

Lectures on
IDENTIFICATION FOR PREDICTION AND DECISION

Charles F. Manski

NAKE, 9-10 June, 2008

Source: C. Manski, Identification for Prediction and Decision, Harvard University Press, 2007.

Lectures 1 and 2: Prediction with Incomplete Data
Chapters 1 through 4

Lecture 3: Analysis of Treatment Response
Chapters 7 through 9

Lectures 4 and 5: Planning under Ambiguity
Chapters 11 and 12

THEME: Partial Identification and Credible Inference

Statistical inference uses sample data to draw conclusions about a population probability distribution of interest.

Data alone do not suffice. Informative inference always requires assumptions about the sampling process. It often requires assumptions about the population.

Identification and Statistical Inference

T. Koopmans (*Econometrica*, 1949, p. 132):

“In our discussion we have used the phrase “a parameter that can be determined from a sufficient number of observations.” We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from “a sufficiently large number of observations” we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.”

It has been common to think of identification as a binary event—a parameter is either identified or it is not.

It has been traditional to combine available data with assumptions strong enough to yield *point identification*. However, these assumptions often are not well motivated, and researchers often debate their validity.

Empirical researchers should be concerned with the credibility of inference.

Credibility is a subjective matter, yet I take there to be wide agreement on a principle that I have called

The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.

This principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain. Stronger assumptions yield inferences that may be tighter but less credible.

Methodological research cannot resolve the dilemma but can clarify its nature.

Studies of Partial Identification

Consider inference when a specified sampling process generates data from a specified population.

First study inference on the population probability distribution when no assumptions are placed on this distribution. The usual finding is a set-valued *identification region* (or *identified set*). When this region is smaller than the set of all logically possible distributions, the distribution is *partially identified*.

Then ask how the identification region shrinks if specified assumptions are imposed. This quantifies the *identifying power* of the assumptions.

Researchers often want to learn particular parameters of a probability distribution. Study of identification of the distribution yields findings on the identification of parameters.

I recommend first considering weaker, more credible assumptions and then stronger, less credible ones.

Conservative analysis enables researchers to learn from available data without imposing untenable assumptions. It enables establishment of a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate. It makes plain the limitations of the available data.

Identification for Prediction and Decisions

The basic prediction problem is to learn the probability distribution $P(y|x)$ of an outcome y conditional on covariates x .

In many decision problems, the relative merits of alternative actions depend on an outcome distribution. I study how a decision maker might reasonably choose an action when the available data and credible assumptions only partially identify this distribution. This is decision making under *ambiguity*.

Note: Some research is not concerned with prediction or decision making. Scientists sometimes motivate research as an effort to “understand” a phenomenon or to determine “causality.” I do not do this.

MISSING OUTCOMES

Inference with missing outcomes is a matter of contemplating all values that the missing data might take. The set of feasible outcome distributions is determined by considering all logically possible configurations of the missing data.

Let each member of the population be characterized by a triple (y, z, x) . Here $y \in \mathcal{Y}$ is the outcome to be predicted and $x \in \mathcal{X}$ are observable covariates. The indicator $z = 1$ if y is observable and $z = 0$ otherwise.

A sampling process draws persons at random from the population. For each $i = 1, \dots, \infty$, outcome y_i is observed if $z_i = 1$ and missing if $z_i = 0$.

The objective is to use the available data to learn about $P(y|x)$.

Application: Nonresponse in Survey Research

Nonresponse is a perennial concern in survey research. Some persons are not interviewed and some who are interviewed do not answer some questions. Longitudinal surveys experience attrition.

Application: Non-Observable Counterfactual Outcomes

Analysis of treatment response aims to predict the outcomes that would occur if alternative treatment rules were applied to a population. A fundamental problem is that one cannot observe the outcomes a person would experience under all treatments. At most one can observe the outcome that a person experiences under the treatment he actually receives. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.

Anatomy of the Problem

By the Law of Total Probability

$$P(y|x) = P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x).$$

The sampling process reveals $P(y|x, z = 1)P(z = 1|x)$ and $P(z = 0|x)$.

The sampling process is uninformative regarding $P(y|x, z = 0)$. Hence, $P(y|x)$ lies in the *identification region*

$$H[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y],$$

where Γ_Y denotes the set of all probability distributions on the set Y .

The identification region is informative when $P(z = 0|x) < 1$ and is the single distribution $P(y|x, z = 1)$ when $P(z = 0|x) = 0$. Hence, $P(y|x)$ is *partially identified* when $0 < P(z = 0|x) < 1$ and is *point-identified* when $P(z = 0|x) = 0$.

Identification of Parameters

The above concerns identification of the entire outcome distribution. A common objective is to infer a parameter of this distribution. For example, one may want to learn the conditional mean $E(y|x)$.

Let $\theta(\cdot)$ be a function mapping probability distributions on Y into the real line. Consider inference on the parameter $\theta[P(y|x)]$. The identification region is the set of all values θ can take when $P(y|x)$ ranges over its feasible values. Thus,

$$H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}.$$

Event Probabilities

By the Law of Total Probability,

$$P(y \in B|x) = P(y \in B|x, z = 1)P(z = 1|x) + P(y \in B|x, z = 0)P(z = 0|x).$$

The sampling process reveals $P(z|x)$ and $P(y \in B|x, z = 1)$, but is uninformative about $P(y \in B|x, z = 0)$. This quantity lies between zero and one. This yields the “worst-case” bound on $P(y \in B|x)$:

$$\begin{aligned} P(y \in B|x, z = 1)P(z = 1|x) &\leq P(y \in B|x) \\ &\leq P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x). \end{aligned}$$

The bound is *sharp*. That is, the lower and upper bounds are the smallest and largest feasible values of $P(y \in B|x)$. The width is $P(z = 0|x)$.

The identification region is the interval between the lower and upper bounds.

The Distribution Function

Let y be real-valued. Let $B = (-\infty, t]$ for a specified t . Then the bound is

$$\begin{aligned} P(y \leq t | x, z = 1)P(z = 1 | x) &\leq P(y \leq t | x) \\ &\leq P(y \leq t | x, z = 1)P(z = 1 | x) + P(z = 0 | x). \end{aligned}$$

The feasible distribution functions are all increasing functions of t that lie within the bound for all values of t .

Quantiles

The bound on the distribution function can be inverted to bound quantiles of $P(y | x)$.

