MORE DATA OR BETTER DATA? A Statistical Decision Problem

Jeff Dominitz
Resolution Economics

and

Charles F. Manski
Department of Economics and Institute for Policy Research, Northwestern University

Revised: October 2016; forthcoming in the *Review of Economic Studies*

Abstract

When designing data collection, crucial questions arise regarding how much data to collect and how much effort to expend to enhance the quality of the collected data. To make choice of sample design a coherent subject of study, it is desirable to specify an explicit decision problem. We use the Wald framework of statistical decision theory to study allocation of a budget between two or more sampling processes. These processes all draw random samples from a population of interest and aim to collect data that are informative about the sample realizations of an outcome. They differ in the cost of data collection and the quality of the data obtained. One may incur lower cost per sample member but yield lower data quality than another. Increasing the allocation of budget to a low-cost process yields more data, while increasing the allocation to a high-cost process yields better data. We initially view the concept of "better data" abstractly and then fix attention on two important cases. In both cases, a high-cost sampling process accurately measures the outcome of each sample member. The cases differ in the data yielded by a low-cost process. In one, the low-cost process has nonresponse and in the other it provides a low-resolution interval measure of each sample member's outcome. In these settings, we study minimax-regret sample design for prediction of a real-valued outcome under square loss; that is, design which minimizes maximum mean square error. The analysis imposes no assumptions that restrict the unobserved outcomes. Hence, the decision maker must cope with both the statistical imprecision of finite samples and partial identification of the true state of nature.

Manski's research was supported in part by National Science Foundation grant SES-1129475. We are grateful to Bruce Spencer for useful discussions and to Max Tabord-Meehan and three anonymous reviewers for constructive comments. We have benefitted from the opportunity to present this work in a seminar at the University of Southern California and a Cemmap conference at University College London.

1. Introduction

When designing data collection, crucial questions arise regarding how much data to collect and how much effort to expend to enhance the quality of the collected data. Suppose, for example, that an agency is designing a new survey of households. The agency must choose how many cases to sample from the population, what questions to ask sampled households, and how intensively to seek responses. The optimal choice depends on the costs of alternative sampling methods, the sensitivity of data quality to alternative methods, and the implications of sample size and data quality for the value of the collected data.

Cochran, Mosteller, and Tukey (1954) provided a notable early example of the design problem. Their report assessing the statistical methodology of the Kinsey study of male sexual behavior reached a strong conclusion regarding the benefits of increased sample size versus increased rates of response when the objective is to estimate the population mean of an outcome. They wrote (p. 282): "Very much greater expenditure of time and money is warranted to obtain an interview from one refusal than to obtain an interview from a new subject."

To reach this conclusion, they considered the mean square error (MSE) of an estimate of the mean. They recognized that, in the absence of knowledge of the process generating nonresponse, obtaining an interview from a new randomly drawn subject only reduces variance but obtaining an interview from a non-respondent sample member reduces both variance and maximum potential bias. They compared the reductions in maximum MSE that can be achieved by (i) increasing sample size while holding the response rate fixed and (ii) increasing the response rate while holding the sample size fixed. Horowitz and Manski (1998, Section 6) and Tetenov (2012) provide further analysis of this question, using the modern framework of partial identification analysis. Coming at the problem from a different perspective, Philipson (1997) studies choice of a schedule of participation incentives to survey respondents that aims to minimize survey cost subject to achievement of a specified sample size and response rate.

The tradeoff between data quantity and quality extends well beyond choice of sample size and

response rate. For instance, one may desire to conduct in-person interviews to reduce measurement error relative to telephone interviews, but doing so may increase the cost of each interview. Data quality may be enhanced by conducting longer in-depth interviews, but doing so again may increase the cost of each interview. Alternatively, one may seek to enhance data quality by supplementing interviews with auxiliary data collection, such as matching survey responses to administrative records.

Another important tradeoff between data quantity and quality arises in studies of treatment response. A classical randomized experiment with complete compliance, no attrition, and accurate measurement of outcomes can provide data enabling highly credible inference on treatment response. However, performance of such an experiment may be highly costly. Lower cost alternatives providing lower quality data include experimental designs with incomplete compliance, some attrition of subjects, and measurement of surrogate outcomes. They also include analysis of observational data collected in settings with self-selected treatments.

To make choice of sample design a coherent subject of study, it is desirable to specify an underlying decision problem that makes data collection potentially informative. The Wald (1950) development of statistical decision theory provides a suitable analytical framework, which jointly considers sample design and how the resulting data will be used.

Wald considered the broad problem of using sample data to make decisions under uncertainty. He posed the task as choice of a sample design and of a statistical decision function, which maps potential data into a choice among the feasible actions. He recommended ex ante evaluation of sample designs and statistical decision functions as procedures, specifying how a decision maker (aka planner) would use whatever data may be realized. Expressing the objective as minimization of a loss function, he proposed that the planner evaluate a design and decision function by the distribution of loss that they yield across realizations of the sampling process. Wald focused attention on mean sampling performance, which he termed risk, and considered use of the minimax decision criterion, which minimizes maximum risk across the feasible states of nature. Researchers have also studied other criteria including minimax regret and

minimization of a subjective mean of the risk function (Bayes risk). See Ferguson (1967) and Berger (1985) for comprehensive expositions of statistical decision theory. Spencer (1985) gives a general discussion of choice of data quality as a decision problem.

We use the Wald framework to study a relatively simple yet subtle class of sample design problems. We consider allocation of a data budget between two or more sampling processes. These processes all draw random samples from a population of interest and aim to collect data that are informative about the sample realizations of a real-valued outcome. They differ in their cost of data collection and the quality of the data obtained. One process may incur lower cost per sample member but yield lower data quality than another. Increasing the allocation of budget to a low-cost process yields more data, while increasing the allocation to a high-cost process yields better data. The principles of our analysis apply to allocation of a budget to data collection by any number of sampling processes. We mainly focus on a setting with two available processes for expositional simplicity.

We initially view the concept of "better data" abstractly and then fix attention on two cases of practical importance. In both cases, the high-cost sampling process accurately measures the outcome of each sample member. The cases differ in the data yielded by the low-cost process. In one case, the low-cost process has nonresponse: it accurately measures the outcomes of some sample members but yields no data for the remaining members. In the second case, the low-cost process provides a low-resolution interval measure of each sample member's outcome.

In much of the analysis, we assume that the budget is predetermined, making our work a study of the cost effectiveness of alternative budget allocations. When the budget is not predetermined, we show how to choose a budget sufficient to achieve an ε-optimal design as defined in Manski and Tetenov (2016); that is, a budget sufficient to make maximum regret less than a specified $\varepsilon > 0$. As did Manski and Tetenov, we argue that ε-optimality provides a more appealing criterion for setting budget size than the statistical power criteria that have traditionally been used.

The statistical decision problem we study is best point prediction of a real-valued outcome under a

specified loss function. Section 2 poses this familiar problem in abstraction and considers use of standard decision criteria---minimax, minimax-regret, and minimization of Bayes risk---to jointly choose a sample design and a predictor. The analysis in later sections of the paper uses maximum regret to evaluate alternative design-predictor pairs. Section 2 provide several reasons why we find this criterion appealing.

Application of the various decision criteria is straightforward in principle but not in practice. They yield analytical solutions only in special cases and computation of numerical solutions often is computationally challenging. In addition to deriving results in cases of practical relevance, we sketch general approaches that may be used to make computation tractable, including Monte Carlo evaluation of risk and discretization of state spaces.

Both analysis and computation simplify in the familiar setting of square loss. With this loss function, the risk of a candidate predictor is the sum of the population variance of the outcome and the MSE of the predictor as an estimate of the mean outcome. The regret of a predictor is its MSE as an estimate of the mean. A minimax-regret predictor minimizes maximum mean square error.

Sections 3 and 4 focus on specific tractable settings. Here we study decision making using the minimax-regret criterion under a square loss function in the two cases mentioned above, where the low-cost sampling process has nonresponse (Section 3) or yields low-resolution interval data on outcomes (Section 4). Focusing on square loss simplifies analysis, but it is still computationally challenging to determine the best joint choice of a sample design and a predictor. To reduce the complexity of the decision problem, we study choice of sample design when the planner commits to use certain reasonable and tractable predictors.

Our analysis in Section 3 of the case when low-cost sampling has nonresponse makes no assumption about the distribution of unobserved outcomes; thus, the population mean outcome is partially identified with low-cost sampling (Manski, 1989, 2003). We assume that a planner having only low-cost data computes the midpoint of the sample analog estimate for the identification region for the population mean and uses this as the predictor. We explain why this midpoint predictor is a reasonable choice.

We show that maximum regret has a simple explicit form when the midpoint predictor is used. Hence, we can easily determine how the planner should act when facing a constrained problem of choice between low-cost and high-cost sampling. Our finding on choice between low-cost and high-cost sampling extends immediately to settings in which the planner chooses one among multiple sampling processes, each with a different sampling cost and response rate.

Having examined constrained choice of a single sampling process, we turn to the unconstrained setting in which the planner may allocate budget to both low-cost and high-cost sampling, obtaining some data with each process. It has been common in applied practice to pool data obtained by low-cost and high-cost sampling, disregarding the specific sampling process that yielded each observation. We derive the maximum regret of the midpoint predictor computed with pooled data.. Given a commitment to use this predictor, the budget allocation that minimizes maximum regret may be found by a straightforward numerical calculation. Performing this calculation for alternative budgets yields a budget sufficiently large to achieve ε-optimality for any specified value of ε. Thus, we show how to choose a budget ensuring that there exists a feasible design-predictor pair whose maximum MSE is less than ε.

Pooling data does not use all available information. Hence, it is of interest to consider predictors that recognize the sampling process yielding each observation. There are many heuristically reasonable options, but we have not found any that have simple forms. Non-pooling predictors can be evaluated numerically using the methods described in Section 2. We report some exploratory work of this type, leaving a deeper analysis for future research.

Our analysis in Section 4 of the case when low-cost sampling yields a low-resolution interval measurement of the outcome parallels the first part of the analysis in Section 3. We make no assumption about the distribution of outcomes within the observed intervals except that it is continuous; thus, the population mean outcome with low-cost sampling is partially identified. We again assume that a planner having only low-cost data computes the midpoint of the sample analog estimate for the identification region and uses this as the prediction. Maximum regret again has a simple explicit form when this predictor is

used, so we can easily determine how the planner should act when facing a constrained problem of choice between low-cost and high-cost sampling. Our finding again extends immediately to settings in which the planner chooses one among multiple sampling processes, each with a different sampling cost and resolution of interval measurement.

We do not study the unconstrained interval-measurement setting in which the planner may allocate budget to both low-cost and high-cost sampling, obtaining some data with each process. We have not found a predictor that reasonably uses data from both sampling processes and has a tractable explicit form for maximum regret. One can evaluate maximum regret numerically for specified predictors using the methods described in Section 2. We leave this as a subject for future research.

As far as we are aware, the analysis in Sections 3 and 4 is new. Although Wald's abstract development of statistical decision theory has very broad scope, applications of the theory have focused on settings in which the true state of nature is point identified, making statistical imprecision the only inferential problem. For example, Manski (2004) used the minimax-regret criterion to study treatment choice with data from a classical randomized experiment, while Manski and Tetenov (2016) used it to study choice of sample size in classical experiments. The few previous applications of the theory to settings in which the true state is partially identified have focused on the use of sample data to choose treatments and have not considered sample design; see Manski (2007), Stoye (2012), and Tetenov (2012).

In principle, the survey research literature on total survey error expresses concern with both sampling and non-sampling error. Groves and Lyberg (2010) describe total survey error as "as an indicator of data quality usually measured by the accuracy or the mean squared error (MSE) of the estimate" (p. 850). Non-sampling error generates identification problems, so the literature should be jointly concerned with statistical imprecision and partial identification. However, in practice the focus has been on statistical imprecision. In their historical synthesis and critique of research on the subject, Groves and Lyberg offer an explanation for this state of affairs, stating (p. 868): "The total survey error format forces attention to both variance and bias terms. . . . . Most statistical attention to surveys is on the variance terms--largely,

we suspect, because that is where statistical estimation tools are best found." See Manski (2015) for further discussion.

A U. S. Census Bureau report evaluating a possible change to the administration of the American Community Survey (ACS) provides a striking example of the prevalent focus on variance rather than bias. The report (Griffin, 2011) considers the implications of making participation in the survey voluntary rather than mandatory. The Griffin report discusses the potential impact on survey "reliability" of removing the mandate to participate. It measures reliability entirely by the variance of estimates obtained using the survey data. The report does not discuss bias nor any other measure of non-sampling error.

