University of Wisconsin-Madison

Institute for Research on Poverty

Discussion Papers

Charles F. Manski

WHAT DO CONTROLLED EXPERIMENTS REVEAL ABOUT OUTCOMES WHEN TREATMENTS VARY?

DP # 1005-93

Institute for Research on Poverty
Discussion Paper no. 1005-93

What Do Controlled Experiments Reveal about Outcomes When Treatments Vary?

Charles F. Manski
Department of Economics
Institute for Research on Poverty
University of Wisconsin-Madison

June 1993

This research was performed while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. Support was provided by NSF Grant SES-9022192 and by a grant from the University of Wisconsin Graduate School.

Abstract

A common concern of evaluation studies is to learn the distribution of outcomes when each member of a population receives a treatment resulting from a specified treatment policy. Many recent studies have used controlled experiments to evaluate policies mandating the same treatment for all members of the population. Policies mandating homogeneous treatment are of interest, but so are policies that make treatment vary across the population. This paper examines the use of experimental evidence to infer the outcomes that would occur when treatment may vary across the population. Experimental evidence from the Perry Preschool Project is used to illustrate the inferential problem and the main findings of the analysis.

What Do Controlled Experiments Reveal about Outcomes When Treatments Vary?

1. Introduction

A common concern of evaluation studies is to learn the distribution of outcomes when each member of a population receives a treatment resulting from a specified treatment policy. Many recent studies have used controlled experiments to evaluate policies mandating the same treatment for all members of the population. In the classical experiment, random samples of the population are drawn and formed into treatment groups, all of whose members are assigned the same treatment. The empirical distribution of outcomes realized by a treatment group is then ostensibly the same (up to random sampling error) as would be observed if the treatment in question were mandatory for the entire population. For example, see Manski and Garfinkel (1992), some of whose chapters describe recent experimental evaluations of mandatory welfare and training programs.

Policies mandating homogeneous treatment are of interest, but so are policies that permit treatment to vary across the population. We often see voluntary treatment policies, calling on persons to select their own treatments. Policies intended to mandate homogeneous treatment sometimes turn out to be voluntary in practice, as compliance with the mandated treatment is not enforced. And resource constraints sometimes prevent universal implementation of desirable treatments.

Consider the following inferential questions:

(S) What do observations of outcomes when treatments vary across the population reveal about the outcomes that would occur if treatment were homogeneous?

(M) What do observations of outcomes when treatment is homogeneous reveal about the outcomes that would occur if treatment were to vary across the population?

Question (S), usually called the *selection* or *switching* problem, has drawn considerable attention and much has been learned; see Maddala (1983), Heckman and Robb (1985), and Manski (1993).

Question (M), which has remained unexplored and unnamed, is the subject of this paper. Formally, question (M) asks what inferences about mixtures of two random variables can be made given knowledge of their marginal distributions. Hence, I refer to the question as the *mixing* problem.

SELECTION AND MIXING PROBLEMS: To formalize these inferential questions, let each member of the population be described by values for $[(y_1,y_0),(z_m,m\epsilon M),x]$. Here x is a vector of covariates, an element of some space X. There are two feasible treatments, labeled 1 and 0.[1] The set M gives the treatment policies of interest. A treatment policy determines which treatment each person receives. The indicator variable z_m denotes the treatment that a given person receives under policy m; $z_m = 1$ if the person receives treatment 1 and $z_m = 0$ otherwise. Associated with the treatments are outcomes (y_1,y_0), a pair of elements of some outcome space Y. The outcome a person realizes under policy m is

(1) $w_m \equiv y_1z_m + y_0(1-z_m)$.

The distribution of outcomes realized by those persons sharing the same value of x is

$$
\begin{aligned}
(2)\ P(w_m \mid x) &\equiv P[y_1z_m+y_0(1-z_m) \mid x] \\
&= P(y_1 \mid x,z_m=1)P(z_m=1 \mid x) + P(y_0 \mid x,z_m=0)P(z_m=0 \mid x).
\end{aligned}
$$

For example, a welfare recipient might be treated by job training or by the "null" treatment of no training intervention. The relevant outcome might be earned income following treatment. One treatment policy might mandate job training for all welfare recipients and enforce the mandate. A second policy might attempt to mandate job training but not be able to enforce compliance. A third policy might permit a person's caseworker to select the treatment expected to yield the larger net benefit, measured as earned income minus training costs.

Suppose one wishes to learn the distribution $P(w_m \mid x)$ of outcomes that would be realized by persons with covariates x if a specified treatment policy m were in effect. Inference is straightforward if one can enact policy m and observe the realized outcomes. The interesting inferential questions concern the feasibility of learning $P(w_m \mid x)$ when one observes realizations under policies other than m.

Selection problems arise when policy m mandates homogeneous treatment, but the available data are realizations under some other policy yielding heterogeneous treatments. Suppose that m makes treatment 1 mandatory, so $P(w_m \mid x) = P(y_1 \mid x)$. Suppose that the observable policy is some $\mu \epsilon M$, so the available data are a random sample of (w_μ, z_μ, x).[2] The sampling process identifies the censored outcome distributions $P(y_1 \mid x, z_\mu = 1)$ and $P(y_0 \mid x, z_\mu = 0)$, as well as the treatment distribution $P(z_\mu \mid x)$. Thus, inferential question (S) formalizes as:

(S) What does knowledge of $[P(y_1 \mid x, z_\mu = 1), P(y_0 \mid x, z_\mu = 0), P(z_\mu \mid x)]$ imply about $P(y_1 \mid x)$?

Mixing problems arise when policy m may yield heterogenous treatments, but the available data are from controlled experiments imposing homogenous treatments on random samples of the population. The classical model of experimentation presumes that experimental evidence is available for both treatments, so the experiments identify $P(y_1 \mid x)$ and $P(y_0 \mid x)$. Thus, question (M) formalizes as:

(M) What does knowledge of $[P(y_1 \mid x), P(y_0 \mid x)]$ imply about $P[y_1 z_m + y_0(1-z_m) \mid x]$?[3]

ORGANIZATION OF THE PAPER: Section 2 uses empirical evidence from a famous social experiment, the Perry Preschool Project, to illustrate the mixing problem and the main findings of this paper. Fifteen years after their participation in this early-childhood educational intervention, 67 percent of an experimental group were high school graduates. At the same time, only 49 percent of a control group were graduates. Our interest is to determine what the experimental evidence and various forms of prior information imply about the rate of high school graduation that would prevail under treatment policies applying the intervention to some children but not to others.

