1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

Treatment Choice under Ambiguity Induced by Inferential Problems

Charles F. Manski

Department of Economics and Institute for Policy Research, Northwestern University

cfmanski@nwu.edu

Abstract

Inferential problems that arise in the empirical analysis of treatment response induce ambiguity about the identity of optimal treatment rules. This paper describes a research program that begins with general themes about decisions under ambiguity, next specializes to problems of treatment choice under ambiguity, and then shows how identification problems and statistical issues induce ambiguity in treatment choice.

Keywords. Identification, ambiguity, treatment response, bounds, statistical treatment rules

1 Introduction

A decision maker with a known choice set but an unknown objective function is said to face a problem of decision under ambiguity. A common source of ambiguity is lack of knowledge of a probability distribution describing a relevant population. Empirical research seeks to draw conclusions about such distributions by combining assumptions with data. Inferential problems occurs when the available prior information and data do not suffice to reveal the distribution of interest. Thus, inferential problems induce ambiguity in decision making. [1]

I describe here my recent research connecting the empirical analysis of treatment response with the normative analysis of treatment choice under ambiguity.

In [25], [26], and [27], I have studied the decision problem faced by a planner must choose a *treatment rule* assigning a treatment to each member of a heterogeneous population of interest. The planner might, for example, be a physician choosing medical treatments for each member of a population of patients or a judge deciding sentences for each member of a population of convicted offenders. The planner observes certain covariates for each person; perhaps demographic attributes, medical or criminal records, and so on. Each member of the population has a *response function* mapping treatments into a real-valued outcome of interest; perhaps a measure of health status in the case of the physician or a measure of recidivism in the case of the judge. The planner wants to choose a treatment rule that maximizes the population mean outcome; that is, the planner wants to maximize a utilitarian social welfare function.

In this setting, an optimal treatment rule assigns to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. It is unusual, however, for planners to have the knowledge needed to implement optimal rules. Identification problems and statistical issues in the empirical analysis of treatment response commonly combine to prevent planners from knowing the relevant conditional mean outcomes. Hence planners commonly face problems of treatment choice under ambiguity.

My concern has been to characterize this ambiguity in settings of practical interest. The program of research on nonparametric analysis of treatment response initiated in [18] and [19] and carried forward in [20], [21], [22], [23], [24], and [28] derives sharp bounds on conditional mean outcomes under alternative treatments, the form of the bounds depending on the available data and the maintained assumptions. These bounds determine the nature of the ambiguity that the planner faces.

Here is the organization of this paper. Sections 2 and 3 lay the normative foundations. Section 2 develops general themes about decisions under ambiguity and Section 3 formalizes the planner's treatment choice problem. Sections 4 and 5 describe how inferential

[1] The term *ambiguity* appears to originate in [7]. Ellsberg's famous experiment required subjects to draw a ball from either of two urns, one with a known distribution of colors and the other with an unknown distribution of colors. [13] and [14] used the term *uncertainty*, but uncertainty has since come to be used to describe optimization problems in which the objective function depends on a known probability distribution. Some modern authors have used *ignorance* as a synonym for ambiguity (e.g., [1] and [30]). Authors writing on decision making with unknown subjective probability distributions may refer to *robust Bayesian analysis* (e.g., [3] or to decision making with *lower/upper probabilities* (e.g., [5], [6], [35]), *subjective probability domains* (e.g., [16]), or *imprecise probabilities* (e.g., [38]).

problems induce ambiguity. Section 4 sets forth the fundamental identification problem arising in empirical analysis of treatment response. Section 5 addresses the statistical problem of induction from finite samples to populations.

2 Decisions Under Ambiguity

2.1 Basic Ideas

We begin with a choice set C and a decision maker who must choose an action from C. The decision maker wants to maximize on C an objective function $f(\cdot)$: $C \to R$ mapping actions into real-valued outcomes. The decision maker faces an optimization problem if he knows the choice set C and the objective function $f(\cdot)$. He faces a problem of decision under ambiguity if he knows the choice set but not the objective function. Instead, he knows only that $f(\cdot) \in F$, where F is some set of functions mapping C into R.

Knowing that $f(\cdot) \in F$, how should the decision maker choose among the feasible actions? Clearly he should not choose a *dominated* action. Action $d \in C$ is said to be dominated (also *inadmissible*) if there exists another feasible action, say c, such that $g(d) \leq g(c)$ for all $g(\cdot) \in F$ and $g(d) < g(c)$ for some $g(\cdot) \in F$.

