welfare has unbounded support with fat tails, quantile welfare is always well-defined.

With sample data, the λ-quantile of welfare in state s of SDF c(·) across realizations of the sampling process is

$$(7) \quad V_{\lambda s}\{w[c(\cdot), s]\} \equiv \min v \in R^1 \text{ s. t. } Q_s\{w[c(\psi), s] \le v\} \ge \lambda.$$

I have learned recently that use of quantile welfare to evaluate SDFs was suggested as early as Blyth (1970), in a brief aside (p. 1040).

Among $\lambda \in (0, 1)$, there is some reason to think that low quantiles of welfare distributions may matter to planners. Writings in finance have shown explicit concern with low quantiles of earnings distributions, commonly the 0.01 and 0.05 quantiles, using the term *value-at-risk*. See, for example, Jorion (2006).

It is generally difficult to analyze or compute the quantile welfare of SDFs. However, Manski and Tetenov (2014) observe that quantile welfare has a simple form in binary choice problems, shown in Section 3 below.

3. Binary Choices and Hypothesis Tests

Section 3.1 shows that SDFs for binary choice problems can always be viewed as hypothesis tests. Section 3.2 observes that statistical decision theory and classical hypothesis testing evaluate tests in fundamentally different ways.

3.1. Equivalence of SDFs and Tests

Let choice set C contain two actions, say C = {a, b}. A SDF $c(\cdot)$ partitions Ψ into two regions that separate the data yielding choice of each action. These regions are $\Psi_{c(\cdot)a} \equiv [\psi \in \Psi: c(\psi) = a]$ and $\Psi_{c(\cdot)b} \equiv [\psi \in \Psi: c(\psi) = b]$.

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 $S_a \equiv [s \in S: w(a, s) \geq w(b, s)]$ and $S_b \equiv [s \in S: w(b, s) > w(a, s)]$.

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.

3.2. Classical and Decision Theoretic Evaluation of Tests

3.2.1. Classical Testing

Let us review the basic practices of classical hypothesis testing. Classical 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 called 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 α, 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, classical testing favors tests 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 ρ be the metric measuring distance on S and let $\rho_a > 0$ be the specified minimum distance from

S_a. In the notation of statistical decision theory, classical testing favors tests that yield small values for max $\{Q_s[c(\psi) = a],\ s \in S_b \text{ s.t. } \rho(s, S_a) \geq \rho_a\}$.

The U. S. Food and Drug Administration (FDA) uses these practices of classical testing to approve new treatments. A firm wanting approval of a new drug or device performs randomized trials that compare the new treatment with an approved one or placebo. An FDA document providing guidance for the design of trials evaluating new medical devices states that the upper bound probability on a Type I error is conventionally set to 0.05 and that the constrained upper bound on the probability of a Type II error ordinarily should not exceed 0.20 (U.S. Food and Drug Administration, 1996). International Conference on Harmonisation (1999) provides similar guidance for the design of trials evaluating drugs.

3.2.2. Decision Theoretic Evaluation 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.

The welfare distribution 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,

$$(8) \qquad \begin{aligned} R_{c(\cdot)s} &= Q_s[c(\psi) = b] \quad \text{if } w(a, s) > w(b, s), \\ &= Q_s[c(\psi) = a] \quad \text{if } w(b, s) > w(a, s). \end{aligned}$$

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,

$$(9) \qquad E_s\{w[c(\cdot), s]\} = R_{c(\cdot)s}\{\min [w(a, s), w(b, s)]\} + [1 - R_{c(\cdot)s}]\{\max [w(a, s), w(b, s)]\}$$

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

The expression $R_{c(\cdot)s}\cdot|w(a, s) - w(b, s)|$ 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.

The above derivation makes plain that statistical decision theory finds it desirable to minimize error probabilities, as does classical testing. In contrast to classical testing, welfare function (9) views Type I and Type II errors symmetrically, and it recognizes that a planner should care about more than error probabilities. A planner should care as well about the magnitudes of the losses to welfare that arise when errors occur. Regret measures the joint effect of error occurrence and magnitude, on average across samples.