This bound is informative on both sides if $P(z = 0 | x) < \min(\alpha, 1 - \alpha)$. Then

$$[\alpha - P(z = 0 | x)]/P(z = 1 | x)\text{-quantile of } P(y | x, z = 1)$$

$$\leq \alpha\text{-quantile of } P(y | x) \leq$$

$$[\alpha/P(z = 1 | x)]\text{-quantile of } P(y | x, z = 1).$$

Means of Functions of y

Consider $E[g(y)]$, where $g(\cdot)$ has range $[g_0, g_1]$.

The Law of Iterated Expectations gives

$$E[g(y)|x] = E[g(y)|x, z = 1]P(z = 1|x) + E[g(y)|x, z = 0]P(z = 0|x).$$

The sampling process reveals $E[g(y)|x, z = 1]P(z|x)$. The data are uninformative about $E[g(y)|x, z = 0]$, which can take any value in the interval $[g_0, g_1]$. Hence, the identification region for $E[g(y)|x]$ is

$$H\{E[g(y)|x]\} = [E[g(y)|x, z = 1]P(z = 1|x) + g_0P(z = 0|x), \\ E[g(y)|x, z = 1]P(z = 1|x) + g_1P(z = 0|x)].$$

This interval has width $(g_1 - g_0)P(z = 0|x)$.

If $g(\cdot)$ is a bounded function, the severity of the identification problem varies directly with the probability of missing data. If $g(\cdot)$ is unbounded, the sampling process per se is uninformative about $E[g(y)|x]$.

Contrast this result with that for $M[g(y)|x]$. There the sampling process is informative when $P(z = 0|x) < \frac{1}{2}$, regardless of whether $g(\cdot)$ is bounded.

Estimation of Identification Regions

The above identification regions are functions of $P(y|x, z = 1)$ and $P(z|x)$. Given sample data, they can be consistently estimated by the sample analogs $P_N(y|x, z = 1)$ and $P_N(z|x)$. Thus, $H[P(y|x)]$ can be consistently estimated by

$$H_N[P(y|x)] \equiv [P_N(y|x, z = 1)P_N(z = 1|x) + \gamma P_N(z = 0|x), \gamma \in \Gamma_Y].$$

Correspondingly, $H\{E[g(y)|x]\}$ can be consistently estimated by

$$H_N\{E[g(y)|x]\} = [E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_0P_N(z = 0|x), \\ E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_1P_N(z = 0|x)].$$

Note: The sample analog of an identification region is a function of the sample data and, hence, is a random set.

Confidence Sets

The statistics literature on point estimation of parameters has used confidence sets to measure sampling imprecision. The standard definition of a confidence set applies to parameters that are partially identified. One can also define confidence sets for identification regions.

Consider $E[g(y)|x]$. A confidence set can be constructed by widening the estimate given above to

$$\begin{aligned} & [E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_0P_N(z = 0|x) - \delta_{0N}, \\ & E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_1P_N(z = 0|x) + \delta_{1N}]. \end{aligned}$$

$\delta_{0N} > 0$ and $\delta_{1N} > 0$ are suitably chosen data-dependent numbers.

DISTRIBUTIONAL ASSUMPTIONS

Inference with no distributional assumptions provides a natural starting point for empirical analysis but ordinarily will not be the ending point. Having determined what can be learned without assumptions, one should then ask what more can be learned if plausible assumptions are imposed.

Missingness at Random

A common practice has been to assume that data are missing at random.

Formally, the assumption is

$$P(y|x, z = 0) = P(y|x, z = 1).$$

It follows that $P(y|x) = P(y|x, z = 1)$.

This assumption is not *refutable* (or *testable*). The available data are uninformative about $P(y|x, z = 0)$. Hence, it is logically possible that this distribution is the same as $P(y|x, z = 1)$.

Refutable and Non-Refutable Assumptions

Assume that $P(y|x, z = 0) \in \Gamma_{0Y}$ for a specified $\Gamma_{0Y} \subset \Gamma_Y$. This assumption is non-refutable. The identification region for $P(y|x)$ is

$$H_0[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{0Y}].$$

Assume that $P(y|x) \in \Gamma_{1Y}$ for a specified $\Gamma_{1Y} \subset \Gamma_Y$. This assumption may be refutable. The data alone reveal that $P(y|x) \in H[P(y|x)]$. Hence, the identification region with the assumption is

$$H_1[P(y|x)] \equiv H[P(y|x)] \cap \Gamma_{1Y}.$$

The assumption is refutable if there exist possible values for $P(y|x, z = 1)$ and $P(z|x)$ such that $H_1[P(y|x)]$ empty. If this set is empty, $P(y|x)$ cannot lie in Γ_{1Y} .

Refutability and Credibility

It is important not to confuse refutability with credibility.

Refutability is a matter of logic. Credibility is a subjective matter.

Refutability is a property of an assumption and the empirical evidence. An assumption is refutable if it is inconsistent with some possible configuration of the empirical evidence. It is non-refutable otherwise.

Credibility is a property of an assumption and the person contemplating it. An assumption is credible to the degree that someone thinks it so.

Wage Regressions

A major problem of missing outcome data in labor economics occurs in efforts to estimate wage regressions, which measure how market wages vary with schooling, work experience, and demographic background.

Surveys provide covariate data for each respondent and wage data for those who work. Surveys do not ordinarily provide wage data for respondents who do not work. Economists consider these wages to be well-defined but unobserved. They are the counterfactual wages that non-workers would earn if they were to work.

The Reservation Wage Model of Labor Supply

Assume each person knows the wage y he would receive if he were to work. The person chooses to work if $y > R$, called the *reservation wage*, and chooses not to work if $y < R$.

Wages are observed when $y > R$ and are missing when $y < R$. Assume that $P(y = R) = 0$. Then the reservation-wage model implies that

$$P(y|x, z = 1) = P(y|x, y > R),$$

$$P(y|x, z = 0) = P(y|x, y < R),$$

The reservation-wage model per se has no identifying power. The model places no assumptions on the distribution $P(y|x, y < R)$. To have identifying power, the model must be augmented by assumptions on the distribution $P(y, R|x)$ of market and reservation wages.

Wages Missing at Random

Wage data are missing at random if $P(y|x, y < R) = P(y|x, y > R)$.

This is logically possible. In particular, it occurs if $R = y + u$, where $u \perp y$.

The question is whether this distributional assumption is credible.