Let D denote the undominated subset of C. How should the decision maker choose among the elements of D? Let c and d be two undominated actions. Then either [$g(c) = g(d)$, all $g(\cdot) \in F$] or there exist $g'(\cdot) \in F$ and $g''(\cdot) \in F$ such that [$g'(c) > g'(d)$, $g''(c) < g''(d)$]. In the former case, c and d are equally good choices and the decision maker is indifferent between them. In the latter case, the decision maker cannot order the two actions. Action c may yield a better or worse outcome than action d; the decision maker cannot say which. Thus the normative question "How should the decision maker choose?" has no unambiguously correct answer.

2.2 Transforming Decisions under Ambiguity into Optimization Problems

Although there is no optimal choice among undominated actions, decision theorists have not wanted to abandon the idea of optimization. So they have proposed various ways of transforming the unknown objective function $f(\cdot)$ into a known function, say $h(\cdot)$: $C \to R$, that can be maximized on D. Three leading proposals – the maximin rule, Bayes decision rules, and imputation rules – are discussed here. Although these proposals differ in their details, they share a key common feature. In each case the solvable optimization problem, $\max_{i \in D} h(\cdot)$, differs from the problem that the decision maker wants to solve, namely $\max_{i \in D} f(\cdot)$. The welfare level that is attained under the solvable optimization problem is $f[\text{argmax}_{i \in D} h(\cdot)]$, not $\max_{i \in D} f(\cdot)$.

The Maximin Rule: Wald [37] proposed that the decision maker should choose an action that maximizes the minimum welfare attainable under the functions in F. Formally,

Maximin Rule: For each $d \in D$, let $h(d) \equiv \inf_{g(\cdot) \in F} g(d)$. Maximize $h(\cdot)$ on D.

The maximin rule has a clear normative foundation in *competitive games*. In a competitive game, the decision maker chooses an action from C. Then a function from F is chosen by an opponent whose objective is to minimize the realized outcome. A decision maker who knows that he is a participant in a competitive game does not face ambiguity. He faces the problem of maximizing the known function $h(\cdot)$ specified in the maximin rule.

There is no compelling reason why the decision maker should or should not use the maximin rule when $f(\cdot)$ is a fixed but unknown objective function. In this setting, the appeal of the maximin rule is a personal rather than normative matter. Some decision makers may deem it essential to protect against worst-case scenarios, while others may not. Wald himself did not contend that the maximin rule is optimal, only that it is "reasonable." Considering the case in which the objective is to minimize rather than maximize $f(\cdot)$, he wrote ([37], page 18): "a minimax solution seems, in general, to be a reasonable solution of the decision problem."

Bayes Decision Rules: Bayesian decision theorists assert that a decision maker who knows only that $f(\cdot) \in F$ should choose an action that maximizes some average of the elements of F. Formally,

Bayes Decision Rule: Place a σ-algebra Σ and some probability measure π on the function space F. Let $h(\cdot) \equiv \int g(\cdot) d\pi$. Maximize $h(\cdot)$ on D.

Bayesian decision theorists recommend that π should express the decision maker's personal beliefs about where $f(\cdot)$ lies within F.

Bayesians offer various rationality arguments for use of Bayes decision rules. The most basic of these is that Bayes decision rules generally yield undominated actions provided that the expectations $\int g(\cdot) d\pi$ are finite ([3], page 253). This and other rationality arguments do not, however, fully answer the decision maker's bottom-line question: how well does the rule perform?

Consider, for example, the famous axiomatic approach of Savage [34]. Savage shows that a decision maker whose choices are consistent with a specified set of axioms can be interpreted as using a Bayes decision rule. Many

decision theorists consider the Savage axioms, or other sets of axioms, to be a priori appealing. Acting in a manner that is consistent with these axioms does not, however imply that chosen actions yield good outcomes. Berger [3] calls attention to this, stating (page 121): "A Bayesian analysis may be 'rational' in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used."

Even use of an "appropriate" prior distribution π does not imply that the decision maker should choose an action that maximizes the π-average of the functions in F. Suppose that π has actually been used to draw $f(\cdot)$ from F; that is, let π describe an objective random process and not just the decision maker's subjective beliefs. Even here, where use of π as the prior distribution clearly is appropriate, Bayesian decision theory does not show that maximizing the π-average of F is superior to other decision rules in terms of the outcome it yields. A decision maker wanting to obtain good outcomes might just as reasonably choose an action that maximizes a π-quantile of F or some other parameter of F that respects stochastic dominance (see [17]).

Imputation Rules: A prevalent practice among applied researchers is to act as if one does know $f(\cdot)$. One admits to not knowing $f(\cdot)$ but argues that pragmatism requires making some "reasonable," "plausible," or "convenient" assumption. Thus one somehow imputes the objective function and then chooses an action that is optimal under the imputed function. Formally,

Imputation Rule: Select an $h(\cdot) \in F$. Maximize $h(\cdot)$ on D.

