Bayesian optimal sample design for surveys with heteroscedasticity

Jonathan Mendelson* Michael R. Elliott †

July 2026

Abstract

We develop a Bayesian optimal sample allocation approach for stratified sampling in heteroscedastic populations. Existing optimal allocation theory typically assumes knowledge of certain design parameters (e.g., strata variances) that may be unknown, leading practitioners to substitute in survey-based estimates when planning samples, often without considering the effects of this substitution on sample efficiency. Bayesian decision theory for optimal experimental design avoids such substitutions and can be applied to sample allocation. Bayesian sample optimization methods were studied heavily from the mid-1960s through early 1980s, but have been overlooked since, in spite of modern computing advances that have facilitated a proliferation of Bayesian methods in other areas of statistics. A limitation of this early Bayesian sample design work is that it did not accommodate heteroscedastic error structures, which underlie commonly used ratio estimation models. Our paper, which optimizes the design under a univariate regression model with heteroscedastic errors, addresses this limitation of earlier work, while illustrating the Bayesian approach to design. We identify the optimal Bayesian allocation under our model, then compare performance of the proposed Bayesian sampling strategy with that of key design-based and model-assisted alternatives across several settings, finding that the proposed methods do as well or better than the alternatives under the scenarios considered. We apply our methods in analyzing revenues of public charities, using publicly available IRS Form 990 data.

Keywords: sample allocation, optimal allocation, stratified sampling, Bayesian decision theory, measure of heteroscedasticity

*Jonathan Mendelson is Research Statistician at the US Bureau of Labor Statistics, Suitland, MD 20746, USA.

†Michael R. Elliott is Professor of Biostatistics and Research Professor of Survey Methodology at the University of Michigan, Ann Arbor, MI 48109, USA.

The authors would like to thank Joseph Sedransk, Eric Slud, Richard Valliant, Partha Lahiri, and Roderick Little, whose suggestions greatly improved the quality of this manuscript. Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

Address correspondence to Jonathan Mendelson, Office of Survey Methods Research, U.S. Bureau of Labor Statistics, 4600 Silver Hill Road, Suitland, MD 20746, USA; E-mail: mendelson.jonathan@bls.gov.

1 Introduction

Survey statisticians have long been interested in optimal designs, wherein optimality refers to minimizing some objective function (e.g., variance of a finite population total) subject to certain constraints (e.g., maximum costs or sample size). Optimizing survey designs has a long history dating back at least to Neyman (1934), who recognized that sampling strata proportional to their standard deviation was more efficient than sampling proportional to their population sample size if strata variances differ. Neyman allocation was subsequently extended to account for differential data collection costs across strata (e.g., Cochran 1977).

Typically, the theoretically optimal sample allocations are conditional on population characteristics that are unknown (e.g., strata variances for Neyman allocation). In practice, such design parameters are often estimated from prior surveys, after which the estimates are substituted in place of the true quantities in determining the design.

Despite vast spending on sample surveys (Office of Management and Budget 2017), there is only a modest amount of recent literature on designing samples in the presence of uncertain design parameters. From a frequentist perspective, Clark (2013) proposed a statistical learning approach for allocating stratified random samples, which led to improvements, although Clark noted that Bayesian methods may be more efficient. There is a sizable Bayesian literature on optimal survey sample design, largely published from the mid-1960s through early 1980s. A common approach entails deriving the posterior loss (e.g., posterior variance) conditional on the data, averaging over the predictive distribution for the data, and then selecting the sample design that minimizes this expected posterior loss (e.g., Draper and Guttman 1968b; Rao and Ghangurde 1972). Given that the Bayesian optimal sample design literature predates recent modern computing advances, this approach typically requires a closed form for the expected posterior loss, which might not be possible for some models. Further, the models employed may sometimes be unrealistic. For example, Draper and Guttman (1968b) consider optimal allocation for stratified random sampling, but employ a normal likelihood with fixed strata means and variances, which does not accommodate heteroscedastic errors. In contrast, Rao and Ghangurde (1972) examine optimal design for categorical data via a Dirichlet-multinomial model, which may be useful in several contexts, but not necessarily for settings in which distributions are continuous specification and possibly skewed.

We examine stratified random sampling for surveys wherein the stratum means and possibly variances are functions of a known population-level auxiliary variable. We assume that some pilot data set is available for estimating parameters needed for designing a main study. Our aim to identify optimal sample allocations for given strata; this is in contrast to a focus on identifying optimal strata boundaries for a size-related measure (e.g., Wright 1983; Dalenius and Hodges 1959; Rivest 2002). Although our focus is on establishment surveys that often have large degrees of heterogenity (Henry and Valliant 2009; Smith, Pont, and Jones 2003), the methods developed here are relevant to any settings with stratified sampling with heterogeneous variances within strata.

We formulate our problem using Bayesian decision theory on optimal experimental design (e.g., Lindley 1972), which is a general and flexible approach that can accommodate analyses such as those of Draper and Guttman (1968b) and Rao and Ghangurde (1972), and

which accounts for uncertain estimates of design parameters. We assume that inferences are made using a Bayesian framework for finite population inference, following Ericson (1969a). We derive approximate analytical results for the expected posterior variance for the finite population mean, then identify an allocation using standard convex optimization methods.

The remainder of this manuscript is organized as follows. Background is provided in Section 2, and includes attention to stratified sample allocation under our assumed model, sampling with uncertain design parameters, and Bayesian optimal experiment design. Section 3 presents analytical results and shows how to identify the Bayesian optimal design for estimating the finite population mean, assuming the proposed model and a squared error loss function. Section 4 outlines simulation methods for comparing the proposed allocations with alternatives across 90 artificial populations. Section 5 presents simulation findings. Section 6 repeats the simulation in an application to the log-revenues of public charities in the United States. Section 7 summarizes results, discusses limitations, and suggests future directions.

2 Background

2.1 Design-based approaches for optimal allocation using ratio estimators

2.1.1 Optimal allocation for the separate ratio estimator

We consider estimation of a population total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$, in a context where we know values for a related variable, $\{X_{hi}\}$, for the population. Our sample is chosen via an stratified simple random sampling (STSRS) design.

Assume that within each stratum, the $\{Y_{hi}\}$ and $\{X_{hi}\}$ are correlated, but with correlations that may vary by strata. For convenience, let $\{y_{hi}, x_{hi}\}$ refer to the values of the $n = \sum_{h=1}^{H} n_h$ sampled units for variables $\{Y_{hi}, X_{hi}\}$. Let $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ and $\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$ denote the stratum h sample means and $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ and $X_h = \sum_{i=1}^{N_h} X_{hi}$ denote the stratum population totals of our outcome and auxiliary variables, respectively. Let $R_h = \frac{X_h}{Y_h}$ be the ratio of stratum population totals. This scenario commonly employs the separate ratio estimator of the population total,

$$\hat{Y}_{Rs} = \sum_{h=1}^{H} \hat{R}_h X_h \tag{1}$$

$$= \sum_{h=1}^{H} \frac{\bar{y}_h}{\bar{x}_h} X_h, \tag{2}$$

termed *separate* (versus *combined*) by virtue of allowing the strata slopes to vary. Cochran

(1977, sec. 6.14) indicated that the optimal allocation for the separate ratio estimator is

$$n_h \propto \frac{N_h S_{dh}}{\sqrt{c_h}} \tag{3}$$

where $S_{dh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} d_{hi}^2$ and $d_{hi} = Y_{hi} - \frac{Y_h}{X_h} X_{hi}$ is the deviation of Y_{hi} from $\frac{Y_h}{X_h} X_{hi}$. This allocation is derived by selecting $\{n_h\}$ to minimize the approximate variance of the separate ratio estimator, $\text{Var}\left(\hat{Y}_{Rs}\right) \doteq \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_{dh}^2}{n_h}$, the latter of which assumes large sample sizes in each stratum.