Departing from Wald's focus on expected welfare, one might alternatively measure performance by quantile welfare. In binary choice problems,

$$(10) \qquad V_{\lambda s}\{w[c(\cdot), s]\} = \min [w(a, s), w(b, s)] \text{ if } R_{c(\cdot)s} \geq \lambda,$$

$$= \max [w(a, s), w(b, s)] \text{ if } R_{c(\cdot)s} < \lambda.$$

Observe that mean and quantile performance both decrease in the error probability, falling from max [w(a, s), w(b, s)] to min [w(a, s), w(b, s)] as $R_{c(\cdot)s}$ increases from 0 to 1. However, the two measures of performance differ in the pattern of decrease. Whereas mean performance varies linearly with the error probability, quantile performance is a step function.

3.2.3. Example

Manski (2019) gives a simple hypothetical medical example illustrating how classical and decision theoretic evaluation of tests differ. I paraphrase here.

Suppose that a typically terminal form of cancer may be treated by a status quo treatment or an innovation. It is known from experience that mean patient life span with the status quo treatment is one year. Prior to use of the innovation, medical researchers see two possibilities for its effectiveness. It may be less effective than the status quo, yielding a mean life span of only 1/3 of a year, or it may be much more effective, yielding a mean life span of 5 years.

A randomized trial is performed to learn the effectiveness of the innovation. The trial data are used to perform a conventional test comparing the innovation and the status quo. The null hypothesis is that the innovation is no more effective than the status quo and the alternative is that the innovation is more effective. The probabilities of Type I and Type II errors are 0.05 and 0.20. The test result is used to choose between the treatments.

If the status quo treatment is superior, a Type I error occurs with sampling probability 0.05 and reduces mean patient life span by 2/3 of a year (1 year minus 1/3 year), so regret is 1/30 of a year. If the innovation is superior, a Type II error occurs with probability 0.20 and reduces mean patient life span by 4 years (5 years minus 1 year), so regret is 4/5 of a year. Use of the test to choose between the status quo and the innovation implies that society is willing to tolerate a large (0.20) chance of a large welfare loss (4 years) when making a Type II error, but only a small (0.05) chance of a small welfare loss (2/3 of a year) when making a Type I error. The theory of hypothesis testing does not motivate this asymmetry.

The maximum regret of the conventional test is 4/5 of a year. Rather than use this test to choose treatment, one could perform a test that has smaller maximum regret. A simple feasible option may be to reverse the conventional probabilities of Type I and Type II errors; thus, one might perform a test with a 0.20 chance of a Type I error and a 0.05 chance of a Type II error. If the status quo treatment is superior, the regret of this unconventional test is 2/15 of a year; that is, a 0.20 chance of a Type I error times a 2/3 of a year reduction in mean life span with improper choice of the innovation. If the innovation is superior, regret is 1/5 of a year; that is, a 0.05 chance of a Type II error times a 4-year reduction in mean life span with improper choice of the status quo. Thus, the maximum regret of the unconventional test is 1/5 of a year.

In this example, the unconventional test delivers much smaller maximum regret than does the conventional test. Other tests may perform even better.

4. As-If Optimization with Point Estimates of the True State

The Introduction mentioned the common practice of using sample data to compute a point estimate of the true state of nature and choice of an action that optimizes welfare as if this estimate is accurate. This section elaborates. Section 4.1 considers the practice in abstraction. Sections 4.2 and 4.3 describe research that has studied as-if optimization from the decision theoretic perspective of maximum regret.

4.1. General Considerations

A point estimate is a function $s(\cdot)$: $\Psi \to S$ that maps data into a state of nature. 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. Then one obtains the SDF $c_{s(\cdot)}(\cdot)$, where

$$(11) \quad c_{s(\psi)}(\psi) \equiv \underset{c \in C}{\operatorname{argmax}}\ w[c, s(\psi)], \qquad \psi \in \Psi.$$

Researchers often motivate use of point estimates in decision making by citing limit theorems of asymptotic theory. They hypothesize a sequence of sampling processes indexed by sample size N and a corresponding sequence of point estimates $s_N(\cdot)$: $\Psi_N \to S$. They show that the sequence is consistent; that is,

$s_N(\psi) \rightarrow s^*$ in probability as $N \rightarrow \infty$. They show that, with suitable regularity conditions on the welfare function and choice set, consistency of $s_N(\cdot)$ implies consistency of as-if optimization. That is, $\max_{c \in C} w[c, s_N(\psi)] \rightarrow \max_{c \in C} w(c, s^*)$ in probability as $N \rightarrow \infty$.