2. Using Statistical Decision Theory to Choose a Sample Design and Predictor

2.1. Best Point Prediction

Best point prediction of a real outcome under a specified loss function has long been a central concern of statistics. Prediction problems are sometimes used as pedagogical devices to motivate interest in certain features of a probability distribution, most famously the mean and median as the best predictors under square and absolute loss. Alternatively, one can have in mind a planner who actually faces a prediction problem. Contemplating an actual planner is particularly appropriate when one uses statistical decision theory.

Planning a major survey such as the ACS, which is used for many purposes, is far more complex than facing an isolated prediction problem. Although study of the classical problem of best point prediction cannot provide a comprehensive approach to design of the ACS and similar major surveys, we think it an appropriate starting point. We will develop constructive ways for survey planners to evaluate designs by the maximum MSE of the estimates they make possible. We think this a considerable advance

relative to the present practice of focusing on variance without consideration of bias.

2.2. Prediction with Sample Data

Consider a planner who must choose a best predictor of real outcomes in a large population J, formalized as a probability space (J, Ω, P) with $P(j) = 0$, all $j \in J$. The set of feasible predictors is T, a subset of the real line. Each $j \in J$ has an outcome denoted y_j. A loss function $L(y - t)$: $T \rightarrow [0, \infty)$ expresses the loss from choosing predictor $t \in T$ when the outcome is y. The minimal logical structure required of a loss function is that $L(0) = 0$ and $L(y - t) \geq 0$ for $t \neq y_j$.

In this setting, the planner may want to choose a predictor that minimizes mean loss. Mean loss with predictor t is $E[L(y - t)] = \int L(y_j - t)dP(j)$. Thus, the planner wants to solve the problem $\min_{t \in T} E[L(y - t)]$. The planner can solve this problem if he knows P(y), the population distribution of outcomes.

Suppose that the planner does not know P(y). However, he can use a positive, finite budget B to draw persons at random and attempt to measure the outcome of each sampled person, after which he uses the data to choose a point prediction. To operationalize the theme of "more data or better data," we assume that two sampling processes are available, denoted process 1 and 2. These processes incur different marginal costs (c_1, c_2) per sample member, with $0 < c_1 < c_2$. They yield data of different quality, where the term "quality" may refer to response rate and/or data accuracy.

The planner faces a joint problem of sample design and choice of a predictor. The design alternatives are feasible integer values for the sample sizes (N_1, N_2) drawn with each sampling process. If the size B of the budget is predetermined, the feasible designs are (N_1, N_2) such that $0 \leq c_1N_1 + c_2N_2 \leq B$.

Given a design, a predictor maps the realized data into a prediction. Suppose that samples of size (N_1, N_2) yield data $\psi = (\psi_{1k}, k = 1, \ldots, N_1; \psi_{2k}, k = 1, \ldots, N_2)$. Let Ψ be the sample space indexing all possible data realizations. Then a predictor is a function $\delta(\cdot)$: $\Psi \rightarrow T$.

The above description of a sample design and a predictor with a predetermined budget restricts consideration to pure strategies that make (N_1, N_2, δ) deterministic choices. We could also consider mixed strategies that use a randomizing device to choose (N_1, N_2, δ) from a specified probability distribution of sample sizes and predictors. Mixed strategies may be attractive in some settings, but we abstract from them so as not to further complicate an already complex decision problem.

2.2.1. The State-Dependent Risk of a Design-Predictor Pair

Wald's statistical decision theory evaluates each design-predictor pair by its risk, the expected value of mean social cost across potential samples. Let $Q(N_1, N_2)$ be the sampling distribution of the data ψ under design (N_1, N_2). The risk of design-predictor pair $[(N_1, N_2), \delta]$ is

$$(1) \qquad r[(N_1, N_2), \delta] = \int E\{L[y - \delta(\psi)]\} dQ(\psi; N_1, N_2) = \iint L[y - \delta(\psi)] dP(y) dQ(\psi; N_1, N_2).$$

The statement of equation (1) assumes that population outcomes y and data realizations ψ are statistically independent. This assumption, commonly made in prediction analysis, holds in our work because the population is a large atomless probability space and the data a finite random sample from this population.

Evaluation of risk is possible if one knows $P(y)$ and $Q(N_1, N_2)$. Knowledge of $P(y)$ enables evaluation of $E\{L[y - \delta(\psi)]\}$ for each realization of ψ. Knowledge of $Q(N_1, N_2)$ enables evaluation of the expectation of $E\{L[y - \delta(\psi)]\}$ across samples.

We want to study decision making when the planner has incomplete knowledge of these distributions. Let the feasible distributions be $(P_s, Q_s, s \in S]$. S is traditionally called the state space in decision theory and the parameter space in statistics; we use the former terminology. In principle, the planner can compute state-dependent risk

(2) $r_s[(N_1, N_2), \delta] = \int E_s\{L[y - \delta(\psi)]\} dQ_s(\psi; N_1, N_2) = \iint L[y - \delta(\psi)] dP_s(y) dQ_s(\psi; N_1, N_2).$

The basic idea is to use the vector $\{r_s[(N_1, N_2), \delta], s \in S\}$ to evaluate each design-predictor pair.

2.2.2. Choosing a Design-Predictor Pair

Given the above, choice of a design-predictor pair has two stages. The first eliminates inadmissible (weakly dominated) options. Pair $[(N_1, N_2), \delta]$ is inadmissible if there exists another pair $[(N_1, N_2), \delta]'$ such that $r_s[(N_1, N_2), \delta] \geq r_s[(N_1, N_2), \delta]'$ for all $s \in S$ and $r_s[(N_1, N_2), \delta] > r_s[(N_1, N_2), \delta]'$ for some $s \in S$.

Let D denote the set of admissible design-predictor pairs. The second stage in decision making is to use some criterion to choose among D. Statistical decision theory gives no consensus prescription, but it suggests various criteria that authors deem “reasonable.” The term “reasonable” seems appropriate because there is no uniquely correct way to choose among admissible rules. Wald (1950), who studied the minimax rule in abstraction, wrote (p. 18): “a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution . . . . does not exist or is unknown to the experimenter.” Ferguson (1967) wrote (p. 29): "A *reasonable* rule is one that is better than just guessing."

Before posing particular decision criteria, we think it prudent to observe that the strength of statistical decision theory is also its vulnerability. The strength of the theory is that it requires one to take an explicit stand on the decision problem to be addressed with sample data and, in return, delivers specific conclusions about what constitutes a good sample design and decision rule. The vulnerability is that findings obtained with a particular decision criterion may not satisfy persons who would rather use a different criterion. Such persons must analyze the design-prediction problem afresh using their preferred criteria. Some may view the dependence of findings on the selected decision criterion to be a deficiency, but we think it a virtue. Statistical decision theory faces up to the reality that one cannot pose and study a well-defined optimization problem without taking a stand on what one wants to optimize.

Leading decision criteria are minimization of Bayes risk (the expectation of risk with respect to a subjective distribution φ on S), minimax, and minimax regret. The quantities to be minimized across D are

Bayes Risk: $\int r_s[(N_1, N_2), \delta] d\varphi(s)$,

Maximum Risk: $\underset{s \in S}{\text{Max}}\, r_s[(N_1, N_2), \delta]$,

Maximum Regret: $\underset{s \in S}{\text{Max}} \left\{ r_s[(N_1, N_2), \delta] - \underset{[(N_1, N_2), \delta]' \in D}{\text{Min}} r_s[(N_1, N_2), \delta]' \right\}$.

It often is difficult to determine the set of admissible options. Given this, researchers applying the Wald theory commonly skip the step of determining admissibility and use a decision criterion to choose among all feasible options, not just those that are admissible. When any of the criteria listed above yields a unique choice, it necessarily is admissible. When a criterion yields a set of equally good choices, the set may include inadmissible options that are strictly dominated only in states that do not affect the value of the optimum. Bayes risk is unaffected by values of risk that occur off the φ-support of S. Maximum risk/regret is unaffected by dominance in states that do not determine the maximum.

2.3. Computation

Although use of statistical decision theory to choose a sample design and predictor is simple in principle, it can be very difficult in practice. Implementation of any of the decision criteria described above requires evaluation of risk $r_s[(N_1, N_2), \delta]$ across designs (N_1, N_2), predictors δ, and states s. We first consider evaluation of risk in a specified state and then across states.

2.3.1. Monte Carlo Evaluation of Risk

Monte Carlo simulation provides a general approach to computation of risk in a specified state.

Risk in state s is the expected value of $L[y - \delta(\psi)]$ over the random variables (y, ψ), which are statistically independent with distributions $P_s(y)$ and $Q_s(\psi; N_1, N_2)$ respectively. Hence, $r_s[(N_1, N_2), \delta]$ can be approximated by drawing multiple realizations of (y, ψ), computing $L[y - \delta(\psi)]$, and averaging the results.

An alternative Monte Carlo approach is to proceed sequentially, taking advantage of the fact that ψ is typically a high-dimensional quantity but $\delta(\psi)$ is real valued. The first step is to draw multiple realizations of y, compute the set of losses $[L(y - t), t \in T]$ for each realization, and average the results to approximate $\{E_s[L(y - t)], t \in T\}$. When T contains infinitely many potential values, one may specify a finite subset that covers T sufficiently closely and compute loss on this subset. The second step is to draw multiple realizations of ψ, compute $\delta(\psi)$ and access the previously computed value of $E_s\{L[y - \delta(\psi)]\}$ for each realization, and average the results.

2.3.2. Risk and Regret under Square Loss

The optimal prediction under square loss is $\mu \equiv E(y)$. Evaluation of the risk and regret of a predictor in a specified state is relatively simple, having been studied as early as Hodges and Lehman (1950). Let $\mu_s = E_s(y)$ and $\lambda_{\delta s} = E_s[\delta(\psi)]$. For each value of the data ψ, decomposition of $E_s[y - \delta(\psi)]^2$ into variance plus squared bias yields

$$(3) \quad E_s[y - \delta(\psi)]^2 = E_s\{(y - \mu_s) + [\mu_s - \delta(\psi)]\}^2 = V_s(y) + [\mu_s - \delta(\psi)]^2.$$

Hence,

$$(4) \quad r_s[(N_1, N_2), \delta] = V_s(y) + \int[\mu_s - \delta(\psi)]^2 dQ_s(\psi; N_1, N_2).$$

The second term on the right-hand side is the MSE of $\delta(\psi)$ as an estimate of μ_s in state s. Decomposition of

this MSE into variance plus squared bias gives

$$(5)\quad \int[\mu_s - \delta(\psi)]^2 dQ_s(\psi; N_1, N_2) = \int\{(\mu_s - \lambda_{\delta s}) + [\lambda_{\delta s} - \delta(\psi)]\}^2 dQ_s(\psi; N_1, N_2) = V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2.$$

Hence,

$$(6)\quad r_s[(N_1, N_2), \delta] = V_s(y) + V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2.$$

Thus, to compute risk it suffices to compute the means and variances of y and $\delta(\psi)$ in state s.

The regret of a design-predictor pair in state s equals its risk minus the smallest risk achievable in that state. Under square loss, the smallest risk in state s is $V_s(y)$. To see this, let δ be the data-invariant rule setting $\delta(\psi) = \mu_s$ for all data realizations. Then $V_s[\delta(\psi)] = 0$ and $(\mu_s - \lambda_{\delta s})^2 = 0$. It follows that the regret of design-predictor pair $[(N_1, N_2), \delta]$ in state s is

$$(7)\quad r_s[(N_1, N_2), \delta] - \min_{[(N_1, N_2), \delta]' \in D} r_s[(N_1, N_2), \delta]' = V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2.$$

Thus, regret is the MSE when δ is used to estimate the mean outcome.

Even though the regret expression $V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2$ is relatively simple, analytical determination of its components $[V_s[\delta(\psi)], \mu_s, \lambda_{\delta s}]$ is feasible only in special cases. Fortunately, Monte Carlo evaluation is straightforward in general.

2.3.3. Cardinality of the Design-Predictor and State Spaces

The above shows that evaluation of risk for one design-predictor pair in one state is generally tractable. The same holds for regret in the case of square loss. The serious computational difficulty is

evaluation of risk across all design-predictor pairs and states.

The number of designs (N_1, N_2) that satisfy the budget constraint is finite but may be large. A planner can limit computational burden by considering a subset of the feasible designs. The set of logically feasible predictors typically is uncountable, making it intractable to consider all possibilities. Again, a planner might limit attention to a finite set of predictors. Thus, to make computation tractable, one might determine a design-predictor pair that optimizes a decision criterion for a constrained set of options rather than one that globally optimizes the criterion. Section 3 and 4 present illustrative cases.

The cardinality of the state space is the size of the set of all possible probability distributions $P_s(y)$ and $Q_s(\psi; N_1, N_2)$. This space typically is uncountable. A standard practice when considering problems with uncountable state spaces is to discretize the space, limiting attention to a finite subset of states that reasonably approximate the full state space.