Cochran did not explicitly provide an estimator for S_{dh}, which he suggested is difficult to speculate about in the context of sample planning (p. 172). Instead, he used the theoretically optimal allocation of Equation (3) to justify approximate solutions tailored to the type of population, based on the relationship between S_{dh} and $\bar{X}_h$. For scenarios in which the ratio estimate is a best linear unbiased estimate (BLUE), S_{dh} will be roughly proportional to $\sqrt{\bar{X}_h}$, suggesting an allocation of $n_h \propto N_h \sqrt{\bar{X}_h} / \sqrt{c_h}$. In populations wherein the d_{hi} are more closely proportional to $\bar{X}_h^2$, Cochran suggested an allocation of $n_h \propto N_h \bar{X}_h / \sqrt{c_h}$.

Although not mentioned by Cochran, if prior data are available, an alternative to these rules of thumb would be to employ a plug-in allocation of the form $n_h \propto N_h \hat{S}_{dh} / \sqrt{c_h}$, where $\hat{S}_{dh}$ is a sample-based estimate from some previous data. One such estimator is

$$\hat{S}_{dh}^2 = \frac{1}{m_h - 1} \sum_{i \in s_{1h}} \left(y_{hi} - \hat{R}_h x_{hi} \right)^2, \tag{4}$$

where $\hat{R}_h = \frac{\sum_{i \in s_{1h}} y_{hi}}{\sum_{i \in s_{1h}} x_{hi}}$ is a ratio estimate from a pilot sample in stratum h of size m_h, and where s_{1h} refers to the pilot sample in stratum h. Although possibly reasonable in some contexts, this estimate of S_{dh}^2 has bias of order $1/m_h$ (Cochran 1977, 32), suggesting worsened performance when using smaller pilot samples.

2.1.2 Heteroscedasticity and optimal sample design

Although our ultimate focus is on Bayesian optimal design, here, we review frequentist methods for handling heteroscedasticity in sample design. Our eventual focus will be on optimal sample allocation for STSRS designs, assuming a stratified population model that follows $Y_{hi} | (X_{hi}, \alpha_h, v_h, b) \sim N\left(\alpha_h X_{hi}, v_h X_{hi}^b\right)$ for known $X_{hi} > 0$, unknown $\{\alpha_h, v_h\}$, and known $b \geq 0$, and where strata are fixed upfront. Although the key texts by Cochran (1977) and Särndal, Swensson, and Wretman (1992) do not directly provide a solution, Särndal, Swensson, and Wretman address optimal design for a related model under somewhat different problem formulations. For instance, Särndal, Swensson, and Wretman provide optimal unit-level selection probabilities for the generalized regression estimator (GREG) assuming heteroscedastic errors; separately, they also describe model-based optimal stratification rules.

As implied by Cochran's rules of thumb, populations can differ with respect to the level of heteroscedasticity. Consider an unstratified superpopulation model of N units, $\{X_i, Y_i : i =$

$1, ..., N\}$, where

$$Y_i = \beta X_i + \varepsilon_i \tag{5}$$
$$\mathrm{E_M}(\varepsilon_i) = 0 \tag{6}$$
$$\mathrm{Var_M}(\varepsilon_i) = \sigma^2 X_i^b \tag{7}$$

for known $X_i > 0$, independent ε_i's, and known $b \geq 0$, and where $\mathrm{E_M}(.)$ and $\mathrm{Var_M}(.)$ denote the expectation and variance with respect to the model. If $b = 1$, S_{dh} will be roughly proportional to $\sqrt{\bar{X}_h}$, which could justify use of Cochran's allocation of $n_h \propto N_h\sqrt{\bar{X}_h}/\sqrt{c_h}$. In contrast, $b = 2$ could lead to Cochran's allocation of $n_h \propto N_h\bar{X}_h/\sqrt{c_h}$. As cited by Henry and Valliant (2009), b has sometimes been referred to as a 'measure of heteroscedasticity' (Foreman 1991) or 'coefficient of heteroscedasticity' (Brewer 2002). The above model, as well as a multivariate generalization thereof, is commonly applicable in establishment contexts, with common values of $0 \leq b \leq 2$ (e.g., Valliant, Dorfman, and Royall 2000, 177).

The level of heteroscedasticity has meaningful implications for sample allocation. Under a probability proportional to size (PPS) design with constant per-unit costs, the optimum unit-level selection probabilities for GREG under this variance structure are $\pi_i \propto X_i^{b/2}$ (e.g., Särndal, Swensson, and Wretman 1992, sec. 12.2). This result, which applies to ratio estimation as a special case, is derived by considering the estimation error $(\hat{Y} - Y)$ for the population total, the latter of which is treated as a random variable (as opposed to the usual design-based treatment of Y as being fixed); Särndal, Swensson, and Wretman define an optimal design as a choice of first-order selection probabilities that minimizes the approximate anticipated variance of $\hat{Y} - Y$ for a given expected sample size, where anticipated variance is defined by Isaki and Fuller (1982) and is the expected variance with respect to the sampling design and the superpopulation model. These results can be further generalized by replacing the assumption of Equation (7) with the weaker assumption that $\mathrm{Var_M}(Y_i) = v_i$ for known and positive v_i, leading to an optimal allocation of $\pi_i \propto \sigma_i$, a result that Valliant, Dever, and Kreuter (2013, 53) attribute to Godambe and Joshi (1965) and Isaki and Fuller (1982) for the univariate and multivariate cases, respectively. These results can also be used to justify equal aggregate stratification rules (e.g., Valliant, Dorfman, and Royall 2000, 189).

Although the model expressed by Equations (5–7) has many applications, we will subsequently loosen the assumptions to accommodate group differences that do not necessarily depend on size. Of particular note, the assumption that model variances are of the form $\mathrm{Var_M}(\varepsilon_i) = \sigma^2 X_i^b$ accommodates heteroscedasticity, but could potentially lead to poor results if the proportionality constant, σ^2, varies by group. We also will subsequently allow for slopes to vary by group. In an establishment survey, the slopes and/or the variance proportionality constant could potentially vary by characteristics other than the size measure, such as industry, firm structure, and/or geography.

2.2 Sampling with uncertainty

There are many works on optimal sample design from a Bayesian perspective, largely published from the mid-1960s through the early 1980s. This includes the development of optimal STSRS designs for estimating population totals (Draper and Guttman 1968b; Ericson 1965,

1967a, 1969b, 1972; Rao and Ghangurde 1972; M. Khan 1976) and proportions (Zacks 1970; Grosh 1972). Some of these articles entailed the use of normally distributed data—for instance, Draper and Guttman (1968b) assumed that the population values within a stratum were normally distributed with separate means and variances, and Ericson (1965) made the somewhat weaker assumption that the strata sample means, conditional on the population means, were normally distributed. An exception is Rao and Ghangurde (1972), who employed a Dirichlet-multinomial model. In aiming to minimize the expected posterior loss (e.g., posterior variance), a common approach is to employ a preposterior analysis, which entails computing the posterior distribution (conditional on the data), but then averaging over unknown quantities, since the data have not yet have been collected (e.g., Draper and Guttman 1968b; Rao and Ghangurde 1972). After this point, optimization methods (e.g., Lagrange multipliers or convex optimization methods) can be used to determine the optimal strata sample sizes. Although much of the literature focuses on optimization for estimation of a single quantity, there have been some extensions to multivariate problems (Draper and Guttman 1968a; M. Z. Khan 1976; Sekkappan 1981a, 1981b; Soland 1967). Bayesian optimal allocation methods have also been extended to more complicated sample designs, such as multi-stage sampling (Ericson 1973, 1975; Rao and Ghangurde 1972), double sampling for nonresponse (Ericson 1967b; Rao and Ghangurde 1972; Singh and Sedransk 1978, 1984), and multi-phase sampling (Smith and Sedransk 1982; Jinn, Sedransk, and Smith 1987). Finally, Ericson (1983a, 1983b) provides some results on optimal allocation between two modes wherein one mode is more expensive but provides higher quality measurements. Summary papers that include some attention to optimal Bayesian sample design are provided by Solomon and Zacks (1970) and Ericson (1985, 1988).