Imputation rules are essentially Bayes rules placing probability one on a single element of F.

2.3 Ambiguity Untransformed

Decision theorists have long sought to transform decisions under ambiguity into optimization problems. Yet the search for an optimal way to choose among undominated actions must ultimately fail. Let us face up to this. What then?

Simply put, normative analysis changes its focus from optimal actions to undominated actions. In optimization problems, the optimal actions and the undominated actions coincide, the decision maker being indifferent among all undominated actions. In decisions under ambiguity, there may be undominated actions that the decision maker cannot order.

3 Treatment Choice Under Ambiguity

I now formalize the problem of treatment choice. I suppose that there is a finite set T of treatments and a planner who must choose a treatment rule assigning a treatment in T to each member of a population J. Each person $j \in J$ has a *response function* $y_j(\cdot): T \to Y$ mapping treatments into real-valued outcomes $y_j(t) \in Y$. A *treatment rule* is a function $\tau(\cdot): J \to T$ specifying which treatment each person receives. Thus, person j's outcome under rule $\tau(\cdot)$ is $y_j[\tau(j)]$. [2]

The planner is concerned with the distribution of outcomes across the population, not with the experiences of particular individuals. With this in mind, I take the population to be a probability space, say (J, Ω, P), where Ω is the σ-algebra on which probabilities are defined and P is the probability measure. Now the population mean outcome under treatment rule $\tau(\cdot)$ is

$$(1)\quad E\{y_j[\tau(j)]\} \equiv \int y_j[\tau(j)]dP(j).$$

I assume that the planner wants to choose a treatment rule that maximizes $E\{y_j[\tau(j)]\}$.

I suppose that the planner observes certain covariates $x_j \in X$ for each member of the population. The planner cannot distinguish among persons with the same observed covariates. Hence he cannot implement treatment rules that systematically differentiate among these persons. With this in mind, I take the feasible rules to be the set of functions mapping the observed covariates into treatments.[3]

To formalize this, let Z denote the space of all functions mapping X into T. Then the feasible rules have the form

$$(2)\quad \tau(j) = z(x_j), \qquad j \in J,$$

where $z(\cdot) \in Z$. Let $P[y(\cdot), x]$ be the probability measure on $Y^T \times X$ induced by P(j) and let $E\{y[z(x)]\} \equiv \int y[z(x)]dP[y(\cdot), x]$ denote the expected value of $y[z(x)]$ with respect to this induced measure. The planner wants to solve the problem

$$(3)\quad \max_{z(\cdot) \in Z} E\{y[z(x)]\}.$$

---

[2] This notation maintains the assumption of "individualistic treatment" made commonly, albeit often only implicitly, in analyses of treatment response. Individualistic treatment means that each person's outcome may depend on the treatment he receives, but not on the treatments received by other persons.

[3] Although the planner cannot systematically differentiate among persons with the same observed covariates, he can randomly assign different treatments to such persons. Thus the set of feasible treatment rules in principle contains not only functions mapping covariates into treatments but also probability mixtures of these functions. Explicit consideration of randomized treatment rules would not substantively change the analysis of this paper, but would complicate the necessary notation. A simple implicit way to permit randomized rules is to include in x a component whose value is randomly drawn by the planner from some distribution. The planner can then make the chosen treatment vary with this covariate component.

The solution to this problem is to assign to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. Let 1[·] be the indicator function taking the value one if the logical condition in the brackets holds and the value zero otherwise. For each $z(\cdot) \in Z$, use the law of iterated expectations to write

$$(4)\quad E\{y[z(x)]\} = E\{E\{y[z(x)]^*x\}\}$$
$$= E\{ \Sigma_{t \in T} E[y(t)^*x] \cdot 1[z(x) = t]\}$$
$$= \int \Sigma_{t \in T} E[y(t)^*x] \cdot 1[z(x) = t]\, dP(x).$$

For each $x \in X$, the integrand $\Sigma_{t \in T} E[y(t)^*x] \cdot 1[z(x) = t]$ is maximized by choosing z(x) to maximize $E[y(t)^*x]$ on $t \in T$. Hence a treatment rule $z^*(\cdot)$ is optimal if, for each $x \in X$, $z^*(x)$ maximizes $E[y(t)^*x]$ on $t \in T$. The optimized population mean outcome is $E\{\max_{t \in T} E[y(t)^*x]\}$.