This asymptotic argument is suggestive, but it does not prove that as-if optimization has desirable properties when used with finite data samples. In general terms, statistical decision theory evaluates the performance of SDF (11) in state s by the sampling distribution $Q_s\{w[c_{s(\psi)}(\psi), s]\}$ that it yields for welfare. In particular, Wald measured performance by expected welfare $E_s\{w[c_{s(\psi)}(\psi), s]\}$. Decision theory calls for study of these finite-sample quantities, not asymptotics.

4.2. Bounds on Maximum Regret for Treatment Choice with the Empirical Success Rule

Decision-theoretic evaluation of as-if optimization has been performed in research that studies use of the empirical success (ES) rule to choose treatments with data from a randomized trial. The ES rule optimizes treatment choice as if the empirical distribution of outcomes observed in the trial equals the population distribution of treatment response. A SDF uses the trial data to allocate population members to treatments. Research on the ES rule has taken the objective of treatment choice to be to maximize the mean outcome of treatment across the population. Maximum regret measures the performance of SDFs.

Manski and Tetenov (2016) use large deviations inequalities for sample averages of bounded outcomes to obtain informative and easily computable upper bounds on the maximum regret of the ES rule applied with any number of treatments. Proposition 1 extends an early finding of Manski (2004) from two to multiple treatments. Proposition 2 derives a new large-deviations bound for multiple treatments.

Let K be the number of treatment arms and let M 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 are

(12) $(2e)^{-½}M(K - 1)n^{-½}$,

(13) $M (\ln K)^{½}n^{-½}$.

Result (12) provides a tighter bound than (13) for two or three treatments, whereas (13) gives a tighter bound for four or more treatments.

4.3. Maximum Regret for Best Point Prediction with Missing Data

Among the earliest statistical decision problems that have been studied is best point prediction of a real-valued outcome under square loss. With this loss function, the risk of a candidate predictor 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. The regret of a predictor is its MSE as an estimate of the mean. A MMR predictor minimizes maximum mean square error.

Hodges and Lehman (1950) derived the MMR predictor with data from a random sample when the outcome has known bounded support. Normalizing the support to be the interval [0, 1], Theorem 6.1 proves that the MMR predictor is $(m\sqrt{N} + ½)(\sqrt{N} + 1)^{-1}$, where N is sample size and m is the sample average outcome. This may be consider an inference-based SDF, using m to estimate the population mean outcome. However, the MMR predictor does not perform as-if optimization. Rather than consider m to be an accurate estimate of the population mean, it recognizes statistical imprecision by shrinking m toward the value ½.

Dominitz and Manski (2017) study best prediction under square loss with data from a random sample with some missing outcome data. The analysis assumes knowledge of the fraction of the population whose outcomes are unobservable, but it makes no assumption about the population distribution of missing outcomes. Thus, the population mean outcome is partially identified. Determination of the MMR predictor is analytically and computationally forbidding in this setting. As-if optimization yields a simple predictor with reasonable properties.

The identification region for the population mean 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.

Let P(z = 1) and P(z = 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. Dominitz and Manski prove that the maximum regret of the sample-analog midpoint predictor is $\frac{1}{4}[P(z = 1)^2/N + P(z = 0)^2]$.

5. As-If Decisions with Set Estimates of the True State

Rather than use sample data to compute a point estimate of the true state of nature, one might compute a set-valued estimate. 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. Practitioners of statistical inference have long computed confidence sets, these being set estimates with specified coverage probabilities. They typically have not used confidence sets or other set estimates to make decisions.

In principle, confidence sets and other set estimates could be used to form inference-based SDFs. Extending the idea of as-if optimization, one could act as if a set estimate is a state space. One could then apply the maximin or minimax-regret criterion posed in Section 2.1. Thus, one could solve the *as-if maximin* problem

$$(14) \qquad \max_{c \in C} \ \min_{s \in S(\psi)} \ w(c, s)$$

or the *as-if MMR* problem

$$(15)\quad \min_{c \in C} \max_{s \in S(\psi)} [\max_{d \in C} w(d, s) - w(c, s)].$$

The adjective "as-if" means that one solves the problem acting as if S(ψ) rather than S is the state space.

There is reason to think that, with judicious choice of the set estimate, (14) and (15) can provide useful SDFs. From a computational perspective, these problems are generally easier to solve than are the actual maximin and MMR problems with sample data, stated in (5) and (6). The as-if problems fix ψ and select one action c, whereas the actual problems require one to consider all potential sample data and choose a decision function c(·). The as-if problems involve computation of welfare values w(c, s), whereas the actual problems involve more complex expected welfare values $E_s\{w[c(\psi), s]\}$.