Since 1990, there has been little attention to Bayesian optimal sample designs. Much of the recent work on Bayesian survey statistics is on inferential topics, such as small area estimation (SAE), accounting for complex sample designs, or handling missing data in analyses. In fact, in examining recent summary articles or texts on Bayesian statistics in the context of surveys (e.g., Ghosh 2009; Rao 2011; Gelman et al. 2014, 229–230, 259), and in conducting literature searches, we did not find mentions of Bayesian methods for optimal sample allocation beyond those cited above, with one exception by O'Malley and Zaslavsky (2016) in the context of SAE. There has also been recent Bayesian literature on adaptive or responsive survey design, but the focus there has been on developing priors (Coffey et al. 2020; West et al. 2021) or improving estimates of unknown design parameters (Schouten et al. 2018; Wagner et al. 2020) rather than directly allocating samples, as we aim to do here.

2.3 Bayesian decision theory for optimal design

Wald (1950) introduced statistical decision theory, which adapted game theory to the formal analysis of making optimal decisions under uncertainty. Following Raiffa and Schlaifer (1961), Lindley (1972, 19–21) provided the Bayesian decision theoretic framework on optimal experimental design. Lindley treats optimal experimental design as a two-part decision: first, the choice of experiment (e.g., sample allocation for an STSRS design), which results in data; second, the choice of how to translate the data into a terminal decision (e.g., compute an estimate), and which, for some given value of the parameter, results in a loss defined by the function $L\left(\hat{\theta}, \theta, e, x\right)$. Under this framework, we are interested in the optimal experiment

and estimate that maximize the expected utility, or equivalently, minimize the expected loss.

In a continuous context, the optimal solution that minimizes the expected value of the loss is:

$$\min_{e} \int_X \min_{\hat{\theta}} \int_\Theta L\left(\hat{\theta}, \theta, e, x\right) p(\theta|x, e)p(x|e)\, d\theta\, dx \tag{8}$$

for loss function $L\left(\hat{\theta}, \theta, e, x\right)$, experiment e, and the sample space X modeled by $\theta \in \Theta$ and estimated by $\hat{\theta}(x) \equiv \hat{\theta}$ for sample $x \in X$. The distribution $p(x|e)$ is a conditional pdf for obtaining a particular sample given the experiment, while $p(\theta|x, e)$ is a posterior pdf for the population parameter given the sample and experiment. The inner minimization step, termed the *terminal analysis*, averages the loss function over Θ (treating the sample as fixed) then optimizes $\hat{\theta}$. The outer minimization, termed the *preposterior analysis*, averages the inner minimization over the predictive distribution for the data then optimizes e.

In this manuscript we assume squared error loss (SEL): $L\left(\hat{\theta}, \theta, e, x\right) = \left(\hat{\theta} - \theta\right)^2$. Then (8) becomes

$$\begin{aligned}
\min_{e} \int_X \min_{\hat{\theta}} \int_\Theta L\left(\hat{\theta}, \theta, e, x\right) p(\theta|x, e)p(x|e)\, d\theta\, dx &= \min_{e} \int_X \min_{\hat{\theta}} \mathop{\mathrm{E}}_{\theta|x,e} (\hat{\theta} - \theta)^2 p(x|e)\, dx \\
&= \min_{e} \int_X \mathrm{Var}(\theta|x, e)p(x|e)\, dx \\
&= \min_{e} \left\{ \mathop{\mathrm{E}}_{x|e} \mathrm{Var}(\theta|x, e) \right\}
\end{aligned} \tag{9}$$

Thus, under SEL, the preposterior analysis entails considering $\mathop{\mathrm{E}}_{x|e} \mathrm{Var}(\theta|x, e)$, which is the expected posterior variance, and then selecting e in order to minimize this quantity.

Hence, we can implement the following algorithm to identify the optimal design:

1. Specify $L\left(\hat{\theta}, \theta, e, x\right)$ (we assume SEL), θ (finite population mean), e (sample allocation to strata).
2. Find the estimate that minimizes the expected loss (posterior of mean under SEL).
3. Find the posterior loss ($V(\theta|x, e)$ in our setting).
4. Conduct preposterior analysis: average over the predictive distribution for the main study data given the pilot data to obtain the expected posterior loss for a given sample size. This requires deriving an approximate closed form for the expected posterior loss for a given allocation via a first-order Taylor series approximation.
5. Find the optimal allocation for minimizing the expected posterior loss by onsidering the space of possible sample allocations using standard numerical methods.

In the next section, we provide the analytical results required to implement this algorithm for stratified random sampling for our specific heteroscedastic model of interest.

3 Analytical results

3.1 Notation

Strata are given by $h = 1, ..., H$. The finite population total and mean within stratum h are given by $Y_h = \sum_{i=1}^{N_h} Y_{hi}$, and $\bar{Y}_h = \frac{Y_h}{N_h}$, respectively; similarly the finite population total and mean for the auxiliary variable are given by $X_h = \sum_{i=1}^{N_h} X_{hi}$ and $\bar{X}_h = \frac{X_h}{N_h}$. The target parameter of interest is the population total $\theta = \bar{Y} = \frac{1}{N} \sum_{h=1}^{H} Y_h$, where $N = \sum_{h=1}^{H} N_h$.

$D1$ and $D2$ refer to the pilot and main study data respectively and m_h and n_h refer to the sample sizes in stratum h for $D1$ and $D2$. Define the sample allocation to the strata as $e_1 = \{n_h : h = 1, ..., H\}$ under the constraint that, e.g., $\sum_{h=1}^{H} n_h = n$ for prespecified n or $\sum_{h=1}^{H} n_h c_h \leq C$ for a constraint on total cost C, and sample size constraints of the form $4 \leq n_h \leq N_h$. The sets of units sampled in stratum h in the pilot and main study are denoted by s_{1h} and s_{2h}, respectively, and s_{1h}^C and s_{2h}^C are the complements of these sets over the set of stratum population units U_h. The sample means of the outcome and auxiliary variables for the main study are given by $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_{2h}} y_{hi}$ and $\bar{x}_h = \frac{1}{n_h} \sum_{i \in s_{2h}} x_{hi}$. Finally we denote the transformed sampled auxiliary variable by $z_{hi} = x_{hi}^b$.

3.2 Heteroscedastic error model

We consider a population model wherein each stratum has its own slope, and the variance of a stratum's error term is proportional to a fixed power of an auxiliary variable. Thus our model of interest is

$$Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi} \tag{10}$$

where $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$ and $X_{hi} > 0$ for stratum $h = 1, ..., H$ with population units $i = 1, ..., N_h$. Assume H strata are predefined, and that each phase uses stratified random sampling (STSRS). We treat $\{\alpha_h, v_h\}$ are unknown and assume that $b \geq 0$ and auxiliary variable $\{X_{hi}\}$ are known for the population. Transforming our scale parameter $g_h = \frac{1}{v_h}$, we further assume a non-informative prior on (α_h, g_h) given by $\pi(\alpha_h, \frac{1}{v_h}) \propto v_h$. This prior leads to a normal-gamma posterior for $(\alpha_h, \frac{1}{v_h})$.

