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

# *The Mixing Problem in Program Evaluation*

## 3.1. The Experimental Evaluation of Social Programs

In the United States, concern with the evaluation of social programs has spread rapidly since the 1960s, when attempts were made to evaluate the impacts of programs proposed as part of the War on Poverty. Evaluation requirements now appear in major federal statutes. One of these is the Family Support Act of 1988, which revised the AFDC program. In Title II of this statute, Congress mandated study of the effectiveness of training programs initiated by the states under the new Job Opportunities and Basic Skills Training Program (JOBS). Congress even stipulated the mode of data collection: "a demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program" (Public Law 100-485, October 13, 1988, section 203, 102 stat. 2380).

Until the early 1980s, evaluations of welfare and training programs generally analyzed data from ongoing programs rather than from controlled social experiments.[1] Some studies analyzed cross-sectional variation in outcomes across cities or states that had different programs. Others analyzed "before-and-after" data; that is, time-series variation within a city or state that altered its program.

More recently, social experiments have come to dominate the evaluations commissioned by the federal government and by the major foundations. Dissatisfaction with the evaluations of job training programs performed in the 1970s led the U.S. Department of Labor to commission an experimental evaluation of the Job Training Partnership Act in the mid-1980s (see Hotz, 1992). And a set of experi-

ments sponsored by the Ford Foundation and executed by the Manpower Demonstration Research Corporation influenced the federal government to choose experimental analysis as the preferred approach to evaluations of AFDC reforms (see Greenberg and Wiseman, 1992).

Controlled social experimentation has become so much the new orthodoxy of evaluation that Jo Anne Barnhart, an Assistant Secretary in the U.S. Department of Health and Human Services during the Bush administration, could write this about the evaluation of training programs for welfare recipients: "In fact, nonexperimental research of training programs has shown such methods to be so unreliable, that Congress and the Administration have both insisted on experimental designs for the Job Training Partnership Act (JTPA) and the Job Opportunities and Basic Skills (JOBS) programs" (letter from Jo Anne B. Barnhart to Eleanor Chelimsky, reproduced as U.S. General Accounting Office, 1992, appendix II).

Barnhart's reference to the unreliability of nonexperimental research reflects the view of some social scientists that the selection problem analyzed in Chapter 2 precludes credible inference on treatment effects from the observation of outcomes when treatments are uncontrolled. These social scientists have recommended that empirical research should focus exclusively on the design and analysis of controlled experiments. See Bassi and Ashenfelter (1986), LaLonde (1986), and Coyle, Boruch, and Turner (1989).

### *The Classical Argument*

Recent arguments for controlled social experiments follow the wonderfully simple, classical lines of Fisher (1935). In a controlled experiment, random samples of persons with specified covariates are drawn and formed into treatment groups. All members of a treatment group 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 applied to all persons with the specified covariates.

Formally, recall the setup of Section 2.5. There are two mutually exclusive treatments, labeled 1 and 0. Each member of the population is described by values for the variables $(y_1, y_0, z, x)$. The outcome $y_1$ is observed if a person receives treatment 1, $y_0$ is observed if the person receives treatment 0, and $z$ indicates which treatment

is received. The sampling process identifies $P(y_1 \mid x, z = 1)$ and $P(y_0 \mid x, z = 0)$. Randomized selection of treatment implies that treatment is statistically independent of the outcomes. Hence $P(y_1 \mid x, z = 1) = P(y_1 \mid x)$ and $P(y_0 \mid x, z = 0) = P(y_0 \mid x)$. A controlled experiment thus reveals $P(y_1 \mid x)$ and $P(y_0 \mid x)$.

### *Extrapolation from Social Experiments*

The classical argument for experimentation does not suffice to conclude that experiments may be used to predict the outcomes of social programs. Experimental evaluation also requires a critical invariance assumption: the experimental version of a program must operate as would an actual program. It is this premise that allows one to extrapolate from the experiment to the real world.

Critiques of social experimentation argue that, for a variety of reasons, experimental versions of social programs may not operate as would actual programs (see Hausman and Wise, 1985, and Manski and Garfinkel, 1992). Four widely recognized concerns are described here. A fifth, previously unexplored, issue will be examined in depth beginning in Section 3.2.