Homogeneous Reservation Wages

Assume that all persons with covariates \mathbf{x} have the same reservation wage $R(\mathbf{x})$. This homogeneous-reservation-wage assumption partially identifies $P(y|\mathbf{x})$.

Let $y^*(\mathbf{x})$ denote the smallest observed wage for persons with covariates \mathbf{x} . Then $y^*(\mathbf{x}) > R(\mathbf{x})$. Hence, for all $t \geq y^*(\mathbf{x})$,

$$P(y \leq t|\mathbf{x}, z = 0) = P[y \leq t|\mathbf{x}, y < R(\mathbf{x})] = 1.$$

It follows that for all such t ,

$$\begin{aligned} P(y \leq t|\mathbf{x}) &= P(y \leq t|\mathbf{x}, z = 1)P(z = 1|\mathbf{x}) + P(y \leq t|\mathbf{x}, z = 0)P(z = 0|\mathbf{x}) \\ &= P(y \leq t|\mathbf{x}, z = 1)P(z = 1|\mathbf{x}) + P(z = 0|\mathbf{x}). \end{aligned}$$

Thus, $P(y \leq t|\mathbf{x})$ is point-identified for $t \geq y^*(\mathbf{x})$.

The α -quantile of $P(y|\mathbf{x})$ is point-identified when $\alpha > P(z = 0|\mathbf{x})$. In particular, $Q_\alpha(y|\mathbf{x}) = Q_a(y|\mathbf{x}, z = 1)$, where $a \equiv [\alpha - P(z = 0|\mathbf{x})]/P(z = 1|\mathbf{x})$.

The question is whether this distributional assumption is credible.

The Normal-Linear Model of Market and Reservation Wages

A common practice has been to restrict $P(y, R|x)$ to a parametric family of distributions. The *normal-linear model* has received considerable attention. This assumes that

$$P(\log y, \log R|x) \sim N[(x\beta_1, x\beta_2), \Sigma].$$

This assumption reduces inference on $P(y|x)$ to inference on $(\beta_1, \beta_2, \Sigma)$.

If the assumption is correct and x has full rank, there exists one value of $(\beta_1, \beta_2, \Sigma)$ that implies the observed $P(y|x, z = 1)$ and $P(z|x)$. Hence, point identification is the norm.

The assumption is refutable. If it is incorrect, typically no value of the parameters generates the observed $P(y|x, z = 1)$ and $P(z|x)$.

The question is whether this distributional assumption is credible.

Selection Models

The normal-linear model of market and reservation wages exemplifies a class of parametric selection models that describe how outcomes and missingness vary with observed and unobserved covariates. The expression “selection model” is used because these models aim to explain missingness as the result of a choice, such as the choice to work in the reservation wage model.

The selection models used in empirical research usually have the form

$$y = \mathbf{x}\mathbf{b} + \delta,$$

$$z = 1[\mathbf{x}\mathbf{c} + \varepsilon > 0],$$

$$P(\delta, \varepsilon | \mathbf{x}) \sim N(0, \Sigma).$$

Parametric Mean Regression with Missing Outcomes

Drop all the assumptions of the parametric selection model except for the mean regression assumption

$$E(y|x) = f(x, b).$$

If y is bounded, this parametric model partially identifies b .

Consider the joint identification region for $[E(y|x = \xi), \xi \in X]$ using the empirical evidence alone. A parameter value $b \in B$ is feasible if and only if the implied values of $[f(\xi, b), \xi \in X]$ lie within this region; that is, if

$$[f(\xi, b), \xi \in X] \in H[E(y|x = \xi), \xi \in X],$$

where
$$H[E(y|x = \xi), \xi \in X] = \times_{\xi \in X} H[E(y|x = \xi)],$$

and

$$H[E(y|x = \xi)] = [E(y|x = \xi, z = 1)P(z = 1|x = \xi) + y_0P(z = 0|x = \xi), \\ E(y|x = \xi, z = 1)P(z = 1|x = \xi) + y_1P(z = 0|x = \xi)].$$

Here y_0 and y_1 are the smallest and largest logically possible values of y .

Thus, b is feasible if and only if

$$\begin{aligned} E(y|x = \xi, z = 1)P(z = 1 | x = \xi) + y_0P(z = 0 | x = \xi) &\leq f(\xi, b) \\ &\leq E(y|x = \xi, z = 1)P(z = 1 | x = \xi) + y_1P(z = 0 | x = \xi) \end{aligned}$$

for all $\xi \in X$.

Let B_0 denote the subset of B for which these inequalities holds. Then B_0 is the identification region for b . It follows that $[f(\xi, b), \xi \in X], b \in B_0$ is the identification region for $[E(y|x = \xi), \xi \in X]$.

The parametric regression model is refutable if it is logically possible for B_0 to be empty. In that event, no value of b yields a function $f(x, b)$ that equals a feasible value of $E(y|x)$. Hence, the model is incorrect.

The set B_0 can be consistently estimated by the *modified-minimum-distance* method.

ANALYSIS OF TREATMENT RESPONSE

Studies of treatment response aim to predict the outcomes that would occur if alternative treatment rules were applied to a population.

One cannot observe the outcomes that a person would experience under all treatments. At most, one can observe a person's *realized outcome*; that is, the one he experiences under the treatment he actually receives. The *counterfactual outcomes* that a person would have experienced under other treatments are logically unobservable.

Example: Suppose that patients ill with a specified disease can be treated by drugs or surgery. The relevant outcome might be life span. One may want to predict the life spans that would occur if all patients of a certain type were to be treated by drugs. The available data may be observations of the realized life spans of patients in a study population, some of whom were treated by drugs and the rest by surgery.

The Selection Problem

Let the set T list all feasible treatments.

Let each member j of a study population have covariates $x_j \in X$.

Let each j have a *response function* $y_j(\cdot): T \rightarrow Y$ that maps the mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Thus, $y_j(t)$ is the outcome that person j would realize if he were to receive treatment t . $y_j(t)$ is a *potential, latent, or conjectural* outcome.

Let $z_j \in T$ denote the treatment received by person j . Then $y_j \equiv y_j(z_j)$ is the realized outcome. The outcomes $[y_j(t), t \neq z_j]$ he would have experienced under other treatments are counterfactual.

Observation may reveal the distribution $P(y, z|x)$ of realized outcomes and treatments for persons with covariates x . Observation cannot reveal the distribution of counterfactual outcomes.

Consider prediction of the outcomes that would occur if all persons with the same observed covariates were to receive the same treatment.