We assume the pilot data are used only for planning the main study, and are not to be used in our ultimate inferences about the population mean. However, our results could be easily extended to allow for a different decision on this matter.

3.3 Find the optimal estimator

As per Lindley (1972), the first step to finding the Bayesian optimal design is to specify a loss function as a function of the parameter (i.e., finite population mean), estimate, experiment (i.e., sample allocation), and sample, after which we find the estimator that minimizes the

expected loss conditional on the data. We specify a loss function of $L(\bar{Y}, \hat{\bar{Y}}, e_1, D2) = \left(\hat{\bar{Y}} - \bar{Y}\right)^2$. As per our Bayesian formulation, $\bar{Y}$ is a parameter, and is random with respect to the parameter space. Let $\hat{\bar{Y}}(D2)$ denote some estimator for the finite population mean as a function of the main study data (D2).

Holding the main sample fixed leads to an expected loss of $\underset{\bar{Y}|D2}{\mathrm{E}}\left\{\left(\bar{Y} - \hat{\bar{Y}}(D2)\right)^2 | D2\right\} = \mathrm{Var}(\bar{Y}|D2) + \left(\mathrm{E}(\bar{Y}|D2) - \hat{\bar{Y}}(D2)\right)^2$. This expression is minimized when estimation is via the posterior mean, leading to an expected loss of $\underset{\bar{Y}|D2}{\mathrm{E}}\left(\bar{Y} - \hat{\bar{Y}}\right)^2 = \mathrm{Var}\left(\bar{Y}|D2\right)$, which is the posterior variance for the finite population mean. Thus, we will ultimately specify our optimization problem as selecting the sample size to minimize the expected posterior variance, from within the space of feasible sample allocations.

3.4 Find the posterior loss

Next, we find the posterior loss, which is the posterior variance for the finite population mean, and which we write as $\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right)$ as to emphasize that we are fixing not only the data but also the sample allocation and coefficient of heteroscedasticity. We can find the posterior loss through an application of results from Ericson (1969a, sec. 5.1) for each stratum. For stratum h, for $h = 1, ..., H$, we assume a diffuse prior on variables $\left(\alpha_h, \frac{1}{v_h}\right)$ with density $\pi\left(\alpha_h, \frac{1}{v_h}\right) \propto v_h$, as per Appendix A.2, and we also assume that the H pairs of variables are independent across strata. Then, we can compute $\mathrm{E}\left(\bar{Y}_h|D2, e_1, b\right)$ and $\mathrm{Var}\left(\bar{Y}_h|D2, e_1, b\right)$ through a direct application of Ericson's Equations (73) and (74), after which we combine results across strata. Assuming that $n_h > 3$ for all h, this results in:

$$\mathrm{E}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N}\left\{n_h\bar{y}_h + \left(N_h\bar{X}_h - n_h\bar{x}_h\right)\left(\frac{\sum_{i\in s_{2h}} \frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}} \frac{x_{hi}^2}{z_{hi}}}\right)\right\} \tag{11}$$

$$\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2}\left\{\frac{1}{n_h - 3}\left[\sum_{i\in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{2h}} \frac{x_{hi}y_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}} \frac{x_{hi}^2}{z_{hi}}}\right]\left[\sum_{i\in(U_h\cap s_{2h}^C)} z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}} \frac{x_{hi}^2}{z_{hi}}}\right]\right\} \tag{12}$$

(See Appendix A.3 for detail.)

3.5 Conduct the preposterior analysis

The posterior variance from Equation (12) above is a function of known population quantities, the main study sample sizes, and the main study sample data. However, the main study sample data have not yet been collected. Therefore, as per Lindley, the next step

is a preposterior analysis, which entails averaging over the predictive distribution $D2|D1$. Specifically, we need to compute $\underset{D2|D1}{\mathrm{E}}\left\{\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right)\right\}$, which is the expected posterior loss for a given sample allocation.

Assuming large $\{n_h\}$, we show in Appendix B that

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) \approx \sum_{h=1}^{H} \frac{\mathrm{E}_P(v_h)}{N^2}\left(\frac{n_h-1}{n_h-3}\right)(N_h-n_h)\left[\bar{Z}_h + \frac{\frac{n_h}{N_h}S_{xh}^2 + (N_h-n_h)\bar{X}_h^2}{n_h\bar{Q}_h^2(1+CV_{qh}^2)}\right] \tag{13}$$

where selected terms above are defined as $\mathrm{E}_P(v_h) = \underset{\alpha_h, v_h|D1}{\mathrm{E}}(v_h)$, $S_{xh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)^2$, $Q_{hi} = \frac{X_{hi}}{\sqrt{Z_{hi}}}$, $\bar{Q}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Q_{hi}$, $CV_{qh} = \frac{S_{qh}}{\bar{Q}_h}$, and $S_{qh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(Q_{hi}-\bar{Q}_h\right)^2$, and where it turns out that $\mathrm{E}_P(v_h) = \frac{1}{m_h-3}\left(\sum_{i\in s_{1h}}\frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{1h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{1h}}\frac{x_{hi}^2}{z_{hi}}}\right)$. The exact result, which does not require large $\{n_h\}$, is $\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H}\frac{1}{N^2}\left(\frac{n_h-1}{n_h-3}\right)\left[\mathrm{E}_P(v_h)\right]\mathrm{E}\left(k(s_{2h})\right)$, where $k(s_{2h}) = \left[\sum_{i\in(U_h\cap s_{2h}^C)} z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]$. The approximate result shown in Equation (13) assumes large $\{n_h\}$ in applying a first-order Taylor series (TS) approximation for $\mathrm{E}\left(k(s_{2h})\right)$.

The TS approximation above was used on account of the rightmost term in $k(s_{2h})$, which is nonlinear. Alternatively, since $k(s_{2h})$ depends solely on the sample indicators and auxiliary variable, one could compute a Monte Carlo (MC) estimate of $\mathrm{E}\left(k(s_{2h})\right)$, as a function of $\{X_{hi}\}$ and b, by averaging over R design-based samples, for large R. If the MC estimator is directly incorporated into the objective function, it may be helpful to construct a high-quality and quickly-evaluated MC emulation function for $\hat{\mathrm{E}}\left(k(s_{2h})|\,\{n_h\}\right)$ for given $\{n_h\}$, as to avoid computational bottlenecks during the optimization process; we describe how to do so in Appendix B.4.

3.6 Obtain optimal allocation

Equation (13) above provides a closed form expression for the approximate expected posterior loss as a function of known population quantities, the pilot sample, and the main study sample allocation. Hence, the optimal allocation, $\underset{e_1}{argmin}\,\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right)$, can be obtained via standard mathematical programming methods (see, e.g., Valliant, Dever, and Kreuter 2013, chap. 5). A common formulation would entail aiming to minimize Equation (13) subject to a constraint on total costs, $\sum_{h=1}^{H} n_h c_h \leq C$, and sample size constraints of the form $4 \leq n_h \leq N_h$. Here, we have assumed that $n_h \geq 4$ on account of the $(n_h-3)^{-1}$ term in Equation (12). Survey designers could also add any other relevant constraints (e.g., precision constraints for domain estimation).

After computing an allocation, if Equation (13) was used, it may be helpful to use MC sampling to assess the appropriateness of the Taylor series approximation for $\mathrm{E}\left(k(s_{2h})\right)$, as suggested in the last paragraph of §3.5.