Sections 3 through 5 present the analysis yielding the empirical results reported in Section 2. To begin, Section 3 examines the mixing problem in the absence of any prior information on the distribution of $[(y_1,y_0),z_m,x]$. The basic finding is a proposition giving sharp bounds on conditional probabilities of the form $P(w_m \epsilon B \mid x)$, $B \subset Y$. When outcomes are real-valued, this finding is easily transformed into sharp bounds on quantiles of $P(w_m \mid x)$, given in a corollary.[4]

Sections 4 and 5 explore the identifying power of several forms of prior information that might plausibly be invoked in empirical studies.[5] Section 4 imposes restrictions on the joint distribution of the outcomes (y_1,y_0). Section 4.1 assumes that y_1 and y_0 are statistically independent, conditional on the covariates x. In contrast, Section 4.2 supposes that the outcomes are shifted versions of one another. Section 4.3 assumes that the outcomes are ordered.

Section 5 imposes restrictions on the treatment policy. Section 5.1 assumes that the treatment received by each person with covariates x is statistically independent of the person's outcomes (y_1,y_0). Section 5.2 considers the polar opposite situation in which treatment is a known function of outcomes; I focus on the case in which outcomes are real-valued and the treatment policy always selects the treatment yielding the smaller outcome (or, symmetrically, the larger outcome). Section 5.3 assumes that the fraction of the population receiving each treatment is known.

Taken one at a time, each of these assumptions on the distribution of outcomes or on the treatment policy implies a distinctive bound on $P(w_m \mid x)$, but none of the assumptions is strong enough to identify the distribution. Combinations of assumptions do identify $P(w_m \mid x)$. Two such are stated in Section 5.4.

IDENTIFICATION AND SAMPLE INFERENCE: The mixing problem, like the selection problem, is a failure of identification rather than a difficulty in sample inference. To keep attention focused on identification, Sections 3 through 5 maintain the assumption that the conditional distributions identified by classical controlled experiments, $[P(y_1 \mid x), P(y_0 \mid x)]$, are known almost everywhere on the covariate space. The identification findings reported in these sections can be translated into consistent sample estimates of identified quantities by replacing $P(y_1 \mid x)$ and $P(y_0 \mid x)$ with consistent nonparametric estimates, as is done in Section 2.

For the sake of simplicity, I often refer to $P(y_1 \mid x)$ and $P(y_0 \mid x)$ simply as the distributions of y_1 and y_0, rather than as the distributions conditional on x. One could similarly shorten the notation by denoting these distributions as $P(y_1)$ and $P(y_0)$. I do not take this step because I want the reader to keep in mind that the analysis of this paper holds for any specification of the covariates x.

CAVEATS ON CLASSICAL EXPERIMENTATION: This paper maintains the classical assumption that experimental regimes operate exactly as would mandatory treatment policies. I have elsewhere discussed some of the many reasons why this central tenet of experimental analysis may fail to hold when applied to welfare and training programs (see the introduction to Manski and Garfinkel, 1992). Experiments may be administered differently from actual programs. Macro feedback effects ranging from information diffusion to norm formation to market equilibration may make the full-scale implementation of a mandatory treatment policy inherently different from the small-scale implementation of an experiment. Strictures on forcing human subjects into experiments may make it impossible to form random treatment groups. The present analysis assumes away all of these very real concerns.

2. An Illustration: The Perry Preschool Experiment

Beginning in 1962, the Perry Preschool Project provided intensive educational and social services to a random sample of black children in Ypsilanti, Michigan. The project investigators also drew a second random sample of such children, but provided them with no special services. Subsequently, a variety of outcomes were ascertained for most members of the experimental and control groups. Among other things, it was found that 67 percent of the experimental group and 49 percent of the control group were high school graduates by age nineteen (see Berrueta-Clement et al., 1984). This and similar findings for other outcomes have been widely cited as evidence that intensive early-childhood educational interventions improve the outcomes of children at risk (see Holden, 1990).

For purposes of discussion, let us accept the Perry Preschool Project as a classical controlled experiment, with

x = black children in Ypsilanti, Michigan

z_m = 1 if early-childhood intervention received, = 0 otherwise

y_1 = 1 if high school graduate by age 19, = 0 otherwise; intervention received

y_0 = 1 if high school graduate by age 19, = 0 otherwise; intervention not received.

Moreover, ignoring attrition and sampling error in the estimation of outcome distributions, let us accept the experimental evidence as showing that the high school graduation rate among children with covariate value x would be .67 if all such children were to receive the intervention, and would be .49 if none of them were to receive the intervention. That is, let us accept the experimental evidence as showing that $P(y_1=1 \mid x) = .67$ and $P(y_0=1 \mid x) = .49$.[6]

What would be the rate of high school graduation if some children with covariates x were to receive the intervention, but not others? Table 1 summarizes the inferences that can be made in a

TABLE 1

The Perry Preschool Project: Implied Rates of High School Graduation under Different Scenarios

Experimental Evidence

$P(y_1=1 \mid x) = .67 \quad P(y_0=1 \mid x) = .49$

Scenario	$P(w_m=1 \mid x)$
no prior information (Proposition 1)	[.16,1]
independent outcomes (Proposition 2)	[.33,.83]
ordered outcomes (Proposition 4)	[.49,.67]
treatment independent of outcomes (Proposition 5)	[.49,.67]
treatment with smaller outcome (Proposition 6A)	[.16,.49]
treatment with larger outcome (Proposition 6B)	[.67,1]
+ independent outcomes (Proposition 9A)	.83
+ ordered outcomes (Proposition 9A)	.67
1/10 population receives treatment 0 (Proposition 7)	[.57,.77]
+ treatment independent of outcomes (Proposition 9B)	.65
5/10 population receives treatment 0 (Proposition 7)	[.17,.99]
+ treatment independent of outcomes (Proposition 9B)	.58
9/10 population receives treatment 0 (Proposition 7)	[.39,.59]
+ treatment independent of outcomes (Proposition 9B)	.51

variety of scenarios. In each case, the table cites a proposition implying the estimate shown. These propositions are developed in Sections 3 through 5.

If the experimental evidence is the only information available, we can conclude that the graduation rate must lie between .16 and 1, but we cannot say more. In other words, there exist treatment policies and distributions of (y_1,y_0) that are consistent with the known values of $P(y_1 \mid x)$ and $P(y_0 \mid x)$ and that imply graduation rates as low as .16 and as high as 1.