2.4. Focus on Minimax Regret

Sections 2.2 and 2.3 considered mathematical and computational aspects of using statistical decision theory to choose a design-predictor pair, without privileging a particular decision criterion. Thus, we defined minimization of Bayes risk, maximum risk, and maximum regret, but we did not comment on their respective merits. Sections 3 and 4 will apply the minimax-regret criterion. Before doing so, we think it important to explain why we focus on minimax regret.

A reason specific to the problem of best point prediction under square loss is that the minimax-regret criterion prescribes minimization of maximum mean square error in this setting. MSE has long been used to measure precision, due to its simplicity and heuristic appeal. The idea of minimizing maximum mean square error is therefore easy to explain without even mentioning its interpretation as the minimax-regret predictor.

Yet we see deeper and more general reasons to focus on minimax regret. The first inclination of

some economists and statisticians is to adopt the Bayesian perspective and, hence, to minimize Bayes risk. Bayesian statisticians have long advocated use of Bayesian statistical decision theory to choose sample sizes for randomized experiments; see, for example, Canner (1970) and Cheng, Su, and Berry (2003).

We think the Bayesian perspective is compelling when a planner feels able to place a credible prior distribution on unknown quantities. However, Bayesian statisticians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. See, for example, the spectrum of views expressed by the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994) in the context of design of randomized trials. The controversy suggests that inability to express a credible prior is common in actual decision settings. Despite the efforts of Bayes statisticians to encourage Bayesian design of randomized trials, use of the frequentist idea of statistical power to choose sample size has remained the norm; see, for example, International Conference on Harmonisation (1999). Similar norms remain in survey design as well; see, for example, Office of Management and Budget (2006).

Bayesian planning is particularly difficult in nonparametric settings such as those we study in Sections 3 and 4. When we examine planning with nonresponse or interval data, the state space will permit the outcome distribution P(y) to lie within large classes of probability distributions. Applications of Bayesian statistics usually place priors on finite-dimensional state spaces, with occasional consideration of certain restricted infinite-dimensional spaces. As far as we are aware, Bayesian theory has not contemplated specification of credible priors that cover the large state spaces we study,

When it is difficult to form a credible subjective prior distribution, we think that a reasonable way to make decisions is to use a criterion that achieves uniformly satisfactory results, whatever the true state of nature may be. There are multiple ways to formalize the idea of uniformly satisfactory results. One prominent idea motivates the minimax-regret (MMR) criterion.

Minimax regret was first suggested as a general principle for decision making under uncertainty by Savage (1951) in an essay commenting on the Wald (1950) development of statistical decision theory. The regret associated with choice of a decision rule in a particular state of nature is the mean loss in welfare that

would occur across repeated samples if one were to choose this rule rather than the one that is best in this state of nature. The actual decision problem requires choice of a decision rule without knowing the true state of nature. The planner can evaluate a rule by the maximum regret that it may yield across all possible states of nature. He can then choose a rule that minimizes the value of maximum regret.

It is important to understand that maximum regret is computed ex ante, before one chooses an action. Maximum regret should not be confused with the psychological notion of regret, which a person may perceive ex post after choosing an action and observing the true state of nature. Manski and Tetenov (2016), observing that the term *maximum regret* means maximum potential distance from optimality, refer to MMR decisions as "near-optimal" decisions.

The minimax-regret criterion yields a decision that is uniformly satisfactory in the sense of yielding the best possible upper bound on regret, whatever the true state of nature may be. MMR is sometimes confused with minimax. Someone using the minimax criterion chooses an action that minimizes the absolute magnitude of the maximum loss that might possibly occur. Whereas minimax considers the worst outcome that an action may yield, MMR considers the worst outcome relative to what is achievable in a given state. Savage (1951) distinguished MMR sharply from minimax, writing that the latter criterion is "ultrapessimistic" while the former is not.

In a literature distinct from statistical decision theory, minimax regret has drawn diverse reactions from axiomatic decision theorists. In a famous early critique, Chernoff (1954) observed that decisions made with the MMR criterion are not always consistent with the choice axiom known as the independence of irrelevant alternatives (IIA). Chernoff considered this a serious deficiency, writing (p. 426):

> "A third objection which the author considers very serious is the following. In some examples, the min max regret criterion may select a strategy d_3 among the available strategies d_1, d_2, d_3, and d_4. On the other hand, if for some reason d_4 is made unavailable, the min max regret criterion will select d_2 among d_1, d_2, and d_3. The author feels that for a reasonable criterion the presence of an undesirable strategy d_4 should not have an influence on the choice among the remaining strategies.

This passage is the totality of Chernoff's argument. He introspected and concluded that any reasonable decision criterion should always adhere to the IIA axiom, but he did not explain why he felt this way. Chernoff's view has been endorsed by some modern axiomatic decision theorists, such as Binmore (2009).

On the other hand, Sen (1993) argued that adherence to axioms such as IIA does not per se provide a sound basis for evaluation of decision criteria. He asserted that consideration of the context of decision making is essential. In this vein, it was argued in Manski (2011) that adherence to the IIA axiom is not a virtue per se. What matters is how violation of the axiom affects welfare. It was observed that the MMR violation of the IIA axiom does not have a determinate effect of welfare. The MMR decision is always undominated when unique and there generically exists an undominated MMR decision when the criterion has multiple solutions. Hence, it was concluded that violation of the IIA axiom is not a sound rationale to dismiss minimax regret.

3. Low-Cost Sampling with Nonresponse

3.1. Background

Unit and item nonresponse are common in survey research. The implications for inference on the outcome distribution P(y) depend on what is known about the statistical association between population outcomes and response behavior. Analysis in Manski (1989) made this transparent by using the Law of Total Probability to decompose P(y) as follows:

$$(8) \qquad P(y) = P(y|z = 1)P(z = 1) + P(y|z = 0)P(z = 0),$$

where z = 1 if a person's outcome is observable and z = 0 if not. Random sampling with nonresponse

point identifies P(y|z = 1), P(z = 1), and P(z = 0), but it is uninformative about P(y|z = 0). Hence, in the absence of other knowledge, a sampling process with nonresponse partially identifies P(y), revealing that it lies in the set of distributions

$$(9) \qquad [P(y|z = 1)P(z = 1) + \gamma P(z = 0), \quad \gamma \in \Gamma],$$

where Γ is the set of all probability distributions on the outcome space.

The Law of Total Probability also shows that P(y) is point identified if combining the data obtained from the sampling process with other knowledge reveals the distribution P(y|z = 0) of missing data. A common practice is to assume that nonresponse is random conditional on specified observable covariates. Formally, this assumes that P(y|x, z = 0) = P(y|x, z = 1), where x denotes the observed covariates. This assumption reveals P(y|z = 0). The assumption underlies the familiar use of sample weights and imputations to produce point estimates in the presence of nonresponse.

Perspectives among survey researchers on the severity of nonresponse as an inference problem depend fundamentally on their views about the availability of knowledge that reveals the distribution of missing data. Cochran, Mosteller, and Tukey (1954) and Manski (1989) took the conservative position that such knowledge may not be available and concluded that nonresponse poses a serious problem. On the other hand, it has been common to assume that data are missing at random, in which case nonresponse reduces effective sample size but has no negative effect on identification.

Whereas much discussion has focused on the nonresponse rate in surveys, Groves (2006) turned the focus from nonresponse rates *per se* to the bias that may result when standard estimation methods that ignore nonresponse are used in the presence of nonresponse. In the same year, the U. S. Office of Management and Budget (OMB) issued a report entitled *Standards and Guidelines for Statistical Surveys* (Office of Management and Budget, 2006) that calls for analysis of nonresponse bias in government sponsored surveys whenever the unit response rate falls below 80% and/or the item response rate falls

below 70%. In both cases, it was implicitly assumed that the objective of data collection is to learn the mean population outcome E(y). Nonresponse bias was defined to be the difference between E(y) and the mean response given by survey respondents, namely E(y|z = 1).

The magnitude of nonresponse bias as defined by Groves (2006) and Office of Management and Budget (2006) can only be known *ex post*—after data have been collected—if at all. Data collected with a survey prone to nonresponse enables estimation of E(y|z = 1). Estimation of E(y) is possible if one also executes a second (high-cost) sampling process with complete response or if one is able to somehow determine the outcome values of nonrespondents to the original survey.

Our concern is with *ex ante* choice of sample design, when there do not yet exist data enabling estimation of E(y|z = 1) or E(y). The decision criteria described in Section 2 addressed this decision problem in principle. The analysis below develops a practical implementation of the minimax-regret criterion applicable under square loss. In this context, the MMR criterion reduces to minimization of maximum MSE.

The design problem simplifies in some special cases of substantial practical significance. Suppose that an agency is designing a new survey to be administered for a predetermined total budget. Suppose further that two vendors submit bids to conduct the survey. One bidder proposes a low-cost sampling process that will generate a known positive rate of unit nonresponse and a large sample size, while the other proposes a high-cost sampling process that has no nonresponse but a smaller sample. We derive the MMR choice between these two sampling processes under the assumption that the planner will use specific reasonable rules to choose a predictor with the sample data. The analysis generalizes easily to MMR comparison of any set of bids that differ only in terms of nonresponse rate and sample size.

The analysis also generalizes to designs that combine low-cost and high-cost sampling processes, under the assumption that the planner will pool the observed outcomes. Pooling the data may not be optimal because it discards information on data quality, but it is a simple practice that occurs frequently. In particular, users of surveys regularly pool data obtained from respondents who were recruited with a

mixture of low-cost and high-cost recruitment methods.

For example, users of the Health and Retirement Study (HRS) and the American Community Survey (ACS) typically pool the data obtained with multiple protocols. The initial wave of the HRS included a "nonresponse experiment" with sampled households assigned to alternative follow-up protocols. Some households were offered an incentive of either $50 or $100 to participate in the study rather than the standard incentive of $10 for single respondents and $15 for couples (Lengacher *et al.*, 1995).

The ACS engages in multi-phase sample recruitment and multi-mode interviewing that begins with contact via mail to sampled addresses (U.S. Census Bureau, 2013). Households may complete a questionnaire that is mailed to them or they may complete it online. After multiple contact attempts via mail, the Census Bureau attempts to conduct follow-up interviews via telephone with those who have not yet completed the survey. Finally, after a period of time in which telephone interviews are attempted, a sample of remaining nonrespondents is selected for in-person interview attempts. Conducting in-person interviews of nonrespondents is estimated to cost $144 per case, versus $14 for mail interviews and $19 for telephone interviews (Griffin, 2011).

3.2. Minimax-Regret Analysis under Square Loss

We assume no a priori knowledge of the two outcome distributions $P(y|z = 1)$ and $P(y|z = 0)$. Being agnostic about these distributions simplifies our analysis and avoids making assumptions that one may not find credible. Nevertheless, we should make clear that minimax-regret analysis may be performed with constrained state spaces. For example, one might find it credible to assume that the distributions of observed and unobserved outcomes are not too different from one another. Formally, one would place a metric ρ on the space of outcome distributions and constrain the state space by assuming that $\rho[P(y|z = 1), P(y|z = 0)] < d$ for a specified $d > 0$. The findings we report on maximum regret without knowledge of the two outcome distributions are upper bounds on the maximum regret that would be achievable with

constrained state spaces.

We maintain several assumptions that simplify analysis. First, we assume that y and t take values in a bounded interval on the real line, normalized to be the unit interval. Second, we assume knowledge of the response rate obtained with low-cost sampling. Third, we assume that with the low-cost sampling method, the constant marginal cost c_1 per sample member is incurred when an outcome is observed rather than when observation of an outcome is attempted. Hence, in this section, the design choice N_1 is the number of outcomes that will be observed.

The second and third assumptions make our analysis most applicable to settings where the dominant response problem is unit rather than item nonresponse. Historical experience often gives survey designers a good sense of the unit response rate to expect with various modes of survey administration (face-to-face, telephone, internet) and with various survey sponsors (federal government, university researchers, private market research firms). The cost per person of attempting to recruit respondents is often low relative to the cost of administering surveys to those who agree to participate.

The high-cost sampling process always yields accurate observations of y. The low-cost process yields accurate outcome data for some sample members but no data for others. The outcome data observed with sample design (N_1, N_2) are $\psi = (y_{1k}, k = 1, \ldots, N_1; y_{2k}, k = 1, \ldots, N_2)$.