4 Simulation study

4.1 Overview

Simulation methods were used to compare performance of the proposed sampling strategy with alternative stratified random sampling (STSRS) designs. For a given population, each simulation entailed drawing a pilot sample and then, for each sampling strategy, computing the sample allocation, drawing a main study sample, and making inferences according to that strategy's estimation methods. Although not limited to establishment surveys, such surveys are a likely setting for implementing these methods. Unfortunately, due to confidentiality-related issues, there is a lack of publicly available establishment survey microdata. Therefore, we compared our proposed sample allocation method with alternatives across a series of artificially generated populations that aimed to reflect selected characteristics of establishment surveys. Specifically, we created $P = 90$ bivariate populations, consisting of three shapes for the auxiliary variable crossed with five coefficients of heteroscedasticity, further crossed with six structures for the slope and correlation for the two variables. We considered a single mechanism for stratification. Then, we compared results primarily in terms of RMSE, relative bias, and coverage and relative width of 95% CIs.

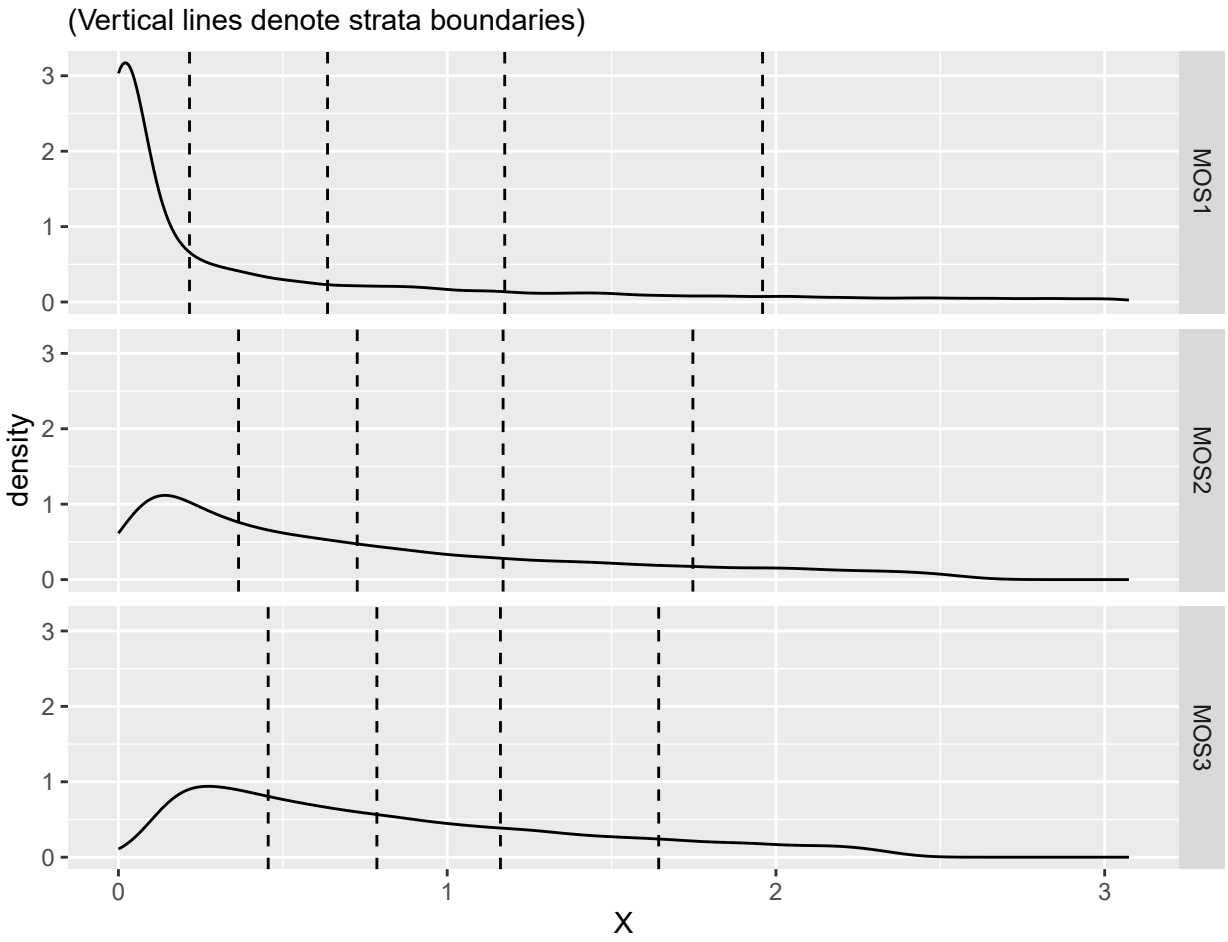
4.2 Data generating mechanism

We generated the covariate X for our population by first generating 12,500 observation from a $GAMMA(k, \theta)$ distribution and then dropping the first and last deciles, resulting in truncated gamma distributions of $N = 10,000$ units. The gamma distribution is relevant to establishment data, which are continuous and heavily skewed; thus we can consider X as a measure of size (MOS). Truncation was done to reduce the numbers of near-zero and extremely large units, the latter of which would presumably be in a take-all stratum under any design and wherein practical considerations would commonly hinge upon reducing non-sampling errors. We considered three different gamma distributions:

- $(k = 0.2, \theta = 5)$: MOS 1, highly skewed
- $(k = 0.6, \theta = 5/3)$: MOS 2, moderately skewed
- $(k = 1, \theta = 1)$: MOS 3, slightly skewed

Next, five strata were then formed based on equalizing the aggregated square roots of X_i. This was operationalized by sorting the $\{X_i : i = 1, ...N\}$ in ascending order, computing the cumulative sum of the first k elements as $c(k) = \sum_{i=1}^{k} \sqrt{X_i}$, for $k = 1, ..., N$, and then using cut points of $\frac{1}{5}\sum_{i=1}^{N} \sqrt{X_i}$, $\frac{2}{5}\sum_{i=1}^{N} \sqrt{X_i}$, $\frac{3}{5}\sum_{i=1}^{N} \sqrt{X_i}$, and $\frac{4}{5}\sum_{i=1}^{N} \sqrt{X_i}$ to group units based on the cumulative sums. Relating our size measures and stratification to what is done in practice, establishment surveys commonly stratify by industry and/or unit size; for instance, several of the Census Bureau's economic surveys have stratified by size within industry (National Academies of Sciences, Engineering, and Medicine 2018). The resulting univariate distributions we generated are displayed in Figure 1.

Figure 1: Distributions of simulated stratified size measures, by MOS



Finally, we generated the outcome variable as $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$ where $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$ for stratum $h = 1, ..., 5$ with population units $i = 1, ..., N_h$. We considered $b = 0$, $b = 0.5$, $b = 1$, $b = 1.5$, and $b = 2$, consistent with the coefficient of heteroscedasticity typically lying within the interval $[0, 2]$ in real applications. For each value of b, we then developed six structures for the remaining parameters, $\{\alpha_h, v_h\}$. We specified a set of slopes, $\{\alpha_h\}$, and a set of target correlations, $\{\rho_h\}$, and solved for $v_h|\alpha_h, \rho_h, b$, such that the correlation between the $\{X_{hi}, Y_{hi}\}$ for a given stratum was approximately equal to ρ_h:

$$v_h = \frac{\alpha_h^2 S_{xh}^2 \left(\frac{1}{\rho_h^2} - 1\right)}{\bar{Z}_h}, \tag{14}$$

where $\bar{Z}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Z_{hi}$, $Z_{hi} = X_{hi}^b$, $S_{xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(X_{hi} - \bar{X}_h\right)^2$, and $\bar{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} X_{hi}$ (see Appendix C.1 for detail). The six scenarios specified were as follows:

1. **Baseline**: assume correlations of $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = 0.7$ and slopes of $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$, where strata 1 through 5 are ordered based on ascending size measures (e.g., stratum 1 refers to the units with the smallest X's).