If prior information about the distribution of outcomes is available, then we can narrow the range of possibilities. If y_1 and y_0 are known to be statistically independent conditional on x, then the graduation rate must lie between .33 and .83; where the graduation rate falls within this range depends on the treatment policy. On the other hand, suppose that receiving the early-childhood intervention can never be harmful to a child; that is, y_1 and y_0 are known to be ordered, with $y_1 = 0 \Rightarrow y_0 = 0$. Then the graduation rate must lie between those observed in the control and experimental groups, namely .49 and .67.

Information restricting the treatment policy offers its own identifying power. If treatment is known to be statistically independent of outcomes, then the graduation rate must again lie between .49 and .67. On the other hand, if a child always receives the treatment yielding the larger outcome, then the graduation rate must lie between .67 and 1.

Suppose that one knows the fraction of the population receiving each treatment. Knowing that 1/10 or 5/10 or 9/10 of the population receives treatment 0 implies that the graduation rate lies in the interval [.57,.77] or [.17,.99] or [.39,.59] respectively. Observe that the first and third intervals are relatively narrow but the second is rather wide, almost as wide as the interval found in the absence of prior information. This pattern of results reflects the fact that the power of treatment policy to determine who receives which treatment is much more constrained when $P(z_m=0 \mid x)$ is fixed at a value near zero or one than it is when $P(z_m=0 \mid x)$ is fixed at 5/10.

The scenarios considered thus far bring to bear enough empirical evidence and prior information to bound the high school graduation rate but not to identify it. If stronger restrictions are imposed, then the high school graduation rate may be identified. For example, if it is known that outcomes are statistically independent and that each child receives the treatment yielding the larger outcome, then the implied high school graduation rate is .83. If it is known that 5/10 of the population receives treatment 0 and that treatment is independent of outcomes, then the implied graduation rate is .58.

The general lesson is that experimental evidence alone permits only weak conclusions to be drawn about the high school graduation rate when treatments vary. Experimental evidence combined with prior information allows stronger conclusions. The nature of these stronger conclusions depends critically on the prior information asserted. This lesson is analogous to the one learned over the past twenty years about the conclusions that can be drawn about mandatory programs from observations of outcomes when treatments vary. Mixing and selection are distinct identification problems, but they are closely related.

3. Identification Using only the Experimental Evidence

Our objective in this section is to characterize the restrictions on $P(w_m \mid x)$ implied by knowledge of $[P(y_1 \mid x), P(y_0 \mid x)]$. No other information is assumed available.

PROBABILITIES OF EVENTS: Consider the probability that the realized outcome w_m falls in some set B, conditional on x; that is, $P(w_m \epsilon B \mid x)$. Given that w_m always equals either y_1 or y_0, one might think that $P(w_m \epsilon B \mid x)$ must lie between $P(y_1 \epsilon B \mid x)$ and $P(y_0 \epsilon B \mid x)$. This is not the case. It turns out that when $P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) \leq 1$, then $P(w_m \epsilon B \mid x)$ must lie in the interval

$[0,P(y_1 \epsilon B \mid x)+P(y_0 \epsilon B \mid x)]$. When $P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) \geq 1$, $P(w_m \epsilon B \mid x)$ must lie in the interval $[P(y_1 \epsilon B \mid x)+P(y_0 \epsilon B \mid x)-1,1]$. Proposition 1 gives the result.

Proposition 1: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Then

$$(3)\quad \max[0,P(y_1 \epsilon B \mid x)+P(y_0 \epsilon B \mid x)-1] \leq P(w_m \epsilon B \mid x) \leq \min[P(y_1 \epsilon B \mid x)+P(y_0 \epsilon B \mid x),1]. \quad \blacksquare$$

PROOF: We first determine the treatment policies that minimize and maximize $P(w_m \epsilon B \mid x)$. Observe that if y_1 and y_0 both fall in the set B, then w_m must fall in B. Moreover, if neither y_1 nor y_0 falls in B, then w_m cannot fall in B. That is,

$$(4a)\quad y_1 \in B \cap y_0 \in B \Rightarrow w_m \in B$$

and

$$(4b)\quad y_1 \notin B \cap y_0 \notin B \Rightarrow w_m \notin B,$$

whatever treatment-policy m may be. The treatment policy is relevant in those cases where one of the two outcomes falls in B and the other does not. The treatment policy minimizes $P(w_m \epsilon B \mid x)$ if it always selects the treatment yielding the outcome not in B; that is, if

$$(5)\quad y_1 \notin B \cap y_0 \in B \Rightarrow z_m = 1$$

$$y_1 \in B \cap y_0 \notin B \Rightarrow z_m = 0.$$

Hence, the smallest possible value of $P(w_m \epsilon B \mid x)$ is $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$. The treatment policy maximizes $P(w_m \epsilon B \mid x)$ if it always selects the treatment yielding the outcome in B; that is, if

$$(6)\quad y_1 \notin B \cap y_0 \in B \Rightarrow z_m = 0$$
$$y_1 \in B \cap y_0 \notin B \Rightarrow z_m = 1.$$

So the largest possible value of $P(w_m \epsilon B \mid x)$ is $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x)$.

The above shows that if $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$ and $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x)$ are known, then

$$(7)\quad P(y_1 \epsilon B \cap y_0 \epsilon B \mid x) \leq P(w_m \epsilon B \mid x) \leq P(y_1 \epsilon B \cup y_0 \epsilon B \mid x)$$

is a sharp bound on $P(w_m \epsilon B \mid x)$. But the only available information is knowledge of $P(y_1 \mid x)$ and $P(y_0 \mid x)$. Therefore, the best computable lower bound on $P(w_m \epsilon B \mid x)$ is the smallest value of $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$ that is consistent with the known $P(y_1 \mid x)$ and $P(y_0 \mid x)$. Similarly, the best computable upper bound is the largest feasible value of $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x)$.

The second step is to determine these best computable bounds. This is simple to do, because Frechet (1951) proved this sharp bound on $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$:[7]

$$(8)\quad \max[0, P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) - 1] \leq P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$$
$$\leq \min[P(y_1 \epsilon B \mid x), P(y_0 \epsilon B \mid x)].$$

It follows immediately from (8) that the best computable lower bound on $P(w_m \mid x)$ is $\max[0, P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) - 1]$. To obtain the best computable upper bound, observe that

(9) $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x) = P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) - P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$.

Applying the Frechet lower bound on $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$ to (9) shows that

(10) $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x) \leq \min[P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x), 1]$.