A planner who knows the conditional mean outcomes $E[y(\cdot)^*x]$, $x \in X$ can implement an optimal treatment rule. The planner faces a problem of treatment choice under ambiguity if he does not know $E[y(\cdot)^*x]$, $x \in X$. Suppose the planner knows only that the population (covariate, response function) distribution $P[x, y(\cdot)]$ lies within a specified set Φ of possible (covariate, response function) distributions. The planner may then partition the feasible treatment rules into dominated and undominated subclasses. A feasible treatment rule $z(\cdot)$ is dominated if there exists another feasible rule, say $z'(\cdot)$, such that

$$(5a)\quad \int y[z(x)]d\phi \leq \int y[z'(x)]d\phi, \qquad \text{all } \phi \in \Phi,$$

$$(5b)\quad \int y[z(x)]d\phi < \int y[z'(x)]d\phi, \qquad \text{some } \phi \in \Phi.$$

A treatment rule $z(\cdot)$ is undominated if no such $z'(\cdot)$ exists. A planner facing a problem of treatment under ambiguity can eliminate dominated rules as sub-optimal but cannot choose optimally among rules that are unordered.

4 Identification of Treatment Response

I now turn to the problem of empirical inference on $E[y(\cdot)^*x]$, $x \in X$. My first concern is identification.

4.1 The Observability of Response Functions

Empirical inference on treatment response faces a fundamental difficulty. Consider any person $j \in J$. By definition, treatments are mutually exclusive. Hence it is logically impossible to observe the vector $[y_j(t), t \in T]$ of outcomes that person j would experience under all treatments. It is at most possible to observe the outcome that j realizes under the treatment that this person actually receives.[4]

Even the realized outcome is observable only retrospectively, after a person's treatment has been chosen. Nothing about response function $y_j(\cdot)$ is observable prospectively, before the treatment decision. Facing this further difficulty, empirical researchers commonly (albeit often only implicitly) assume the existence of two populations having the same distribution of covariates and response functions, or at least the same conditional mean response functions. One is the *population of interest*, which I have denoted J. The other is a *treated population*, say K, in which treatments have previously been chosen and outcomes realized.

Let $s(\cdot): K \to T$ denote the *status quo* treatment rule; that is, the rule actually applied in the treated population. Then the realized (covariate, treatment, outcome) triples $\{x_k, s(k), y_k[s(k)]; k \in K\}$ are observable. Under the assumption that populations J and K are distributionally identical, observation of the treated population can reveal the distribution $P[x, s, y(s)]$ of (covariate, treatment, outcome) triples that would be realized in the population of interest if treatment rule $s(\cdot)$ were to be applied there.

In this section, which focuses on identification, I maintain the idealized assumption that the planner observes the entire treated population, or at least an infinite random sample, and so knows the distribution $P[x, s, y(s)]$. Section 5 addresses the problem of statistical inference that arises when a finite sample from this population is available.

4.2 All Feasible Treatment Rules Are Undominated

What is the set of undominated treatment rules given empirical knowledge of $P[x, s, y(s)]$ but no maintained assumptions about the process generating realized treatments and outcomes? A straightforward extension of the analysis of [18], [19] shows that this question has a simple but unpleasant answer: All feasible treatment rules are undominated.

Let K_0 and K_1 denote the lower and upper endpoints of the logical range of the response functions. If outcomes

[4] The mutual exclusivity of treatments has been a central theme of empirical research on the analysis of treatment response. Mutual exclusivity of treatments is the reason why the term *experiment* is generally taken to mean a *randomized* experiment, in which each person receives one randomly chosen treatment [8]. A different perspective is found in the economic theory literature on revealed preference analysis. Here, it is sometimes assumed that treatments are not mutually exclusive or, equivalently, that persons receiving different treatments have the same response function. Varian [36], for example, supposes that an analyst observes multiple realized (treatment, outcome) pairs for a given individual j. He investigates how these observations may be used to learn about j's response function $y_j(\cdot)$.

are binary, for example, then $K_0 = 0$ and $K_1 = 1$. If outcomes can take any non-negative value, then $K_0 = 0$ and $K_1 = \infty$. For each $t \in T$ and $x \in X$, use the law of iterated expectations to write

(6) $E[y(t)*x] = E[y(t)*x, s = t]\cdot P(s = t*x) + E[y(t)*x, s \neq t]\cdot P(s \neq t*x).$

Empirical knowledge of P[x, s, y(s)] implies knowledge of $E[y(t)*x, s = t]$, $P(s = t*x)$, and $P(s \neq t*x)$ but reveals nothing about $E[y(t)*x, s \neq t]$. We know only that the last quantity lies in the interval $[K_0, K_1]$. Hence $E[y(t)*x]$ lies within this sharp bound:

(7) $E[y(t)*x, s = t]\cdot P(s = t*x) + K_0\cdot P(s \neq t*x) \leq E[y(t)*x] \leq E[y(t)*x, s = t]\cdot P(s = t*x) + K_1\cdot P(s \neq t*x).$

