TECHNICAL WORKING PAPER SERIES

STATISTICAL TREATMENT RULES FOR HETEROGENEOUS POPULATIONS: WITH APPLICATION TO RANDOMIZED EXPERIMENTS

Charles F. Manski

Technical Working Paper 242
http://www.nber.org/papers/T0242

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 1999

This research was supported in part by National Science Foundation grant SBR-9726846. I have benefitted from the opportunity to present this work in seminars at New York University, Northwestern University, the University of Bristol, and the University of Virginia. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Statistical Treatment Rules for Heterogeneous
Populations: With Application to Randomized Experiments
Charles F. Manski
NBER Technical Working Paper No. 242
May 1999

ABSTRACT

This paper uses Wald's concept of the risk of a statistical decision function to address the question: How should sample data on treatment response be used to guide treatment choices in a heterogeneous population? *Statistical treatment rules* (STRs) are statistical decision functions that map observed covariates of population members and sample data on treatment response into treatment choices. I propose evaluation of STRs by their *expected welfare* (negative risk in Wald's terms), and I apply this criterion to compare two STRs when the sample data are generated by a classical randomized experiment.

The rules compared both embody the reasonable idea that persons should be assigned the treatment with the best empirical success rate, but they differ in their use of covariate information. The *conditional success* (CS) rule selects treatments with the best empirical success rates conditional on specified covariates and the *unconditional success* (US) rule selects a treatment with the best unconditional empirical success rate. 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 Hoeffding, yield explicit sample-size and distributional conditions under which the CS Rule is superior to the US rule.

Charles F. Manski
Department of Economics
Northwestern University
Evanston, IL 60208
and NBER
cfmanski@nwu.edu

1. Introduction

This paper uses the Wald (1950) concept of the risk of a statistical decision function to integrate statistical analysis of treatment response with normative analysis of treatment choice. I begin from the premise that empirical studies of treatment response should aim to improve treatment choices. The question I address is: How should sample data on treatment response be used to guide treatment choices in a heterogeneous population?

Bayesian decision theory coherently integrates statistical and normative analysis, but rests on a subjective probabilistic foundation that has inhibited applications. Prevailing frequentist research practices connect statistical analysis to normative objectives only loosely, if at all. Many empirical studies test the null hypothesis that some treatment effect is zero, with no reference to a decision problem that might motivate the test. Many studies report estimates of treatment effects, but do not evaluate the estimation methods used from the perspective of a decision problem. Viewing this situation, I concluded that Wald's frequentist approach, once well-appreciated but not much used today, warrants renewed attention.

The class of treatment choice problems considered here combines some realism and some simplicity. As in some of my recent research focused on identification problems (Manski, 1997a, 1998, 1999), I assume that 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. These covariates determine the set of non-randomized treatment rules that are feasible to implement: the set of feasible such rules is the set of all functions mapping the observed covariates into treatments. 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. I assume that the planner wants to choose a feasible treatment rule that maximizes

the population mean outcome; in economic terms, the planner wants to maximize a utilitarian social welfare function. An early discussion of treatment choice from this perspective appears in Stafford (1985). Manski and Nagin (1998) uses this framework to analyze judges' sentencing decisions and Dehejia (1999) uses it to study treatment choices for welfare recipients. Heckman, Smith, and Clements (1997) propose alternatives to the idea of maximizing population mean outcome.

Under the above assumptions, an optimal treatment rule assigns to each member of the population a treatment that maximizes mean outcome conditional on the observed covariates. Planners rarely, however, have the knowledge of treatment response needed to implement optimal rules. Empirical analysis of treatment response seeks to provide this knowledge, but identification problems and statistical issues stand in the way. Over time, a body of research has illuminated the identification problems (e.g., Angrist, Imbens, and Rubin, 1996; Balke and Pearl, 1997; Heckman and Robb, 1985; Hotz, Mullins, and Sanders, 1997; Imbens and Angrist, 1994; Manski, 1990, 1995, 1997b; Manski, Sandefur, McLanahan, and Powers, 1992; Robins, 1989; Rosenbaum, 1995; Rosenbaum and Rubin, 1983) and some work has addressed the statistical issues from a Bayesian perspective (e.g., Dehejia, 1999; Rubin, 1978). Frequentist statistical inference, however, has remained focused on hypothesis testing and on the asymptotic behavior of estimates. Frequentist statistical inference requires re-orientation to better inform treatment choice. Wald's approach is well-suited to this task.

Section 2 sets out the treatment-choice problem formally. After specifying the planner's choice set and objective, I present the optimal treatment rule and characterize the value of covariate information. I then define statistical treatment rules, which are rules that map covariates and sample data into treatments. The term *statistical treatment rule*, or *STR* for short, recalls Wald (1950), who used the term *statistical decision function* to describe functions that map sample data into decisions. I also introduce the class of STRs that *condition on covariates*.

We immediately confront a conceptual question about the evaluation of STRs, and of statistical decision functions more generally. Should such rules be evaluated ex ante, using the sampling distribution of

data yet to be realized, or ex post, conditioning on the actual sample data realized? Wald's frequentist decision theory adopts the former perspective, while Bayesian decision theory largely adopts the latter. I evaluate STRs ex ante and I focus attention on the *expected welfare* (negative risk in Wald's terms) achieved by a rule. To partly close the gap between ex ante and ex post evaluation, I present a Lemma giving conditions under which the two approaches yield the same conclusions with high sampling probability.

The heart of this paper is Section 3, which applies the expected welfare criterion to evaluate two specific STRs when the sample data are generated by a classical randomized experiment. I consider a randomized experiment in order to focus cleanly on the problem of statistical inference, unencumbered with concerns about identification. I evaluate two simple rules that embody the reasonable idea that persons should be assigned the treatment with the best empirical success rate, but that differ in their use of covariate 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. The analysis in Section 3 characterizes this tension and assesses 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 Hoeffding (1963), yield explicit sample-size and distributional conditions under which the CS Rule is superior to the US rule. I use a numerical illustration to give a quantitative sense of these conditions.

Section 3 also draw implications for the reporting of covariate information in research articles describing randomized experiments. A prevalent practice has been to report estimates of conditional mean

outcomes only if a classical hypothesis test rejects the null hypothesis of zero treatment effect. This reporting criterion has no clear connection to treatment choice and may prevent implementation of rules that condition on covariates.

Section 4 closes the paper with a question that may have no fully satisfactory answer. Applying Wald's frequentist approach to statistical decision theory to the problem of treatment choice, we find that the expected-welfare ranking of alternative STRs depends on the population distribution of treatment response. The use of STRs to make treatment choices, however, arises when planners have only sample data on treatment response, not knowledge of the response distribution. How then should a planner use the expected welfare criterion to guide choice of a treatment rule? The frequentist and Bayesian literatures suggest pragmatic answers, but neither is complete. I use the CS and US rules to illustrate.