Hence, $\min[P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x), 1]$ is the best computable upper bound on $P(w_m \mid x)$.

Q.E.D.

QUANTILES: Suppose that Y is the real line. Let $u \epsilon R^1$ and $B = (-\infty, u]$. By Proposition 1,

(11) $\max[0, P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - 1] \leq P(w_m \leq u \mid x)$

$\leq \min[P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x), 1]$.

Let $\alpha \epsilon (0,1)$ and let $q_m(\alpha \mid x)$ denote the α-quantile of w_m, conditional on x. Corollary 1.1 inverts the bound (11) to obtain a sharp bound on $q_m(\alpha \mid x)$.

<u>Corollary 1.1</u>: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let Y be the real line. Let

$r_1(\alpha \mid x) \equiv \inf_u$ s.t. $P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) \geq \alpha$

$s_1(\alpha \mid x) \equiv \inf_u$ s.t. $P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - 1 \geq \alpha$.

Then

(12) $r_1(\alpha \mid x) \leq q_m(\alpha \mid x) \leq s_1(\alpha \mid x)$. ■

PROOF: By the upper bound on $P(w_m \leq u \mid x)$ in (11),

$$\begin{aligned} u < r_1(\alpha \mid x) &\Rightarrow P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) < \alpha \\ &\Rightarrow P(w_m \leq u \mid x) < \alpha \\ &\Rightarrow q_m(\alpha \mid x) > u. \end{aligned}$$

Hence, $r_1(\alpha \mid x) \leq q_m(\alpha \mid x)$. By the lower bound on $P(w_m \leq u \mid x)$ in (11),

$$\begin{aligned} u \geq s_1(\alpha \mid x) &\Rightarrow P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - 1 \geq \alpha \\ &\Rightarrow P(w_m \leq u \mid x) \geq \alpha \\ &\Rightarrow q_m(\alpha \mid x) \leq u. \end{aligned}$$

Hence, $q_m(\alpha \mid x) \leq s_1(\alpha \mid x)$. These bounds on $q_m(\alpha \mid x)$ are sharp because the bounds in (11) are sharp.

Q.E.D.

It is of interest to note that these bounds on quantiles of $P(w_m \mid x)$ are always informative both above and below. This is so even though the bound on $P(w_m \leq u \mid x)$ used to derive Corollary 1.1 is only informative above or below, the informative direction depending on the value of u.

4. Restrictions on the Outcome Distribution

In the course of proving Proposition 1, we showed that if $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x)$ and $P(y_1 \epsilon B \cup y_0 \epsilon B \mid x)$ are known and if no restrictions are imposed on the treatment policy m, then inequality (7) provides a sharp bound on $P(w_m \epsilon B \mid x)$. One may sometimes have prior information that, when combined with empirical knowledge of $[P(y_1 \mid x), P(y_0 \mid x)]$, makes the bound (7) computable. This section presents three leading cases.

4.1. INDEPENDENT OUTCOMES

Suppose it is known that the outcomes y_1 and y_0 are statistically independent, conditional on x. Then

(13) $P(y_1 \epsilon B \cap y_0 \epsilon B \mid x) = P(y_1 \epsilon B \mid x) P(y_0 \epsilon B \mid x).$

Our second proposition follows immediately:

Proposition 2: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let it be known that y_1 and y_0 are statistically independent, conditional on x. Then

(14) $$P(y_1 \epsilon B \mid x) P(y_0 \epsilon B \mid x) \leq P(w_m \epsilon B \mid x)$$
$$\leq P(y_1 \epsilon B \mid x) + P(y_0 \epsilon B \mid x) - P(y_1 \epsilon B \mid x) P(y_0 \epsilon B \mid x). \quad \blacksquare$$

Whereas the bound obtained in Proposition 1 was generically one-sided, the present bound is generically two-sided. The new lower bound on $P(w_m \mid x)$ is informative whenever $P(y_1 \epsilon B \mid x) > 0$

and $P(y_0 \epsilon B \mid x) > 0$. The upper bound is informative whenever $P(y_1 \epsilon B \mid x) < 1$ and $P(y_0 \epsilon B \mid x) < 1$.

Suppose that Y is the real line. By Proposition 2,

$$(15)\quad P(y_1 \leq u \mid x)P(y_0 \leq u \mid x) \leq P(w_m \leq u \mid x)$$
$$\leq P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - P(y_1 \leq u \mid x)P(y_0 \leq u \mid x)$$

for all $u \epsilon R^1$. Corollary 2.1 inverts (15) to obtain sharp bounds on quantiles of $P(w_m \mid x)$. The proof uses the same argument as was applied to prove Corollary 1.1, and so is omitted.

<u>Corollary 2.1</u>: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let it be known that y_1 and y_0 are statistically independent, conditional on x. Let Y be the real line. Let

$$r_2(\alpha \mid x) \equiv \inf_u \text{ s.t. } P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - P(y_1 \leq u \mid x)P(y_0 \leq u \mid x) \geq \alpha$$

$$s_2(\alpha \mid x) \equiv \inf_u \text{ s.t. } P(y_1 \leq u \mid x)P(y_0 \leq u \mid x) \geq \alpha$$

Then

$$(16)\quad r_2(\alpha \mid x) \leq q_m(\alpha \mid x) \leq s_2(\alpha \mid x). \blacksquare$$

4.2. SHIFTED OUTCOMES

Evaluation studies often assume that y_1 and y_0 are not only statistically dependent but functionally dependent. It is especially common to assume that real-valued outcomes are shifted versions of one another; that is,

$$(17)\quad P(y_1 = y_0 + \delta \mid x) = 1,$$

for some $\delta \epsilon R^1$. For example, see Heckman and Robb (1985) or Robinson (1989).

Suppose that (17) holds.[8] Knowledge of $P(y_1 \mid x)$ and $P(y_0 \mid x)$ implies knowledge of δ. So the joint distribution $P(y_1, y_0 \mid x)$ is known and the bound (7) is computable. Thus, we have

<u>Proposition 3:</u> Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let Y be the real line. Let it be known that $P(y_1 = y_0 + \delta \mid x) = 1$, for some $\delta \in R^1$. Then δ is identified and

$$(18)\quad P[(y_0+\delta)\epsilon B \cap y_0 \epsilon B \mid x] \leq P(w_m \epsilon B \mid x)$$
$$\leq P[(y_0+\delta)\epsilon B \mid x] + P(y_0 \epsilon B \mid x) - P[(y_0+\delta)\epsilon B \cap y_0 \epsilon B \mid x]. \quad \blacksquare$$

When $B = (-\infty, u]$, this bound takes a very simple form. Assume, without loss of generality, that $\delta \geq 0$. Then (18) becomes

$$(19)\quad P(y_0 \leq u - \delta \mid x) \leq P(w_m \leq u \mid x) \leq P(y_0 \leq u \mid x)$$

or, equivalently,

$$(19')\quad P(y_1 \leq u \mid x) \leq P(w_m \leq u \mid x) \leq P(y_0 \leq u \mid x).$$

Corollary 3.1 inverts (19') to obtain sharp bounds on quantiles of $P(w_m \mid x)$.