Now let us compare two treatment rules. Under one rule, all persons with covariates x receive treatment t'. Under the other rule, all such persons receive a different treatment, say t''. In the absence of any empirical evidence on treatment response, we would be able to say only that $E[y(t'')*x] - E[y(t')*x] \in [K_0 - K_1, K_1 - K_0]$. With the available empirical evidence, (7) yields a narrower bound on $E[y(t'')*x] - E[y(t')*x]$. The sharp lower (upper) bound is the lower (upper) bound on $E[y(t'')*x]$ minus the upper (lower) bound on $E[y(t')*x]$. Thus

(8) $E[y(t'')*x, s = t'']\cdot P(s = t''*x) + K_0\cdot P(s \neq t''*x) - E[y(t')*x, s = t']\cdot P(s = t'*x) - K_1\cdot P(s \neq t'*x) \leq E[y(t'')*x] - E[y(t')*x] \leq E[y(t'')*x, s = t'']\cdot P(s = t''*x) + K_1\cdot P(s \neq t''*x) - E[y(t')*x, s = t']\cdot P(s = t'*x) - K_0\cdot P(s \neq t'*x).$

This bound is a subset of the interval $[K_0 - K_1, K_1 - K_0]$. Its width is $(K_1 - K_0)\cdot[P(s \neq t''*x) + P(s \neq t'*x)]$, which can be no smaller than $(K_1 - K_0)$. Hence the bound (8) necessarily contains the value zero. Thus the empirical evidence alone does not reveal which treatment, t' or t", yields the larger mean outcome. The same reasoning holds for all pairs of treatments and for all values of x. Hence all feasible treatment rules are undominated.

It is important to understand that this negative finding does not imply that the planner should be paralyzed, unwilling and unable to choose a treatment rule. What it does imply is that, using empirical evidence alone, the planner cannot claim optimality for whatever treatment rule he does choose. The planner might, for example, apply the maximin rule. This calls for each person with covariates x to receive the treatment that maximizes the lower bound in (7). The planner cannot claim that this rule is optimal, but he may find some solace in the fact that it fully protects against worst-case scenarios.

4.3 Credibility of Identifying Assumptions

Although there are fundamental limits to the observability of response functions, there are no limits other than internal consistency to the assumptions that one can impose. Further conclusions about the mean response functions $E[y(\cdot)*x]$, $x \in X$ can be deduced, and ambiguity in treatment choice reduced, if empirical knowledge of P[x, s, y(s)] is combined with maintained assumptions.

The prevailing practice in the literature on treatment response has been to combine observations of realized (covariates, treatments, outcomes) with assumptions strong enough to identify mean response functions. Researchers applying these strong assumptions, however, have commonly found it difficult to justify them. There is a need to face up to the fact that imposing assumptions that are not credible does not really eliminate ambiguity in treatment choice. I discuss the three main approaches below.

Exogenous Treatment Selection: Certainly the most well known and often used way to identify mean response functions is to impose the non-testable assumption [5]

(9) $E[y(t)*x] = E[y(t)*x, s = t].$

Empirical knowledge of P[x, s, y(s)] implies knowledge of the right side of (9); hence $E[y(t)*x]$ is identified. Researchers asserting assumption (9) may say that treatment selection is *exogenous* or *random* or *ignorable* conditional on x. See [10], [15], and [33].

The assumption of exogenous treatment selection is well-motivated in classical randomized experiments [8]. Here the status quo treatment rule $s(\cdot)$ involves a planner who randomly assigns treatments to the members of the treated population, all of whom comply with the assigned treatment. Hence s is necessarily statistically independent of $[x, y(\cdot)]$. Equation (9) is an immediate consequence of this statistical independence.

The assumption is typically difficult to motivate in experiments that deviate from the classical ideal and in non-experimental settings, especially those in which the status quo treatments are self-selected by the members of the treated population (see[22] and [9]). In these cases, the assumption is often no more than an imputation rule (see Section 2.2).

---

[5] Assumption (5) is not testable because $E[y(t)*x, s \neq t]$ is not observable. Hence there is no empirical basis for refutation of the hypothesis $E[y(t)*x, s \neq t] = E[y(t)*x, s = t]$, which implies (9).

Latent-Variable Models: When status quo treatments are self-selected, it is easier to argue that treatment selection is not exogenous than to find a credible alternative assumption that identifies mean outcomes. Some researchers have proposed *latent-variable models* that jointly explain treatment and response. These models make assumptions about the form of P[s, y(·)*x]. If the assumptions are sufficiently strong, combining them with empirical knowledge of P[x, s, y(s)] identifies the mean outcomes E[y(t)*x]. See, for example, [4], [10], and [15].