*Program administration.* Experiments with randomized treatments may be administered differently from actual programs mandating homogeneous treatment of the population. Social experiments generally cannot be performed using the double-blind protocols of medical trials, in which neither experimenters nor subjects know who is in each treatment group. Program administrators inevitably know who is in each group and cannot be prevented from using this information to influence outcomes.

*Macro feedback effects.* Full-scale programs may change the environment in ways that influence outcomes. Possible feedbacks range from labor market equilibration to information diffusion to norm formation. The scale of the typical social experiment is too small to discern these macro effects, which may become prominent when a program is actually implemented (see Garfinkel, Manski, and Michalopolous, 1992).

*Site selection.* The classical experimental paradigm calls for random selection of treatment sites, but evaluators generally do not

have the power to compel sites to cooperate. Hence experiments are typically conducted at sites selected jointly by evaluators and local officials. Hotz (1992) describes how the JTPA evaluators originally sought to select sites randomly but, being unable to secure the agreement of the randomly drawn sites, were ultimately required to provide large financial incentives to nonrandomly chosen sites in order to obtain their cooperation.

*Program participation.* A classical social experiment randomly assigns the participants in a social program to groups receiving different treatments. Analyis of the experimental data presumes that the existence of the experiment does not alter the population of persons participating in the program. Heckman (1992) and Moffitt (1992b) observe that it is not plausible to assume an invariant population of participants, because the value to a person of participating in a program with randomized treatment is not the same as that of participating in a program with known treatment. Heckman makes the point well when he observes that social experimentation is intrinsically different from experimentation in the biological sciences and agriculture: "Plots of ground do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated" (p. 215).

None of the foregoing concerns implies that experimental evidence is uninformative. These concerns do imply that one should not expect the distribution of outcomes in a randomly selected treatment group to coincide with the outcomes that would be realized in an actual social program. Extrapolation from experimental data, as from nonexperimental data, requires that the empirical evidence be combined with prior information.

## 3.2. Variation in Treatment

Let us abstract from the very real concerns just expressed and accept the classical argument that the outcomes realized by a randomly selected treatment group are the same (up to random sampling error) as would be observed if the treatment were applied to all persons with specified covariates. Policies mandating homogeneous treatment of

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

Consider the following inferential questions:

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

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

The first question poses the selection problem. The second question, which has remained unexplored and unnamed, is the subject of this chapter. Formally, the question asks what inferences about mixtures of two random variables can be made given knowledge of their marginal distributions. Hence I refer to it as the *mixing* problem.[2]

To state the mixing problem formally, let us enhance slightly the description of the population given in Section 2.5 by making explicit the *treatment policy,* or assignment rule, being implemented. Let each member of the population be described by values for $[(y_1, y_0), z_m, x]$. As before, there are two mutually exclusive treatments, labeled 1 and 0, and $(y_1, y_0)$ are the outcomes associated with the two treatments.

A treatment policy, now denoted $m$, determines which treatment each person receives. The indicator variable $z_m$ denotes the treatment received under policy $m$; $z_m = 1$ if the person receives treatment 1, and $z_m = 0$ otherwise. The outcome a person realizes under policy $m$ is $y_1$ if $z_m = 1$ and $y_0$ otherwise. Let $y_m$ denote the realized outcome; that is,

$$y_m \equiv y_1 z_m + y_0(1 - z_m). \tag{3.1}$$

The distribution of outcomes realized by persons with covariates $x$ is

$$(3.2)\quad P(y_m \mid x) \equiv P[y_1 z_m + y_0(1 - z_m) \mid x]$$
$$= P(y_1 \mid x, z_m = 1)P(z_m = 1 \mid x)$$
$$+ P(y_0 \mid x, z_m = 0)P(z_m = 0 \mid x).$$

A welfare recipient, for example, might be treated by job-specific training or by basic education. The relevant outcome might be earned income following treatment. One treatment policy might mandate the job training treatment for all welfare recipients and enforce the mandate. A second policy might attempt to mandate the basic education treatment 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 treatment costs.

The problem of interest is to learn about the distribution $P(y_m \mid x)$ 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 outcomes. The interesting inferential questions concern the feasibility of learning $P(y_m \mid x)$ when one observes outcomes under policies other than $m$. The selection problem and the mixing problem both concern the feasibility of extrapolating from observed treatment policies to unobserved ones.

