MORE DATA OR BETTER DATA? A Statistical Decision Problem

Jeff Dominitz
Resolution Economics

and

Charles F. Manski
Northwestern University

Review of Economic Studies, 2017

Summary

When designing data collection, questions arise regarding how much data to collect and how much effort to expend to enhance the quality of the 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 predetermined budget between two or more sampling processes.

All processes draw random samples from a population of interest and aim to collect data that are informative about the sample realizations of an outcome.

The processes differ in the cost of data collection and quality of the data obtained.

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, the high-cost 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.

In the other, it provides a low-resolution interval measure of outcomes.

We study minimax-regret prediction of a real outcome under square loss when the decision maker imposes no assumptions that restrict the unobserved outcomes.

Background

Cochran, Mosteller, and Tukey (1954) considered an instance of the design decision.

They assessed the methodology of the Kinsey study of male sexual behavior.

They compared 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: “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.”

They considered the mean square error of an estimate of the mean.

They recognized that, with no knowledge of the process generating nonresponse, obtaining an interview from a new randomly drawn subject only reduces variance.

Obtaining an interview from a non-respondent sample member reduces both variance and maximum potential bias.

Horowitz and Manski (1998) and Tetenov (2012) provide further analysis, using the modern framework of partial identification analysis.

Using Statistical Decision Theory to Choose a Sample Design and Predictor

Optimal Prediction (Optimal Choice of a Standard Treatment)

Consider a planner who must choose a standard treatment for members of population J, formally a probability space (J, Ω, P) with $P(j) = 0$, all $j \in J$.

The set of feasible treatments is T, a subset of the real line.

Each $j \in J$ has an optimal treatment value, denoted y_j.

A loss function $L_j(y_j - t)$ gives the social cost of assigning treatment t to person j.

The unconstrained optimal plan assigns each person his optimal treatment.

The planner faces a constrained problem, in which every person must receive the same treatment.

The mean cost that results if everyone is assigned treatment t is $E[L(y - t)]$.

The planner wants to solve the problem $\min_{t \in T} E[L(y - t)]$.

He can solve this problem if he knows $P(y)$.

Prediction with Sample Data

The planner does not know P(y).

He can use a budget B to draw persons at random and attempt to measure the outcome of each sampled person. He then uses the data to choose a point prediction.

Let two sampling processes be available, labelled 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.

The planner faces a joint problem of sample design and choice of a predictor.

The design alternatives are feasible values for the sample sizes (N_1, N_2).

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.

Samples of size (N_1, N_2) yield data $\psi = (\psi_{1k}, k = 1, \ldots, N_1; \psi_{2k}, k = 1, \ldots, N_2)$.
Ψ denotes the sample space indexing all possible data realizations.

A predictor is a function $\delta(\cdot)$: $\Psi \rightarrow T$.

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 under design (N_1, N_2).

The risk of design-predictor pair $[(N_1, N_2), \delta]$ is

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

$$= \int \int L[y - \delta(\psi)] dP(y) dQ(\psi; N_1, N_2).$$

Evaluation of risk is possible if one knows $P(y)$ and $Q(N_1, N_2)$.

We study decision making with incomplete knowledge of these distributions.

Let the feasible distributions be $(P_s, Q_s, s \in S]$. S is called the state space in decision theory and the parameter space in statistics.

In principle, one can compute state-dependent risk

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

Choosing a Design-Predictor Pair

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

One first eliminates inadmissible (weakly dominated) options.

One then uses some criterion to choose among the set D of admissible options.

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

Note: One might skip the step of determining admissibility and use a decision criterion to choose among all feasible options, not just those that are admissible.

Computation

Use of statistical decision theory to choose a sample design and predictor is simple in principle, but it can be difficult in practice.

Monte Carlo simulation enables computation of risk in a specified state.

Risk in state s is the expected value of $L[y - \delta(\psi)]$ over (y, ψ), which are statistically independent with distributions $P_s(y)$ and $Q_s(\psi; N_1, N_2)$.

Risk can be approximated by drawing multiple realizations of (y, ψ), computing $L[y - \delta(\psi)]$, and averaging the results.

Risk and Regret under Square Loss

The optimal prediction with square loss is E(y). Let $\mu_s = E_s(y)$ and $\lambda_{\delta s} = E_s[\delta(\psi)]$.