Prediction of outcomes under a policy mandating treatment t for persons with covariates x requires inference on $P[y(t)|x]$.

The problem of identification of $P[y(t)|x]$ from knowledge of $P(y, z|x)$ is called the *selection problem*. This expression refers to the fact that treatment selection determines which potential outcome is observable.

Prediction Using the Empirical Evidence Alone

The selection problem has the same structure as the missing-outcomes problem.

$$\begin{aligned} P[y(t)|x] &= P[y(t)|x, z = t]P(z = t|x) + P[y(t)|x, z \neq t]P(z \neq t|x) \\ &= P(y|x, z = t)P(z = t|x) + P[y(t)|x, z \neq t]P(z \neq t|x). \end{aligned}$$

The first equality is the Law of Total Probability. The second holds because $y(t)$ is the outcome realized by persons who receive treatment t .

The identification region for $P[y(t)|x]$ using the data alone is

$$H\{P[y(t)|x]\} = \{P(y|x, z = t)P(z = t|x) + \gamma P(z \neq t|x), \gamma \in \Gamma_Y\}.$$

This has the same form as the identification region for $P(y|x)$ when outcome data are missing. The outcome there was y and $\{z = 1\}$ indicated observability of y . The outcome here is $y(t)$ and $\{z = t\}$ indicates observability of $y(t)$.

Average Treatment Effects

Researchers often compare policies mandating alternative treatments, say t and t' . It is common to use data on realized treatments and outcomes to infer the *average treatment effect* $E[y(t)|x] - E[y(t')|x]$.

Let Y have smallest and largest elements y_0 and y_1 respectively. The identification region for the average treatment effect is the interval

$$H\{E[y(t)|x] - E[y(t')|x]\} =$$

$$[E(y|x, z = t)P(z = t|x) + y_0P(z \neq t|x) \\ - E(y|x, z = t')P(z = t'|x) - y_1P(z \neq t'|x),$$

$$E(y|x, z = t)P(z = t|x) + y_1P(z \neq t|x) \\ - E(y|x, z = t')P(z = t'|x) - y_0P(z \neq t'|x)].$$

This interval necessarily contains the value zero. Its width is

$$(y_1 - y_0)[P(z \neq t|x) + P(z \neq t'|x)] = (y_1 - y_0)[2 - P(z = t|x) - P(z = t'|x)].$$

Let t and t' be the only feasible treatments in the study population. Then $P(z = t|x) + P(z = t'|x) = 1$. Hence, the interval has width $(y_1 - y_0)$.

Treatment at Random

The analog to the assumption of outcomes missing at random is

$$P[y(\cdot)|x, z = t] = P[y(\cdot)|x, z \neq t].$$

This is credible in classical randomized experiments, where an explicit randomization mechanism has been used to assign treatments and all persons comply with their treatment assignments. Its credibility in other settings is almost invariably a matter of controversy.

Illustration: Sentencing and Recidivism

Consider how the sentencing of offenders may affect recidivism.

Data are available on the outcomes experienced by offenders given the sentences that they receive. However, researchers have long debated the counterfactual outcomes that offenders would experience if they were to receive other sentences. Moreover, the sentencing rules that judges actually use are largely unknown. Thus, predicting the response of criminality to sentencing might reasonably be studied using the empirical evidence alone.

Manski and Nagin (1998) analyzed data on the sentencing and recidivism of males in the state of Utah who were born from 1970 through 1974 and who were convicted of offenses before they reached age 16. We compared recidivism under the two main sentencing options available to judges: confinement in residential facilities ($t = b$) and sentences that do not involve residential confinement ($t = a$). The outcome of interest was taken to be a binary measure of recidivism, with $y = 1$ if an offender is not convicted of a subsequent crime in the two-year period following sentencing, and $y = 0$ if the offender is convicted of a subsequent crime

The data reveal that

Probability of residential treatment: $P(z = b) = 0.11$

Recidivism probability in sub-population receiving residential treatment:

$$P(y = 0 | z = b) = 0.77$$

Recidivism probability in sub-population receiving nonresidential treatment: $P(y = 0 | z = a) = 0.59$.

Consider two policies, one mandating residential treatment for all offenders and the other mandating non-residential treatment. The recidivism probabilities under these policies are $P[y(b) = 0]$ and $P[y(a) = 0]$ respectively.

Assuming treatment at random,

$$P[y(b) = 0] = P(y = 0 | z = b) = 0.77$$

$$P[y(a) = 0] = P(y = 0 | z = a) = 0.59.$$

Using the data alone,

$$H\{P[y(b) = 0]\} = [0.08, 0.97] \quad H\{P[y(a) = 0]\} = [0.53, 0.64].$$

The identification region for the average treatment effect is

$$H\{P[y(b) = 0] - P[y(a) = 0]\} = [-0.56, 0.44].$$

Assumptions Linking Outcomes Across Treatments

Thus far, analysis of treatment response is simply an instance of missing outcomes. The problem takes a distinctive character when one poses assumptions that link outcomes across treatments. Then observation of the realized outcome y_j can be informative about the counterfactual outcomes $y_j(t)$, $t \neq z_j$.

The *perfect-foresight optimization* model assumes $y_j(t) \leq y_j$ for all t and j .

Monotone treatment response assumes that treatments are ordered and $y_j(\cdot)$ is monotone in the treatment. Hence,

$$t > z \Rightarrow y_j(t) \geq y_j. \quad t < z \Rightarrow y_j(t) \leq y_j.$$

Homogeneous Linear Response

Let treatments be real-valued. Assume that each response function is linear in t , with slope that is homogeneous across the population. Thus,

$$y_j(t) = \beta t + \varepsilon_j.$$

The homogeneous-linear-response assumption per se reveals nothing about the slope parameter β . For each person j , only (y_j, z_j) are observable. This observable pair satisfies the equation

$$y_j = \beta z_j + \varepsilon_j.$$

Given any conjectured value for β , this equation is satisfied by setting the unobserved value of ε_j equal to $y_j - \beta z_j$. Hence, the assumption and the empirical evidence are uninformative about β .