2. The Planner's Problem

I set out basic concepts and assumptions here. Section 2.1 follows the description in Manski (1999) of the planner's choice set and objective function. Section 2.2 derives the optimal treatment rule and the value of covariate information. Section 2.3 defines statistical treatment rules and considers the ex ante and ex post evaluation of such rules. Section 2.4 introduces the class of STRs that condition on covariates and presents a lemma connecting the ex ante and ex post evaluation of such rules.

2.1. The Choice Set and Objective Function

I suppose that there is a finite set T of mutually exclusive and exhaustive treatments. A planner must choose a treatment rule assigning a treatment in T to each member of a population J. Treatment assignment

is sometimes referred to as *intention-to-treat*.

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 is assigned. Thus person j's outcome under rule $\tau(\cdot)$ is $y_j[\tau(j)]$. This notation maintains the assumption of individualistic treatment made commonly in analyses of treatment response. That is, a person's outcome may depend on the treatment he is assigned, but not on the treatments assigned to others.

The planner is concerned with the distribution of outcomes across the population, not with the outcomes of particular persons. Hence the population is taken 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 well-defined as

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

I assume that the planner wants to choose a treatment rule that maximizes $E\{y_j[\tau(j)]\}$. This criterion function has both normative and analytical appeal. Maximization of a population mean outcome, or perhaps some weighted average outcome, is the standard utilitarian criterion of the public economics literature on social planning. The linearity of the expectation operator yields substantial analytical simplifications, particularly through use of the law of iterated expectations.

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 and so cannot implement treatment rules that systematically differentiate among these persons. Hence the feasible non-randomized rules are functions mapping the observed covariates into treatments. I do not explicitly consider randomized treatment rules, but there is a simple implicit way to permit such rules. Let x include a component whose value is randomly drawn by the planner from some distribution. Then the planner can make the chosen treatment vary with this

covariate component.

To formalize the planner's problem, let Z denote the space of all functions mapping X into T. Let $z(\cdot) \in Z$. Then the feasible treatment rules have the form

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

Let $P[y(\cdot), x]$ be the probability measure on $Y^T \times X$ induced by P(j). 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. Then the planner wants to solve the problem

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

In practice, institutional constraints may restrict the feasible treatment rules to some proper subset of the space Z. In particular, the planner may be precluded from using certain covariates (say race or gender) to assign treatments. The analysis in this paper continues to hold if x is defined to be the covariates that the planner is permitted to consider, rather than the full vector of covariates that the planner observes.

2.2. Optimal Treatment Rules and the Value of Covariate Information

The solution to the planner's problem is to assign to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. Let $1[\cdot]$ 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

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

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 $x \in X$, $z^*(x)$ solves the problem

$$(5)\quad \max_{t \in T} E[y(t)^*x].$$

The optimized population mean outcome is $E\{\max_{t \in T} E[y(t)^*x]\}$.

The set of feasible treatment rules grows as more covariates are observed. Hence the optimal mean outcome achievable by the planner cannot fall, and may rise, as more covariates are observed. The value of covariate information is appropriately measured by the difference between the optimal mean outcome achievable with and without use of this information. This is

$$(6)\quad V(X) \equiv E\{\max_{t \in T} E[y(t)^*x]\} - \max_{t \in T} E[y(t)].$$

Inspection of (6) shows that covariate information has no value if there exists a common optimal treatment; that is, a $t^* \in T$ such that $z^*(x) = t^*$, almost everywhere on X. Covariate information does have value if optimal treatments vary with x.

More generally, we may compare the value of observing distinct covariate vectors, say x and w. A planner who knows the conditional mean outcomes $E[y(\cdot)|x]$ and $E[y(\cdot)|w]$ should prefer observation of x to w if and only if $E\{\max_{t \in T} E[y(t)^*x]\} \geq E\{\max_{t \in T} E[y(t)^*w]\}$. In words, the planner should prefer x to w if x better separates persons who differ in their optimal treatments.

Note that the present criterion for comparison of covariates x and w differs from the prediction

criterion familiar in statistical decision theory. The prediction criterion supposes that, for each $t \in T$, one wants to predict y(t) as well as possible in the sense of minimizing expected square loss. The best predictors conditional on x and w are E[y(t)*x] and E[y(t)*w] respectively. A statistician who knows $E[y(t)|x]$ and $E[y(t)|w]$ and wants to predict y(t) as well as possible should prefer x to w if and only if $E\{y(t) - E[y(t)*x]\}^2 \leq E\{y(t) - E[y(t)*w]\}^2$.

2.3. Statistical Treatment Rules

A planner who does not know the conditional mean outcomes $E[y(\cdot)|x]$, $x \in X$ generally cannot implement an optimal treatment rule. Suppose, however, that the planner has sample data that enable statistical inference on $E[y(\cdot)*x]$, $x \in X$. Then the planner may use these data to choose a treatment rule.

Considering this in some abstraction, 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, following Wald (1950), 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 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.

I evaluate statistical treatment rules from the ex ante perspective. The (ex ante random) population mean outcome under a specified rule ζ is

$$(7)\quad E\{y[\zeta(x, \psi)]\} = \int \sum E[y(t)*x]\cdot 1[\zeta(x, \psi) = t]\, dP(x) .$$

$t \in T$

The Q-expected value of $E\{y[\zeta(x, \psi)]\}$ is

$$(8)\quad W(P, Q, \zeta) \equiv \int E\{y[\zeta(x, \psi)]\}\, dQ(\psi) = \int \Big[\int \sum_{t \in T} E[y(t)|x] \cdot 1[\zeta(x, \psi) = t]\, dP(x) \Big]\, dQ(\psi)$$

$$= \int \sum_{t \in T} E[y(t)|x] \cdot Q[\zeta(x, \psi) = t]\, dP(x),$$

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 and I use $W(P, Q, \cdot)$ to compare alternative rules. This criterion follows Wald except that he described decision makers as minimizing risk rather than as maximizing expected welfare. The loss under rule ζ is $-E\{y[\zeta(x, \psi)]\}$ and the risk is $-W(P, Q, \zeta)$.

I can offer two substantive and one technical reason for adopting the ex ante perspective in general and for focusing on expected welfare in particular. The first substantive reason, which is commonly given by statisticians performing ex ante evaluation of statistical decision functions, is that one may want to understand how such decision functions perform as *procedure*s in repeated applications (e.g., Berger, 1985, Section 1.6.2). In the present setting, this argument is appealing if one is a statistician recommending a treatment rule to be applied repeatedly in treatment choice problems with statistically independent sample data. Focusing on expected welfare in particular is appropriate if the statistician's objective is to maximize a utilitarian social welfare function aggregating outcomes across repetitions of the choice problem.