It can be shown that risk and regret in state s are

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

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

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

Low-Cost Sampling with Nonresponse

Background

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

where $z = 1$ if a person's outcome is observable and $z = 0$ if not. A sampling process with nonresponse partially identifies $P(y)$, revealing that

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

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

Suppose that an agency is designing a new survey to be administered for a predetermined total budget.

Two vendors submit bids to conduct the survey.

One proposes a low-cost sampling process with a known positive rate of nonresponse and a large sample size. The other proposes a high-cost sampling process with full response but a smaller sample.

We derive the MMR choice between these sampling processes under the assumption that the planner will use specific reasonable predictors.

The analysis generalizes 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.

Minimax-Regret Analysis under Square Loss

We maintain several assumptions that simplify analysis.

The planner knows the response rate P(z = 1) with low-cost sampling.

y and t take values in the unit interval.

With low-cost sampling, the cost c_1 per sample member is incurred when an outcome is observed rather than when observation of an outcome is attempted. Hence, N_1 is the number of outcomes that will be observed.

The outcome data observed with sample design (N_1, N_2) are

$$\psi = (y_{1k}, k = 1, \ldots, N_1; y_{2k}, k = 1, \ldots, N_2).$$

The state space is $[P_s(y|z = 1), P_s(y|z = 0), s \in S] = \Gamma \times \Gamma$.

Each pair $[P_s(y|z = 1), P_s(y|z = 0)]$ determines a unique population outcome distribution $P_s(y)$.

In this setting, the MMR predictor is known when only high-cost data are available; that is, when $N_1 = 0$ and $N_2 > 0$. (Hodges and Lehman, 1950).

The MMR predictor does not have a known explicit form 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.

We first consider these polar cases and then consider the general design decision, where the planner chooses a (N_1, N_2) pair that satisfies the budget constraint.

Prediction with Only High-Cost Sampling

Let m_2 denote the sample average value of the N_2 observations of y. Hodges and Lehmann (1950) prove that

the MMR prediction is $(m_2\sqrt{N_2} + ½)/(\sqrt{N_2} + 1)$,

the minimax value of regret is $1/[4(\sqrt{N_2} + 1)^2]$.

The conventional estimate of a population mean under random sampling is the sample average m_2. It has maximum regret $1/(4N_2)$.

Prediction with Only Low-Cost Sampling

The MMR predictor and minimax value of regret do not have known forms.

We study a simple predictor, the midpoint of a sample estimate of the interval that forms the identification region for the optimal prediction E(y).

This is a reasonable choice because, if the identification interval were known rather than estimated, its midpoint would be the minimax-regret prediction.

Low-cost sampling reveals that E(y) lies in the interval

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

Let m_1 be the average of the observed N_1 outcomes. The interval estimate is

$$[m_1P(z = 1),\quad m_1P(z = 1) + P(z = 0)].$$

We consider use of $m_1P(z = 1) + ½P(z = 0)$ as the predictor.

It can be shown that the maximum regret of this predictor is

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

Choosing Between the Low-Cost and High-Cost Designs

We first examine the constrained setting in which the planner must choose between one of the two polar designs, intermediate options not being feasible.

With marginal sampling costs (c_1, c_2) and budget B, the feasible sample sizes are $N_1 = \text{INT}(B/c_1)$ for low-cost sampling and $N_2 = \text{INT}(B/c_2)$ 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 design from the MMR perspective is the one that minimizes maximum regret when 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 $m_1 P(z = 1) + \frac{1}{2} P(z = 0)$ for low-cost sampling and m_2 for high-cost sampling.

With these predictors, the feasible low-cost and high-cost designs yield maximum regret $\frac{1}{4}[P(z = 1)^2(c_1/B) + P(z = 0)^2]$ and $\frac{1}{4}(c_2/B)$, respectively.

Hence, the low-cost or high-cost design yields smaller maximum regret if

$$P(z = 1)^2 c_1 + BP(z = 0)^2 < c_2,$$
$$P(z = 1)^2 c_1 + BP(z = 0)^2 > c_2.$$

The threshold budget is

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

This finding generalizes to choice among multiple sampling processes that differ in their costs and response rates.

Let a set Q of sampling processes be 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 minimizes $P_q(z = 1)^2(c_q/B) + P_q(z = 0)^2$.