The selection problem arises when policy $m$ mandates homogeneous treatment, but the available data are realizations under some other policy that may yield heterogeneous treatments. Suppose that $m$ makes treatment 1 mandatory for all persons with covariates $x$, so $P(z_m = 1 \mid x) = 1$ and $P(y_m \mid x) = P(y_1 \mid x)$. Suppose that the observable policy is some $\mu \neq m$.[3] 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 $P(z_\mu \mid x)$. Then the formal statement of the selection problem is:

> 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 $P(y_1 \mid x)$?

The mixing problem arises when policy $m$ may yield heterogeneous treatments, but the available data are realizations under policies imposing homogeneous treatments. In particular, the classical model of experimentation presumes that experimental evidence is available for both treatments, so the experiments identify $P(y_1 \mid x)$ and $P(y_0 \mid x)$. Therefore the formal statement of the mixing problem is:[4]

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

Section 3.3 uses empirical evidence from a famous social experiment, the Perry Preschool Project, to illustrate the mixing problem and the main findings of the chapter. Fifteen years after their participation in an 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 assumptions 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 3.4 through 3.7 present the analysis that yields the empirical results reported in Section 3.3. In studying the selection problem in Chapter 2, we found it productive to begin by determining what can be learned when the sampling process provides the only information available to the researcher. We then examined the identifying power of various forms of prior information that might plausibly be invoked in empirical studies. The present analysis uses the same approach.

Section 3.4 investigates the mixing problem when knowledge of the two marginal distributions $P(y_1 \mid x)$ and $P(y_0 \mid x)$ is the only information available. Then Sections 3.5 through 3.7 explore the identifying power of assumptions that restrict the determinants of outcomes.[5] Section 3.5 examines the implications of assumptions restricting the joint distribution of the outcomes $(y_1, y_0)$. Section 3.6 examines assumptions restricting the treatment policy. Section 3.7 cites some combinations of assumptions that identify $P(y_m \mid x)$.

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 3.4 through 3.7 maintain the assumption that the distributions $[P(y_1 \mid x), P(y_0 \mid x)]$ are known. 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 estimates, as is done in Section 3.3.

## 3.3. The Perry Preschool Project

Beginning in 1962, the Perry Preschool Project provided intensive educational and social services to a random sample of disadvantaged 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 19 (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 from disadvantaged backgrounds (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,
$y_1$ = outcome if a child were to be assigned to the experimental group ($y_1 = 1$ if high school graduate by age 19, $= 0$ otherwise), and
$y_0$ = outcome if a child were to be assigned to the control group ($y_0 = 1$ if high school graduate by age 19, $= 0$ otherwise).

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 probability among children with covariates $x$ would be .67 if all such children were to receive the intervention, and .49 if none of them was 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 probability of high school graduation under a treatment policy in which some children with covariates $x$ receive the intervention, but not others? Table 3.1 summarizes the inferences that can be made given the experimental evidence and varying forms of prior information about the outcome distribution and the treatment policy. The remainder of this section discusses the empirical findings. Sections 3.4 through 3.7 present the analysis underlying these findings.

### *Identification Using Only the Experimental Evidence*

It might be conjectured that $P(y_m = 1 \mid x)$ must lie between the graduation probabilities of the control and treatment groups, namely, between .49 and .67. This conjecture is correct for special outcome distributions and treatment policies. It holds if (a) the outcomes $(y_1, y_0)$

*Table 3.1* The Perry Preschool Project

Experimental Evidence

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

| Prior Information | $P(y_m = 1 \mid x)$ |
|---|---|
| No prior information | [.16, 1] |
| Independent outcomes | [.33, .83] |
| Ordered outcomes | [.49, .67] |
| Treatment independent of outcomes | [.49, .67] |
| Treatment maximizing graduation probability | [.67, 1] |
| + independent outcomes | .83 |
| + ordered outcomes | .67 |
| Treatment minimizing graduation probability | [.16, .49] |
| 1/10 population receives treatment 1 | [.39, .59] |
| + treatment independent of outcomes | .51 |
| 5/10 population receives treatment 1 | [.17, .99] |
| + treatment independent of outcomes | .58 |
| 9/10 population receives treatment 1 | [.57, .77] |
| + treatment independent of outcomes | .65 |

*Source:* Manski (1994b), table 1.

are ordered, with $y_1 \geq y_0$ for all children, or if (b) the treatment policy makes $z_m$ statistically independent of the outcomes $(y_1, y_0)$.