We assume that the planner knows the response rate $P(z = 1)$ yielded by the low-cost process, but he has no a priori knowledge of the outcome distributions $P(y|z = 1)$ and $P(y|z = 0)$. The state space comprises all possible pairs of these distributions; thus, $[P_s(y|z = 1), P_s(y|z = 0), s \in S] = \Gamma \times \Gamma$. By (8), each pair $[P_s(y|z = 1), P_s(y|z = 0)]$ determines a unique population outcome distribution $P_s(y)$. If the planner were not to know the response rate, the state space would be the larger set $[P_s(y|z = 1), P_s(y|z = 0), P_s(z = 1), s \in S] = \Gamma \times \Gamma \times [0, 1]$.

In this setting, the MMR predictor has long been known in the polar case when only high-cost data are available; that is, when $N_1 = 0$ and $N_2 > 0$. The MMR predictor does not have a known explicit form

in the opposing polar case when only low-cost data are available; that is, when $N_1 > 0$ and $N_2 = 0$. However, we are able to easily derive the maximum regret of a reasonable choice. In what follows, we first consider these polar cases and then consider the sample design decision, where the planner chooses a (N_1, N_2) pair that satisfies a budget constraint.

3.3. Prediction with Only High-Cost Sampling

Consider the polar case in which only high-cost data are available. Hodges and Lehmann (1950), Theorem 6.1, prove that the MMR prediction is $(m_2\sqrt{N_2} + ½)(\sqrt{N_2} + 1)^{-1}$, where m_2 denotes the average value of the N_2 observations of y. They show (p. 190) that the minimax value of regret is $¼(\sqrt{N_2} + 1)^{-2}$.

The Hodges and Lehmann theorem provides an early example of a "shrinkage" estimator, predating the more famous Stein (1956) analysis of admissible estimation of the mean of a multivariate normal distribution. The MMR predictor is a weighted average of m_2 and ½; thus, it shrinks m_2 towards ½. The weight placed at ½ is $(\sqrt{N_2} + 1)^{-1}$, which goes to zero as sample size increases. This is sensible because m_2 becomes increasingly informative as sample size increases.

Shrinkage towards ½ occurs because, in the absence of data, the MMR prediction would be ½; that is, the midpoint of the interval of support. To see this, observe that in the absence of data, a predictor δ is a fixed number. Its variance being zero, the MSE of δ in state s is its squared bias $(\mu_s - \delta)^2$. With an unrestricted state space, μ_s can take any value in [0, 1]; hence, maximum bias is max $(\delta, 1 - \delta)$. Maximum bias is minimized by setting $\delta = ½$. The minimum value of maximum square error is ¼.

The conventional estimate of a population mean under random sampling is the sample average. Suppose that the planner selects m_2 as the prediction. The estimate is unbiased, so $\lambda_{\delta s} = \mu_s$ in each state s. It follows that regret is

$$(10) \quad V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2 \;=\; V_s(m_2) \;=\; V_s(y)/N_2.$$

Variance $V_s(y)$ is maximized in states where y is Bernoulli with $P_s(y = 1) = ½$, which yields $V_s(y) = ¼$. Hence, the maximum regret of the sample-average predictor is $1/(4N_2)$. This exceeds the minimax-regret value $¼(\sqrt{N_2} + 1)^{-2}$ by an amount that is negligible except when sample size is very small. In both cases, maximum regret goes to zero with increasing sample size at rate $1/N_2$.

3.4. Prediction with Only Low-Cost Sampling

Next consider the polar case in which only low-cost data are available. The MMR predictor and the minimax value of regret do not have known explicit forms in this case. We study a simple choice, this being the midpoint of the sample analog estimate of the interval that forms the identification region for the optimal prediction E(y). This is a reasonable choice. As the size of the low-cost sample goes to infinity, the midpoint predictor converges to $E(y|z = 1)P(z = 1) + ½P(z = 0)$, which is the MMR predictor when $E(y|z = 1)$ is known rather than estimated. The maximum regret of the midpoint predictor has a known explicit form. It thus provides an easily computable upper bound on the value of minimax regret.

3.4.1. Prediction when $E(y|z = 1)$ is Known

First consider prediction when $E(y|z = 1)$ is known. The Law of Iterated Expectations shows that

(11) $$E(y) = E(y|z = 1)P(z = 1) + E(y|z = 0)P(z = 0).$$

The sampling process point identifies $E(y|z = 1)$ but is uninformative about $E(y|z = 0)$, which can take any value in the unit interval. Hence, low-cost sampling reveals that E(y) lies in the interval

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

The MMR predictor given knowledge of (12) is $E(y|z = 1)P(z = 1) + ½P(z = 0)$, a weighted average that shrinks $E(y|z = 1)$ towards ½.. To see this, let predictor δ be any real number. The MSE of δ in state s is its squared bias $(\mu_s - \delta)^2$. Given that μ_s can take any value in interval (12), maximum bias is max $[\delta - E(y|z = 1)P(z = 1), E(y|z = 1)P(z = 1) + P(z = 0) - \delta]$. Maximum bias is minimized by setting $\delta = E(y|z = 1)P(z = 1) + ½P(z = 0)$.

3.4.2. Maximum Regret of the Midpoint Predictor

Now consider prediction when $E(y|z = 1)$ is not known. Let m_1 be the sample average of the observed N_1 outcomes. These outcomes are a random sample from $P(y|z = 1)$, so m_1 is the sample analog estimate of $E(y|z = 1)$. This yields $[m_1P(z = 1), m_1P(z = 1) + P(z = 0)]$ as an estimate of interval (12). We will determine maximum regret when the planner selects the midpoint of this interval estimate, $m_1P(z = 1) + ½P(z = 0)$, to be the prediction.

The general form of regret under square loss is $V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2$. In this case,

(13a) $\mu_s = E_s(y|z = 1)P(z = 1) + E_s(y|z = 0)P(z = 0)$,

(13b) $\lambda_{\delta s} = E_s(y|z = 1)P(z = 1) + ½P(z = 0)$.

Hence,

(14a) $V_s[\delta(\psi)] = E_s\{[m_1 - E_s(y|z = 1)]^2\}P(z = 1)^2 = [V_s(y|z = 1)/N_1]P(z = 1)^2$.

(14b) $(\mu_s - \lambda_{\delta s})^2 = [E_s(y|z = 0) - ½]^2P(z = 0)^2$.

Variance $V_s[\delta(\psi)]$ is maximized in states where $P_s(y|z = 1)$ is Bernoulli with $P_s(y = 1|z = 1) = ½$, which yields $V_s(y|z = 1) = ¼$. Squared bias $(\mu_s - \lambda_{\delta s})^2$ is maximized in states where $P_s(y|z = 0)$ is degenerate with

$P_s(y = 1|z = 0) = 1$ or $P_s(y = 0|z = 0) = 1$, which yields $[E_s(y|z = 0) - ½]^2 = ¼$. Both quantities can be simultaneously maximized, because $P_s(y|z = 1)$ and $P_s(y|z = 0)$ are logically separate distributions in the absence of assumptions that relate the distributions of observed and unobserved outcomes. Hence, maximum regret is

$$(15) \qquad \text{Max}_{s \in S} V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2 \;=\; ¼[P(z = 1)^2/N_1 + P(z = 0)^2].$$

Maximum variance goes to zero with increasing sample size at rate $1/N_1$, but maximum bias does not vary with sample size. The limiting value of (15) is $¼P(z = 0)^2$, which is the MMR value of maximum MSE when $E(y|z = 1)$ is known.

3.4.3. Benchmarking the Midpoint Predictor

It would be desirable to assess the performance of the midpoint predictor relative to the MMR predictor. We cannot do so in generality because we have no explicit expression for the MMR predictor and numerical calculation appears difficult. Nevertheless, we can say that we find the midpoint predictor appealing for multiple reasons.

First, it is simple to compute. Second, as sample size increases, its maximum MSE converges to the MMR value achievable when $E(y|z = 1)$ is known. Third, holding sample size fixed, the predictor behaves sensibly as the response rate $P(z = 1)$ approaches zero or one. When $P(z = 1) \to 0$, the midpoint predictor converges to ½, the MMR predictor without data. When $P(z = 1) \to 1$, it converges to m_1, the standard predictor used in settings with no missing data.

When there are missing data, it has been common to continue to use m_1 to estimate $E(y)$. We believe that this is reasonable when nonresponse is negligible or close to random, but not otherwise. A further appeal of the midpoint predictor is that it always yields smaller maximum MSE than m_1 when the distribution of unobserved outcomes is unknown. The discrepancy can be large in magnitude.

To see that the midpoint predictor always yields smaller maximum MSE than m_1, observe that in any state of nature s, the variance of m_1 is $V_s(y|z = 1)/N_1$ and its squared bias is $[E_s(y|z = 1) - E_s(y|z = 0)]^2P(z = 0)^2$. Consider a state in which $P_s(y|z = 1)$ is Bernoulli(½) and $E_s(y|z = 0)$ equals zero or one. Then $V_s(y|z = 1) = ¼$ and $[E_s(y|z = 1) - E_s(y|z = 0)]^2 = ¼$. Hence, m_1 has MSE $¼[1/N_1 + P(z = 0)^2]$ in this state. Comparison with (15) shows that the MSE of m_1 in this state exceeds the maximum MSE of the midpoint predictor over all possible states.

To see that the discrepancy can be large, consider a state in which $E_s(y|z = 1) = 1$ and $E_s(y|z = 0) = 0$, or vice versa. Then $V_s(y|z = 1) = 0$ and $[E_s(y|z = 1) - E_s(y|z = 0)]^2 = 1$. Hence, m_1 has MSE $P(z = 0)^2$ in this state. As $N_1 \rightarrow \infty$, the ratio of the MSE of m_1 in this state to the maximum MSE of the midpoint predictor converges to four.

Further analysis, presented in Supplementary Section 3.4, shows that the above states --- $[E_s(y|z = 1) = 1, E_s(y|z = 0) = 0]$ and vice versa --- maximize MSE whenever $P(z = 0)^2 \geq 1/(2N_1)$. Hence, maximum MSE is $P(z = 0)^2$ in these cases. When $P(z = 0)^2 < 1/(2N_1)$, maximum MSE is larger than $P(z = 0)^2$.

3.5. Choosing Between the Low-Cost and High-Cost Designs with a Predetermined Budget

To start consideration of sample design, we first examine the constrained setting in which the planner must choose between one of the two designs, intermediate options not being feasible. With marginal sampling costs (c_1, c_2) and predetermined budget B, the feasible sample sizes are $N_1 = INT(B/c_1)$ for low-cost sampling and $N_2 = INT(B/c_2)$ for high-cost sampling. We henceforth ignore for simplicity the fact that sample sizes must be integers and take the feasible sample sizes to be $N_1 = B/c_1$ and $N_2 = B/c_2$.

The best polar design from the MMR perspective minimizes maximum regret using the MMR predictor for that design. We have an explicit expression for the maximum regret predictor with high-cost sampling but not with low-cost sampling. To level the playing field, we consider choice of a design when the planner commits to use the simple predictors $m_1P(z = 1) + ½P(z = 0)$ for low-cost sampling and m_2 for

high-cost sampling; that is, the low-cost sample interval midpoint and the high-cost sample average.

With these predictors, the feasible low-cost and high-cost designs yield maximum regret ¼[P(z = 1)^2(c_1/B) + P(z = 0)^2] and ¼(c_2/B), respectively. Hence, the low-cost design is better from the MMR perspective when the budget is less than a particular threshold and the high-cost design is better when it is above the threshold. The threshold budget is

(16) $B = [c_2 - P(z = 1)^2 c_1]/P(z = 0)^2$.

Supplementary Section 3.5 gives numerical illustrations showing how MMR choice between a low-cost and high-cost design varies with the budget, the sampling costs, and the low-cost response rate.

The above analysis generalizes easily to choice among multiple sampling processes that differ in their costs and response rates. Suppose that a set Q of sampling processes are feasible, each $q \in Q$ having sampling cost c_q and response rate $P_q(z = 1)$. Given the predetermined budget B, the maximum regret of process q is ¼[$P_q(z = 1)^2(c_q/B) + P_q(z = 0)^2$]. If the planner is constrained to choose among these processes, the best design from the MMR perspective is one that minimizes $P_q(z = 1)^2(c_q/B) + P_q(z = 0)^2$. In general, the best design depends on the size B of the available budget.

3.6. Allocation of Budget to Both Sampling Processes, with Commitment to Data Pooling

Now suppose that it is feasible to allocate budget to both a low-cost and a high-cost sampling process, subject only to the overall budget constraint $c_1N_1 + c_2N_2 \leq B$. There are many prima facie reasonable ways to choose a predictor combining the data from both samples. However, we have found computation of maximum regret to be burdensome in general. We therefore focus on a particular predictor for which MMR computation is tractable.

Specifically, we suppose that the planner pools the observed outcomes across the two samples and

then proceeds as if the data were drawn entirely by low-cost sampling. We observed earlier that data users often pool data in this manner, so study of prediction with pooled data has practical importance. Pooling is easy to study because we can apply the results obtained above to the pooled sample.