“The” Instrumental Variable Estimator

Let v be a real covariate, called an *instrumental variable*. Assume that

$$\text{Cov}(v, \varepsilon) = 0, \quad \text{Cov}(v, z) \neq 0.$$

Then

$$\begin{aligned} 0 &= \text{Cov}(v, \varepsilon) = \text{Cov}(v, y - \beta z) \\ &= E[v(y - \beta z)] - E(v)E(y - \beta z) \\ &= E(vy) - \beta E(vz) - E(v)E(y) + \beta E(v)E(z) = \text{Cov}(v, y) - \beta \text{Cov}(v, z). \end{aligned}$$

Hence, β is point-identified, with

$$\beta = \text{Cov}(v, y)/\text{Cov}(v, z).$$

Researchers sometimes call the sample analog of $\text{Cov}(v, y)/\text{Cov}(v, z)$ “the” instrumental variables estimator. This designation has historical foundation but no compelling scientific basis. There are many distributional assumptions using instrumental variables. Each such assumption may be combined with a variety of restrictions on the shapes of response functions.

Randomized Experiments

Many researchers argue that the assumption of treatment-at-random should have primacy among all of the assumptions that might be brought to bear in analysis of treatment response. That is,

$$P[y(\cdot)|x, z] = P[y(\cdot)|x].$$

The rationale for giving this assumption special status is the strong credibility that it enjoys in classical randomized experiments.

The classical argument for experiments with randomly assigned treatments is generally attributed to Fisher (1935) and can be phrased as follows:

Let random samples of persons be drawn from the population of interest and formed into treatment groups. Let all members of a treatment group be assigned the same treatment and suppose that each subject complies with the assigned treatment. Then the distribution of outcomes experienced by the members of a treatment group will be the same (up to random sampling error) as would be observed if the treatment in question were received by all members of the population.

The argument applies both to *controlled experiments*, in which a researcher purposefully randomizes treatment assignments, and to so-called *natural experiments*, in which randomization is a consequence of some process external to the research project. The randomization mechanism is irrelevant. What matters is that randomization makes it credible to assume that z is statistically independent of $y(\cdot)$.

Experiments in Practice

Classical randomized experiments have clear appeal, but they typically cannot be performed in practice.

(1) The classical argument supposes that subjects are drawn at random from the population of interest. Yet participation in experiments ordinarily cannot be mandated in democracies. Hence, experiments in practice usually draw subjects at random from a pool of persons who volunteer to participate. So one learns about treatment response within the population of volunteers rather than within the population of interest.

(2) The argument supposes that all participants in the experiment comply with their assigned treatments. In practice, subjects often do not comply.

(3) The argument supposes that one observes the realized treatments, outcomes, and covariates of all participants in the experiment. In practice, experiments may have missing data. A particularly common problem is missing outcome data when researchers lose contact with participants before their outcomes can be recorded.

Hence, identification problems typically occur in experiments in practice.

PLANNING UNDER AMBIGUITY

An important objective of empirical studies of treatment response is to provide decision makers with information useful in choosing treatments.

Often the decision maker is a planner who must choose treatments for a heterogeneous population. The planner may want to choose treatments whose outcomes maximize the welfare of this population.

Partial Identification and Ambiguity

Partial identification of treatment response may generate ambiguity about the identity of optimal treatment rules.

Economists have usually regarded social planning as an optimization problem, where the planner knows the distribution of treatment response and chooses a treatment rule to maximize social welfare. For example, studies of optimal income taxation assume the planner knows how the tax schedule affects the distribution of labor supply (e. g., Mirrlees, 1971).

An actual planner may have only partial knowledge of the distribution of treatment response. As a consequence, he may not be able to determine an optimal policy. He then faces a problem of planning under ambiguity.

Criteria for Choice Under Ambiguity

Consider a choice set C and a decision maker who must choose an action from this set. The decision maker wants to maximize an objective function $f(\cdot): C \rightarrow \mathbb{R}$.

The decision maker faces an optimization problem if he knows C and $f(\cdot)$. He faces a problem of choice under ambiguity if he knows C but knows only that $f(\cdot) \in F$, where F is a set of possible objective functions.

How should the decision maker behave? Clearly he should not choose a *dominated* action. Action $d \in C$ is dominated if there exists another feasible action, say c , that is at least as good as d for all objective functions in F and strictly better for some function in F .

How should he choose among the undominated actions? One idea is to average the elements of F and maximize the resulting function. This yields Bayes rules. Another is to choose an action that, in some sense, works uniformly well over all elements of F . This yields the maximin and minimax-regret criteria.

A Static Planning Problem with Individualistic Treatment

There are two treatments, labeled a and b. Let $T \equiv \{a, b\}$.

Each member j of population J has a response function $y_j(\cdot): T \rightarrow Y$ that maps treatments $t \in T$ into outcomes $y_j(t) \in Y$.

$P[y(\cdot)]$ is the population distribution of treatment response. The population is large, in the sense that $P(j) = 0$ for all $j \in J$.

The task is to allocate the population to treatments. An allocation is a number $\delta \in [0, 1]$ that randomly assigns a fraction δ of the population to treatment b and the remaining $1 - \delta$ to treatment a.

Assume that the planner wants to choose an allocation that maximizes mean welfare.

Let $u_j(t) \equiv u_j[y_j(t), t]$ be the net contribution to welfare that occurs if person j receives treatment t and realizes outcome $y_j(t)$.

Let $\alpha \equiv E[u(a)]$ and $\beta \equiv E[u(b)]$. Welfare with allocation δ is

$$W(\delta) = \alpha(1 - \delta) + \beta\delta = \alpha + (\beta - \alpha)\delta.$$

Treatment Choice Under Ambiguity

$\delta = 1$ is optimal if $\beta \geq \alpha$ and $\delta = 0$ if $\beta \leq \alpha$. The problem is treatment choice when (α, β) is partially known.

Let S index the feasible states of nature. The planner knows that (α, β) lies in the set $[(\alpha_s, \beta_s), s \in S]$. Assume that this set is bounded. Let

$$\alpha_L \equiv \min_{s \in S} \alpha_s, \quad \beta_L \equiv \min_{s \in S} \beta_s,$$

$$\alpha_U \equiv \max_{s \in S} \alpha_s, \quad \beta_U \equiv \max_{s \in S} \beta_s.$$

The planner faces ambiguity if $\alpha_s > \beta_s$ for some values of s and $\alpha_s < \beta_s$ for other values.