The above reasoning is not germane to a single planner concerned only with his own treatment choice problem, a point made with compelling logic by Bayesian critics of frequentist statistical theory (e.g., Berger, 1985, Section 1.6.3). A second substantive rationale may be relevant however. Suppose that institutional constraints require a planner to commit to an STR before observing the relevant sample data. The institutional constraint may, for example, reflect public distrust of the planner and a desire to limit his discretion. A planner

who is required to pre-commit must evaluate $\zeta(\cdot, \psi)$ ex ante rather than $\zeta(\cdot, \psi^0)$ ex post. If the planner is risk neutral, his objective will be to maximize $W(P, Q, \cdot)$.

The technical reason for focusing attention on expected welfare is its status as an approximate sufficient statistic for ex post evaluation of STRs. In some settings, the Q-distribution of $E\{y[\zeta(x, \psi)]\}$ can be shown to be tightly concentrated near $W(P, Q, \zeta)$, implying that ex ante and ex post evaluation of STRs yields the same conclusions with high Q-probability. Section 2.4 develops this idea formally.

2.4. Statistical Treatment Rules That Condition on Covariates

The complexity of expected welfare $W(P, Q, \zeta)$ stands in the way of a constructive general analysis of statistical treatment rules. This being the case, I now focus on a class of STRs that is amenable to interesting analysis. These are the rules that *condition on covariates*.

Assume that the covariate space X is finite, with $P(x) > 0$, all $x \in X$. Assume that the sampling process Q generates separate, statistically independent data for persons with different values of the covariates x; that is, a data sample ψ is composed of a set of statistically independent sub-samples $(\psi_x, x \in X)$. In this setting, I shall say that an STR ζ conditions on x if the treatment that ζ selects for persons with covariates x depends on the sample data ψ only through ψ_x. With some flexibility of notation, I henceforth write $\zeta(x, \psi_x)$ to indicate such a rule. The *conditional success* rule introduced in Section 1 conditions on x. The *unconditional success* rule does not.

Expected welfare under an STR that conditions on x is

$$(9) \quad W(P, Q, \zeta) = \sum_{x \in X} P(x) \sum_{t \in T} E[y(t)|x] \cdot Q[\zeta(x, \psi_x) = t].$$

I indicated in Section 2.3 that interest in $W(P, Q, \zeta)$ can be motivated by the status of this quantity as an approximate sufficient statistic for ex post evaluation of ζ. I now formalize this idea.

Consider the Q-variance of $E\{y[\zeta(x, \psi_x)]\}$. By (7) and the assumed statistical independence of the sub-samples $(\psi_x, x \in X)$,

$$(10) \quad Var_Q \{E\{y[\zeta (x, \psi_x)]\}\} = Var_Q \{ \sum_{x \in X} P(x) \sum_{t \in T} E[y(t)^* x] \cdot 1[\zeta (x, \psi_x) = t]\} = \sum_{x \in X} P(x)^2 \cdot C_{\zeta x},$$

where $C_{\zeta x} \equiv Var_Q \{ \sum_{t \in T} E[y(t)|x] \cdot 1[\zeta (x, \psi_x) = t]\}$. The following Lemma establishes upper bounds on $Var_Q \{E\{y[\zeta(x, \psi_x)]\}\}$:

Lemma: Let $\alpha \equiv \max [P(x), x \in X]$. For $x \in X$, let $M_x \equiv \max_{t \in T} E[y(t)|x]$, $m_x \equiv \min_{t \in T} E[y(t)|x]$, and $\delta_x \equiv M_x - m_x$. Let $\beta_{\zeta x}$ be the Q-probability that rule $\zeta(x, \psi_x)$ selects a treatment that maximizes $E[y(t)|x]$. Let $\gamma_{\zeta x} \equiv \max(1/2, \beta_{\zeta x})$. Then

$$(11) \quad Var_Q \{E\{y[\zeta(x, \psi_x)]\}\} \leq \alpha \sum_{x \in X} P(x) \cdot \gamma_{\zeta x}(1 - \gamma_{\zeta x})\delta_x^2 \leq (\alpha/4) \sum_{x \in X} P(x) \cdot \delta_x^2 . \quad \blacksquare$$

Proof: For each $x \in X$, $P(x)^2 \leq \alpha P(x)$. Hence $Var_Q \{E\{y[\zeta (x, \psi_x)]\}\} \leq \alpha \sum_{x \in X} P(x) \cdot C_{\zeta x}$. A universally valid upper bound on $C_{\zeta x}$ is $(1/4)\delta_x^2$, this being the variance under a treatment rule that selects an optimal treatment with Q-probability ½ and a worst treatment with Q-probability ½. A tighter upper bound on $C_{\zeta x}$ is available if $\beta_{\zeta x} > ½$. Then an upper bound on $C_{\zeta x}$ is $\beta_{\zeta x}(1 - \beta_{\zeta x})\delta_x^2$, this being the variance under a treatment rule that selects an optimal treatment with Q-probability $\beta_{\zeta x}$ and a worst treatment with Q-probability $1 - \beta_{\zeta x}$. Hence $C_{\zeta x} \leq \gamma_{\zeta x}(1 - \gamma_{\zeta x})\delta_x^2$.

Q.E.D.

The Lemma shows that $Var_Q \{E\{y[\zeta (x, \psi_x)]\}\}$ is small if either of two sufficient conditions holds. One sufficient condition is that the quantities $\gamma_{\zeta x}$, $x \in X$ be near one. This holds if the sample size is large and if rule ζ is consistent. The other sufficient condition is that the quantity α be near zero. This holds if the covariate space X is large, with no dominant value of x. If, for example, P(x) is uniform, then $\alpha = 1/|X|$, where $|X|$ is the cardinality of X. If either condition holds, the Lemma and Chebychev's inequality imply that the Q-distribution of $E\{y[\zeta (x, \psi_x)]\}$ is concentrated near $W(P, Q, \zeta)$.