Let m_{12} be the pooled sample average of the observed $N_1 + N_2$ outcomes. Let $\pi \equiv P(z = 1)$ be the response rate with low-cost data. Assuming for simplicity that the sample realized response rate equals the population response rate, let N_1/π be the total size of the low-cost sample that must be drawn to obtain N_1 responses. Then the response rate in the pooled sample is $(N_1 + N_2)(N_1/\pi + N_2)^{-1}$. The resulting predictor is the interval-estimate midpoint $m_{12}(N_1 + N_2)(N_1/\pi + N_2)^{-1} + \frac{1}{2}(N_1/\pi - N_1)(N_1/\pi + N_2)^{-1}$.

This pooled midpoint predictor is a simple extension of the one using only low-cost data. That predictor was a weighted average of the low-cost sample average m_1 and ½. This one analogously shrinks m_{12} towards ½. We earlier found that the low-cost midpoint predictor outperforms m_1 in maximum MSE. The pooled midpoint predictor similarly outperforms m_{12}.

By (15), maximum regret for this predictor with a given sample design (N_1, N_2) is

$$(17) \quad \tfrac{1}{4}\{[(N_1 + N_2)(N_1/\pi + N_2)^{-1}]^2(N_1 + N_2)^{-1} + [(N_1/\pi - N_1)(N_1/\pi + N_2)^{-1}]^2\}$$
$$= \tfrac{1}{4}(N_1/\pi + N_2)^{-2}[(N_1 + N_2) + (N_1/\pi - N_1)^2].$$

Given the commitment to data pooling, the design minimizing maximum regret chooses (N_1, N_2) to solve

$$(18) \quad \min_{(N_1, N_2):\ 0 \le c_1N_1 + c_2N_2 \le B} \tfrac{1}{4}(N_1/\pi + N_2)^{-2}\,[(N_1 + N_2) + (N_1/\pi - N_1)^2].$$

The optimal design will exhaust the budget. Hence, we can plug in $N_2 = (B - c_1N_1)/c_2$ and rewrite (18) as

$$(19) \quad \min_{N_1:\ 0 \le N_1 \le B/c_1} \tfrac{1}{4}[N_1/\pi + (B - c_1N_1)/c_2]^{-2}\{[N_1 + (B - c_1N_1)/c_2] + (N_1/\pi - N_1)^2\}.$$

The solution to problem (19) depends on the low-cost response rate π, the sampling costs (c_1, c_2), and the budget B. Supplementary Section 3.6 provides a numerical illustration.

3.7. Allocation of Budget to Both Sampling Processes, Using an Intersection Estimator as Predictor

We have shown that it is straightforward to compare alternative sample designs when the planner commits to data pooling and to prediction using the midpoint of the analog interval estimate of the identification region for the optimal predictor μ. We showed that the maximum regret of a design using this predictor has the simple form (17). Numerical solution of the one-dimensional extremum problem (19) then determined the design that minimizes maximum regret.

Pooling low-cost and high-cost data discards available information on data quality. It is reasonable to ask whether a predictor that uses this information may outperform one that pools the data. While comprehensive comparison of alternative rules appears computationally prohibitive, we can make progress by considering particular alternative rules. We report some exploratory analysis here, leaving a deeper analysis for future research.

We focus on rules that use "intersection estimates" as the predictor. As earlier, let m_1 and m_2 be the sample average values of y observed using the low-cost and high-cost sampling processes. The two samples yield analog interval and point estimates of μ, namely $[m_1P(z = 1), m_1P(z = 1) + P(z = 0)]$ and m_2. The question is how to combine the two estimates to form a predictor for use with a specified sample design, recognizing the imprecision in m_1 and m_2.

One possibility begins with confidence intervals for m_1 and m_2. Conventional confidence intervals have the form $[m_1 - b_1/\sqrt{N_1}, m_1 + b_1/\sqrt{N_1}] \cap [0, 1]$ and $[m_2 - b_2/\sqrt{N_2}, m_2 + b_2/\sqrt{N_2}] \cap [0, 1]$, where $b_1 > 0$, $b_2 > 0$. These intervals quantify sampling imprecision in the analog estimates of μ. This done, a heuristically reasonable predictor is a point in the intersection of the two confidence intervals for μ, namely

$$[(m_1 - b_1/\sqrt{N_1})P(z = 1),\ (m_1 + b_1/\sqrt{N_1}]P(z = 1) + P(z = 0)] \ \cap \ [m_2 - b_2/\sqrt{N_2},\ m_2 + b_2/\sqrt{N_2}] \ \cap \ [0, 1].$$

Similar intersection estimates have been considered in the literature on partial identification with missing data; see Manski (1990, 2003), Manski and Pepper (2000, 2009), Krieder and Pepper (2007), and Chernozhukov, Lee, and Rosen (2013).

A class of predictors that vary in their attention to sampling imprecision is obtained by considering alternative values for the constants b_1 and b_2. Any predictor of this class is consistent, converging to the optimal prediction μ as $N_1 \to \infty$ and $N_2 \to \infty$. With a finite sample, there is positive probability that the intersection of the two interval estimates is null. If this occurs, one must use an auxiliary criterion to determine the prediction. The probability of a null intersection goes to zero as N_1 and N_2 go to infinity.

With a square loss function, the state-dependent regret of a design-predictor pair that uses an intersection estimate as the predictor has the form $V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2$ given in equation (7). The component terms $\{V_s[\delta(\psi)], \mu_s, \lambda_{\delta s}\}$ do not generally have simple explicit forms, but they can be computed relatively easily by Monte-Carlo simulation. This done, maximum regret across states can be approximated by discretizing the state space, as discussed in Section 2.3.3.

Supplementary Section 3.7 gives numerical illustrations. These calculations cannot be definitive, but they suggest that intersection estimates are well-behaved. We find that the maximum MSE of intersection estimates is usually smaller than that of pooled estimates, especially with larger budget and correspondingly larger sample sizes.

3.8. Budgeting for Near Optimality Rather than Statistical Power

We have thus far studied allocation of a predetermined budget between low-cost and high-cost sampling processes. Now suppose that budget is a choice variable. In principle, the planner should perform a benefit-cost analysis. Devoting a larger budget to data collection improves prediction of

outcomes but diverts resources from other uses. The planner must resolve this tension.

Benefit-cost analysis of survey sampling has been rare in practice; see Spencer (1985). Instead, it has been standard to use statistical power calculations to evaluate proposed budgets. Implicitly, one presumes that the purpose of data collection is to test specified null hypotheses against certain alternatives and one assumes that nonresponse is random. Then one determines whether the sample size generated by a proposed budget suffices to make probabilities of Type II errors smaller than specified thresholds. Consider, for example, the standards for survey samples requiring OMB approval given in Office of Management and Budget (2006). Guideline 1.2.2 states that sample designs should include (p. 7) "power analyses to determine sample sizes and effective sample sizes for key variables."

Studying choice of sample size in classical randomized trials, Manski and Tetenov (2016) argue against use of power calculations, citing multiple deficiencies of the practice. They suggest instead that sample size be selected to enable ε-optimal treatment decisions; ε-optimality means that there exists a treatment rule whose maximum regret is no larger than a specified $\varepsilon > 0$. The specific planning problem studied by Manski and Tetenov differs from that studied in this paper, but the general idea of choosing a design to enable ε-optimal decisions is applicable to our survey sampling problem.

For a specified value of ε, suppose that one wants to choose a design-predictor pair such that the maximum MSE of the predictor is no larger than ε. If one is constrained to choose between low-cost and high-cost sampling, the analysis of Section 3.5 shows that a budget of size B suffices to achieve this objective if $\min\{¼[P(z = 1)^2(c_1/B) + P(z = 0)^2], ¼(c_2/B)\} \leq \varepsilon$. If one can allocate budget to both sampling processes and commits to using the pooled midpoint predictor, a budget of size B suffices to achieve ε-optimality if the value of the minimand in (19) is no larger than ε.

These budget sizes are sufficient for ε-optimality but may not be necessary. The smallest budgets that enable ε-optimal prediction occur when one uses MMR predictors rather than the tractable predictors studied in Sections 3.5 and 3.6. In the absence of knowledge of the MMR predictors, we can provide sufficient budget sizes but not necessary ones.

Although choice of budget size to achieve ε-optimal prediction does not fulfill the ideal of a comprehensive benefit-cost analysis, it is closer to that ideal than is conventional use of power calculations. Implementation of the idea requires specification of a value for ε. The need to choose an effect size of interest already arises in conventional practice, where a survey planner must specify the alternative hypotheses to be compared with the specified nulls. Office of Management and Budget (2016) directs sample designers to specify a *minimum substantively significant effect size*, defined to be (p. 32) "the smallest effect, that is, the smallest departure from the null hypothesis, considered to be important for the analysis of key variables."

Designers may similarly be able to specify substantively meaningful values for maximum mean square error in prediction. Consider, for example, the discussion of ACS reliability with voluntary response in Griffin (2011). The report states (page 6):

> "If the ACS was a voluntary survey and no additional funding was provided, we estimate that sampling variances would be increased by 45 percent. This would raise questions about whether or not these estimates should be released to the public. . . . . The ACS was designed to produce 5-year estimates at the tract-level and such deterioration in sample sizes and reliability would compromise our ability to accomplish that goal."

While the statement refers to variance rather than mean square error, it suggests that the Census Bureau may be able to specify the maximum MSE it finds acceptable in estimates using the ACS.

4. Low-Cost Sampling with Interval Measurement of Outcomes

4.1. Low-Resolution Interval Measurement

Again let outcomes and potential predictions take values in the unit interval and let the high-cost

measurement method always yield errorless observations of y. Let the low-cost method yield an interval measurement. That is, for $k = 1, \ldots, N_1$, one observes a sub-interval of [0, 1] that contains y_k.

It appears very difficult to characterize the maximum regret of design-predictor pairs when low-cost sampling produces general forms of interval measurement. However, progress is possible in special cases of practical importance. One is the nonresponse case studied in Section 3. Nonresponse is interval measurement in which only two types of intervals occur: the point interval $[y_k, y_k]$ when the outcome is observable and the trivial interval [0, 1] when the outcome is unobservable.

This section considers a different special case. We suppose that low-cost sampling uses a low-resolution measurement device that locates each value of a continuously distributed outcome within a predetermined finite set of $M \geq 2$ intervals. These intervals, denoted $(I_1, I_2, \ldots, I_M)$, collectively cover the unit interval and overlap at most at their endpoints. We focus on the particularly simple case of equal-length closed intervals, each of length 1/M. Thus, the intervals are $I_m = [(m - 1)/M, m/M]$, $m = 1, \ldots, M$. High-cost sampling yielding precise outcome data is the limit as $M \to \infty$.

Low-resolution interval measurement differs from nonresponse, where one either observes an outcome perfectly or not all. Here, in contrast, data collection yields partial information about every outcome by placing it within one of the M intervals. The analysis of this section also differs in the accrual of sampling costs. In Section 3, we assumed that the marginal cost c_1 for low-cost sampling is incurred only when an outcome is observed. Here it is incurred for every sample member.

4.2. Prediction with Low-Cost Sampling

The polar case with only high-cost data collection is the same as discussed in Section 3 and needs no elaboration. Consider the polar case in which only low-cost data are collected. The MMR predictor and the minimax value of regret do not have known explicit forms. Given this, we study a simple predictor, the midpoint of the sample analog estimate of the identification region for the optimal prediction E(y).

To begin, the Law of Iterated Expectations and the assumption that y is continuous imply that

(20) $E(y) = \sum_m E(y|y \in I_m)P(y \in I_m)$.

For each m = 1, . . . , M, the sampling process point-identifies $P(y \in I_m)$ but is uninformative about $E(y|y \in I_m)$, which can take any value in the open interval $((m-1)/M, m/M)$. Hence, low-cost sampling reveals that E(y) lies in the open interval

(21) $(\sum_m [(m-1)/M]P(y \in I_m), \sum_m (m/M)P(y \in I_m))$.

Given a sample of size N_1, let p_{1m} be the sample frequency with which the outcome is observed to lie in interval I_m. This frequency is the sample analog estimate of $P(y \in I_m)$. This yields

(22) $(\sum_m [(m-1)/M]p_{1m}, \sum_m (m/M)p_{1m})$

as the sample analog estimate of interval (21). We will suppose that the planner commits to use the midpoint of this interval estimate, namely $\sum_m [(m-½)/M]p_{1m}$, to be the predictor.

The general form of regret under square loss is $V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2$. In this case,

(23a) $\mu_s = \sum_m E_s(y|y \in I_m)P_s(y \in I_m)$.