The use of latent-variable models to identify treatment effects has been quite controversial. Some researchers have regarded these models as ill-motivated imputation rules whose functional form and distributional assumptions lack foundation. Others have viewed them as credible assumptions.

Instrumental Variable Assumptions and Constant Treatment Effects: In situations where outcomes are continuous, mean outcomes can be identified by combining an *instrumental variable assumption* with the assumption of *constant treatment effects*. The classical econometric research on linear response models that began in the 1920s and crystallized by the early 1950s invokes these assumptions.

An instrumental variable assumption holds that mean response is constant across sub-populations defined by different values of some covariate.[6] Let $x \equiv (w, v)$. Covariate v, taking values in a space V, is said to be an instrumental variable if, for $t \in T$, each value of w, and all $(u, u') \in (V \times V)$,

$$(10)\quad E[y(t)|w, v = u'] = E[y(t)|w, v = u].$$

The constant-treatment-effect assumption is that the response functions $y_j(\cdot), j \in J$ are parallel to one another. That is, there exists a function $g(\cdot)$: $T \to R$ and a set of real constants α_j, $j \in J$, such that

$$(11)\quad y_j(t) = g(t) + \alpha_j.$$

The controversy surrounding latent-variable models re-appears in applications that assume constant treatment effects. Whereas applied researchers sometimes feel that they can plausibly assert an instrumental variable assumption, the assumption of constant treatment effects usually strains credibility. In particular, this assumption implies that it is optimal to assign the same treatment to every member of the population, namely the treatment that maximizes $g(\cdot)$ on T.

4.4 Instrumental Variable Bounds

I have thus far described two polar informational cases. In the absence of prior information, observation of a treated population reveals something about mean treatment response but not enough to conclude that any rule is dominated (Section 4.2). Empirical knowledge combined with strong assumptions can identify mean response and enable optimization, but these strong assumptions are only rarely credible (Section 4.3).

Consideration of intermediate cases opens new inferential possibilities, with implications for treatment choice. Empirical knowledge combined with weak assumptions may imply non-overlapping bounds for mean outcomes under some alternative treatments. When this happens, the planner can partially order the feasible rules.

In the past ten years, a literature deriving bounds under various assumptions has begun to take form.. To illustrate the possibilities, I describe here the simple sharp bound under the instrumental variable (IV) assumption (10) obtained in [19].[7] This bound characterizes the identifying power of an IV assumption alone, not combined with the constant treatment effects assumption or any other prior restriction. Thus, it speaks to the situation of a planner who finds an IV assumption credible but does not want to predicate his choice of treatment rule on other assumptions.

The starting point for determination of the identifying power of an IV assumption is the no-assumptions bound on $E[y(t)|w, v]$ given in equation (7). Under the IV assumption, $E[y(t)|w, v = u]$ is constant across $u \in V$. It follows that the common value of $E[y(t)|w, v = u]$, $u \in V$ lies in the intersection of the bounds (7) across the elements of V Any point in this intersection is feasible. Thus, for all $u \in V$, we obtain the common sharp bound

$$(12)\quad \sup_{u' \in V} [E(y|w, v = u', s = t)\cdot P(s = t|w, v = u') + K_0\cdot P(s \neq t|w, v = u')] \leq E[y(t)|w, v = u] \leq \inf_{u' \in V} [E(y|w, v = u', s = t)\cdot P(s = t|w, v = u') + K_1\cdot P(s \neq t|w, v = u')].$$

This is also the sharp bound on $E[y(t)|w]$. The IV bound (12) is necessarily a subset of the no-assumptions bound

[6] Consider, for example, the literature in labor economics on the returns to schooling. Here treatments are different levels or forms of schooling that a child may receive. The outcome of interest is the net benefit of each schooling treatment, commonly measured by life-cycle earnings. Labor economists often use attributes of a person's parents as instrumental variables. The argument is that, although parental attributes may affect the schooling treatment that children receive, they should not, on average, affect the life-cycle earnings that children would experience if they were to receive a given schooling treatment.

[7] Bounds under related assumptions are reported in [2], [12], [29], [31], and [32].

(7). It is a proper subset for some $u \in V$ if and only if the no-assumptions bounds for $u \in V$ do not all coincide.

Now compare two treatment rules. All persons with covariates w receive treatment t' under one rule, and all such persons receive a different treatment t'' under the other rule. We may use (12) to obtain a sharp bound on the average treatment effect E[y(t'')*w] - E[y(t')*w], just as we did in deriving (8) from (7). This bound, however, need not cover zero. If the lower IV bound on E[y(t'')*w] exceeds the upper IV bound on E[y(t')*w], we may conclude that the rule mandating treatment t'' dominates the rule mandating t', and vice versa.