The sufficient condition on α is of particular interest. Let ζ and ζ' denote any two STRs that condition on x. The bound $(\alpha/4) \sum_{x \in X} P(x) \cdot \delta_x^2$ holds for both rules. Hence

$$(12) \quad Var_Q \{E\{y[\zeta(x, \psi_x)]\} - E\{y[\zeta'(x, \psi_x)]\}\} \leq \alpha \sum_{x \in X} P(x) \cdot \delta_x^2.$$

Suppose that $W(P, Q, \zeta) < W(P, Q, \zeta')$. Then (12) and Chebychev's inequality imply that

$$(13) \quad Q\{E\{y[\zeta(x, \psi_x)]\} \geq E\{y[\zeta'(x, \psi_x)]\} \leq [\alpha \sum_{x \in X} P(x) \cdot \delta_x^2]/[W(P, Q, \zeta) - W(P, Q, \zeta')]^2.$$

Thus, if α is small, the ex ante and ex post rankings of rules ζ and ζ' are the same with high Q-probability.

3. Statistical Treatment Rules Using Data From Randomized Experiments

I now apply the expected welfare criterion to compare two STRs, the CS and US rules, when the sample data are generated by a randomized experiment. Section 3.1 describes the sampling process generating the data. Section 3.2 formalizes the CS and US rules and examines their expected welfare. Section 3.3 develops the main finding, a proposition giving bounds on expected welfare under the two rules. A corollary

gives explicit sample-size and distributional conditions under which the CS Rule is superior to the US rule, and a numerical illustration gives a quantitative sense of the conditions. Section 3.4 draws implications for reporting covariate information in research articles describing randomized experiments. Section 3.5 poses variations on and extensions to the present analysis that seem worthy of study.

3.1. The Sampling Process

I assume that, for each $x \in X$ and $t \in T$, the sampling process draws N_{xt} persons at random from the subpopulation with covariates x and assigns them to treatment t. Each sample of subjects, denoted N(x, t), then realizes outcomes y_j, $j \in N(x, t)$. I assume that the experiment is classical in all respects: subjects comply with their assigned treatments, they do not interact with one another, and the planner observes their covariates, treatments, and outcomes. Thus, for each $x \in X$, the planner observes $\psi_x \equiv [y_j, j \in N(x, t), t \in T]$.

For simplicity, I restrict attention to treatment-choice problems with two feasible treatments, denoted $t = 0$ and $t = 1$. I also assume that the planner knows the covariate distribution P(x). I evaluate expected welfare under the assumption that the sampling process Q repeats the randomized experiment with the sample sizes $(N_{xt}, t \in T, x \in X)$ held fixed. This is the natural sampling process to consider if the experimenter specifies these sample sizes a priori. Experiments are sometimes carried out under another protocol in which, for each $t \in T$, the experimenter specifies a number of subjects, say N_t, to be drawn at random from the entire population and assigned to treatment t. Under this protocol, $(N_{xt}, t \in T, x \in X)$ varies across repetitions of the experiment, necessitating a more complex analysis than that performed here.

3.2. Expected Welfare Under the CS and US Rules

In Section 1, I described the conditional success (CS) rule as one that selects treatments with the best empirical success rates conditional on the observed covariates, and the unconditional success (US) rule as one that selects a treatment with the best unconditional empirical success rate. I now formalize these rules.

Let $\bar{y}_{xt} \equiv (1/N_{xt}) \sum_{j \in N(x,t)} y_j$ be the sample average outcome among subjects with covariates x assigned to treatment t. Let $\bar{y}_t \equiv \sum_{x \in X} \bar{y}_{xt} \cdot P(x)$ be the population-weighted average outcome among all subjects assigned to treatment t. For each $x \in X$, the CS rule selects a treatment that maximizes $\bar{y}_{xt}$ on $t \in T$. The US rule selects a treatment that maximizes $\bar{y}_t$ on $t \in T$. Each rule requires a tie-breaking convention to be used when multiple treatments maximize the relevant average outcome. I use the convention that treatment 1 is chosen when both treatments yield the same average outcome.

With these definitions, the CS rule yields expect welfare

$$(14) \quad W(P, Q, CS) = \sum_{x \in X} P(x) \{E[y(1)|x] \cdot Q[\bar{y}_{x1} \geq \bar{y}_{x0}] + E[y(0)|x] \cdot Q[\bar{y}_{x1} < \bar{y}_{x0}]\}.$$

The US rule yields expected welfare

$$(15) \quad W(P, Q, US) = E[y(1)] \cdot Q[\bar{y}_1 \geq \bar{y}_0] + E[y(0)] \cdot Q[\bar{y}_1 < \bar{y}_0].$$

Applying the expected welfare criterion, we shall say that the CS rule is superior or inferior to the US rule if W(P, Q, CS) is larger or smaller than W(P, Q, US).

It is easy to see that the CS rule asymptotically yields the optimal population mean outcome and that this rule is asymptotically superior to the US rule if the value of covariate information is positive. Let $n \equiv \min$

(N_{xt}, $t \in T$, $x \in X$) denote the smallest experimental sample. The strong law of large numbers implies that as $n \to \infty$, $W(P, Q, CS) \to E[\max\{E[y(1)|x], E[y(0)|x]\}]$, which is the optimal mean outcome. Moreover,

$$(16) \quad \lim_{n\to\infty} W(P, Q, CS) - W(P, Q, US) \overset{a.s.}{=} E[\max\{E[y(1)|x], E[y(0)|x]\}] - \max\{E[y(1)], E[y(0)]\}.$$

The right side of (16) is the value of covariate information defined in equation (6). Thus $W(P, Q, CS) > W(P, Q, US)$ a.s. if n is sufficiently large and if the value of covariate information is positive.

Asymptotic theory may be suggestive, but a planner comparing the CS and US rules must be concerned with their performance in finite samples. A simple example illustrates the subtlety of the matter:

Example: Let the covariate space have two elements, with X = (a, b) and P(x = a) = P(x = b) = ½. Let the experimental design be balanced with one subject in each sample, so $N_{a1} = N_{a0} = N_{b1} = N_{b0} = 1$. Let the response distributions $P[y(0)|x = a]$ and $P[y(0)|x = b]$ be degenerate with mass points λ_a and λ_b respectively, where $0 < \lambda_a < 1$, $0 < \lambda_b < 1$, and $1 < \lambda_a + \lambda_b$. Let the response distributions $P[y(1)|x = a]$ and $P[y(1)|x = b]$ be Bernoulli with means μ_a and μ_b respectively, where $0 < \mu_a < 1$ and $0 < \mu_b < 1$.