(23b) $\lambda_{\delta s} = \sum_m [(m-½)/M]P_s(y \in I_m)$.

Hence, squared bias in state s is

(24) $(\mu_s - \lambda_{\delta s})^2 = \{ \sum_m [E_s(y|y \in I_m) - (m - ½)/M]P_s(y \in I_m)\}^2$.

Given any values for $[P_s(y \in I_m), m = 1, \ldots, M]$, squared bias increases across states as $[E_s(y|y \in I_m) \rightarrow (m - 1)/M$, all m] or as $[E_s(y|y \in I_m) \rightarrow m/M$, all m], with the same limit value $¼M^{-2}$. Hence, the supremum of squared bias across all states with continuous outcome distributions is $¼M^{-2}$.

To analyze the variance $V_s[\delta(\psi)]$, observe that the prediction $\sum_m [(m - ½)/M]p_{1m}$ is the average across a random sample of size N_1 of a multinomial random variable with (mass point, probability) pairs $[(m - ½)/M, P_s(y \in I_m)]$, $m = 1, \ldots, M$. Among all multinomial distributions with these mass points, the one with the largest variance is the Bernoulli distribution with $P_s(y \in I_1) = P_s(y \in I_M) = ½$. This distribution has mean ½ and variance $½[1/(2M) - ½]^2 + ½[(M - ½)/M - ½]^2 = ¼ M^{-2}(M - 1)^2$. Hence, $\max_s V_s[\delta(\psi)] = ¼M^{-2}(M - 1)^2/N_1$.

Combining the above findings shows that the supremum of regret across all states occurs when $P_s(y \in I_1) = P_s(y \in I_M) = ½$ and when either $[E_s(y|y \in I_1) \rightarrow 0, E_s(y|y \in I_M) \rightarrow (M - 1)/M]$ or $[E_s(y|y \in I_1) \rightarrow 1/M, E_s(y|y \in I_M) \rightarrow 1]$. The value of the supremum is

(25) $\sup_{s \in S} V_s[\delta(\psi)] + (\mu_s - \lambda_{\delta s})^2 = ¼M^{-2}(M - 1)^2/N_1 + ¼M^{-2} = ¼M^{-2}[(M - 1)^2/N_1 + 1]$.

Observe that the variance component of supremum regret goes to zero as $N_1 \rightarrow \infty$, but the bias component does not vary with N_1. The bias component goes to zero as $M \rightarrow \infty$. Holding N_1 fixed, the variance component increases with M, rising from $1/(16N_1)$ at M = 2 to $1/(4N_1)$ as $M \rightarrow \infty$.

4.3. Choosing Between the Low-Cost and High-Cost Designs

Suppose that the planner must choose between one of the low-cost and high-cost designs,

intermediate options not being feasible. With marginal sampling costs (c_1, c_2) and budget B, the feasible sample sizes are $N_1 = INT(B/c_1)$ for low-cost sampling and $N_2 = INT(B/c_2)$ for high-cost sampling. As in Section 3, we ignore for simplicity the fact that sample sizes must be integers and take the feasible sample sizes to be $N_1 = B/c_1$ and $N_2 = B/c_2$.

The best design from the MMR perspective is the one that minimizes maximum regret using the MMR predictor for that design. As in Section 3, we have an explicit expression for the maximum regret predictor with high-cost sampling but not with low-cost sampling. To level the playing field, we consider choice of a design when the planner commits to use the simple predictors $\sum_m [(m - ½)/M]p_{1m}$ for low-cost sampling and m_2 for high-cost sampling; that is, once again, the low-cost sample interval midpoint and the high-cost sample average. With these predictors, the feasible low-cost and high-cost designs yield maximum regret $¼M^{-2}[(M - 1)^2(c_1/B) + 1]$ and $¼(c_2/B)$, respectively. Hence, the low-cost design is better from the MMR perspective when the budget is less than a particular threshold and the high-cost design is better when it is above the threshold. The threshold budget is

$$(26) \qquad B = c_2M^2 - (M - 1)^2c_1.$$

This finding generalizes easily to choice among multiple sampling processes that differ in costs and in the resolution of interval measurement. Suppose that a set Q of sampling processes are feasible, each $q \in Q$ having sampling cost c_q and number M_q of intervals. Given the predetermined budget B, the maximum regret of process q is $¼(M_q)^{-2}[(M_q - 1)^2c_q/B + 1]$. The best design from the MMR perspective minimizes $(M_q)^{-2}[(M_q - 1)^2c_q/B + 1]$. In general, the best design depends on the size of the budget.

4.4. Allocation of Budget to Both Sampling Processes

As in Section 3, it is of interest to consider the unconstrained setting where it is feasible to allocate

budget to both a low-cost and a high-cost sampling process, subject only to the overall budget constraint $c_1N_1 + c_2N_2 \leq B$. Again there are many prima facie reasonable ways to choose a predictor combining the data from both samples. However, we have not found one for which maximum regret has a tractable explicit form. We leave further study of sampling with interval measurement for future research.

5. Conclusion

This paper demonstrates how statistical decision theory may be usefully applied to data collection design problems in two specific tractable settings, one with nonresponse and the other with interval measurement. Interactions between these two phenomena yield another application. For example, household surveys often elicit interval data on real-valued outcomes in order to reduce rates of item nonresponse that arise for questions about sensitive topics, such as income and assets. In this case, marginal sampling cost may be invariant to the resolution of interval measurement, but increasing the resolution could incur a higher cost in the form of increased nonresponse, a phenomenon discussed by Philipson (2001). It would be of interest to extend Sections 3 and 4 to choice among multiple sampling processes that differ in their resolution of interval measurement and rate of item nonresponse.

We believe that statistical decision theory should play an important role in data collection design. This paper has developed tractable methods for doing so, in a setting where the planner is concerned with both statistical imprecision and partial identification. Beyond its specific contributions, we hope that this paper will encourage increased use of statistical decision theory to inform the design of data collection more generally when data quality is a decision variable.

Supplementary Section 3.4: Maximum Mean Square Error of the Sample Average Estimate

In state s, the sample average m_1 has variance

$$V_s(m_1) = V_s(y|z = 1)/N_1 = [E_s(y^2|z = 1) - E_s(y|z = 1)^2]/N_1$$

and square bias

$$(\mu_s - \lambda_{\delta s})^2 = [E_s(y|z = 1) - E_s(y|z = 0)]^2 P(z = 0)^2.$$

Our analysis of maximum mean square error uses the fact that our state space is unrestricted, placing no constraints on the distributions of observed and unobserved outcomes.

To obtain the maximum MSE, first fix $P_s(y|z = 1)$ and maximize squared bias over $E_s(y|z = 0) \in [0, 1]$. The maximum occurs when $E_s(y|z = 0)$ equals zero or one. Hence, maximum MSE thus far is

$$[E_s(y^2|z = 1) - E_s(y|z = 1)^2]/N_1 + \max\{E_s(y|z = 1)^2, [1 - E_s(y|z = 1)]^2\} P(z = 0)^2.$$

Next fix $E_s(y|z = 1)$ and maximize over the feasible values of $E_s(y^2|z = 1)$. The maximum occurs when y is Bernoulli with mean $E_s(y|z = 1)$. To show this, observe that $y^2 = y$ when y is Bernoulli; hence, $E_s(y^2|z = 1) = E_s(y|z = 1)$. If $P_s(y|z = 1)$ is not Bernoulli, there is positive probability that $0 < y < 1$, in which case $y^2 < y$. Hence, $E_s(y^2|z = 1) < E_s(y|z = 1)$. Therefore, maximum MSE now is

$$[E_s(y|z = 1) - E_s(y|z = 1)^2]/N_1 + \max\{E_s(y|z = 1)^2, [1 - E_s(y|z = 1)]^2\} P(z = 0)^2.$$

It remains to maximize over $E_s(y|z = 1) \in [0, 1]$. To simplify notation, let $\theta \in [0, 1]$ denote a possible value of $E_s(y|z = 1)$. Consider $\theta \in [0, ½]$ and $\theta \in [½, 1]$. The maximization problems in these domains are

$$\text{Max}_{\theta \in [0, ½]} \quad (\theta - \theta^2)/N_1 \quad + \quad (1 - \theta^2)P(z = 0)^2,$$

$$\text{Max}_{\theta \in [½, 1]} \quad (\theta - \theta^2)/N_1 \quad + \quad \theta^2 P(z = 0)^2.$$

These problems have symmetric forms, with $\theta \in [½, 1]$ solving the second problem if and only if $1 - \theta$ solves the first. Hence, it suffices to consider the second problem.

Rewrite the maximand as

$$(\theta - \theta^2)/N_1 \quad + \quad \theta^2 P(z = 0)^2 \quad = \quad \theta/N_1 \quad + \quad \theta^2[P(z = 0)^2 - 1/N_1].$$

When $P(z = 0)^2 = 1/N_1$, the maximand reduces to θ/N_1. Hence, maximum occurs at at $\theta = 1$. The value of maximum MSE is $P(z = 0)^2$.

When $P(z = 0)^2 \neq 1/N_1$, the maximand is quadratic. Examination of the first order condition shows that it has global extremum at $- ½[N_1 P(z = 0)^2 - 1]^{-1}$. If $P(z = 0)^2 > 1/N_1$, the extremum is a minimum and occurs at a negative value of θ. Hence, maximum MSE occurs at $\theta = 1$. The value of maximum MSE is $P(z = 0)^2$.

Finally, suppose that $P(z = 0)^2 < 1/N_1$. Then the extremum is a maximum and occurs at a positive value of θ. Given that $N_1 P(z = 0)^2 \geq 0$, the maximum occurs at some $\theta \geq ½$. If $1/(2N_1) \leq P(z = 0)^2 < 1/N_1$, the global maximum occurs at $\theta \geq 1$. Hence, as above, maximum MSE occurs at $\theta = 1$ and the value of maximum MSE is $P(z = 0)^2$. If $P(z = 0)^2 < 1/(2N_1)$, the global maximum occurs at a $\theta \in [½, 1)$ and maximum MSE equals the MSE at this value of θ.

Supplementary Section 3.5: Choice between a Low-Cost and a High-Cost Internet Survey

To illustrate choice between a low-cost and high-cost design, consider an internet survey that poses a set of questions with the aim of estimating the population mean outcome for each question. Each question has a bounded real range of possible responses, normalized to the [0, 1] interval. The designer of the survey wants to minimize the maximum mean square error per outcome when the midpoint predictor is used to estimate each mean outcome.

For specificity, let the internet survey contain 20 questions. Let the cost of survey administration to a person with internet access be $100 per respondent, of which part is a subject payment and part is the cost of fielding the survey and processing the findings. Let the unit response rate be 0.6, comprising persons who are willing to participate and who have internet access enabling them to do so in practice. Let it be known that offering free internet to persons who lack access would increase the unit response rate to 0.70, with cost $500 per additional respondent for providing a computer/tablet and internet service. Given a specified survey budget B, the survey designer must decide whether to offer free internet to persons who lack access.