The suggestive asymptotic arguments often made for as-if optimization can be extended to as-if decisions with set estimates. Consider a sequence of sampling processes indexed by sample size N and a corresponding sequence of set estimates $S_N(\cdot): \Psi_N \rightarrow 2^S$. Let $H(s^*)$ denote the identification region for the true state s^*; that is, the subset of S that the sampling process would ideally reveal with observable population rather than sample data. Let the sequence of set estimates be consistent; that is, $S_N(\psi) \rightarrow H(s^*)$ in probability as $N \rightarrow \infty$. With suitable regularity conditions on the welfare function and choice set, it should be possible to show that consistency of $S_N(\cdot)$ implies consistency of as-if maximin and MMR decision making. That is,

$$(16)\quad \max_{c \in C} \min_{s \in S_N(\psi)} w(c, s) \rightarrow \max_{c \in C} \min_{s \in H(s^*)} w(c, s)$$

and

$$(17)\quad \min_{c \in C} \max_{s \in S_N(\psi)} [\max_{d \in C} w(d, s) - w(c, s)] \rightarrow \min_{c \in C} \max_{s \in H(s^*)} [\max_{d \in C} w(d, s) - w(c, s)]$$

in probability as $N \rightarrow \infty$.

Of course, this asymptotic argument is only suggestive. It does not prove that as-if maximin or MMR decision making has desirable properties when used in practice with finite data samples. As far as I am aware, there has been no decision theoretic study of as-if maximin and MMR to date.

When research on this subject is undertaken, I see no reason to restrict attention to set estimates that satisfy the coverage properties of confidence sets. A confidence set with uniform coverage probability α is a set estimate $S(\cdot)$ such that $Q_s[\psi\colon s \in S(\psi)] \geq \alpha$ for all $s \in S$. High coverage probability is neither necessary nor sufficient for a set estimate to have good decision theoretic properties.

A pragmatic reason to focus attention on confidence sets is that applied statisticians routinely report confidence sets rather than other forms of set estimates. Nevertheless, I encourage statisticians and econometricians to broaden their attention beyond confidence sets and learn what forms of set estimates may be particularly useful in decision making.

4. Conclusion

Neyman (1962) remarked (p. 16): "In my opinion, the inferential theory solves no problem but, given a degree of earnestness of the authors, may create the illusion that it does." Inference makes no explicit appearance in the general form of Wald's statistical decision theory. Nevertheless, some inference-based SDFs have desirable theoretic properties and are computationally tractable. When this is so, inference-based SDFs are worthy of study.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Blyth, C. (1970), “On the Inference and Decision Models of Statistics,” *The Annals of Mathematical Statistics*, 41, 1034-1058.

Dominitz, J. and C. Manski (2017), "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583-1605.

Hirano, K. and J. Porter (2009), “Asymptotics for Statistical Treatment Rules,” *Econometrica*, 77,1683-1701.

International Conference on Harmonisation (1999), “ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline,” *Statistics in Medicine*, 18, 1905-1942.

Jorion, P. (2006), *Value at Risk: The New Benchmark for Managing Financial Risk* (3rd ed.). New York: McGraw-Hill.

Manski, C. (1988), "Ordinal Utility Models of Decision Making Under Uncertainty," *Theory and Decision*, 25, 79-104.

Manski, C. (1989), “Anatomy of the Selection Problem,” *Journal of Human Resources*, 24, 343-360.

Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2007), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press.

Manski, C. (2019), "Treatment Choice with Trial Data: Statistical Decision Theory should Supplant Hypothesis Testing," *The American Statistician*, 2019, forthcoming.

Manski, C. and A. Tetenov (2014), "The Quantile Performance of Statistical Treatment Rules Using Hypothesis Tests to Allocate a Population to Two Treatments," cemmap working paper CWP44/14.

Manski, C. and A. Tetenov (2016), “Sufficient Trial Size to Inform Clinical Practice,” *Proceedings of the National Academy of Sciences*, 113, 10518-10523.

Neyman, J. (1962), “Two Breakthroughs in the Theory of Statistical Decision Making,” *Review of the International Statistical Institute*, 30, 11-27.

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

U.S. Food and Drug Administration (1996), *Statistical Guidance for Clinical Trials of Nondiagnostic Medical Devices*,
http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm106757.htm.

Wald A. (1950), *Statistical Decision Functions*, New York: Wiley.