Bayes Rules

A Bayesian planner places a subjective probability distribution π on the states of nature, computes the subjective mean value of social welfare under each treatment allocation, and chooses an allocation that maximizes this subjective mean. Thus, the planner solves the optimization problem

$$\max_{\delta \in [0, 1]} E_{\pi}(\alpha) + [E_{\pi}(\beta) - E_{\pi}(\alpha)]\delta.$$

$E_{\pi}(\alpha) = \int \alpha_s d\pi$ and $E_{\pi}(\beta) = \int \beta_s d\pi$ are the subjective means of α and β .

The Bayes decision assigns everyone to treatment b if $E_{\pi}(\beta) > E_{\pi}(\alpha)$ and everyone to treatment a if $E_{\pi}(\alpha) > E_{\pi}(\beta)$. All treatment allocations are Bayes decisions if $E_{\pi}(\beta) = E_{\pi}(\alpha)$.

Bayesian planning is conceptually straightforward, but it may not be straightforward to form a credible subjective distribution on the states of nature. The allocation chosen by a Bayesian planner depends on the subjective distribution used. The Bayesian paradigm is appealing only when a decision maker is able to form a subjective distribution that really expresses his beliefs.

The Maximin Criterion

To determine the maximin allocation, one first computes the minimum welfare attained by each allocation across all states of nature. One then chooses an allocation that maximizes this minimum welfare. Thus, the criterion is

$$\max_{\delta \in [0, 1]} \min_{s \in S} \alpha_s + (\beta_s - \alpha_s)\delta.$$

The solution has a simple form if (α_L, β_L) is a feasible value of (α, β) . Then the maximin allocation is $\delta = 0$ if $\alpha_L > \beta_L$ and $\delta = 1$ if $\alpha_L < \beta_L$.

The Minimax-Regret Criterion

The regret of allocation δ in state of nature s is the difference between the maximum achievable welfare and the welfare achieved with allocation δ .

Maximum welfare in state of nature s is $\max(\alpha_s, \beta_s)$. Hence, the minimax-regret criterion is

$$\min_{\delta \in [0, 1]} \max_{s \in S} \max(\alpha_s, \beta_s) - [\alpha_s + (\beta_s - \alpha_s)\delta].$$

Let $S(a) \equiv \{s \in S: \alpha_s > \beta_s\}$ and $S(b) \equiv \{s \in S: \beta_s > \alpha_s\}$.

Let $M(a) \equiv \max_{s \in S(a)} (\alpha_s - \beta_s)$ and $M(b) \equiv \max_{s \in S(b)} (\beta_s - \alpha_s)$. Then

$$\delta_{MR} = \frac{M(b)}{M(a) + M(b)}.$$

Special Case: Let (α_L, β_U) and (α_U, β_L) be feasible values of (α, β) . Then

$$\delta_{MR} = \frac{\beta_U - \alpha_L}{(\alpha_U - \beta_L) + (\beta_U - \alpha_L)}.$$

Proof: The maximum regret of rule δ on all of S is $\max [R(\delta, a), R(\delta, b)]$, where

$$R(\delta, a) \equiv \max_{s \in S(a)} \alpha_s - [(1 - \delta)\alpha_s + \delta\beta_s] = \max_{s \in S(a)} \delta(\alpha_s - \beta_s) = \delta M(a),$$

$$R(\delta, b) \equiv \max_{s \in S(b)} \beta_s - [(1 - \delta)\alpha_s + \delta\beta_s] = \max_{s \in S(b)} (1 - \delta)(\beta_s - \alpha_s) = (1 - \delta)M(b),$$

are maximum regret on $S(a)$ and $S(b)$. Both treatments are undominated, so $R(1, a) = M(a) > 0$ and $R(0, b) = M(b) > 0$. As δ increases from 0 to 1, $R(\cdot, a)$ increases linearly from 0 to $M(a)$ and $R(\cdot, b)$ decreases linearly from $M(b)$ to 0. Hence, the MR rule is the unique $\delta \in (0, 1)$ such that $R(\delta, a) = R(\delta, b)$. This yields the result.

In contrast to Bayes decisions and the maximin rule, the minimax-regret rule is *fractional*.

Illustration: Sentencing Juvenile Offenders in Utah

Consider a Utah judge who observes the treatments and outcomes of the study population and who must choose sentences for a new cohort of convicted offenders. The judge believes that the study population and the new cohort have the same distribution of treatment response. The judge does not feel that any other assumptions are credible.

This is an instance of the two-treatment problem. The distribution of realized treatments and outcomes in the study population is

$$P(z = b) = 0.11 \quad P(y = 1 | z = b) = 0.23 \quad P(y = 1 | z = a) = 0.41.$$

A Bayes rule sets $\delta = 1$ if $0.03 + (0.89)q(b) > 0.36 + (0.11)q(a)$ and $\delta = 0$ if the inequality is reversed.

The maximin rule sets $\delta = 0$.

The minimax-regret rule sets $\delta = 0.55$.

The Ethics of Fractional Treatment Rules

Research on social planning has usually considered only singleton rules, which treat observationally identical people identically. This is inconsequential in many cases where the planner knows the distribution of treatment response. Then there often exists an optimal singleton rule.

Considering only singleton rules is consequential in settings with partial knowledge of treatment response. The minimax-regret rule is always fractional when there are two undominated treatments.

Implementation of a fractional minimax-regret rule enables society to diversify a risk that is privately indivisible. An individual cannot diversify; a person receives either treatment a or b. Yet society can diversify by having positive fractions of the population receive each treatment.

A possible ethical objection is that fractional rules violate the ethical principle calling for “equal treatment of equals.” Fractional rules are consistent with this principle in the *ex ante* sense that all observationally identical people have the same probability of receiving a particular treatment. Fractional rules violate the principle in the *ex post* sense that observationally identical persons ultimately receive different treatments.

Choosing Treatments for X-Pox

Suppose that a new viral disease, x-pox, is sweeping the world. Medical researchers have proposed two mutually exclusive treatments, a and b, which reflect alternative hypotheses, H_a and H_b , about the nature of the virus. If H_t is correct, all persons who receive treatment t survive and all others die. It is known that one of the two hypotheses is correct, but it is not known which; thus, there are two states of nature, $\gamma = H_a$ and $\gamma = H_b$. The objective is to maximize the survival rate of the population.

There are two singleton rules in this setting, one giving treatment a to the entire population and the other giving b. Each rule provides equal treatment of equals in the ex post sense. Each also equalizes realized outcomes. The entire population either survives or dies.