The conjecture does not hold more generally. In fact, the experimental evidence only implies that the graduation probability must lie between .16 and 1. That is, there exist outcome distributions and treatment policies that are consistent with the known values of $P(y_1 \mid x)$ and $P(y_0 \mid x)$ and that imply graduation probabilities as low as .16 and as high as 1.

This result is easily understood once one considers precisely what the experimental evidence reveals. Observing the outcomes of the treatment group reveals (ignoring sampling error) that $y_1 = 1$ for 67 percent of the population and $y_1 = 0$ for the remaining 33 percent. Observing the outcomes of the control group reveals that $y_0 = 1$ for 49 percent of the population and $y_0 = 0$ for the remaining 51 percent.

The experimental evidence does not reveal how $y_1$ and $y_0$ are related within the population, nor how policy $m$ assigns treatments. The impact of treatment policy on the graduation rate is most pronounced when $y_1$ and $y_0$ are most negatively associated. Among all distributions of $(y_1, y_0)$ that are consistent with the experimental evidence, the one with the greatest negative association between $y_1$ and $y_0$ is this:

$$P(y_1 = 0, y_0 = 0 \mid x) = .00 \qquad P(y_1 = 0, y_0 = 1 \mid x) = .33$$
$$P(y_1 = 1, y_0 = 0 \mid x) = .51 \qquad P(y_1 = 1, y_0 = 1 \mid x) = .16.$$

Given this distribution of outcomes, the graduation rate is maximized by adopting a treatment policy that provides the intervention only to those children with $y_1 = 1$. The result is a 100 percent graduation rate. At the other extreme, the graduation probability is minimized by adopting a treatment policy that provides the intervention only to those children with $y_1 = 0$. The result is a 16 percent graduation rate.

### *Prior Information*

The interval [.16, 1] is a worst-case bound on the graduation probability, computed in the absence of any prior information restricting the outcome distribution or the treatment policy. A researcher who pos-

sesses such information may be able to narrow the range of possible graduation probabilities.

Imagine that one has no information about the treatment policy but does have information about the outcome distribution. One might think that being treated by the preschool intervention can never harm a child's schooling prospects; that is, outcomes are ordered with $y_1 \geq y_0$ for all children. If so, then the graduation probability must lie between those observed in the control and treatment groups, namely, between .49 and .67. A more neutral assumption might be that $y_1$ and $y_0$ are statistically independent conditional on $x$. This assumption implies that the graduation probability must lie between .33 and .83; where the probability falls within this range depends on the treatment policy.

Next imagine that one has no information about the outcome distribution but does have information about the treatment policy. One might think that treatment decisions will be made by omniscient parents who choose for each child the treatment yielding the better outcome. This assumption implies that the graduation rate must lie between .67 and 1; where the rate falls within this range depends on the outcome distribution. Or one might think that assignments to treatments are statistically independent of outcomes, as they would be if an explicit random assignment rule were used. Then the graduation rate must lie between the .49 and .67 observed in the control and treatment groups.

Finally, imagine that resource constraints limit implementation of the intervention to part of the population. Suppose that one knows the fraction of the population receiving the intervention, but does not know the composition of the treated and untreated subpopulations. As Table 3.1 shows, knowing that 1/10 or 5/10 or 9/10 of the population receives the intervention implies that the graduation rate must lie in the interval [.39, .59] or [.17, .99] or [.57, .77], 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 $P(z_m = 1 \mid x)$ is fixed at a value near zero or one than it is when $P(z_m = 1 \mid x)$ 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, suppose it is known that the outcomes $y_1$ and $y_0$ are statistically independent of one another and that each child receives the treatment yielding the better outcome. Then the implied high school graduation rate is .83. Or suppose it is known that 5/10 of the population receives the intervention and that the treatment $z$ is independent of the outcomes $(y_1, y_0)$, as defined in Section 2.5. 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 implies 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.

## 3.4. Identification of Mixtures Using Only Knowledge of the Marginals

This section characterizes the restrictions on $P(y_m \mid x)$ implied by knowledge of $[P(y_1 \mid x), P(y_0 \mid x)]$ in the worst-case situation where no other information is available. Here and throughout the remainder of the chapter, I focus on the basic problem of inferring conditional probabilities of events. Manski (1994b) shows how these findings can be used to study the identification of quantiles of $P(y_m \mid x)$.