Allocation of Budget to Both Sampling Processes, with Data Pooling

Suppose 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 reasonable ways to choose a predictor combining the data from both samples, but computation of maximum regret is burdensome.

We focus on a particular predictor for which MMR computation is tractable.

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.

Pooling is easy to study because the results obtained above apply 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.

Assume for simplicity that the realized response rate equals the population response rate.

N_1/π is the total size of the low-cost sample drawn to obtain N_1 responses.

The response rate in the pooled sample is $(N_1 + N_2)/(N_1/\pi + N_2)$.

The predictor is the interval-estimate midpoint

$$m_{12}[(N_1 + N_2)/(N_1/\pi + N_2)] + \tfrac{1}{2}[(N_1/\pi - N_1)/(N_1/\pi + N_2).$$

Maximum regret with a given sample design (N_1, N_2) is

$$[2(N_1/\pi + N_2)]^{-2}[(N_1 + N_2) + (N_1/\pi - N_1)^2].$$

The design that minimizes maximum regret chooses (N_1, N_2) to solve the problem

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

The optimal design exhausts the budget. Hence, we can set $N_2 = (B - c_1N_1)/c_2$ and rewrite this as the one-dimensional minimization problem

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

Numerical calculations are instructive.

Figure 1. Pooled Sampling:
Maximum Regret by High-Cost Sample Budget Share and Budget
π=.85, c_1=1, c_2=10