In this setting, $Q[\bar{y}_{a1} \geq \bar{y}_{a0}] = P[y(1) = 1|x = a] = \mu_a$, $Q[\bar{y}_{b1} \geq \bar{y}_{b0}] = P[y(1) = 1|x = b] = \mu_b$, and $Q[\bar{y}_1 \geq \bar{y}_0] = P[y(1) = 1|x = a] \cdot P[y(1) = 1|x = b] = \mu_a \mu_b$. Hence

$$\begin{aligned} &W(P, Q, CS) - W(P, Q, US) \\ &\quad = \tfrac{1}{2}\{[\mu_a^2 + \lambda_a(1 - \mu_a) + \mu_b^2 + \lambda_b(1 - \mu_b)] - [(\mu_a + \mu_b)\mu_a\mu_b + (\lambda_a + \lambda_b)(1 - \mu_a\mu_b)]\} \\ &\quad = \tfrac{1}{2}\{\mu_a(\mu_a - \lambda_a) + \mu_b(\mu_b - \lambda_b) - \mu_a\mu_b(\mu_a - \lambda_a) - \mu_a\mu_b(\mu_b - \lambda_b)\} \\ &\quad = \tfrac{1}{2}\{\mu_a(\mu_a - \lambda_a)(1 - \mu_b) + \mu_b(\mu_b - \lambda_b)(1 - \mu_a)\}. \end{aligned}$$

Thus the CS rule is superior or inferior to the US rule, depending on (μ_a, μ_b, λ_a, λ_b). Observe that the CS rule is superior if ($\mu_a > \lambda_a$, $\mu_b > \lambda_b$) and the US rule is superior if the inequalities are reversed. ■

In general, the expected-welfare ranking of the CS and US rules depends on the distributional and sample-size features of the treatment-choice problem. We would like to characterize the circumstances in which the planner should prefer one rule to the other. Direct analysis of the expressions for expected welfare in (14) and (15) is difficult because the treatment-selection probabilities $Q[\bar{y}_{x1} \geq \bar{y}_{x0}]$ and $Q[\bar{y}_1 \geq \bar{y}_0]$ typically are complex functions of the response distributions $\{P[y(\cdot)|x], x \in X\}$, the sample sizes ($N_{xt}$, $t \in T$, $x \in X$), and the covariate distribution P(x). Fortunately, a large-deviations theorem of Hoeffding (1963) for averages of bounded random variables yields relatively simple bounds on $Q[\bar{y}_{x1} \geq \bar{y}_{x0}]$ and $Q[\bar{y}_1 \geq \bar{y}_0]$. In Section 3.3, I use Hoeffding's theorem to develop bounds on expected welfare under the CS and US rules. These bounds imply explicit sample-size and distributional conditions under which the CS Rule is superior to the US rule.

3.3. Bounds on Expected Welfare Under the CS and US Rules

Here is the Hoeffding theorem that forms the basis for my findings:

Large Deviations Theorem (Hoeffding, 1963, Theorem 2): Let $w_1, w_2, \cdots, w_n$ be independent real random variables, with bounds $a_i \leq w_i \leq b_i$, $(i = 1, 2, \cdots, n)$. Let $\bar{w} \equiv (1/n) \Sigma^n_{i=1} w_i$ and $\mu \equiv E(\bar{w})$. Then, for $v > 0$, $Pr(\bar{w} - \mu \geq v) \leq \exp[-2n^2v^2/\Sigma^n_{i=1} (b_i - a_i)^2]$. ■

Hoeffding's Theorem 2 is a very broad, powerful result. The only distributional assumptions are that the random variables $w_1, w_2, \cdots, w_n$ be independent and have bounded supports. The derived upper bound on $Pr(\bar{w} - \mu \geq v)$ has no nuisance parameters and is of order $\exp(-nv^2)$ in the sample size n and the distance v. I

would note that Hoeffding (1963), Theorem 1gives tighter but more complicated bounds on $Pr(\bar{w} - \mu \geq v)$ that hold if $w_1, w_2, \cdots, w_n$ have the same range. It may be that these alternative bounds can be used to improve on my Proposition below. I leave this as an open question.

I now use Hoeffding's Theorem 2 to obtain finite-sample bounds on expected welfare under the CS and US rules. The Proposition developed here requires that the outcome variable y be bounded but otherwise is entirely general. (The Proposition assumes that outcomes take values in the unit interval but, given boundedness, this may always be achieved by appropriate normalization of location and scale.) The proof of the Proposition is in an Appendix.

Proposition: Let $T = \{0, 1\}$ and $0 \leq y(t) \leq 1, t \in T$. For $x \in X$, let $M_x \equiv \max\{E[y(1)|x], E[y(0)|x]\}$ and $\delta_x \equiv |E[y(1)|x] - E[y(0)|x]|$. Let $M \equiv \max\{E[y(1)], E[y(0)]\}$ and $\delta \equiv |E[y(1)] - E[y(0)]|$. Then

$$(17) \quad \sum_{x \in X} P(x) M_x - \sum_{x \in X} P(x)\delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] \leq W(P, Q, CS) \leq \sum_{x \in X} P(x) M_x.$$

$$(18) \quad M - \delta \cdot \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}] \leq W(P, Q, US) \leq M. \quad \blacksquare$$

Expected welfare under the CS rule necessarily exceeds that under the US rule if the sample sizes are sufficiently large that the CS lower bound exceeds the US upper bound. The Corollary below states this immediate implication of the Proposition.

Corollary: Let the sample sizes $(N_{x1}, N_{x0}; x \in X)$ be such that

$$(19) \quad \sum_{x \in X} P(x)\delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] < \sum_{x \in X} P(x) M_x - M.$$

Then W(P, Q, CS) > W(P, Q, US). ■

The right side of (19) is the value of covariate information, which is necessarily non-negative and is positive if optimal treatments vary with x (Section 2.2). The left side of (19) bounds from above the damage that sampling variation may cause to expected welfare under the CS rule. This quantity falls to zero as the sample sizes ($N_{x1,}$ N_{x0}; $x \in X$) grow. Hence the Corollary reiterates our earlier finding (Section 3.1) that the CS rule is superior to the US rule if the samples are sufficiently large and if optimal treatments vary with x. The important new contribution of the Corollary is that its sufficient condition for superiority of the CS rule is a simple explicit function of the sample sizes (N_{x1}, N_{x0}; $x \in X$), the covariate distribution P(x), and the conditional mean outcomes $\{E[y(\cdot)|x], x \in X\}$. Moreover, this sufficient condition supposes only that outcomes are bounded. No other distributional assumptions are imposed.