Consider the probability $P(y_m \in B \mid x)$ that the realized outcome $y_m$ falls in a specified set $B$, conditional on $x$. Given that $y_m$ always equals either $y_1$ or $y_0$, one might think that $P(y_m \in B \mid x)$ must lie between $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$. This is not the case. It turns out that when $P(y_1 \in B \mid x) + P(y_0 \in B \mid x) \leq 1$, then $P(y_m \in B \mid x)$ must lie in the interval $[0, P(y_1 \in B \mid x) + P(y_0 \in B \mid x)]$. When

$P(y_1 \in B \mid x) + P(y_0 \in B \mid x) \geq 1$, $P(y_m \in B \mid x)$ must lie in the interval $[P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - 1, 1]$. That is, when $P(y_1 \mid x)$ and $P(y_0 \mid x)$ are known but no information restricting the distribution of $[(y_1, y_0), z_m, x]$ is available, we can conclude only that

$$\max[0, P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - 1] \leq P(y_m \in B \mid x) \leq \min[P(y_1 \in B \mid x) + P(y_0 \in B \mid x), 1]. \tag{3.3}$$

This central finding may appear unintuitive, so I shall prove it here. We first determine the treatment policies that minimize and maximize $P(y_m \in B \mid x)$. Observe that if $y_1$ and $y_0$ both fall in the set $B$, then $y_m$ must fall in $B$. Moreover, if neither $y_1$ nor $y_0$ falls in $B$, then $y_m$ cannot fall in $B$. That is,

$$y_1 \in B \cap y_0 \in B \Rightarrow y_m \in B \tag{3.4a}$$

and

$$y_1 \notin B \cap y_0 \notin B \Rightarrow y_m \notin B, \tag{3.4b}$$

whatever treatment policy $m$ may be.

The treatment policy is relevant in those cases in which one of the two outcomes falls in $B$ and the other does not. The treatment policy minimizes $P(y_m \in B \mid x)$ if it always selects the treatment yielding the outcome that is not in $B$; that is, if

$$\begin{aligned} y_1 \notin B \cap y_0 \in B &\Rightarrow z_m = 1 \\ y_1 \in B \cap y_0 \notin B &\Rightarrow z_m = 0. \end{aligned} \tag{3.5}$$

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

$$\begin{aligned} y_1 \notin B \cap y_0 \in B &\Rightarrow z_m = 0 \\ y_1 \in B \cap y_0 \notin B &\Rightarrow z_m = 1. \end{aligned} \tag{3.6}$$

Therefore the largest possible value of $P(y_m \in B \mid x)$ is $P(y_1 \in B \cup y_0 \in B \mid x)$.

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

$$(3.7) \quad P(y_1 \in B \cap y_0 \in B \mid x) \leq P(y_m \in B \mid x) \leq P(y_1 \in B \cup y_0 \in B \mid x)$$

is the sharp bound on $P(y_m \in B \mid x)$. 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 $P(y_m \in B \mid x)$ is the smallest value of $P(y_1 \in B \cap y_0 \in B \mid x)$ 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 $P(y_1 \in B \cup y_0 \in B \mid x)$.

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

$$(3.8) \quad \max[0, P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - 1] \leq P(y_1 \in B \cap y_0 \in B \mid x) \leq \min[P(y_1 \in B \mid x), P(y_0 \in B \mid x)].$$

It follows immediately from (3.8) that the best computable lower bound on $P(y_m \mid x)$ is $\max[0, P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - 1]$. To obtain the best computable upper bound, observe that

$$(3.9) \quad P(y_1 \in B \cup y_0 \in B \mid x) = P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - P(y_1 \in B \cap y_0 \in B \mid x).$$

Applying the Frechet lower bound on $P(y_1 \in B \cap y_0 \in B \mid x)$ to (3.9) shows that

$$(3.10) \quad P(y_1 \in B \cup y_0 \in B \mid x) \leq \min[P(y_1 \in B \mid x) + P(y_0 \in B \mid x), \; 1]$$

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

## 3.5. Restrictions on the Outcome Distribution

In the preceding section, we showed that if $P(y_1 \in B \cap y_0 \in B \mid x)$ and $P(y_1 \in B \cup y_0 \in B \mid x)$ are known and if no restrictions are imposed on the treatment policy $m$, then inequality (3.7) provides the sharp bound on $P(y_m \in B \mid x)$. One may sometimes have prior information that makes the bound (3.7) computable. This section presents three cases.