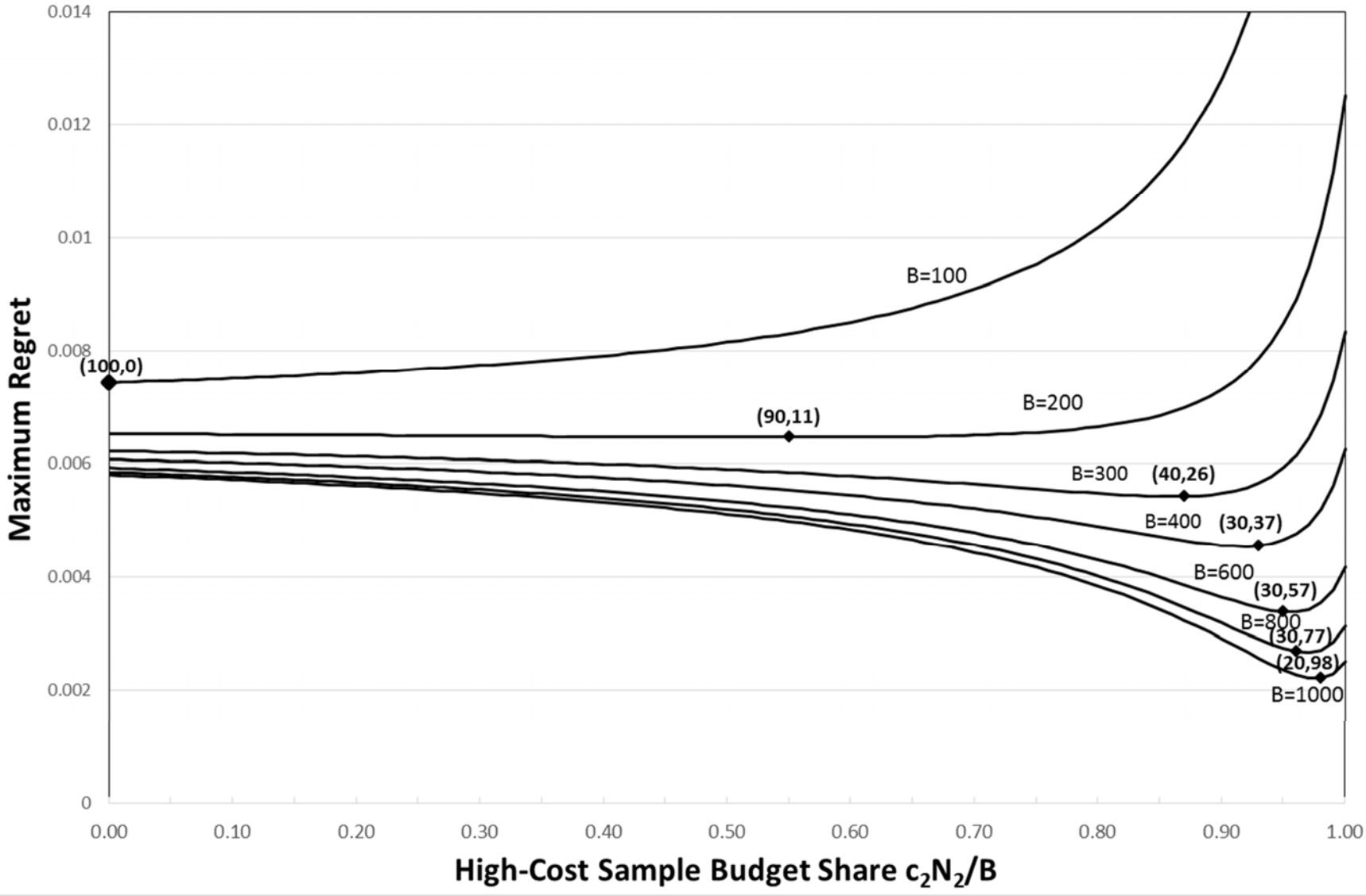


Figure 2. Pooled Sampling:
Maximum Regret by High-Cost Sample Budget Share and Marginal Cost
π=.85, c_1=1, B=600