<u>Corollary 3.1</u>: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let Y be the real line. Let it be known that $P(y_1 = y_0 + \delta \mid x) = 1$, for some $\delta \geq 0$. Let

$$r_3(\alpha \mid x) \equiv \inf_u \text{ s.t. } P(y_0 \leq u \mid x) \geq \alpha$$

$$s_3(\alpha \mid x) \equiv \inf_u \text{ s.t. } P(y_1 \leq u \mid x) \geq \alpha.$$

Then

(20) $r_3(\alpha \mid x) \leq q_m(\alpha \mid x) \leq s_3(\alpha \mid x)$. ■

4.3. ORDERED OUTCOMES

Outcomes y_1 and y_0 are said to be ordered with respect to a given set B if y_0 almost always falls in B when y_1 does; that is,

(21) $P(y_0 \epsilon B \mid x, y_1 \epsilon B) = 1$.

For example, let the outcomes be binary, taking the value 0 or 1. If $P(y_1=0 \mid x, y_0=0) = 1$, then the outcomes are ordered with respect to the set $B = \{0\}$. As another example, suppose that the outcomes are real-valued and that

(22) $P(y_1 \geq y_0 \mid x) = 1$.

Then y_1 and y_0 are ordered with respect to the sets $B = (-\infty, u]$, as $y_1 \leq u \Rightarrow y_0 \leq u$.[9]

The assumption of ordered outcomes might be invoked in analyzing a preschool educational intervention such as the Perry Preschool project. Let $z_m = 1$ if a child receives the intervention and $z_m = 0$ otherwise. Let the outcome of interest be high school graduation (1 = yes, 0 = no). One may believe that receiving the intervention cannot possibly diminish a child's prospects for graduation. If so, then any child who receives the intervention and does not graduate would not graduate in the absence of the intervention. That is, $P(y_0=0 \mid x, y_1=0) = 1$. So the outcomes are ordered with respect to the set $B = \{0\}$.

The assumption might also be invoked in analyzing a cancer treatment such as chemotherapy. Let $z_m = 1$ if a patient is treated by chemotherapy and $z_m = 0$ if by placebo. Let the outcome of interest be life-span following each treatment. If chemotherapy is never harmful, then $y_1 \geq y_0$ for all patients.

If y_1 and y_0 are ordered with respect to B, then

$$(23)\quad P(y_1 \epsilon B \cap y_0 \epsilon B \mid x) = P(y_1 \epsilon B \mid x),$$

so the bound (7) is computable. In particular, we have

<u>Proposition 4</u>: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let it be known that $P(y_0 \epsilon B \mid x, y_1 \epsilon B) = 1$. Then

$$(24)\quad P(y_1 \epsilon B \mid x) \leq P(w_m \epsilon B \mid x) \leq P(y_0 \epsilon B \mid x). \blacksquare$$

An interesting result emerges when (24) is applied to real-valued outcomes satisfying (22). Letting $B = (-\infty, u]$, we find that (24) coincides with the bound (19') that holds when outcomes are known to be shifted. Thus, it turns out that assumptions (17) and (22) have the same identifying power.[10]

5. Restrictions on the Treatment Policy

To prove Proposition 1, we constructed two extreme treatment policies, one minimizing $P(w_m \epsilon B \mid x)$ and one maximizing it (see equations 5 and 6). In this section, we examine the identifying power of prior information implying that m is not one of these extreme policies.

5.1. TREATMENT INDEPENDENT OF OUTCOMES

Suppose it is known that, under policy m, the treatment z_m received by each person with covariates x is statistically independent of the person's outcomes (y_1,y_0). That is,

$$(25a) \quad P(y_1 \mid x) = P(y_1 \mid x, z_m=1) = P(y_1 \mid x, z_m=0)$$

and

$$(25b) \quad P(y_0 \mid x) = P(y_0 \mid x, z_m=1) = P(y_0 \mid x, z_m=0).$$

Then equation (2) reduces to

$$(26) \quad P(w_m \mid x) = P(y_1 \mid x)P(z_m=1 \mid x) + P(y_0 \mid x)P(z_m=0 \mid x).$$

If the fraction of the population receiving each treatment (i.e., $P(z_m \mid x)$) is known, then $P(w_m \mid x)$ is identified. Our present concern, however, is with the situation in which (25) is the only prior information available. In this case, the only restriction on the treatment distribution is that $P(z_m=1 \mid x)$ and $P(z_m=0 \mid x)$ must lie in the unit interval and add up to one. Hence, Proposition 5 follows immediately:

Proposition 5: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let it be known that z_m is statistically independent of (y_1,y_0), conditional on x. Then

$$(27)\quad \min[P(y_1 \epsilon B \mid x), P(y_0 \epsilon B \mid x)] \;\le\; P(w_m \epsilon B \mid x) \;\le\; \max[P(y_1 \epsilon B \mid x), P(y_0 \epsilon B \mid x)]. \;\blacksquare$$

Observe that the present bound on $P(w_m \epsilon B \mid x)$ is a subset of the bound reported in Proposition 2, which assumed that y_1 and y_0 are statistically independent. This fact has a simple explanation. Equation (26) shows that, if z_m is statistically independent of (y_1,y_0), then $P(w_m \mid x)$ depends on the distribution of (y_1,y_0) only through the two marginal distributions $P(y_1 \mid x)$ and $P(y_0 \mid x)$. Hence, if one knows that z_m is independent of (y_1,y_0), then knowing that y_1 and y_0 are statistically independent adds no identifying power.