In this setting, the regular (low cost) response rate is $P_1(z = 1) = 0.6$ and the survey cost per question and respondent is $c_1 = 100/20 = 5$. The enhanced (high cost) response rate with free internet access is $P_2(z = 1) = 0.7$, of whom 0.1 (that is, 1/7 of the respondents) are persons given internet access. The enhanced survey cost per question and respondent is $c_2 = c_1 + (1/7)\cdot(500/20) = 8.57$. Let $b = B/20$ denote the available budget per question. Then the maximum mean square error per question with low-cost and high-cost sampling are

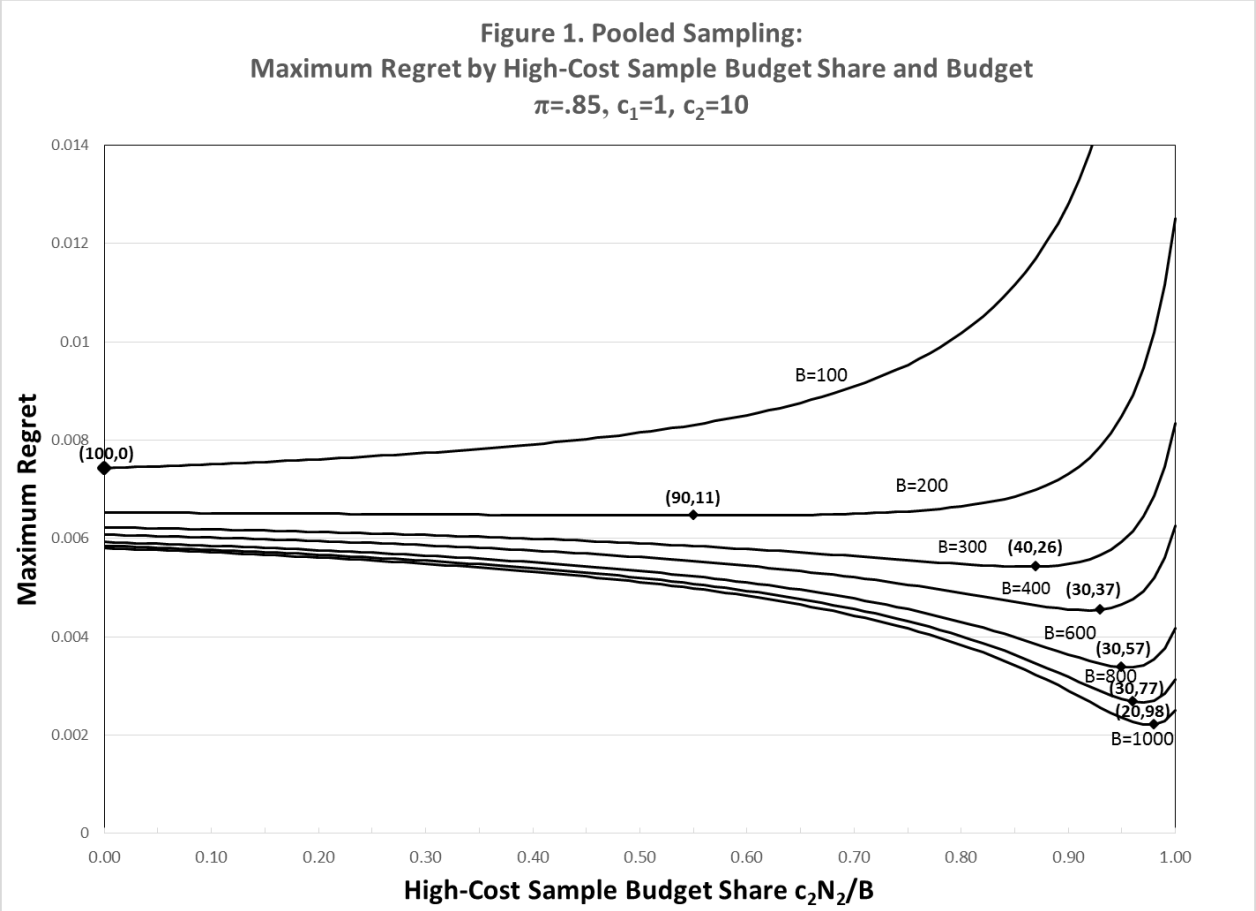
$$P_1(z = 1)^2(c_1/b) + P_1(z = 0)^2 \quad = \quad (1.8)/b + 0.16,$$

$P_2(z = 1)^2(c_2/b) + P_2(z = 0)^2 \quad = \quad (4.2)/b + 0.09.$

Comparison of the low-cost and high-cost maximum mean square errors shows that it is better not to offer free internet access if $b < 34.29$ and to offer it if $b > 34.29$. In terms of the overall survey budget, the threshold above which free internet should be offered is $B = 686$. This is a remarkably small value, sufficient to enroll about 7 respondents who have internet or 1 who does not. Thus, offering free internet access is preferable with essentially any non-trivial budget.

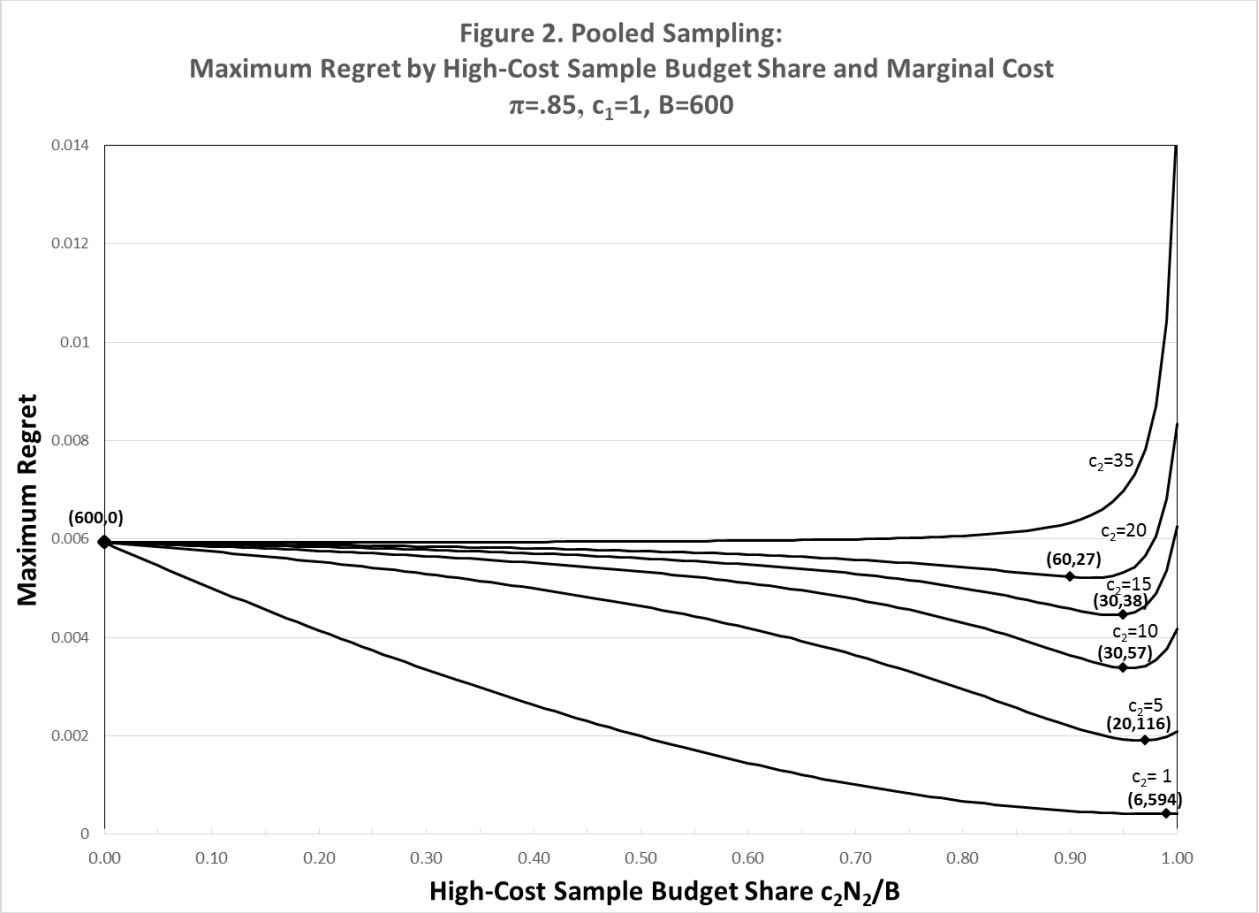
Supplementary Section 3.6: Numerical Illustrations of Allocation of Budget to Both Sampling Processes, with Data Pooling

We first allow the budget B to vary while holding the remaining parameters fixed at the values ($\pi = .85$, $c_1 = 1$, $c_2=10$). With this specification of (π, c_1, c_2), equation (17) shows that the threshold budget for binary choice between the low-cost and high-cost designs is about 412. That is, low-cost sampling would be chosen for $B < 412$, whereas high-cost sampling would be chosen for $B > 412$. Figure 1 graphically shows the optimal allocation of budget to the two sampling processes when the planner is not constrained to choose between the two designs.



Note: The maximum regret curves depicted here ignore the fact that sample sizes N_1 and N_2 must be integers. However, the depicted MMR sample designs, identified by diamond markers with labelled pairs (N_1, N_2), impose this requirement.

Figure 1 plots maximum regret on the vertical axis and the high-cost sample budget share c_2N_2/B on the horizontal axis. Maximum regret falls as the budget increases, holding the high-cost share fixed. Allocating the entire budget to low-cost sampling (B, 0) minimizes maximum regret for any budget less than about 176. For B > 176, the optimal high-cost sample share increases as B increases. The size of the low-cost sample N_1 declines from 100 to 30 as B approaches the binary-choice threshold budget B = 412. Above that threshold budget, the MMR choice of N_1 declines slowly and the high-cost sample share goes to 1 as B goes to ∞.

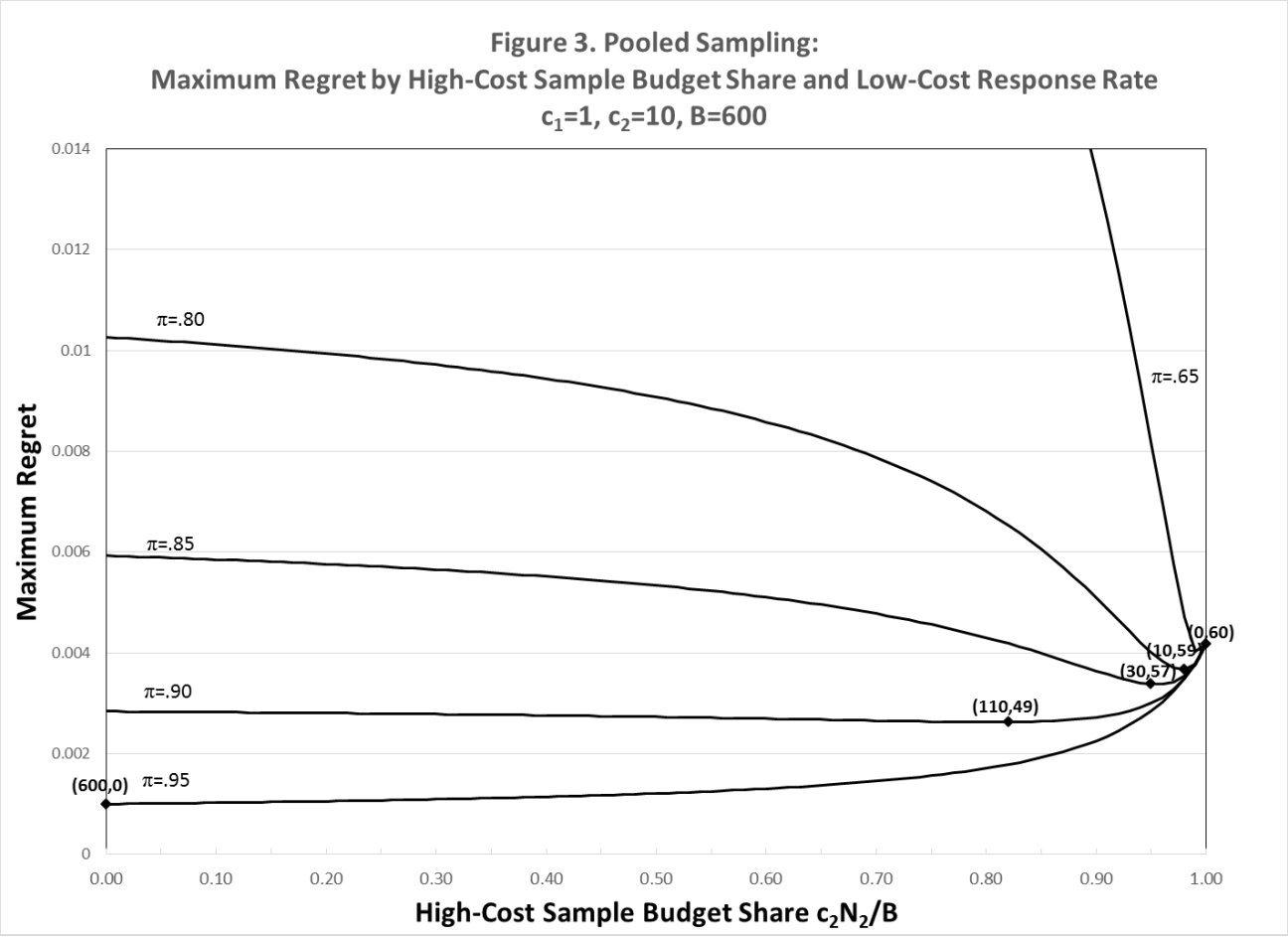


Note: The maximum regret curves depicted here ignore the fact that sample sizes N_1 and N_2 must be integers. However, the depicted MMR sample designs, identified by diamond markers with labelled pairs (N_1, N_2), impose this requirement.

Figure 2 allows the marginal cost of high-cost sampling c_2 to vary while holding the other parameters fixed at ($\pi = .85$, $c_1 = 1$, $B = 600$). Maximum regret falls as c_2 decreases, holding the high-cost sample budget share fixed. Allocation of the entire budget to low-cost sampling (600, 0) minimizes maximum regret for any $c_2 > 33$. The minimax-regret design increasingly allocates budget to high-cost sampling as c_2 decreases below 33.