I know of no general way to determine a priori whether a given IV assumption will be sufficiently powerful to yield non-overlapping bounds. It seems that one must compute the bound case-by-case. There is, however, an important special case in which an IV assumption has full identifying power. This is the case of exogenous treatment selection discussed in Section 4.3. Exogenous treatment selection is an IV assumption in which the instrumental variable v is the realized treatment s.

5 Statistical Treatment Rules

In Section 4, I supposed that the planner knows the distribution P[x, s, y(s)]. In practice, planners may observe only a finite sample of the treated population. The problem of statistical induction from sample to population then arises.

In [27], I apply Wald's concept of statistical decision functions to analyze the problem of treatment choice using sample data. Let Q denote a sampling process and let Ψ denote the associated *sample space*; that is, Ψ is the set of data samples that may be drawn under Q. Let $\mathbf{Z}$ denote the space of functions mapping $X \times \Psi$ into T. Then each function $\zeta(\cdot, \cdot) \in \mathbf{Z}$ defines a *statistical treatment rule*, or *STR*. Thus, an STR is a feasible rule whose identity depends in some way on the sample drawn.

One's perspective on a statistical treatment rule depends on whether one evaluates it before or after the sampling process is engaged. Let $\psi \in \Psi$ denote a sample that may potentially be drawn under Q and let $\psi^0 \in \Psi$ denote the sample that is actually drawn. Ex ante ψ is a random variable, so $\zeta(\cdot, \psi)$ is a random function of X. Ex post ψ^0 is a determinate element of Ψ, so $\zeta(\cdot, \psi^0)$ is a determinate function of X. Thus an STR is ex ante a random member of the set Z of feasible rules and ex post a determinate member of Z.

As did Wald, I evaluate statistical treatment rules from the ex ante perspective. In particular, I analyze the expected value under Q of the (ex ante random) population mean outcome

$$(13)\quad E\{y[\zeta(x, \psi)]\} = \int \sum_{t \in T} E[y(t)^* x] \cdot 1[\zeta(x, \psi) = t]\, dP(x).$$

This is

$$(14)\quad W(P, Q, \zeta) \equiv \int E\{y[\zeta(x, \psi)]\}\, dQ(\psi)$$
$$= \int [\int \sum_{t \in T} E[y(t)|x] \cdot 1[\zeta(x, \psi) = t]\, dP(x)]\, dQ(\psi)$$
$$= \int \sum_{t \in T} E[y(t)|x] \cdot Q[\zeta(x, \psi) = t]\, dP(x),$$

where $Q[\zeta(x, \psi) = t] \equiv \int 1[\zeta(x, \psi) = t] dQ(\psi)$ denotes the Q-probability of the event $[\zeta(x, \psi) = t]$. I refer to $W(P, Q, \zeta)$ as the *expected welfare* under rule ζ. In the literature on statistical decision theory, $-W(P, Q, \zeta)$ would be called the *risk* of statistical decision function ζ.[8]

One may in principle apply the expected welfare criterion to evaluate alternative statistical treatment rules ζ under varying assumptions and sampling processes. In [27], I apply the criterion in a simple setting of considerable practical interest. I evaluate two statistical treatment rules when the sample data are generated by a classical randomized experiment. Both rules embody the reasonable idea that persons should receive the treatment with the best empirical success rate, but they differ in their use of covariate and sample information.

The *conditional success* (CS) rule selects treatments with the best empirical success rates conditional on specified covariates. The *unconditional success* (US) rule selects a treatment with the best unconditional empirical success rate. Whereas the US Rule constrains the planner to choose the same treatment for all persons, the CS Rule permits the planner to treat persons with different covariates differentially. Whereas the US Rule has the planner compare success rates using the entire available sample, the CS Rule requires that the planner compare success rates in sub-samples.

There is an evident tension between use of covariate information and available sample size. I use the expected welfare criterion to characterize this tension and assess the implications for treatment choice. The main finding is a proposition giving finite-sample bounds on expected welfare under the two rules. The bounds, which rest on a large-deviations theorem of [11], yield explicit sample-size and distributional conditions under which the CS Rule dominates the US rule.

I also briefly consider the situation of a planner who can

[8] The convention in statistical decision theory has been to describe the planner as minimizing expected loss rather than as maximizing expected welfare. The loss associated with rule ζ is $-E\{y[\zeta(x, \psi)]\}$ and the risk is $-W(P, Q, \zeta)$.

choose what covariate information to observe. The planner should, ideally, want to observe covariates that best separate persons who differ in their optimal treatments.

Acknowledgments

This research was supported in part by United States National Science Foundation grant SBR-9726846.