Suppose that Y is the real line. The bound obtained in Proposition 5 can be inverted to produce the following bound on quantiles of $P(w_m \mid x)$:

Corollary 5.1: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let it be known that z_m is statistically independent of (y_1,y_0), conditional on x. Let Y be the real line. Let

$$r_5(\alpha \mid x) \;\equiv\; \inf_u \text{ s.t. } \max[P(y_1 \le u \mid x), P(y_0 \le u \mid x)] \;\ge\; \alpha.$$

$$s_5(\alpha \mid x) \;\equiv\; \inf_u \text{ s.t. } \min[P(y_1 \le u \mid x), P(y_0 \le u \mid x)] \;\ge\; \alpha.$$

Then

$$(28)\quad r_5(\alpha \mid x) \;\le\; q_m(\alpha \mid x) \;\le\; s_5(\alpha \mid x). \;\blacksquare$$

5.2. TREATMENT A KNOWN FUNCTION OF THE OUTCOMES

Suppose that, under policy m, the treatment z_m received by each person with covariates x is a known function of the person's outcomes (y_1,y_0). That is,

$$(29)\quad z_m = z_m(y_1,y_0)$$

for some known function $z_m(.,.)$: $YxY \rightarrow \{0,1\}$. Assumption (29) is essentially the polar opposite of the independence assumption just examined in Section 5.1. It follows from (29) that the realized outcome w_m is a known function of (y_1,y_0), namely

$$(30)\quad w_m = y_1 z_m(y_1,y_0) + y_0[1-z_m(y_1,y_0)].$$

Hence, $P(w_m \mid x)$ is completely determined by the joint distribution of (y_1,y_0).

Clearly, $P(w_m \mid x)$ is identified if assumption (29) is combined with knowledge of the joint distribution of (y_1,y_0). For example, (29) might be combined with the assumption invoked in Proposition 2, where it was assumed that y_1 and y_0 are statistically independent, or with the assumption invoked in Proposition 3, where it was assumed that y_1 and y_0 are shifted outcomes. Our present concern, however, is with the situation in which (29) is the only prior information available.

It appears difficult to characterize the identifying power of assumption (29) in general terms. Consider the event probability $P(w_m \epsilon B \mid x)$. By (30),

$$(31)\quad P(w_m \epsilon B \mid x) = P[y_1 z_m(y_1,y_0) + y_0\{1-z_m(y_1,y_0)\} \epsilon B \mid x].$$

The restrictions on this probability implied by knowledge of $[P(y_1 \mid x), P(y_0 \mid x)]$ depend on the form of the set B and of the function $z_m(.,.)$ determining treatment selection.

On the other hand, it is easy to analyze two symmetric special cases often assumed in empirical studies. The remainder of this section focuses on these cases.

SELECTION OF THE TREATMENT WITH THE SMALLER/LARGER OUTCOME: One case is the *competing risks* model applied widely in survival analysis (see Kalbfleisch and Prentice, 1980). Here, Y is the real line and the treatment yielding the smaller outcome is selected, so

(32) $w_m = \min(y_1, y_0)$.

At the other extreme, economic analyses of voluntary treatment policies often assume that the treatment yielding the larger outcome is selected, so

(33) $w_m = \max(y_1, y_0)$.

In the labor-economics literature on occupation choice, assumption (33) is often called the *Roy* model (see Heckman and Honore, 1990).

Proposition 6 gives sharp bounds for event probabilities of the form $P(w_m \leq u \mid x)$, assuming that (32) or (33) holds:

Proposition 6: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let Y be the real line.

A. Let it be known that $w_m = \min(y_1, y_0)$. Then

(34) $\max[P(y_1 \leq u \mid x), P(y_0 \leq u \mid x)] \leq P(w_m \leq u \mid x) \leq \min[P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x), 1]$.

B. Let it be known that $w_m = \max(y_1, y_0)$. Then

(35) $\max[0, P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - 1] \leq P(w_m \leq u \mid x)$

$\leq \min[P(y_1 \leq u \mid x), P(y_0 \leq u \mid x)]$. ■

PROOF:

A. In this case,

$$P(w_m \leq u \mid x) = P[\min(y_1, y_0) \leq u \mid x] = P(y_1 \leq u \cup y_0 \leq u \mid x)$$

$$= P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - P(y_1 \leq u \cap y_0 \leq u \mid x).$$

The Frechet bound on $P(y_1 \leq u \cap y_0 \leq u \mid x)$, given in (8), is

$$\max[0, P(y_1 \leq u \mid x) + P(y_0 \leq u \mid x) - 1] \leq P(y_1 \leq u \cap y_0 \leq u \mid x)$$

$$\leq \min[P(y_1 \leq u \mid x), P(y_0 \leq u \mid x)].$$

The result follows.

B. Here

$$P(w_m \leq u \mid x) = P[\max(y_1, y_0) \leq u \mid x] = P(y_1 \leq u \cap y_0 \leq u \mid x).$$

So the result is an immediate application of Frechet bound (8).

Q.E.D.

It is interesting to compare these bounds with those reported earlier under other assumptions. The lower bound on $P(w_m \leq u \mid x)$ under the assumption that $w_m = \min(y_1, y_0)$ coincides with the upper bound under the assumption that treatment is independent of the outcomes (see Proposition 5). The new upper bound coincides with the upper bound in the absence of prior information (see Proposition 1). So the competing-risks model strongly constrains $P(w_m \leq u \mid x)$ from below but does not constrain it from above.

Conversely, the lower bound on $P(w_m \leq u \mid x)$ under the assumption that $w_m = \max(y_1, y_0)$ coincides with the lower bound in the absence of prior information (see Proposition 1). The new upper bound coincides with the lower bound under the assumption that treatment is independent of the outcomes (see Proposition 5). So the Roy model does not constrain $P(w_m \leq u \mid x)$ from below but strongly constrains it from above.

Corollary 6.1 inverts the bounds in Proposition 6 to produce bounds on quantiles of $P(w_m \mid x)$.

Corollary 6.1: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let Y be the real line.

A. Let it be known that $w_m = \min(y_1, y_0)$. Then

(36) $r_1(\alpha \mid x) \leq q_m(\alpha \mid x) \leq r_5(\alpha \mid x)$.

B. Let it be known that $w_m = \max(y_1, y_0)$. Then

(37) $s_5(\alpha \mid x) \leq q_m(\alpha \mid x) \leq s_1(\alpha \mid x)$. ∎

5.3. KNOWN TREATMENT DISTRIBUTION

Suppose that under policy m, a known fraction p of the persons with covariate value x receive treatment y_0, the remaining fraction (1-p) receiving treatment y_1. That is,

$$(38)\quad P(z_m = 0 \mid x) \ = \ p,$$

where p is known. Also suppose that no information is available on the composition of the subpopulation receiving each treatment.