### *Independent Outcomes*

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

$$(3.11) \quad P(y_1 \in B \cap y_0 \in B \mid x) = P(y_1 \in B \mid x) P(y_0 \in B \mid x).$$

With $P(y_1 \in B \cap y_0 \in B \mid x)$ known, the sharp bound on $P(y_m \in B \mid x)$ is

$$\begin{aligned} (3.12) \quad P(y_1 \in B \mid x) P(y_0 \in B \mid x) &\leq P(y_m \in B \mid x) \\ &\leq P(y_1 \in B \mid x) + P(y_0 \in B \mid x) \\ &\quad - P(y_1 \in B \mid x) P(y_0 \in B \mid x). \end{aligned}$$

Whereas the worst-case bound obtained in Section 3.4 was informative only in one direction, the present bound is generally informative both above and below. The new lower bound on $P(y_m \mid x)$ is informative when $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$ are positive. The upper bound is informative when $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$ are both less than 1.

*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 outcomes are shifted versions of one another; that is,[8]

$$y_1 = y_0 + k, \tag{3.13}$$

for some constant $k$. The assumption of shifted outcomes was discussed in Section 2.6.

Suppose it is known that (3.13) holds. Knowledge of $P(y_1 \mid x)$ and $P(y_0 \mid x)$ implies knowledge of $k$. So the joint distribution $P(y_1, y_0 \mid x)$ is known. With $P(y_1 \in B \cap y_0 \in B \mid x)$ known, the sharp bound on $P(y_m \in B \mid x)$ is

$$\begin{aligned} P[(y_0 + k) \in B \cap y_0 \in B \mid x] &\leq P(y_m \in B \mid x) \\ &\leq P[(y_0 + k) \in B \mid x] + P(y_0 \in B \mid x) \\ &\quad - P[(y_0 + k) \in B \cap y_0 \in B \mid x]. \end{aligned} \tag{3.14}$$

When $B$ is the set of all real numbers below some cutoff point $t$, this bound takes a simple form. Assume, without loss of generality, that $k \geq 0$. Then (3.14) becomes

$$P(y_0 \leq t - k \mid x) \leq P(y_m \leq t \mid x) \leq P(y_0 \leq t \mid x) \tag{3.15}$$

or, equivalently,

$$P(y_1 \leq t \mid x) \leq P(y_m \leq t \mid x) \leq P(y_0 \leq t \mid x). \tag{3.15'}$$

*Ordered Outcomes*

Outcomes $y_1$ and $y_0$ are said to be ordered with respect to a set $B$ if $y_0$ always falls in $B$ when $y_1$ does; that is,[9]

$$y_1 \in B \Rightarrow y_0 \in B. \tag{3.16}$$

For example, let the outcomes be binary, taking the value 0 or 1. If $y_1 = 0 \Rightarrow y_0 = 0$, then the outcomes are ordered with respect to the set $B = \{0\}$. Another example was given in Section 2.6, which considered the case

$$(3.17) \quad y_1 \geq y_0.$$

Here $y_1$ and $y_0$ are ordered with respect to all sets of the form $B = (-\infty, t]$, as $y_1 \leq t \Rightarrow y_0 \leq t$.

The assumption of ordered outcomes has earlier been discussed in the context of the Perry Preschool Project. One may believe that receiving the intervention cannot possibly diminish a child's prospects for high school graduation. If so, then any child who receives the intervention and does not graduate is a child who would not graduate in the absence of the intervention. That is, $y_1 = 0 \Rightarrow y_0 = 0$.

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

$$(3.18) \quad P(y_1 \in B \cap y_0 \in B \mid x) = P(y_1 \in B \mid x).$$

With $P(y_1 \in B \cap y_0 \in B \mid x)$ known, the sharp bound on $P(y_m \in B \mid x)$ is

$$(3.19) \quad P(y_1 \in B \mid x) \leq P(y_m \in B \mid x) \leq P(y_0 \in B \mid x).$$

An interesting result emerges when (3.19) is applied to outcomes satisfying (3.17). Letting $B = (-\infty, t]$, we find that (3.19) coincides with the bound (3.15′) that holds when outcomes are known to be shifted. Thus it turns out that, in the context of the mixing problem, assumptions (3.13) and (3.17) have the same power to identify $P(y_m \leq t \mid x)$. Section 2.6 showed that these two assumptions have different identifying power in the context of the selection problem.