Illustration: A numerical illustration gives a quantitative sense of the Proposition and Corollary. Let X = (a, b), with P(x = a) = P(x = b) = ½. Let the design be balanced, with $N_{a1} = N_{a0} = N_{b1} = N_{b0} = n$, where n is a specified positive integer. Let E[y(1)] = E[y(0)] = ½. Then the CS and US bounds are

$$½(M_a + M_b) - ½\delta_a \cdot \exp(-n\delta_a^2) - ½\delta_b \cdot \exp(-n\delta_b^2) \leq W(P, Q, CS) \leq ½ (M_a + M_b)$$

$$½ \leq W(P, Q, US) \leq ½ .$$

The table below evaluates the CS bound when E[y(1)|x = a], E[y(0)|x = a], and n have specified values, namely E[y(1)|x = a] = .4, E[y(0)|x = a] $\in$ (.4, .5, .6, .7, .8), and n $\in$ (1, 10, 25, 50). The quantities E[y(t)|x = b] cannot be varied freely because E[y(t)] = E[y(t)|x = a]P(x = a) + E[y(t)|x = b]P(x = b). Hence the terms of the illustration require that ½ = ½E[y(t)|x = a] + ½ E[y(t)|x = b], implying that E[y(t)|x = b] =

1 - E[y(t)|x = a]. Thus, in the table, treatment 0 is always optimal for persons with x = a and treatment 1 is always optimal for persons with x = b.

The entries in the column titled "n′" give the smallest value of n such that the lower CS bound exceeds ½, the expected welfare under the US rule; that is, n′ is the smallest integer n such that $(M_a + M_b) - \delta_a \cdot \exp(-n\delta_a^2) - \delta_b \cdot \exp(-n\delta_b^2) > 1$. When n exceeds n′, the CS rule definitely yields higher expected welfare than does the US rule. When n is smaller than n′, Proposition 1 does not yield a definite ranking of the two rules. The column titled "n″" will be explained in Section 3.4.

The CS Bound

E[y(0)|x = a]	n = 1	n = 10	n = 25	n = 50	n′	n″
.4	[.50, .50]	[.50, .50]	[.50, .50]	[.50, .50]	∞	∞
.5	[.45, .55]	[.46, .55]	[.47, .55]	[.49, .55]	70	196
.6	[.41, .60]	[.47, .60]	[.53, .60]	[.57, .60]	18	48
.7	[.38, .65]	[.53, .65]	[.62, .65]	[.65, .65]	8	20
.8	[.37, .70]	[.62, .70]	[.69, .70]	[.70, .70]	5	5

The first row of the table describes the boundary case in which treatments 0 and 1 yield the same conditional mean outcomes, so the planner is indifferent between the CS and US rules. The other rows show the tension between use of covariate data and sample size. The value of covariate information increases as E[y(0)|x = a] moves away from E[y(1)|x = a], with treatment 0 becoming increasingly better for persons with x = a and, symmetrically, treatment 1 becoming increasingly better for persons with x = b. Hence the upper CS bound increases monotonically. The behavior of the lower CS bound is more complex. As the value of covariate information increases, so does the loss to the planner if covariate data are used to make sub-optimal treatment choices. The result is that, holding sample size fixed, the lower CS bound first falls and then rises

as $E[y(0)|x = a]$ moves away from $E[y(1)|x = a]$.

Although the lower CS bound varies non-monotonically with $E[y(0)|x = a]$, the sample size n' at which the lower bound first exceeds ½ falls monotonically. Observe how small the values of n' are. If $E[y(0)|x = a] = .5$, the CS rule is superior to the US rule in samples of 70 observations or more. If $E[y(0)|x = a] = .8$, the CS rule is superior in samples of 5 observations or more. ■

3.4. Implications for Reporting Covariate Information in Research Articles

The foregoing analysis carries implications for reporting covariate information in research articles describing randomized experiments. Planners often have extensive covariate information on the population of interest. However, research articles reporting the findings of randomized experiments often present estimates of mean outcomes with little accompanying covariate information. As a result, planners often have only limited ability to apply CS rules.

Consider, for example, a physician who must choose treatments for a population of heterogeneous patients. Physicians often observe many covariates – medical histories, diagnostic test findings, and demographic attributes – for the patients that they treat. Research articles often report the outcomes of randomized clinical trials evaluating alternative treatments. These articles, however, rarely report much covariate information for the subjects of the experiment. Articles reporting on clinical trials usually describe outcomes only within broad risk-factor groups.

There seem to be several reasons why research articles report little covariate information. (I say "seem to" because these reasons are rarely stated explicitly.) Sometimes researchers seem to assume that there exists a common optimal treatment across the population of interest; then covariate information has no value (see Section 2.2). Sometimes concern for the confidentiality of subjects' identities inhibits researchers from reporting covariates that may be related to treatment outcomes. Sometimes sampling variability inhibits

researchers from reporting estimates of treatment effects conditional on covariates.

The merits of the first two reasons must be assessed on a case-by-case basis, but the third reason is subject to a general critique. Researchers often perform randomized experiments with samples of subjects that are large enough to yield statistically precise findings for unconditional treatment effects but not large enough to yield precise findings for treatment effects conditional on covariates. Findings conditional on covariates commonly go unreported if they do not meet conventional criteria for statistical precision. A prevalent practice is to report estimates of $E[y(\cdot)|x]$ only if a classical hypothesis test rejects the null hypothesis $\{H_0: E[y(1)|x] = E[y(0)|x]\}$. In particular, researchers often use the t-statistic criterion

(20) Report $(\bar{y}_{x1}, \bar{y}_{x0})$ if $(\bar{y}_{x1} - \bar{y}_{x0})/[SVar\,(\bar{y}_{x1} - \bar{y}_{x0})]^{½} > 2$,

where $SVar\,(\bar{y}_{x1} - \bar{y}_{x0})$ is the conventional sample estimate of the variance of $(\bar{y}_{x1} - \bar{y}_{x0})$.

Reporting criteria based on statistical precision bear no clear connection to treatment choice. I think it would be better if researchers describing randomized experiments would report treatment effects conditional on covariates whenever (i) there is a priori reason to think that optimal treatments may vary with these covariates and (ii) reporting is consistent with maintenance of confidentiality of subjects' identities.

Illustration: The illustration in Section 3.3 gives a quantitative sense of the implications of conventional reporting criteria. Consider the idealized t-statistic criterion

(21) Report $(\bar{y}_{x1}, \bar{y}_{x0})$ if $E_Q(\bar{y}_{x1} - \bar{y}_{x0})/[Var_Q(\bar{y}_{x1} - \bar{y}_{x0})]^{½} > 2$.

I refer to this as an "idealized" criterion because the operational t-statistic rule given in (20) makes reporting a function of the sample drawn, hence a random variable, whereas the idealized t-statistic given in (21) makes

reporting a function of population characteristics specified in the illustration. Let y be binary, so $E[y(\cdot)|x] = P[y(\cdot) = 1|x]$. Let $P_{tx} \equiv P[y(t) = 1|x]$, t = 0, 1. Then the idealized criterion becomes

(22) Report $(\bar{y}_{x1}, \bar{y}_{x0})$ if $(P_{1x} - P_{0x})/[P_{1x}(1 - P_{1x})/n + P_{0x}(1 - P_{0x})/n]^{1/2} > 2$.