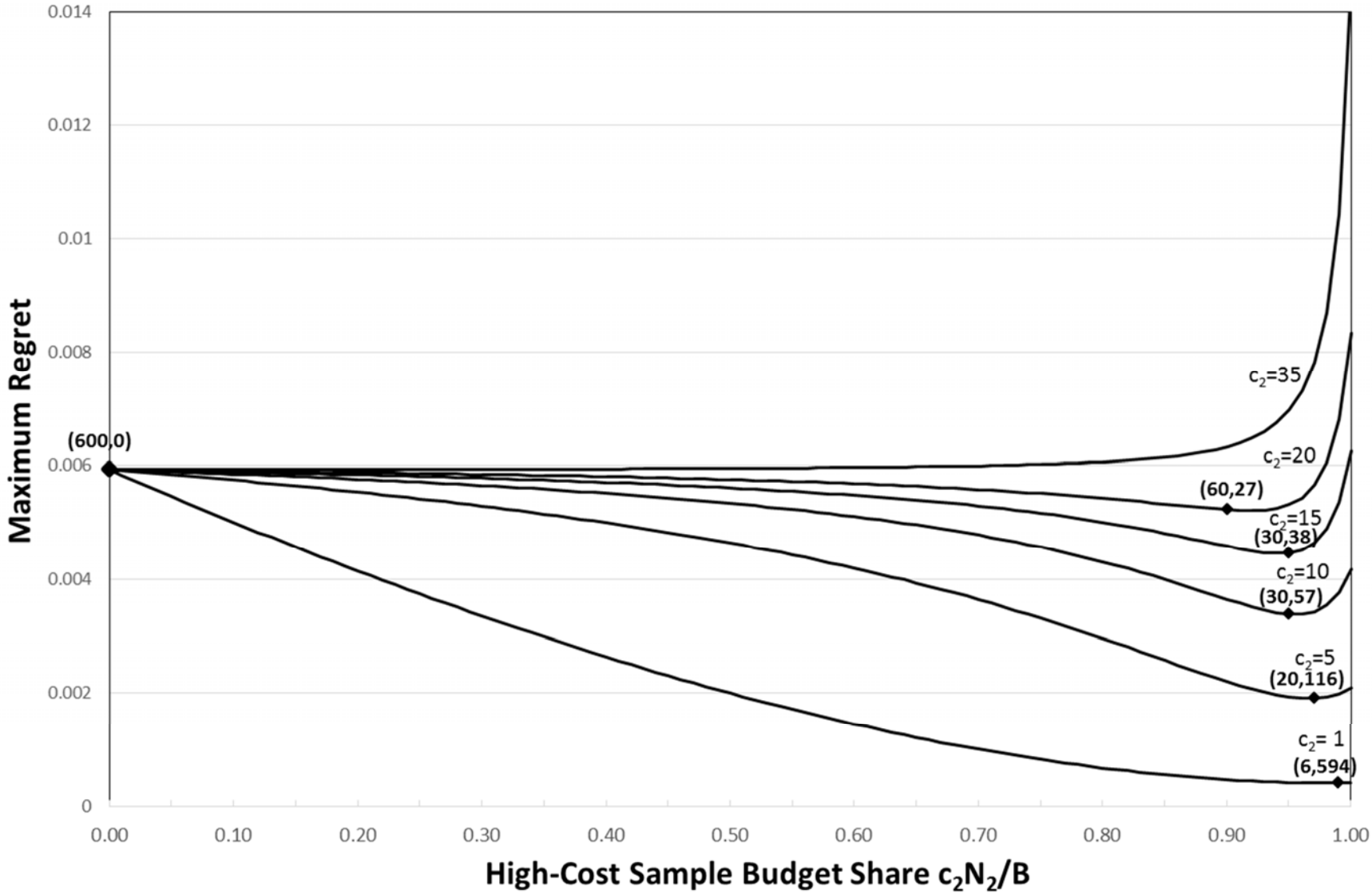
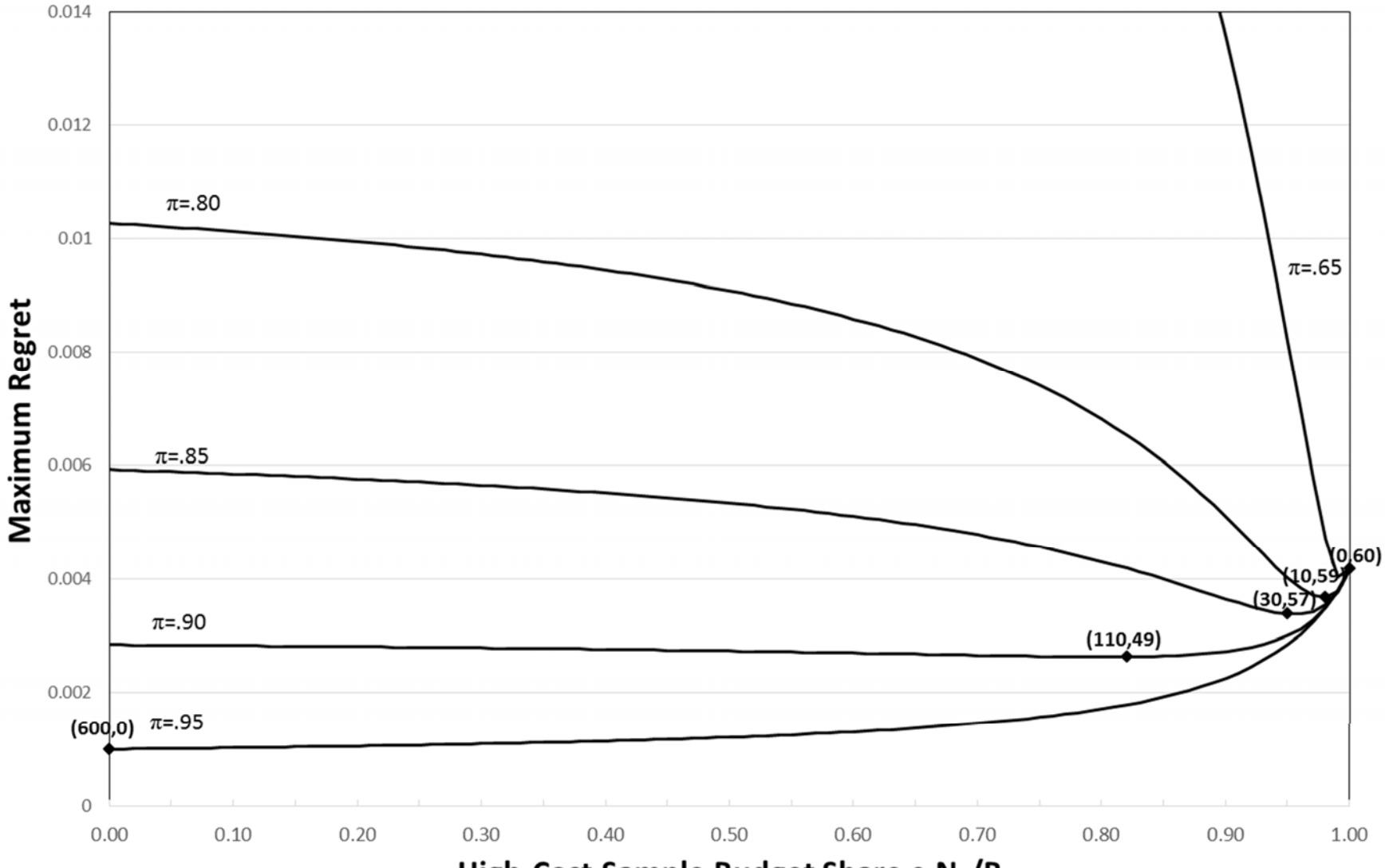