Consider the rule in which a fraction $\delta \in [0, 1]$ of the population receives treatment b and the remaining $1 - \delta$ receives treatment a. Under this rule, the fraction who survive is

$$\delta \cdot 1[\gamma = H_b] + (1 - \delta) \cdot 1[\gamma = H_a].$$

The maximin and minimax-regret rules both set $\delta = 1/2$. These rules treat everyone equally ex ante, each person having a 50 percent chance of receiving each treatment. They do not treat people equally ex post. Nor do they equalize outcomes. Half the population lives and half dies.

PLANNING WITH SAMPLE DATA

Statistical Induction

The ambiguity in treatment choice studied thus far arose purely out of identification problems. In practice, a planner may observe only a random sample of the study population. This generates further ambiguity.

The standard practice is to estimate point-identified population features by their sample analogs. Econometricians appeal to asymptotic theory to justify the practice.

Asymptotic theory gives at most approximate guidance to a planner who must make treatment choices using sample data. The limit theorems of asymptotic theory describe the behavior of estimates as sample size increases to infinity. They do not reveal how estimates perform in specific finite-sample settings. A planner's objective is not to obtain estimates with good asymptotic properties but rather to choose a good treatment rule with the data available.

The Wald (1950) development of statistical decision theory addresses treatment choice with sample data directly, without recourse to asymptotic approximations. It seamlessly integrates the study of identification and statistical inference.

Wald developed the principles of statistical decision theory in the 1930s and 1940s. Important extensions and applications followed in the 1950s, but this period of rapid development came to a close by the 1960s.

Why did statistical decision theory lose momentum long ago? One reason may have been the technical difficulty of the subject. Wald's ideas are fairly easy to describe in the abstract, but applying them tends to be analytically and numerically demanding.

Another reason may have been diminishing interest in decision making as the motivation for statistical analysis. Modern statisticians and econometricians tend to view their objectives as estimation and hypothesis testing rather than decision making.

A third contributing factor may have been the criticism of Wald's thinking put forward by those decision theorists who espouse the *conditionality principle* as a sine qua non of statistical decision making.

Wald's Development of Statistical Decision Theory

Wald considered the broad problem of using sample data to make a decision. His world view eliminates the common separation of activities between empirical research and decision making. Thus, the researcher and the decision maker are one and the same person.

Wald posed the task as choice of a *statistical decision function*, which maps the available data into a choice among the feasible actions. In a treatment-choice setting, a statistical decision function is a rule for using the data to choose a treatment allocation. I call such a rule a *statistical treatment rule*.

No statistical decision function that makes non-trivial use of sample data can perform best in every realization of a sampling process. Hence, Wald recommended evaluation of statistical decision functions as *procedures* applied as the sampling process is engaged repeatedly to draw independent data samples.

The idea of a procedure transforms the original statistical problem of induction from a single sample into the deductive problem of assessing the probabilistic performance of a statistical decision function across realizations of the sampling process.

Admissibility

Wald suggested comparison of statistical decision functions by their mean performance across realizations of the sampling process. Wald termed this criterion *risk*. Here, where the goal is maximization of a social welfare function, I call it *expected welfare*.

The most basic prescription of Wald's statistical decision theory is that a decision maker should not choose an action that is dominated in risk. Such an action is called *inadmissible*. An action that is not dominated in risk is called *admissible*.

In the treatment choice setting, a statistical treatment rule is inadmissible if there exists another feasible rule that yields at least the same expected welfare in all feasible states of nature and larger expected welfare in some state of nature.

Admissibility is a desirable property, but its operational implications are limited for two main reasons. First, when it is possible to determine which statistical decision functions are admissible, there often turn out to be many such functions. Second, there are many settings of practical interest where analysis of admissibility is technically challenging.

Implementable Criteria for Treatment Choice

To develop implementable criteria for decision making with sample data, statistical decision theorists have studied the same broad ideas that were discussed earlier, but now applied to risk. The Bayesian prescription uses a statistical decision function that works well on average across the feasible states of nature. The maximin and minimax-regret prescriptions use a decision function that, in one of two senses, works well uniformly over Γ .

Using a Randomized Experiment to Evaluate an Innovation

Outcomes are binary, there are no observed covariates, and there are two treatments, one being the status quo and the other being an innovation. The planner knows the response distribution of the status quo treatment, but not that of the innovation. To learn about the innovation, a classical randomized experiment is performed. The problem is to use the experimental data to inform treatment choice.

The Setting

Let $t = a$ be the status quo treatment and $t = b$ be the innovation. The planner knows the success probability $\alpha \equiv P[y(a) = 1]$ of the status quo treatment but does not know the success probability $\beta \equiv P[y(b) = 1]$ of the innovation. The planner wants to choose treatments to maximize the success probability.

An experiment is performed to learn about outcomes under the innovation, with N subjects randomly drawn from the population and assigned to treatment b . There is full compliance with the assigned treatment. n subjects realize $y = 1$ and the remaining $N - n$ realize $y = 0$. These outcomes are observed.

This classical experiment point-identifies β . The planner only faces a problem of statistical inference.

The sample size N indexes the sampling process and the number n of experimental successes is a sufficient statistic for the sample data. The feasible statistical treatment rules are functions $\delta(\cdot): [0, \dots, N] \rightarrow [0, 1]$ that map the number of experimental successes into a treatment allocation. For each value of n , rule $\delta(\cdot)$ randomly allocates a fraction $\delta(n)$ of the population to treatment b and the remaining $1 - \delta(n)$ to treatment a .

The expected welfare of rule δ is

$$W(\delta, P, N) = \alpha \cdot E[1 - \delta(n)] + \beta \cdot E[\delta(n)] = \alpha + (\beta - \alpha) \cdot E[\delta(n)].$$

The number of experimental successes is distributed binomial $\mathbf{B}[\beta, N]$, so

$$E[\delta(n)] = \sum_{i=0}^N \delta(i) \cdot f(n=i; \beta, N),$$

where $f(n=i; \beta, N) \equiv N!/[i! \cdot (N-i)!] \beta^i (1-\beta)^{N-i}$ is the Binomial probability of i successes.

The only unknown determinant of expected welfare is β . Hence, Γ indexes the feasible values of β . Let $\beta_\gamma \equiv P_\gamma[y(b) = 1]$. Let $(\beta_\gamma, \gamma \in \Gamma)$ contain values that are smaller and larger than α ; otherwise, the choice problem is trivial.

The Admissible Treatment Rules

It is reasonable to conjecture that admissible treatment rules should be ones in which the fraction of the population allocated to treatment b increases with n . The admissible treatment rules are a simple subclass of these rules.