The column titled "n″" in the table of Section 3.3 gives the minimal value of n at which this reporting criterion is met. Comparison of the entries for n′ and n″ shows that $n' \leq n''$ in every case and that n″ is much larger than n′ when $P_{0x} \in (.5, .6, .7)$. Thus, use of a reporting criterion based on statistical precision may prevent use of the CS rule when that rule is superior to the US rule. ■

3.5. Variations on and Extensions to the Analysis

Many variations on and extensions to this analysis seem worthy of study. These include

Measurement of Empirical Success: The versions of the CS and US rules analyzed here measure the empirical success of treatment t by the sample averages $\bar{y}_{xt}$ and $\bar{y}_t$. These averages are natural nonparametric estimates of $E[y(t)|x]$ and $E[y(t)]$, but it may be that other estimates yield higher expected welfare. More generally, there is a large open question about optimal estimation of $E[y(\cdot)|x]$ for use in the CS rule. What nonparametric estimate should be used when x is not discrete, given specified smoothness restrictions on $E[y(\cdot)|x]$ as a function of x? What estimate should be used when the planner has prior parametric or semiparametric information restricting the form of $E[y(\cdot)|x]$? These are familiar questions in the literature on efficient estimation of regressions, but the traditional objective has been to minimize mean square error in predicting outcomes. Here the objective is to choose treatments that maximize expected welfare.

Hybrid CS-US Rules: The CS and US rules are polar cases, one always and the other never using the available covariate information. It may be that hybrid CS-US rules, in which the use of covariate information depends on sample size, are more effective. The literature on prediction suggests shrinkage estimators to minimize mean square error. It may be that similar approaches improve expected welfare.

Experimental Design: The analysis in this paper takes the design as given. Experimental design has received extensive study from the perspective of hypothesis testing. In particular, there is a longstanding practice of selecting sample sizes that achieve specified power when testing a null hypothesis of no treatment effect against a specified alternative. The expected welfare criterion may be used to study experimental design from the perspective of treatment choice.

4. Using the Expected-Welfare Criterion to Choose a Treatment Rule

This paper has used Wald's statistical decision theory to evaluate statistical treatment rules in the setting of a randomized experiment. Wald's approach is capable of yielding important findings on the performance of STRs. At the same time, it is incomplete.

The difficulty is that the expected-welfare ranking of alternative STRs depends on the population distribution of treatment response. The use of STRs to make treatment choices, however, arises when planners have only sample data on treatment response, not knowledge of the response distribution. Hence it is not clear how a planner should use the expected welfare criterion to guide choice of a treatment rule. The CS and US rules illustrate the conundrum. Expected welfare under the CS rule is a function of $\{E[y(\cdot)|x], Q[\bar{y}_{x1} \geq \bar{y}_{x0}], x \in X\}$ and under the CS rule is a function of $\{E[y(\cdot)], Q[\bar{y}_1 \geq \bar{y}_0]\}$. The planner, however, only has sample data on treatment response.

The frequentist and Bayesian literatures suggest alternative pragmatic solutions. The frequentist literature suggests that the planner use the available sample data to estimate the relevant parameters of the distribution of treatment response, and then "plug-in" these estimates to evaluate expected welfare. For example, one might estimate W(P, Q, CS) and W(P, Q, US) by

$$(23) \quad w(P, Q, CS) \equiv \sum_{x \in X} P(x) \{\bar{y}_{x1} \cdot q[\bar{y}_{x1} - \bar{y}_{x0} \geq 0] + \bar{y}_{x0} \cdot q[\bar{y}_{x1} - \bar{y}_{x0} < 0]\}$$

$$(24) \quad w(P, Q, CS) \equiv \bar{y}_1 \cdot q[\bar{y}_1 - \bar{y}_0 \geq 0] + \bar{y}_0 \cdot q[\bar{y}_1 - \bar{y}0 < 0].$$

Here $q[\cdot]$ denotes a sample estimate of the corresponding treatment-selection probability $Q[\cdot]$; for example, the empirical distribution of the sample data might be used to generate a bootstrap estimate of $Q[\cdot]$. Then one might choose the CS rule if $w(P, Q, CS) > w(P, Q, US)$ and the US rule otherwise.

The plug-in prescription is easy to explain and implement, but its theoretical foundation is incomplete. Lacking finite-sample theory, frequentist statisticians commonly cite asymptotic theory showing that sample estimates of population parameters have well-behaved limiting distributions. The planner's objective, however, is not to obtain estimates of expected welfare with good asymptotic properties. It is to choose a treatment rule maximizing expected welfare when applied to samples of specified finite size. Using asymptotic theory to guide a finite-sample statistical decision problem requires a leap of faith.

The Bayesian literature suggests that the planner should assert a subjective prior distribution on the space of treatment-response distributions, use the available sample data to update the prior, and apply the resulting posterior subjective distribution to evaluate alternative treatment rules. Bayesian decision theory provides a coherent finite-sample approach to evaluation of treatment rules by reaching beyond frequentist statistics to introduce a new concept, the subjective prior distribution on the space of treatment-response

distributions. Bayesian conclusions about the performance of alternative treatment rules inevitably depend on the prior invoked, but the Bayesian paradigm is silent on the critical question of how the prior should be specified. Thus, in practice, the Bayesian prescription is incomplete.

It may be that a fully satisfactory approach to evaluation of treatment rules using sample data is unachievable. We can, however, expand the set of available options. The bounds on expected welfare under the CS and US rules developed in Section 3.3 demonstrate one approach. Evaluation of the bounds does not require all of the distributional information needed to evaluate expected welfare. Consider the CS rule. Whereas expected welfare is a function of $\{E[y(\cdot)|x], Q[\bar{y}_{x1} \geq \bar{y}_{x0}]\}$, the bound depends only on $E[y(\cdot)|x]$. A frequentist contemplating plug-in estimation of the bound need not address the subtle problem of estimating $Q[\bar{y}_{x1} \geq \bar{y}_{x0}]$. A Bayesian contemplating use of the bound need not articulate a full subjective posterior distribution on the space of treatment-response distributions; the posterior for $E[y(\cdot)|x]$ suffices. Of course these benefits are achieved with an accompanying cost. The CS and US bounds may overlap, in which case the ranking of these treatment rules is indeterminate.