## 3.6. Restrictions on the Treatment Policy

This section examines the restrictions on $P(y_m \mid x)$ implied by a set of polar treatment policies, in the absence of prior information about the outcome distribution. We first suppose that treatment is statisti-

cally independent of outcomes, as in random assignment policies. We then suppose that treatment minimizes or maximizes the probability that the realized outcome $y_m$ falls in specified sets $B$, as in competing-risks models and in the Roy model. The section also examines the quite different problem of inference when the fraction of the population receiving each treatment is known, but nothing is known about the composition of the subpopulations receiving each treatment.

### *Treatment Independent of Outcomes*

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

$$(3.20) \quad P[(y_1, y_0) \mid x, z_m] = P[(y_1, y_0) \mid x].$$

Then equation (3.2) reduces to

$$(3.21) \quad P(y_m \mid x) = P(y_1 \mid x)P(z_m = 1 \mid x) + P(y_0 \mid x)P(z_m = 0 \mid x).$$

If the fractions $P(z_m \mid x)$ of the population receiving each treatment are known, then $P(y_m \mid x)$ is identified. Our present concern, however, is with the situation in which (3.20) 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 (3.21) implies that $P(y_m \in B \mid x)$ must lie between $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$. That is,

$$(3.22) \quad \min[P(y_1 \in B \mid x), P(y_0 \in B \mid x)] \leq P(y_m \in B \mid x) \leq \max[P(y_1 \in B \mid x), P(y_0 \in B \mid x)].$$

The bound (3.22) is contained within each of the bounds reported in Section 3.5, which left the treatment policy unspecified and imposed restrictions on the outcome distribution. This fact has a simple explanation. Equation (3.21) shows that, if treatment is independent of the outcomes, then $P(y_m \mid x)$ 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 treatment is independent of the outcomes, then restrictions on the joint distribution of $(y_1, y_0)$ have no identifying power.

### *Optimizing Treatments*

To derive the worst-case bound (3.3), we constructed two extreme treatment policies, one minimizing $P(y_m \in B \mid x)$ and the other maximizing it. The former policy satisfies equation (3.5), while the latter satisfies (3.6). Suppose that one of these optimizing policies is implemented. What can be learned about $P(y_m \in B \mid x)$ in the absence of prior restrictions on the outcome distribution?

The derivation of (3.3) showed that the treatment policy minimizing $P(y_m \in B \mid x)$ makes $P(y_m \in B \mid x) = P(y_1 \in B \cap y_0 \in B \mid x)$, while the policy maximizing $P(y_m \in B \mid x)$ makes $P(y_m \in B \mid x) = P(y_1 \in B \cup y_0 \in B \mid x)$. Applying the Frechet bound (3.8) shows that when the treatment policy is known to minimize $P(y_m \in B \mid x)$, we obtain the bound

$$(3.23)\quad \max[0,\ P(y_1 \in B \mid x) + P(y_0 \in B \mid x) - 1] \leq P(y_m \in B \mid x) \leq \min[P(y_1 \in B \mid x),\ P(y_0 \in B \mid x)].$$

When the treatment policy is known to maximize $P(y_m \in B \mid x)$, we obtain the bound

$$(3.24)\quad \max[P(y_1 \in B \mid x),\ P(y_0 \in B \mid x)] \leq P(y_m \in B \mid x) \leq \min[P(y_1 \in B \mid x) + P(y_0 \in B \mid x),\ 1].$$

It is interesting to compare these bounds with those under other assumptions. In (3.23), the lower bound coincides with the lower bound in the absence of prior information, while the upper bound coincides with the lower bound under the assumption that treatment is independent of the outcomes. In (3.24), the lower bound coincides with the upper bound under the assumption that treatment is independent of the outcomes, while the upper bound coincides with the upper bound in the absence of prior information. Thus the three treat-

ment policies examined here imply that $P(y_m \in B \mid x)$ lies in mutually exclusive intervals, and these three intervals partition the range of values that is feasible in the absence of prior information.