Karlin and Rubin (1956) define a *monotone treatment rule* to be one of the form

$$\delta(n) = 0 \quad \text{for } n < n_0,$$

$$\delta(n) = \lambda \quad \text{for } n = n_0,$$

$$\delta(n) = 1 \quad \text{for } n > n_0,$$

where $0 \leq n_0 \leq N$ and $0 \leq \lambda \leq 1$ are constants specified by the planner.

Let $0 < \alpha < 1$ and let the feasible set $(\beta_\gamma, \gamma \in \Gamma)$ exclude the values 0 and 1. Karlin and Rubin (1956, Theorem 4) shows that the collection of monotone treatment rules is the set of admissible rules.

Some Monotone Rules

The collection of monotone treatment rules is a mathematically “small” subset of the space of all feasible treatment rules, but it still contains a broad range of rules. Here are some of them.

Data-Invariant Rules: These are the rules $\delta(\cdot) = 0$ and $\delta(\cdot) = 1$, which assign all persons to treatment a or b respectively, whatever the realization of n may be.

Empirical Success Rules: An optimal treatment rule allocates all persons to treatment a if $\beta < \alpha$ and all to treatment b if $\beta > \alpha$. An empirical success rule emulates the optimal rule by replacing β with its sample analog, the empirical success rate n/N . Thus, an empirical success rule has the form

$$\begin{aligned}\delta(n) &= 0 && \text{for } n < \alpha N, \\ \delta(n) &= \lambda && \text{for } n = \alpha N, \quad \text{where } 0 \leq \lambda \leq 1, \\ \delta(n) &= 1 && \text{for } n > \alpha N.\end{aligned}$$

Bayes Rules: The form of the Bayes rule depends on the prior subjective distribution placed on β . Consider the class of Beta priors, which form the conjugate family for a Binomial likelihood. Let $(\beta_\gamma, \gamma \in \Gamma) = (0, 1)$ and let the prior be Beta with parameters (c, d) . Then the posterior mean for β is $(c + n)/(c + d + N)$. The resulting Bayes rule is

$$\delta(n) = 0 \quad \text{for } (c + n)/(c + d + N) < \alpha,$$

$$\delta(n) = \lambda \quad \text{for } (c + n)/(c + d + N) = \alpha, \quad \text{where } 0 \leq \lambda \leq 1,$$

$$\delta(n) = 1 \quad \text{for } (c + n)/(c + d + N) > \alpha.$$

As (c, d) tend to zero, the Bayes rule approaches an empirical success rule. The class of Bayes rules includes the data-invariant rules $\delta(\cdot) = 0$ and $\delta(b, \cdot) = 1$. The former occurs if the parameters (c, d) of the Beta prior distribution satisfy $(c + N)/(c + d + N) < \alpha$. The latter occurs if $c/(c + d + N) > \alpha$.

Statistical Significance Rules: These rules use a one-sided hypothesis test to choose between the status quo treatment and the innovation. The null hypothesis is that both treatments yield the same social welfare; that is, $\beta = \alpha$. The alternative is that treatment b is superior to treatment a; that is, $\beta > \alpha$. Treatment b is chosen if the null is rejected, and treatment a otherwise. Thus, the rule is

$$\begin{aligned}\delta(n) &= 0 && \text{for } n \leq d(\alpha, s, N), \\ \delta(n) &= 1 && \text{for } n > d(\alpha, s, N),\end{aligned}$$

where s is the specified size of the test and $d(\alpha, s, N)$ is the associated critical value. With n is binomial, $d(\alpha, s, N) = \min i: f(n > i; \alpha, N) \leq s$.

Although statistical significance rules are monotone treatment rules, the conventional practice of hypothesis testing is remote from the problem of treatment choice with sample data. If the null hypothesis [$\beta = \alpha$] is correct, all feasible treatment rules yield the same expected welfare. If not, alternative rules may yield different expected welfare. A statistical test indicates only whether the sample data are inconsistent (in the usual sense of having low probability of being realized under the null) with the hypothesis that all feasible rules yield the same expected welfare.

The Maximin Rule: The minimum expected welfare for rule δ is

$$\min_{\gamma \in \Gamma} W(\delta, P_\gamma, N) = \alpha + \min_{\gamma \in \Gamma} (\beta_\gamma - \alpha) E_\gamma[\delta(n)],$$

By assumption, $(\beta_\gamma, \gamma \in \Gamma)$ contains values that are smaller than α . Moreover, $E_\gamma[\delta(n)] > 0$ for all $\beta_\gamma > 0$ and all monotone treatment rules except for $\delta(\cdot) = 0$, the rule that always chooses treatment a. Hence, the maximin rule is the data-invariant rule $\delta(\cdot) = 0$.

The Minimax-Regret Rule: The regret of rule δ in state of nature γ is

$$\begin{aligned} & \max(\alpha, \beta_\gamma) - \{\alpha + (\beta_\gamma - \alpha) \cdot E_\gamma[\delta(n)]\} \\ &= (\beta_\gamma - \alpha) \{1 - E_\gamma[\delta(n)]\} \cdot 1[\beta_\gamma \geq \alpha] + (\alpha - \beta_\gamma) E_\gamma[\delta(n)] \cdot 1[\alpha \geq \beta_\gamma]. \end{aligned}$$

Thus, regret is the mean welfare loss when a member of the population is assigned the inferior treatment, multiplied by the expected fraction of the population assigned this treatment.

The minimax-regret rule does not have an analytical solution but it can be determined numerically.

Savage on the Maximin and Minimax-Regret Criteria

Treatment choice using the minimax-regret rule differs fundamentally from treatment choice using the maximin rule. Savage (1951), whose review of Wald (1950) first explicitly distinguished between these criteria for decision making, argued strongly against application of the minimax (here maximin) criterion, writing (p. 63):

Application of the minimax rule is indeed ultra-pessimistic; no serious justification for it has ever been suggested, and it can lead to the absurd conclusion in some cases that no amount of relevant experimentation should deter the actor from behaving as though he were in complete ignorance.

Our finding that the maximin treatment rule is data-invariant illustrates this “absurd conclusion.” Savage emphasized that although the minimax criterion is “ultra-pessimistic,” the minimax-regret criterion is not. Our finding that the minimax-regret rule approximates the empirical success rule illustrates that the minimax-regret criterion is not particularly pessimistic.