One might expect the entire budget to be allocated to high-cost sampling when $c_2 = c_1 = 1$, the two processes then having equal costs but different response rates. Yet the figure shows the minimax-regret design in this case to be (6, 594), close to but not equal to the polar allocation (0, 600). The explanation is

the commitment of the planner to use the sample average m_2 as the predictor with design (0, 600) rather than the MMR estimate $(m_2\sqrt{600} + ½)/(\sqrt{600} + 1)$. The sample average is unbiased and does not minimize mean square error. The MMR estimate has a small bias towards ½ and does minimize mean square error. The pooled estimate with sample design (6, 594) also has a small bias towards ½ and yields a smaller MSE than does the sample average with design (0, 600).



Note: The maximum regret curves depicted here ignore the fact that sample sizes N_1 and N_2 must be integers. However, the depicted MMR sample designs, identified by diamond markers with labelled pairs (N_1, N_2), impose this requirement.

Figure 3 allows the low-cost sampling response rate π to vary while holding the other parameters fixed at ($c_1 = 1$, $c_2 = 10$, $B = 600$). Maximum regret increases as the response rate decreases, holding the high-cost sample budget share fixed. Allocating the entire budget to low-cost sampling (600, 0) minimizes

maximum regret for any response rate $\pi > 0.92$. The optimal high-cost sample budget share increases as π decreases below 0.92. If we impose the requirement that sample sizes must be integers, allocation of the entire budget to high-cost sampling (0, 60) minimizes maximum regret for any $\pi \leq 0.68$.

Supplementary Section 3.7: Numerical Illustrations of Allocation of Budget to Both Sampling Processes, Using an Intersection Estimator as Predictor

To illustrate, we suppose that the outcome y takes the values {0, ½, 1}. We consider the predictor that sets the prediction equal to (i) the midpoint of the intersection interval when the two intervals intersect and (ii) the midpoint between the lesser upper bound and the greater lower bound when they do not intersect.

For any specified sample design (N_1, N_2), response rate P(z = 1), and outcome distributions [$P_s(y|z = 1)$, $P_s(y|z = 0)$], we use Monte-Carlo simulation to compute the regret of this predictor. To perform each replication of the simulation, we draw $N_1 + N_2 \cdot P(z = 1)$ observations from $P_s(y|z = 1)$ and $N_2 \cdot P(z = 0)$ observations from $P_s(y|z = 0)$. Using STATA software, an observation from $P_s(y|z = j)$, $j \in \{0, 1\}$, is generated by drawing a standard uniform pseudorandom variable x and letting y = 0 if $x < P_s(y = 0|z = j)$, y = ½ if $P_s(y = 0|z = j) \leq x < P_s(y = 0|z = j) + P_s(y = ½|z = j)$, and y = 1 otherwise. We discretize the state space by restricting the values of $P_s(y = 0|z = j)$ and $P_s(y = 0|z = j) + P_s(y = ½|z = j)$ to multiples of 0.1.

For each replication, we use the N_1 draws from $P_s(y|z = 1)$ to calculate the sample average m_1 and set b_1 equal to 1.96 times the sample standard deviation of y. Similarly, we use the $N_2 \cdot P(z = 1)$ draws from $P_s(y|z = 1)$ and $N_2 \cdot P(z = 0)$ draws from $P_s(y|z = 0)$ to calculate the average mean m_2 and set b_2 equal to 1.96 times the standard deviation of y in this sample. We then calculate the intersection estimate of μ.

We perform 500 replications for each specification of (N_1, N_2), P(z = 1), and [$P_s(y|z = 1)$, $P_s(y|z = 0)$]. This done, we compute the Monte Carlo estimates of means and variances needed to form the Monte Carlo estimate of regret in state s. We determine maximum regret over all states in the discretized state space. The 500 Monte Carlo replications and discretization of the state space by a grid with spacing 0.1

suffices to yield acceptably small approximation errors in our numerical analysis. With additional computational effort, we could reduce approximation error further by increasing the number of Monte Carlo replications and refining the grid of the discretized state space.

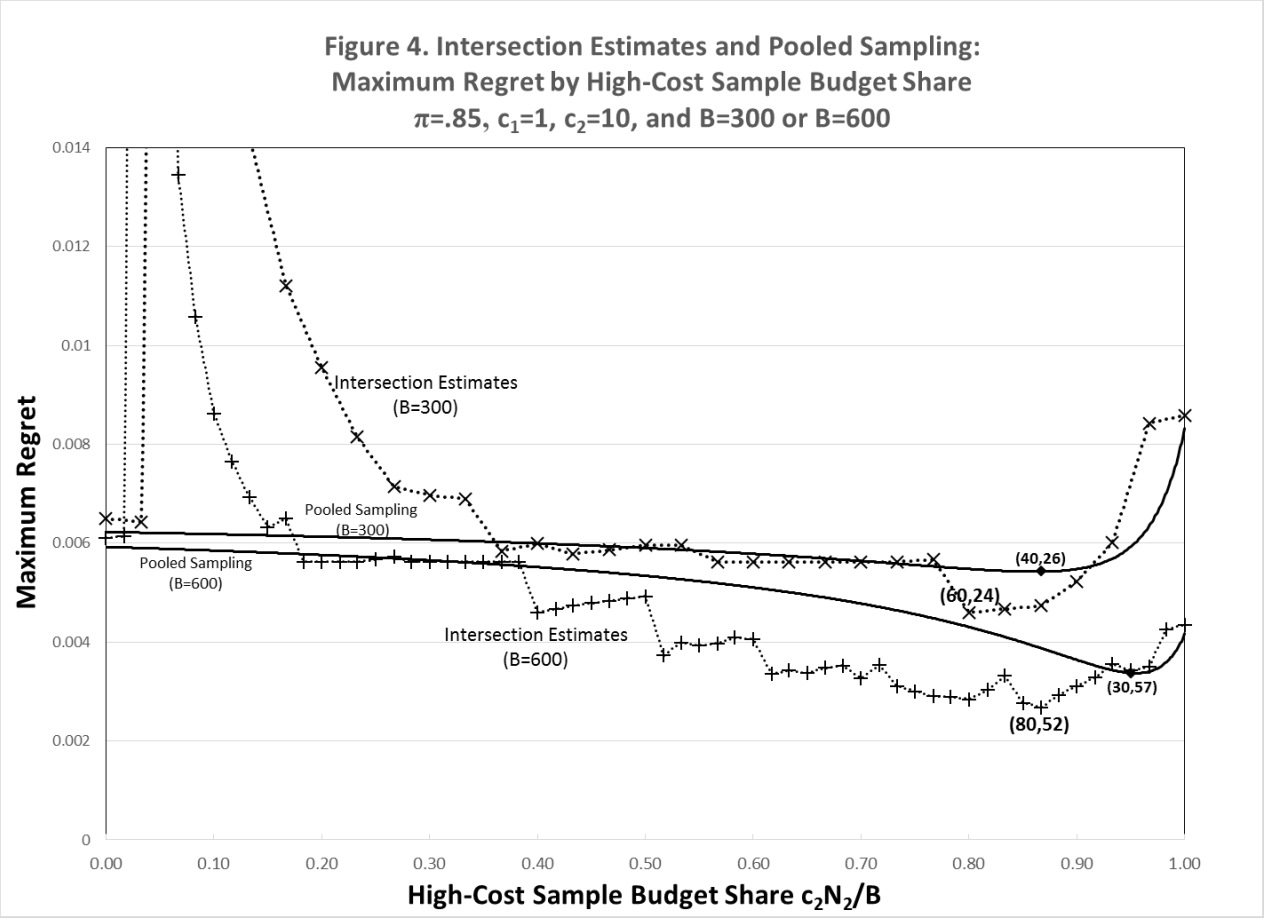


Figure 4 numerically compares the maximum regret of some design-predictor pairs that use data pooling and intersection estimates to make predictions. As earlier, we consider a setting where (π = .85, c_1 = 1, c_2 = 10) and perform calculations with two budgets examined previously, B = 300 and B = 600. We find that maximum regret based on intersection estimates of μ is usually smaller than maximum regret based on pooled estimates, especially with the larger budget and correspondingly larger sample sizes. For each budget, the maximum regret of the best sample design is lower when the intersection estimate is used to make the prediction than when the pooled estimate is used.

Consider the budget B = 300. Results using intersection estimates, marked with an “X”, are presented for each feasible sample design. The results using pooled estimates reproduce curves presented earlier. The MMR sample design using the intersection estimate is (60, 24) and the associated value of maximum regret is 0.0046. The MMR design using the pooled estimate is (40, 26) and its value of maximum regret is 0.0054.

Next consider B = 600. Results using intersection estimates, marked with a “+”, are presented for each feasible sample design. The MMR sample design using the intersection estimates is (80, 52) and yields maximum regret value 0.0027. The MMR design using pooled estimates is (30, 57) and yields maximum regret 0.0034.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, New York: Springer.

Binmore, K. (2009), *Rational Decisions*, Princeton: Princeton University Press.

Canner, P. (1970), "Selecting One of Two Treatments when the Responses are Dichotomous," *Journal of the American Statistical Association*, 65, 293-302.

Cheng, Y., F. Su, and D. Berry (2003), "Choosing Sample Size for a Clinical Trial Using Decision Analysis," *Biometrika*, 90, 923-936.

Chernoff, H. (1954), "Rational Selection of Decision Functions," *Econometrica*, 22, 422-443.

Chernozhukov, V., S. Lee, and A. Rosen (2013), "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 668-737.

Cochran, W., F. Mosteller, and J. Tukey (1954), *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*, Washington, DC: American Statistical Association.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, San Diego: Academic Press.

Griffin, D. (2011), "Cost and Workload Implications of a Voluntary American Community Survey: Final Report," 2011 American Community Survey Research and Evaluation Report Memorandum Series #ACS11-RER-01 American Community Survey Office, https://www.census.gov/content/dam/Census/library/working-papers/2011/acs/2011_Griffin_01.pdf, accessed September 24, 2016.

Groves, R. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 646–675.

Groves, R. and L. Lyberg (2010), "Total Survey Error: Past, Present, and Future," *Public Opinion Quarterly*, 74, 849-879.

Groves, R. and L. Magilavy (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," *Public Opinion Quarterly*, 50, 251-266.

Groves, R. and E. Peytcheva (2008), "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly*, 72, 167-189.

Hodges, E. and E. Lehmann (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics*, 21, 182-197.

Horowitz, J. and C. Manski (1998), "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37-58.

International Conference on Harmonisation (1999), "ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline," *Statistics in Medicine*, 18, 1905-1942.

Kreider, B. and J. Pepper (2007), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432-441.

Lengacher J., C. Sullivan, M. Couper, and R. Groves (1995), "Once Reluctant, Always Reluctant? Effects of Differential Incentives on Later Survey Participation in a Longitudinal Study," paper presented at the Annual Conference of the American Association for Public Opinion Research; Fort Lauderdale, FL. Manuscript accessed at https://www.amstat.org/sections/srms/proceedings/papers/1995_179.pdf.

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

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

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer.

Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2007), "Minimax-Regret Treatment Choice with Missing Outcome Data," *Journal of Econometrics*, 139, 105-115.

Manski, C. (2011), "Actualist Rationality," *Theory and Decision*, 71, 195-210.

Manski, C. (2015), "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern," *Journal of Economic Literature*, 53, 631-653.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables: with an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

Manski, C. and J. Pepper (2009), "More on Monotone Instrumental Variables," *The Econometric Journal*, 12, S200-S216.

Manski, C. and A. Tetenov (2016), "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, doi: 10.1073/pnas.1612174113.

Office of Management and Budget (2006). *Standards and Guidelines for Statistical Surveys* (September 2006), http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf, accessed September 16, 2016.

Philipson, T. (1997), "Data Markets and the Production of Surveys," *Review of Economic Studies*, 64, 47-72.

Philipson, T. (2001), "Data Markets, Missing Data, and Incentive Pay," *Econometrica*, 69, 1099-1111.

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46, 55-67.

Sen, A. (1993), "Internal Consistency of Choice," *Econometrica*, 61, 495-521.

Spencer, B. (1985), “Optimal Data Quality” *Journal of the American Statistical Association* 80, 564-573.

Spiegelhalter D, L. Freedman, and M. Parmar (1994), “Bayesian Approaches to Randomized Trials (with discussion),” *Journal of the Royal Statistical Society Series A*, 157, 357-416.

Stein, C. (1956), “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution,” Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, 1, 197-206.

Stoye, J. (2012), "Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 157-165.

Tetenov, A. (2012), “Measuring Precision of Statistical Inference on Partially Identified Parameters,” Collegio Carlo Alberto, Moncalieri (Torino), Italy.

U.S. Census Bureau (2013), *American Community Survey: Information Guide*, Economics and Statistics Administration, https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf, accessed September 24, 2016.

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