For example, one treatment may be universally preferred to the other, but the available supply of the preferred treatment may suffice to treat only a fraction of the persons with covariates x. This fraction may be known, but the rule used to allocate the supply of the preferred treatment may not be known.

Given (38), $P(w_m \mid x)$ may be written

$$(39)\quad P(w_m \mid x) \ = \ P(y_1 \mid x, z_m = 1)(1-p) \ + \ P(y_0 \mid x, z_m = 0)p.$$

The distributions $[P(y_1 \mid x), P(y_0 \mid x)]$ identified by the experimental evidence may be written

$$(40a)\quad P(y_1 \mid x) \ = \ P(y_1 \mid x, z_m = 1)(1-p) \ + \ P(y_1 \mid x, z_m = 0)p$$

and

$$(40b)\quad P(y_0 \mid x) \ = \ P(y_0 \mid x, z_m = 1)(1-p) \ + \ P(y_0 \mid x, z_m = 0)p.$$

Knowledge of $P(y_1 \mid x)$ and p restricts $P(y_1 \mid x,z_m=1)$ and $P(y_1 \mid x,z_m=0)$ to pairs of distributions that satisfy (40a); similarly, knowledge of $P(y_0 \mid x)$ and p restricts $P(y_0 \mid x,z_m=1)$ and $P(y_0 \mid x,z_m=0)$ to pairs of distributions that satisfy (40b). Examination of the feasible pairs shows that $P(y_1 \mid x,z_m=1)$ and $P(y_0 \mid x,z_m=0)$ must lie in the following sets of distributions:

(41a) $P(y_1 \mid x,z_m=1) \in \Psi_{11}(p) \equiv \Psi \cap [\{P(y_1 \mid x)-p\psi\}/(1-p): \psi \in \Psi]$

and

(41b) $P(y_0 \mid x,z_m=0) \in \Psi_{00}(p) \equiv \Psi \cap [\{P(y_0 \mid x) - (1-p)\psi\}/p: \psi \in \Psi]$,

where Ψ denotes the set of all distributions on Y. It follows that $P(w_m \mid x)$ is a (1-p,p) mixture of a distribution in $\Psi_{11}(p)$ and one in $\Psi_{00}(p)$. That is,

(42) $P(w_m \mid x) \in [(1-p)\psi_{11} + p\psi_{00}: (\psi_{11},\psi_{00}) \in \Psi_{11}(p)\times\Psi_{00}(p)]$.

Relation (42) completely characterizes the restrictions on $P(w_m \mid x)$ implied by knowledge of $[P(y_1 \mid x),P(y_0 \mid x),P(z_m \mid x)]$, but the characterization is not transparent. Horowitz and Manski (1992) have analyzed the sets $\Psi_{11}(p)$ and $\Psi_{00}(p)$ in their recent study of the *contaminated sampling* problem, whose formal structure is similar to the problem studied here. In particular, their Corollary 1.2 proves the following sharp bounds on $P(y_1 \in B \mid x,z_m=1)$ and $P(y_0 \in B \mid x,z_m=0)$:

(43a) $\max[0,\{P(y_1 \in B \mid x)-p\}/(1-p)] \leq P(y_1 \in B \mid x,z_m=1) \leq \min[P(y_1 \in B \mid x)/(1-p),1]$

and

(43b) $\max[0,\{P(y_0\epsilon B \mid x)-(1-p)\}/p] \;\leq\; P(y_0\epsilon B \mid x,z_m=0) \;\leq\; \min[P(y_0\epsilon B \mid x)/p,1]$.

This and (39) imply Proposition 7:

Proposition 7: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let $P(z_m=0 \mid x) = p$, for known p. Then

(44) $$\max[0,P(y_1\epsilon B \mid x)-p] + \max[0,P(y_0\epsilon B \mid x)-(1-p)] \;\leq\; P(w_m\epsilon B \mid x)$$
$$\leq\; \min[1-p,P(y_1\epsilon B \mid x)] + \min[p,P(y_0\epsilon B \mid x)]. \blacksquare$$

Inverting this bound yields Corollary 7.1.

Corollary 7.1: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known. Let $P(z_m=0 \mid x) = p$, for known p. Let Y be the real line. Let

$$r_{7p}(\alpha \mid x) \;\equiv\; \inf_u \text{ s.t. } \min[1-p,P(y_1\leq u \mid x)] + \min[p,P(y_0\leq u \mid x)] \geq \alpha$$
$$s_{7p}(\alpha \mid x) \;\equiv\; \inf_u \text{ s.t. } \max[0,P(y_1\leq u \mid x)-p] + \max[0,P(y_0\leq u \mid x)-(1-p)] \geq \alpha.$$

Then

(45) $$r_{7p}(\alpha \mid x) \;\leq\; q_m(\alpha \mid x) \;\leq\; s_{7p}(\alpha \mid x). \blacksquare$$

EXPERIMENTAL EVIDENCE ON ONE TREATMENT: Throughout this paper, I have assumed that experimental evidence is available for both treatments. Suppose now that such evidence is available for only one treatment, say treatment 1; so $P(y_1 \mid x)$ is known and $P(y_0 \mid x)$ is unrestricted. In the absence of information on the fraction of the population receiving each treatment, nothing can be learned about $P(w_m \mid x)$. After all, $P(z_m=0 \mid x) = 1$ might hold, in which case $P(w_m \mid x) = P(y_0 \mid x)$. On the other hand, inference on $P(w_m \mid x)$ is possible if the treatment distribution is known. Proposition 8 and Corollary 8.1 provide the results.

Proposition 8: Let $P(y_1 \mid x)$ be known. Let $P(z_m=0 \mid x) = p$, for known p. Then

$$(46) \quad \max[0, P(y_1 \epsilon B \mid x) - p] \leq P(w_m \epsilon B \mid x) \leq \min[P(y_1 \epsilon B \mid x) + p, 1]. \blacksquare$$

PROOF: The present problem is a mirror image of the *corrupted sampling* problem studied by Horowitz and Manski (1992). There the concern was to characterize the restrictions on $P(y_1 \mid x)$ implied by knowledge of $P(w_m \mid x)$ and p. Inspection of (39) and (40a) shows that the two problems are formally equivalent. So Corollary 1.2 of Horowitz and Manski (1992) gives the result.

Q.E.D.