References

[1] K. Arrow and L. Hurwicz. An Optimality Criterion for Decision-Making Under Ignorance. in D. Carter and J. Ford (editors),*Uncertainty and Expectations in Economics*, Oxford: Blackwell, 1972.

[2] A. Balke and J. Pearl. Bounds on Treatment Effects from Studies With Imperfect Compliance. *Journal of the American Statistical Association*, 92:1171-1177, 1997.

[3] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.

[4] A. Björklund and R. Moffitt. Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *Review of Economics and Statistics*, 69:42-49,1987.

[5] A. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Annals of Mathematical Statistics*, 38:325-339, 1967.

[6] A. Dempster. A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society*, Series B, 30:205-232, 1968.

[7] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75:643-669, 1961.

[8] R. Fisher. *The Design of Experiments*, London: Oliver and Boyd, 1935.

[9] R. Gronau. Wage Comparisons - A Selectivity Bias. *Journal of Political Economy*, 82:1119-1143, 1974.

[10] J. Heckman and R. Robb. Alternative Methods for Evaluating the Impact of Interventions. in J. Heckman and B. Singer (editors), *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press, 1985.

[11] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58:13-30, 1963.

[12] J. Hotz, C. Mullins, and S. Sanders. Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing. *Review of Economic Studies*, 64:575-603, 1997.

[13] J. Keynes. *A Treatise on Probability*. MacMillan, 1921.

[14] F. Knight. *Risk, Uncertainty, and Profit*. Boston: Houghton-Mifflin, 1921.

[15] G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, U.K.: Cambridge University Press, 1983.

[16] C. Manski. Learning and Decision Making When Subjective Probabilities Have Subjective Domains. *Annals of Statistics*, 9:59-65, 1981.

[17] C. Manski. Ordinal Utility Models of Decision Making Under Uncertainty. *Theory and Decision*, 25:79-104, 1988.

[18] C. Manski. Anatomy of the Selection Problem. *Journal of Human Resources*, 24:343-360, 1989.

[19] C. Manski. Nonparametric Bounds on Treatment Effects. *American Economic Review Papers and Proceedings* 80:319-323, 1990.

[20] C. Manski. The Selection Problem. in C. Sims (editor), *Advances in Econometrics, Sixth World Congress*, Cambridge, England: Cambridge University Press, 1994.

[21] C. Manski. *Identification Problems in the Social Sciences*. Cambridge, Mass.: Harvard University Press, 1995.

[22] C. Manski. Learning about Treatment Effects from Experiments with Random Assignment of Treatments. *Journal of Human Resources*, 31:707-733, 1996.

[23] C. Manski. The Mixing Problem in Programme Evaluation. *Review of Economic Studies*, 64:537-553, 1997.

[24] C. Manski. Monotone Treatment Response. *Econometrica*, 65:1311-1334, 1997.

[25] C. Manski. Treatment Choice in Heterogeneous Populations Using Experiments Without Covariate Data. in G. Cooper and S. Moral (editors), *Uncertainty in Artificial Intelligence, Proceedings of the Fourteenth Conference*, San Francisco, CA: Morgan Kaufmann, 379-385, 1998.

[26] C. Manski. Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice. *Journal of Econometrics*, forthcoming.

[27] C. Manski. Statistical Treatment Rules for

Heterogeneous Populations. Department of Economics, Northwestern University, 1999.

[28] C. Manski and D. Nagin. Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism. *Sociological Methodology*, 28:99-137, 1998.

[29] C. Manski and J. Pepper. Monotone Instrumental Variables: With an Application to the Returns to Schooling. Department of Economics, Northwestern University, 1999.

[30] E. Maskin. Decision-Making Under Ignorance with Implications for Social Choice. *Theory and Decision*, 11:319-337, 1979.

[31] J. Robins. The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies. in Sechrest, L., H. Freeman, and A. Mulley (editors), *Health Service Research Methodology: A Focus on AIDS*, NCHSR, U.S. Public Health Service, 1989.

[32] J. Robins and S. Greenland. Comment on Angrist, Imbens, and Rubin's 'Identification of Causal Effects Using Instrumental Variables'. *Journal of the American Statistical Association*, 91:456-458, 1996.

[33] P. Rosenbaum and D. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70:41-55, 1983.

[34] L. Savage. *The Foundations of Statistics*, New York: Wiley, 1954.

[35] G. Shafer, *A Mathematical Theory of Evidence*, Princeton, NJ: Princeton University Press, 1976.

[36] H. Varian. The Nonparametric Approach to Demand Analysis. *Econometrica*, 50:945-973, 1982.

[37] A. Wald. *Statistical Decision Functions*, New York: Wiley, 1950.

[38] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, London: Chapman & Hall, 1991.