2. **Lower correlations**: assume correlations of $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = 0.5$ and slopes of $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$.

3. **Higher correlations**: assume correlations of $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = 0.9$ and slopes of $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$.

4. **Increasing correlations, fixed slopes**: assume monotonically increasing correlations of $\rho_1 = 0.5$, $\rho_2 = 0.6$, $\rho_3 = 0.7$, $\rho_4 = 0.8$, and $\rho_5 = 0.9$ and constant slopes of $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$.

5. **Fixed correlations, decreasing slopes**: assume correlations of $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = 0.7$ and monotonically decreasing slopes of $\alpha_1 = 1.4$, $\alpha_2 = 1.2$, $\alpha_3 = 1.0$, $\alpha_4 = 0.8$, and $\alpha_5 = 0.6$.

6. **Increasing correlations, decreasing slopes**: assume monotonically increasing correlations of $\rho_1 = 0.5$, $\rho_2 = 0.6$, $\rho_3 = 0.7$, $\rho_4 = 0.8$, and $\rho_5 = 0.9$, and monotonically decreasing slopes of $\alpha_1 = 1.4$, $\alpha_2 = 1.2$, $\alpha_3 = 1.0$, $\alpha_4 = 0.8$, and $\alpha_5 = 0.6$.

The first three scenarios assumed equal slopes and correlations across strata, with slopes fixed at 1 and correlations fixed at 0.5, 0.7, or 0.9. The last three scenarios allowed the correlations to increase and/or slopes to decrease with respect to increasing size measures, which could possibly be motivated by the up-or-out dynamic of young firms. On the latter point, Haltiwanger, Jarmin, and Miranda (2013) observed a strong inverse relationship between firm size and net employment growth using microdata from the Census Bureau's Longitudinal Business Database; the authors attributed this relationship to the role of young firms, which tended to be smaller and more volatile than older firms. Thus, our latter scenarios allowed for strata with larger units to have larger correlations between the X's and Y's (e.g., larger correlation between previous and current year's revenue) and/or smaller slopes (e.g., lower revenue growth). Figure 2 displays six bivariate populations that used MOS1 and where $b = 1$, for a randomly selected ten-percent sample of each population; each panel reflects a different slope/correlation scenario.

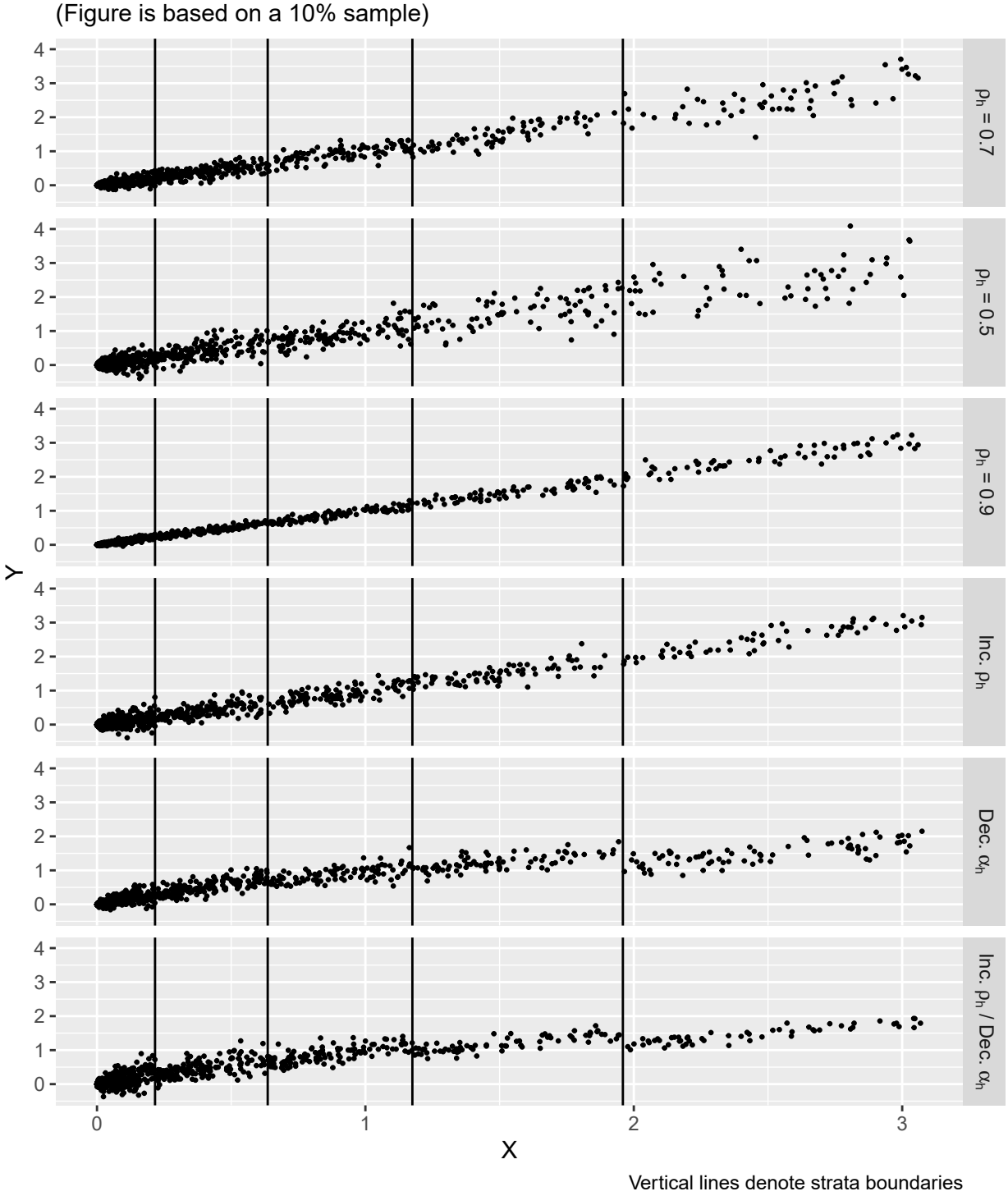
4.3 Pilot design

An equally allocated pilot sample $D1$ of size 75 was drawn for each simulation, so that $m_1 = \cdots = m_5 = 15$; simple random sampling was conducted within each stratum. Although, in practice, the allocation for a pilot sample will depend on the specific study, equal allocation can be justified in some situations, for instance, in scenarios where a sample designer has no reason, a priori, to prioritize certain strata.

4.4 Main study allocations and estimators

Next, we applied competing strategies for each given pilot data set, wherein a strategy comprises an allocation method and an estimator. For the pth population and rth simulation, we used the pilot data set $d_1^{(p,r)}$ to compute A different allocations (i.e., one per strategy),

Figure 2: Distributions of simulated stratified size measures: MOS1, b=1 populations



then drew a single design-based sample for each allocation, and computed the estimated population total along with variance estimates and associated 95% equal-tail CI (confidence interval or credible interval) for each sample. We fixed the total sample size at $n = 500$ units. A total of 1000 simulations were generated to evaluate each data generating mechanism.