The idea of optimizing treatments has important applications in economics and in survival analysis. Economic analyses of voluntary treatment policies often assume that the treatment yielding the larger outcome is selected, so

$$y_m = \max(y_1, y_0). \tag{3.25}$$

An example is the Roy model of occupation choice, discussed previously in Section 2.6. For any $t$, treatment policy (3.25) makes $P(y_m \leq t \mid x) = P(y_1 \leq t \cap y_0 \leq t \mid x)$. So this policy minimizes $P(y_m \leq t \mid x)$. We may therefore apply (3.23) to show that

$$\begin{aligned} &\max[0, P(y_1 \leq t \mid x) + P(y_0 \leq t \mid x) - 1] \\ &\quad \leq P(y_m \leq t \mid x) \leq \min[P(y_1 \leq t \mid x), P(y_0 \leq t \mid x)]. \end{aligned} \tag{3.26}$$

The competing-risks model of survival analysis assumes that the treatment yielding the smaller outcome is selected, so

$$y_m = \min(y_1, y_0). \tag{3.27}$$

For any $t$, this treatment policy maximizes $P(y_m \leq t \mid x)$. So (3.24) shows that

$$\begin{aligned} &\max[P(y_1 \leq t \mid x), P(y_0 \leq t \mid x)] \leq P(y_m \leq t \mid x) \\ &\quad \leq \min[P(y_1 \leq t \mid x) + P(y_0 \leq t \mid x), 1]. \end{aligned} \tag{3.28}$$

*Known Treatment Distribution*

The restrictions on treatment policy examined so far in this section specify the rule used to make treatment assignments, but do not con-

strain the fraction of the population receiving each treatment. It is also of interest to consider the reverse situation, where one knows the fraction receiving each treatment but does not know the rule used to make treatment assignments. For example, we noted earlier that resource constraints could limit implementation of the Perry Preschool treatment to part of the eligible population. Knowledge of the budget constraint and the cost of preschool would suffice to determine the fraction of the population receiving the treatment. It may be more difficult to learn how school officials, social workers, and parents interact to determine which children receive the treatment.

Thus suppose that under policy $m$, a known fraction $p$ of the persons with covariates $x$ receive treatment 0 and the remaining fraction $1 - p$ receive treatment 1. So

$$P(z_m = 0 \mid x) = p, \tag{3.29}$$

where $p$ is known. No information is available on the rule used to make treatment assignments that satisfy (3.29).

Given (3.29), $P(y_m \mid x)$ may be written

$$P(y_m \mid x) = P(y_1 \mid x, z_m = 1)(1 - p) + P(y_0 \mid x, z_m = 0)p. \tag{3.30}$$

The distributions $[P(y_1 \mid x), P(y_0 \mid x)]$ may be written

$$P(y_1 \mid x) = P(y_1 \mid x, z_m = 1)(1 - p) + P(y_1 \mid x, z_m = 0)p \tag{3.31a}$$

and

$$P(y_0 \mid x) = P(y_0 \mid x, z_m = 1)(1 - p) + P(y_0 \mid x, z_m = 0)p. \tag{3.31b}$$

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 (3.31a); 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 (3.31b). Through examination of the feasible pairs, it can be shown that $P(y_m \in B \mid x)$ satisfies the following sharp bound (see Manski, 1994b):

$$\begin{aligned}(3.32)\quad \max[0, P(y_1 \in B \mid x) - p] + \max[0, P(y_0 \in B \mid x) - (1 - p)] \\ \leq P(y_m \in B \mid x) \\ \leq \min[1 - p, P(y_1 \in B \mid x)] \\ + \min[p, P(y_0 \in B \mid x)].\end{aligned}$$

## 3.7. Identifying Combinations of Assumptions

Taken one at a time, the assumptions examined in Sections 3.5 and 3.6 improve the worst-case bound of Section 3.4, but are not strong enough to identify the outcome distribution under policy $m$. What assumptions do identify this distribution?

Suppose one combines the assumption that treatment is independent of outcomes with prior knowledge of the fraction of the population receiving each treatment. These two assumptions together imply that

$$(3.33)\quad P(y_m \mid x) = P(y_1 \mid x)(1 - p) + P(y_0 \mid x)p,$$

where $p$ is the known fraction of the population receiving treatment 0. All the quantities on the right side are identified, so $P(y_m \mid x)$ is identified.

Alternatively, suppose that the outcomes $y_1$ and $y_0$ are statistically independent of one another and that the treatment with the larger outcome is always selected. Then

$$\begin{aligned}(3.34)\quad P(y_m \leq t \mid x) &= P[\max(y_1, y_0) \leq t \mid x] \\ &= P(y_1 \leq t \mid x)P(y_0 \leq t \mid x)\end{aligned}$$

is identified for all values of $t$. Thus $P(y_m \mid x)$ is identified.