Figure 3. Pooled Sampling:
Maximum Regret by High-Cost Sample Budget Share and Low-Cost Response Rate
c_1=1, c_2=10, B=600



Allocation of Budget to Both Processes with an Intersection Estimate as Predictor

Pooling 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.

We focus on rules that use “intersection estimates” as the predictor.

As earlier, let m_1 and m_2 be the average values of y observed using the low-cost and high-cost sampling processes.

The two samples yield interval and point estimates of μ, namely

$$[m_1 P(z = 1),\ m_1 P(z = 1) + P(z = 0)] \quad \text{and} \quad m_2.$$

Conventional confidence intervals for m_1 and m_2 have the form

$$[m_1 - b_1/\sqrt{N_1},\ m_1 + b_1/\sqrt{N_1}] \cap [0, 1] \quad \text{and} \quad [m_2 - b_2/\sqrt{N_2},\ m_2 + b_2/\sqrt{N_2}] \cap [0, 1],$$

where $b_1 > 0$, $b_2 > 0$. This suggests an interval estimate of μ of the form

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

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

We consider the predictor that equals

(i) the midpoint of the intersection interval when the two intervals intersect,

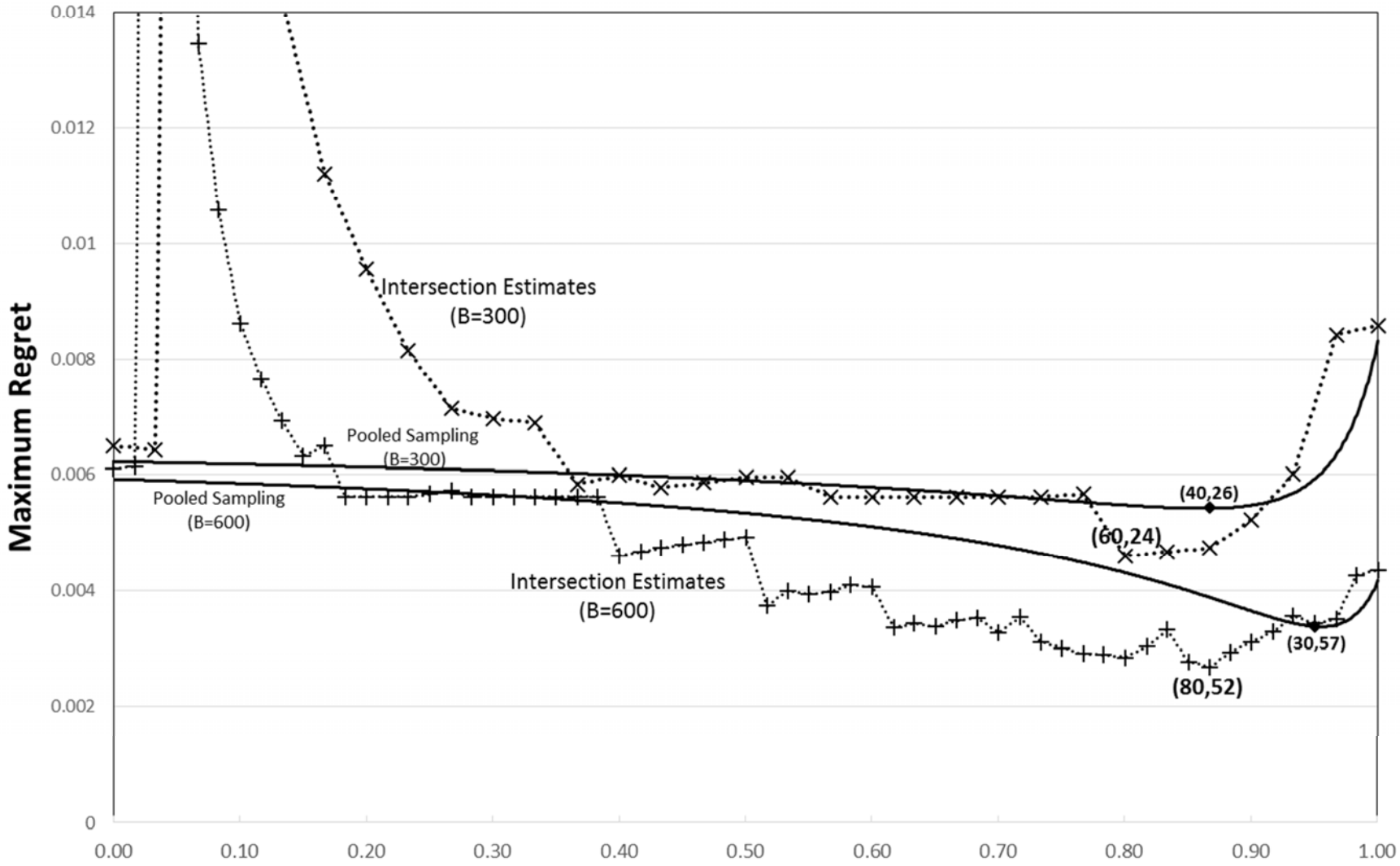
(ii) the midpoint between the lesser upper bound and the greater lower bound when the intervals do not intersect.