For our B-P (Bayesian allocation/prediction estimation) strategy, the sample allocation was computed to minimize the expected posterior loss subject to constraints of $4 \leq n_h \leq N_h$. The expected posterior loss, expressed in Equation (13), incorporates a TS approximation for $\mathrm{E}(k_{s_{2h}})$; as an alternative, we computed Monte Carlo (MC) estimates of $\mathrm{E}(k_{s_{2h}})$, using the methods described in Appendix B.4, and based on 10,000 random permutations of the population. Sample optimization used a COBYLA-based algorithm (Powell 1994), as implemented in NLopt (Johnson 2014), and inferences were made using the prediction estimator from Equation (11). Given that the MC- and TS-based implementations of B-P produced nearly identical results, we report on the TS-based methods, which are simpler to implement and are reasonable to use for large $\{n_h\}$.

In addition to B-P, we consider five alternative stratified random sampling (STSRS) designs, assuming constant per-unit costs.

1. The fully design-based Neyman allocation, in conjunction with the HT-estimator (i.e., N-HT strategy).
2. Cochran's model-assisted strategy for separate ratio (SR) estimation (C-SR strategy; see §2.1.1). The allocation is $n_h \propto N_h \hat{S}_{dh}$, where $\hat{S}_{dh}$ given by Equation (4) is a plug-in estimator for S_{dh}, the standard deviation of a residual term from a SR model; inferences are by the SR estimator.
3. A Cochran "rule of thumb" variation of 2 (RT0-SR) given by $n_h \propto N_h$.
4. A Cochran "rule of thumb" variation of 2 (RT1-SR) given by $n_h \propto N_h \sqrt{\bar{X}_h}$.
5. A Cochran "rule of thumb" variation of 2 (RT2-SR) given by $n_h \propto N_h \bar{X}_h$.

Strategies 3–5 may sometimes be appropriate when $b = 0$, $b = 1$, or $b = 2$, respectively, and when the v_h terms are constant across strata.

For all strategies, sample allocations were rounded, incorporating adjustments when necessary to ensure total sample sizes of 500. The latter involved multiplying a given allocation by $\left(1 + \epsilon^{(r)}_{(p,a)}\right)$, before rounding, for some $\epsilon^{(r)}_{(p,a)}$ close to zero that was chosen to ensure the intended total sample size.

The above strategies are summarized in Table 1.

4.4.1 Variance estimation

We used classical variance estimators for the HT and SR estimators. HT variances were estimated via $\mathrm{var}(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$. SR variances were estimated

Table 1: Overview of main sampling strategies for simulation

	Strategy	Inferential approach	Allocation	Estimator	Comments
N-HT	Neyman/HT	Design-based	Neyman with assumption of $S_h = \hat{S}_h$	HT estimator	
C-SR	Cochran plug-in/ SR estimator	Design-based/ model-assisted	Cochran (1977), §6.14, with assumption of $S_{dh} = \hat{S}_{dh}$	SR estimator	$\hat{S}_{dh}$ not directly provided by Cochran
RT0-SR	Proportional allocation/ SR estimator	Design-based/ model-assisted	$n_h \propto N_h$	SR estimator	Implied by Cochran if S_{dh} is roughly constant
RT1-SR	$b = 1$ rule of thumb/ SR estimator	Design-based/ model-assisted	$n_h \propto N_h\sqrt{\bar{X}_h}$	SR estimator	Suggested by Cochran if S_{dh} is roughly proportional to $\sqrt{\bar{X}_h}$
RT2-SR	$b = 2$ rule of thumb/ SR estimator	Design-based/ model-assisted	$n_h \propto N_h\bar{X}_h$	SR estimator	Suggested by Cochran if S_{dh} is roughly proportional to $\bar{X}_h^2$
B-P	Bayesian allocation/ prediction estimator	Bayesian	Minimize Eqn. (13) assuming that $b = b^{(p)}$	$\mathrm{E}(Y\|D2, e_1, b)$ via Eqn. (11) and $b = b^{(p)}$	

via $v_2(\hat{Y}_{Rs}) = \sum_{h=1}^{H}\left(\frac{\bar{X}_h}{\bar{x}_h}\right)^2 \frac{N_h^2(1-f_h)}{n_h}\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \hat{R}_h x_{hi})^2$, where $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$ is the ratio of sample means in stratum h. For B-P, the posterior variance was computed via Equation (12).

4.5 Performance comparison

Using the above procedures, let $\hat{Y}_{(p,a)}$ denote the estimator for population p and estimation strategy a, and let $\hat{Y}_{(p,a)}^{(r)}$ denote the corresponding estimate for the rth simulation, for $p = 1, ..., P$, $a = 1, ..., A$, and $r = 1, ..., R$. Let $Y_{(p)}$ denote the true population total for population p. We computed four key metrics for a given population p and strategy a:

1. **RMSE**. Root mean square error was estimated as $rmse\left(\hat{Y}_{(p,a)}\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{(p,a)}^{(r)} - Y_{(p)}\right)^2}$.

2. **Relative bias.** Bias was estimated as $bias\left(\hat{Y}_{(p,a)}\right) = \frac{1}{R}\left(\sum_{r=1}^{R}\hat{Y}_{(p,a)}^{(r)}\right) - Y_{(p)}$ and relative bias was estimated as $relbias(\hat{Y}_{(p,a)}) = \frac{bias(\hat{Y}_{(p,a)})}{Y_{(p)}}$.

3. **Coverage of 95% CIs**. Compute a 95% CI for each $\hat{Y}_{(p,a)}^{(r)}$. Let $I^{(r)(p,a)}$ be an indicator that is 1 if the CI contains $Y_{(p)}$ and 0 otherwise. Coverage was estimated as $coverage(p, a) = \frac{1}{R}\sum_{r=1}^{R} I_{(p,a)}^{(r)}$.

4. **Relative width of 95% CIs**. Estimate the average width of a 95% CI via $width(p,a) = \frac{1}{R}\sum_{r=1}^{R} W_{(p,a)}^{(r)}$, where $W_{(p,a)}^{(r)}$ is the width of the 95% CI for population p, simulation r, strategy a. The estimated average relative width is $relwidth(p,a) = \frac{width(p,a)}{Y_{(p)}}$.

For each metric, we also computed Monte Carlo error, defined as the standard deviation of a simulation estimator with respect to repeated implementations of the simulation (e.g., Koehler, Brown, and Haneuse 2009). This error was estimated via $sd(\hat{\theta}) = \sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2}$, where $\hat{\theta}_r$ is an estimate obtained from simulation r, and $\hat{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$.

For allocation methods that incorporated pilot sample information, we also considered the distribution of sample allocations with respect to repeated pilot sampling. For population p and strategy a, we computed the average allocation for stratum h as

$$\bar{n}_{h(p,a)} = \frac{1}{R}\sum_{r=1}^{R} n_{h(p,a)}^{(r)}, \tag{15}$$

where $n_{h(p,a)}^{(r)}$ is the stratum h allocation for the rth simulation. Likewise, we computed the standard deviation via

$$sd_{h(p,a)} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(n_{h(p,a)}^{(r)} - \bar{n}_{h(p,a)}\right)^2}. \tag{16}$$

4.6 Sensitivity analysis for misspecification of heteroscedasticity

Although the B-P strategy assumes that the sample designer knows the heteroscedasticity measure b upfront and uses it to design an efficient sample, this may not always be realistic. Therefore, we conducted a sensitivity analysis to examine the implications of misspecification of b at the sample allocation stage. For each population, we evaluated four variations of B-P, wherein b was misspecified as 0, 0.5, 1, 1.5, or 2 at the sample allocation stage, but where the true value was used for data generation.