Corollary 8.1: Let $P(y_1 \mid x)$ be known. Let $P(z_m=0 \mid x) = p$, for known p. Let Y be the real line. Let

$$r_{8p}(\alpha \mid x) \equiv \inf_u \text{ s.t. } \min[P(y_1 \leq u \mid x) + p, 1] \geq \alpha$$

$$s_{8p}(\alpha \mid x) \equiv \inf_u \text{ s.t. } \max[0, P(y_1 \leq u \mid x) - p] \geq \alpha.$$

Then

(47) $r_{8p}(\alpha \mid x) \leq q_m(\alpha \mid x) \leq s_{8p}(\alpha \mid x)$. ■

The lower bound on $P(w_m \epsilon B \mid x)$ is informative if $P(y_1 \epsilon B \mid x) > p$; the upper bound is informative if $P(y_1 \epsilon B \mid x) < 1-p$. When the bound is informative both above and below, it restricts $P(w_m \epsilon B \mid x)$ to an interval of width 2p, centered at $P(y_1 \epsilon B \mid x)$. In contrast to the case when experimental evidence is available for both treatments, the present bounds on quantiles are not always informative. The lower bound on $q_m(\alpha \mid x)$ is informative if $p < \alpha$; the upper bound is informative if $p < 1 - \alpha$.

5.4. ASSUMPTIONS IDENTIFYING THE OUTCOME DISTRIBUTION

Propositions 1 through 8 assume enough empirical evidence and prior information to bound event probabilities $P(w_m \epsilon B \mid x)$, but not enough to identify them. In Section 5.1 and 5.2, we noted in passing some assumptions that do suffice to identify $P(w_m \epsilon B \mid x)$. Proposition 9 presents these simple findings formally.

Proposition 9: Let $P(y_1 \mid x)$ and $P(y_0 \mid x)$ be known.

A. Let $z_m = z_m(y_1,y_0)$ for some known function $z_m(.,.)$: $YxY \rightarrow \{0,1\}$. Let it be known that y_1 and y_0 are either statistically independent, shifted, or ordered outcomes, conditional on x. Then

(48) $P(w_m \epsilon B \mid x) = P[y_1 z_m(y_1,y_0) + y_0\{1-z_m(y_1,y_0)\} \epsilon B \mid x]$

is identified.

B. Let it be known that z_m is statistically independent of (y_1,y_0), conditional on x. Let $P(z_m=0 \mid x) = p$, for known p. Then

(49) $P(w_m \epsilon B \mid x) = P(y_1 \epsilon B \mid x)(1-p) + P(y_0 \epsilon B \mid x)p$

is identified. ■

Notes

[1]In practice there often are multiple feasible treatments, but this paper restricts attention to the two-treatment case assumed in most of the literature. It is common to call one of these the "treatment" or "experiment," and the other the "control."

[2]Of course, one might observe realizations under more than one policy. Most work on selection problems has focused on the case in which only one policy is observed.

[3]The mixing problem should not be confused with the converse problem: What does knowledge of $P[y_1z_m+y_0(1-z_m) \mid x]$ imply about $[P(y_1 \mid x),P(y_0 \mid x),P(z_m \mid x)]$? The latter is sometimes referred to as a *mixture* problem.

[4]A more demanding technical challenge, not addressed here, is to determine the identifiability of the conditional mean $E(w_m \mid x)$.

[5]Not all forms of prior information have identifying power. Restrictions on $[P(y_1 \mid x),P(y_0 \mid x)]$ are superfluous, as these distributions are identified by the experimental evidence. Such restrictions may improve the precision of sample estimates of $[P(y_1 \mid x),P(y_0 \mid x)]$. This usage is distinct from the identification concerns of the present paper.

[6]It would be easy to take sampling error into account. All of the estimated bounds on $P(w_m=1 \mid x)$ reported in Table 1 are smooth functions of the estimates of $[P(y_1=1 \mid x),P(y_0=1 \mid x)]$, which are based on sample sizes of fifty-eight and sixty-three respectively. Conventional sampling confidence bands can be placed around the estimates in Table 1. Manski et al. (1992) presents confidence bands of this kind in an empirical study concerned with the selection problem.

[7]See Ord (1972) for a brief exposition of the Frechet bounds, and Ruschendorf (1981) for a rather general analysis.

[8]Experimental evidence makes the shifted-outcome assumption a testable hypothesis. If (17) holds, the known distributions $P(y_1 \mid x)$ and $P(y_0 \mid x)$ must be the same up to a translation of location. In contrast, the statistical independence assumption of the preceding section is not testable, as it implies no restrictions on $P(y_1 \mid x)$ and $P(y_0 \mid x)$.

[9]Experimental evidence makes the ordered-outcomes assumption a testable hypothesis. If (21) holds, $P(y_0 \epsilon B \mid x)$ must be at least as large as $P(y_1 \epsilon B \mid x)$.

[10]Assumptions (17) and (22) are not formally equivalent; they just have the same identifying power in the present setting. In the context of the selection problem, the two assumptions have different implications. See Manski (1993).

References

Berrueta-Clement, J., Schweinhart, L., Barnett, W., Epstein, A., and Weikart, D. (1984), *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19*, Ypsilanti, Michigan: High/Scope Press.

Frechet, M. (1951), "Sur Les Tableaux de Correlation Donte les Marges sont Donnees," *Annals de Universite de Lyon A*, Series 3, 14, 53-77.

Heckman, J. and Honore, B. (1990), "The Empirical Content of the Roy Model," *Econometrica* 58, 1121-1149.

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

Holden, C. (1990), "Head Start Enters Adulthood," *Science* 247, 1400-1402.

Horowitz, J. and Manski, C. (1992), "Identification and Robustness in the Presence of Errors in Data," Social Systems Research Institute Paper 9209, University of Wisconsin-Madison.

Kalbfleisch, J. and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Maddala, G. S. (1983), *Qualitative and Limited Dependent Variable Models in Econometrics*, Cambridge, England: Cambridge University Press.

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

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

Manski, C. (1993), "The Selection Problem," in C. Sims (editor), *Advances in Econometrics, Sixth World Congress of the Econometric Society*, Cambridge, UK: Cambridge University Press, forthcoming.

Manski, C. and Garfinkel, I. (editors) (1992), *Evaluating Welfare and Training Programs*, Cambridge, Mass.: Harvard University Press.

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

Ord, J. (1972), *Families of Frequency Distributions*, Griffin's Statistical Monographs & Courses No. 30, New York: Hafner.

Robinson, C. (1989), "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy* 97, 639-667.

Ruschendorf, L. (1981), "Sharpness of Frechet-Bounds," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* 57, 293-302.