Appendix: Proof of the Proposition

CS Bound: The upper bound follows from (14), so the task is to prove the lower bound. For $x \in X$, I write $\bar{y}_{x1} - \bar{y}_{x0}$ as the average of independent random variables and apply Hoeffding's theorem to show that

$$(A1)\quad M_x - \delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] \leq E[y(1)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} \geq 0] + E[y(0)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} < 0].$$

Let $N_x \equiv N_{x1} + N_{x0}$. For each $x \in X$,

(A2) $\bar{y}_{x1} - \bar{y}_{x0} = (1/N_{x1}) \sum_{j \in N(x,1)} y_j - (1/N_{x0}) \sum_{j \in N(x,0)} y_j$

$= (1/N_x)[\sum_{j \in N(x,1)} (y_j \cdot N_x/N_{x1}) + \sum_{j \in N(x,0)} (-y_j \cdot N_x/N_{x0})]$.

Thus $\bar{y}_{x1} - \bar{y}_{x0}$ is the average of N_x independent random variables. The first N_{x1}have range $[0, N_x/N_{x1}]$ and the remaining N_{x0} have range $[-N_x/N_{x0}, 0]$.

Consider $x \in X$ such that $E[y(1)|x] < E[y(0)|x]$. Then $E(\bar{y}_{x1} - \bar{y}_{x0}) = -\delta_x$. Hoeffding's theorem yields

(A3) $Q[\bar{y}_{x1} - \bar{y}_{x0} \geq 0] \leq \exp[-2N_x^2 \delta_x^2/\{N_{x1} \cdot (N_x/N_{x1})^2 + N_{x0} \cdot (N_x/N_{x0})^2\}] = \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$.

Hence

(A4) $E[y(1)|x] \cdot Q[\bar{y}_{x1} - \bar{y}_{x0} \geq 0] + E[y(0)|x] \cdot Q[\bar{y}_{x1} - \bar{y}_{x0} < 0]$

$\geq E[y(1)|x] \cdot \exp[-2\delta_x^2 / (N_{x1}^{-1} + N_{x0}^{-1})] + E[y(0)|x] \cdot \{1 - \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$

$= M_x - \delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$.

So (A1) holds. Next consider $x \in X$ such that $E[y(1)|x] > E[y(0)|x]$. For such x, $E(\bar{y}_{x0} - \bar{y}_{x1}) = -\delta_x$. Application of Hoeffding's theorem yields

(A5) $Q[\bar{y}_{x1} - \bar{y}_{x0} < 0] = Q[\bar{y}_{x0} - \bar{y}_{x1} > 0] \leq Q[\bar{y}_{x0} - \bar{y}_{x1} \geq 0] \leq \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$.

Thus (A1) continues to hold by an argument analogous to (A4). Finally consider $x \in X$ such that $E[y(1)|x] = E[y(0)|x]$. For such x, $\delta_x = 0$. Hence (A1) holds as an equality.

US Bound: The upper bound follows from (15). The lower bound holds as an equality if $E[y(1)] = E[y(0)]$.

The task is to show that the lower bound holds otherwise. As in the proof of the CS bound, I write $\bar{y}_1 - \bar{y}_0$ as the average of independent random variables and then apply Hoeffding's theorem.

Let $N \equiv \Sigma_{x \in X} (N_{x1} + N_{x0})$. Then

$$\text{(A6)} \quad \bar{y}_1 - \bar{y}_0 = \sum_{x \in X} P(x) \; (1/N_{x1}) \sum_{j \in N(x,1)} y_j - \sum_{x \in X} P(x) \; (1/N_{x0}) \sum_{j \in N(x,0)} y_j$$

$$= (1/N)[\sum_{x \in X} \sum_{j \in N(x,1)} (y_j \cdot P(x)N/N_{x1}) + \sum_{x \in X} \sum_{j \in N(x,0)} (-y_j \cdot P(x)N/N_{x0})].$$

Thus $\bar{y}_1 - \bar{y}_0$ averages N independent random variables with ranges $[0, P(x)N/N_{x1}]$ and $[-P(x)N/N_{x0}, 0]$, $x \in X$.

Let $E[y(1)] < E[y(0)]$. Then $E(\bar{y}_1 - \bar{y}_0) = -\delta$. Application of Hoeffding's theorem yields

$$\text{(A7)} \quad Q[\bar{y}_1 - \bar{y}_0 \geq 0] \leq \exp[-2N^2 \delta^2/\{\Sigma_{x \in X} N_{x1}(P(x) \cdot N/N_{x1})^2 + N_{x0}(P(x) \cdot N/N_{x0})^2\}]$$

$$= \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}].$$

Hence

$$\text{(A8)} \quad E[y(1)] \cdot Q[\bar{y}_1 - \bar{y}_0 \geq 0] + E[y(0)] \cdot Q[\bar{y}_1 - \bar{y}_0 < 0] \geq M - \delta \cdot \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}].$$

The same result holds when $E[y(1)] > E[y(0)]$.

Q. E. D.

References

Angrist, J., G. Imbens, and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-455

Balke, A. and J. Pearl(1997), "Bounds on Treatment Effects from Studies With Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1177.

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

Dehejia, R. (1999), "Program Evaluation as a Decision Problem," National Bureau of Economic Research Working Paper 6954.

Heckman, J., J. Smith, and N. Clements (1997), "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," Review of Economic Studies 64, 487-535.

Heckman, J. and R. Robb (1985), "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.

Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," Journal of the American Statistical Association, 58, 13-30.

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

Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467-476.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," American Economic Review Papers and Proceedings 80, 319-323.

Manski, C. (1995), Identification Problems in the Social Sciences, Cambridge, MA.: Harvard University Press.

Manski, C. (1997a), "The Mixing Problem in Programme Evaluation," Review of Economic Studies , 64, 537-553.

Manski, C. (1997b), "Monotone Treatment Response," Econometrica, 65, 1311-1334.

Manski, C. (1998), "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.

Manski, C. (1999), "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," Journal of Econometrics, forthcoming.

Manski, C. and D. Nagin (1998), “Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism,” Sociological Methodology, 28, 99-137.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), “Alterrnative Estimates of the Effect of Family Structure During Adolescence on High School Graduation," Journal of the American Statistical Association, 87, 25-37.

Robins, J. (1989), "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. eds. Health Service Research Methodology: A Focus on AIDS, NCHSR, U.S. Public Health Service.

Rosenbaum, P. (1995), Observational Studies, New York: Springer-Verlag.

Rosenbaum, P., and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika, 70, 41-55.

Rubin, D. (1978), “Bayesian Inference for Causal Effects: The Role of Randomization,” The Annals of Statistics, 6, 34-58.

Stafford, F. (1985), "Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor-Market Studies." in J. Hausman and D. Wise (editors), Social Experimentation. Chicago: University of Chicago Press.

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