For any specified design (N_1, N_2), response rate $P(z = 1)$, and outcome distributions [$P_s(y|z = 1)$, $P_s(y|z = 0)$], Monte-Carlo simulation can be used to compute the regret of this predictor in state s.

Maximum regret can be approximated by discretizing the state space.

To illustrate, we suppose that y has support {0, ½, 1}. We set b_1 and b_2 equal to 1.96 times the respective sample standard deviation of y.

Figure 4. Intersection Estimates and Pooled Sampling: Maximum Regret by High-Cost Sample Budget Share
π=.85, c_1=1, c_2=10, and B=300 or B=600



Low-Cost Sampling with Interval Measurement of Outcomes

Low-Resolution Interval Measurement

Let y and t take values in the unit interval.

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, . . . , N_1, one observes a sub-interval of [0, 1] that contains y_k.

It appears difficult to characterize the maximum regret of design-predictor pairs when low-cost sampling produces general forms of interval measurement.

Progress is possible in special cases of practical importance.

One is data with nonresponse. Here only two types of intervals occur: $[y_k, y_k]$ when the outcome is observable and $[0, 1]$ when the outcome is unobservable.

Another uses a low-resolution measurement device that locates each value of a continuous outcome within a given 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 case of equal-length closed intervals, each of length 1/M.

Thus, the intervals are $I_m = [(m - 1)/M, m/M]$, $m = 1, \ldots, M$.

Example: Low-resolution measurement of a patient's optimal drug dose relative to a specified maximum dose may place it in one of four intervals [0, ¼], [¼, ½], [½, ¾], [¾, 1].

Prediction with Low-Cost Sampling

We earlier 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.

We study a simple predictor, the midpoint of the sample analog estimate of the identification region for the optimal prediction E(y). This is $\sum_m [(m - ½)/M]p_{1m}$.

It can be shown that the maximum regret of this predictor is

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

Choosing Between the Low-Cost and High-Cost Designs

Suppose that one must choose between the low and high-cost designs, using predictor $\sum_m [(m - ½)/M]p_{1m}$ or m_2.

Then maximum regret is $¼M^{-2}[(M - 1)^2(c_1/B) + 1]$ or $¼(c_2/B)$.

The low-cost or high-cost design yields smaller maximum regret if

$$M^{-2}[(M - 1)^2 c_1 + B] < c_2, \qquad M^{-2}[(M - 1)^2 c_1 + B] > c_2.$$

The threshold budget is $B = c_2M^2 - (M - 1)^2 c_1$.

Conclusion

We hope that this paper will encourage use of statistical decision theory to inform the design of data collection when data quality is a decision variable.

We have considered two settings: nonresponse and interval measurement.

We note that these phenomena may interact.

Surveys often elicit interval data on real outcomes in order to reduce rates of item nonresponse that arise for questions about sensitive topics. Increasing the resolution may yield increased nonresponse.