5 Simulation results

5.1 Main strategies

First, we consider the N-HT, C-SR, and B-P strategies, which are potentially applicable for any level of b. All three methods were approximately unbiased across populations (Appendix C.2, Table 2), with empirical relative bias that was either indistinguishable from zero or not of practical significance. Further, all three achieved near-nominal coverage for 95% CIs (Appendix C.2, Table 3), with any deviations appearing to be attributable to Monte Carlo

error. Given the strategies' approximate unbiasedness and near-nominal CI coverage, we focus on comparing strategies in terms of RMSE, which we generally consider relative to the B-P strategy. Increases in relative RMSE were paralleled by very similar proportional increases in CI relative width; therefore, we do not report the latter.

Tables 2 and 3 display the relative increase in RMSE of N-HT and C-SR, respectively, compared with B-P, for each of the 90 populations. Rows reflect different levels of heteroscedasticity used in generating the populations, panels reflect the three size measures used, and columns within a panel reflect the six scenarios for strata correlations and slopes.

Across all populations, N-HT consistently underperformed B-P, increasing the RMSE by 11%–175% for individual populations (Table 2). The increases in relative RMSE were greatest for the *higher correlations* scenarios (122%–175%) and tended to be smallest for the *lower correlations* scenarios (11%–44%). Similar patterns were observed when comparing N-HT with C-SR, although the magnitudes were sometimes smaller. The advantages of C-SR and B-P over N-HT were not surprising, since the latter does not incorporate the auxiliary variable in estimation.

Further, B-P performed about as well or better than C-SR, but with marked differences across populations, with B-P showing the greatest advantage for a subset of MOS1 scenarios, wherein the true value of b was equal to 2 or 0 (Table 3). Specifically, C-SR resulted in relative increases in RMSE of 21%–29% for the MOS1, $b = 2$ populations, and 10%–40% for the MOS1, $b = 0$ populations. In contrast, C-SR only yielded a 3.1% average increase in RMSE across the MOS2 populations and 2.0% for the MOS3 populations, which were not as heavily skewed.

For the populations where $b \geq 0.5$, B-P tended to produce more stable sample allocations than C-SR with respect to repeated pilot sampling, and increasingly so for higher levels of b and/or increased skewness of X. The differences were starkest for the MOS1, $b = 2$ populations, as displayed in Table 4, where the CVs for the B-P allocations were appreciably lower than those of the alternatives. Note that the B-P allocation incorporates more detailed population information than does C-SR, the latter of which is solely based on estimated population residuals and does not directly incorporate b.

Table 2: Relative increase in RMSE from Neyman-HT versus B-P (%)

	MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
b = 0	65.4	31.8	175.3	72.2	69.4	77.3	37.7	24.9	142.1	45.6	54.3	37.8	43.6	18.4	133.0	39.3	41.4	41.3
b = 0.5	44.3	18.4	138.4	43.7	50.8	36.1	40.8	21.8	133.8	41.4	42.1	36.9	39.7	21.1	129.9	32.0	30.7	29.6
b = 1	46.0	24.5	127.4	34.9	40.5	27.8	45.0	14.1	135.2	36.7	42.3	32.0	35.0	18.9	126.2	34.2	34.2	36.0
b = 1.5	51.9	24.4	139.5	41.0	50.9	38.2	37.2	11.5	131.5	41.1	45.1	31.3	40.0	21.5	134.2	41.1	41.3	27.9
b = 2	63.9	43.7	158.7	74.6	67.9	61.7	47.4	13.0	121.9	40.0	46.5	40.5	42.0	19.2	133.1	41.0	39.2	41.6

Table 3: Relative increase in RMSE from Cochran-SR versus B-P (%)

	MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
b = 0	19.4	9.7	18.1	32.4	28.5	40.5	-0.8	7.1	5.6	13.5	7.3	2.6	2.2	0.8	-0.8	1.6	4.8	4.3
b = 0.5	7.1	3.0	0.2	2.1	11.6	7.5	1.6	-2.4	8.4	2.6	1.2	3.6	-0.5	7.5	2.4	-0.2	2.7	-1.7
b = 1	5.8	7.4	3.2	1.0	4.3	3.4	-1.3	-1.2	1.4	-2.3	-0.3	7.2	-3.4	3.6	-0.2	1.1	-3.6	3.9
b = 1.5	11.0	9.1	5.1	5.3	5.9	5.3	3.6	0.8	2.9	5.2	0.5	6.0	2.3	-1.7	2.4	3.9	2.3	0.5
b = 2	22.7	21.1	23.0	24.6	24.2	28.5	3.1	4.2	1.6	1.2	4.0	6.9	1.8	4.6	5.8	2.6	2.0	7.9

Table 4: Allocation summary statistics by scenario, stratum, and strategy: populations 25 to 30 (MOS1, b=2)

		Neyman-HT		Cochran-SR		Bayesian-P	
Scenario	h	mean	sd	mean	sd	mean	sd
1. Baseline	1	148.7	47.9	134.9	50.4	111.6	18.8
	2	90.4	19.9	92.2	21.8	98.1	16.7
	3	75.5	16.4	80.3	18.8	85.2	15.1
	4	84.5	16.8	85.1	18.9	91.2	15.3
	5	100.9	21.6	107.4	24.2	113.9	19.1
2. Lower corrs	1	147.1	50.4	135.5	52.0	114.0	18.5
	2	91.8	21.6	94.8	22.4	97.9	16.2
	3	75.7	18.0	77.7	18.3	83.3	14.5
	4	82.8	18.2	84.2	19.3	89.7	15.3
	5	102.7	22.3	107.8	23.7	115.1	18.4
3. Higher corrs	1	154.3	40.4	133.4	51.1	112.2	19.1
	2	87.5	16.9	93.3	22.3	96.8	15.7
	3	74.5	13.8	78.6	17.8	83.5	14.2
	4	84.0	15.4	90.6	20.8	96.5	15.6
	5	99.7	17.2	104.1	21.1	110.9	16.5
4. Inc. corrs	1	189.6	56.5	206.8	64.7	183.2	25.2
	2	96.5	24.2	111.9	31.0	118.9	20.0
	3	72.8	18.4	77.7	22.7	84.4	15.6
	4	67.7	16.6	59.4	17.5	64.8	11.9
	5	73.4	17.3	44.2	13.0	48.7	9.2
5. Dec. slopes	1	193.7	53.1	174.8	59.9	154.7	23.0
	2	103.8	24.6	108.4	28.7	113.8	19.7
	3	79.4	18.4	82.1	20.3	87.2	14.3
	4	63.4	15.4	68.5	17.5	73.6	13.2
	5	59.8	14.4	66.2	17.1	70.7	12.6
6. Inc. corrs, dec. slopes	1	232.5	60.2	242.9	66.7	220.7	26.4
	2	108.8	29.5	122.0	37.6	129.8	21.9
	3	65.2	18.7	67.3	22.0	73.2	14.0
	4	51.5	14.9	42.8	14.1	47.6	9.5
	5	41.9	11.5	25.0	8.5	28.8	5.7

5.2 Rule-of-thumb allocations

Next, we examined performance of rule-of-thumb strategies in populations where they are potentially appropriate (i.e., RT0-SR, RT1-SR, and RT2-SR for the $b = 0$, $b = 1$, and $b = 2$ populations, respectively). The C-SR and B-P strategies, which incorporate pilot data when computing allocations, consistently performed as well or better than the RT-SR strategies,

which do not. For example, for the MOS1, $b = 2$ populations, RT2-SR led to 82%–204% increases in RMSE compared with B-P (Table 5, first panel, third row), with the worst performance for the scenario with decreasing slopes and increasing correlations. In contrast, RT1-SR performed similarly or only modestly worse than the other strategies in the nine $b = 1$ populations with constant correlations and slopes (i.e., MOS1, MOS2, and MOS3 populations under baseline, lower, or higher correlation scenarios).

Table 5: Relative increase in RMSE from RT-SR strategies versus B-P (%)

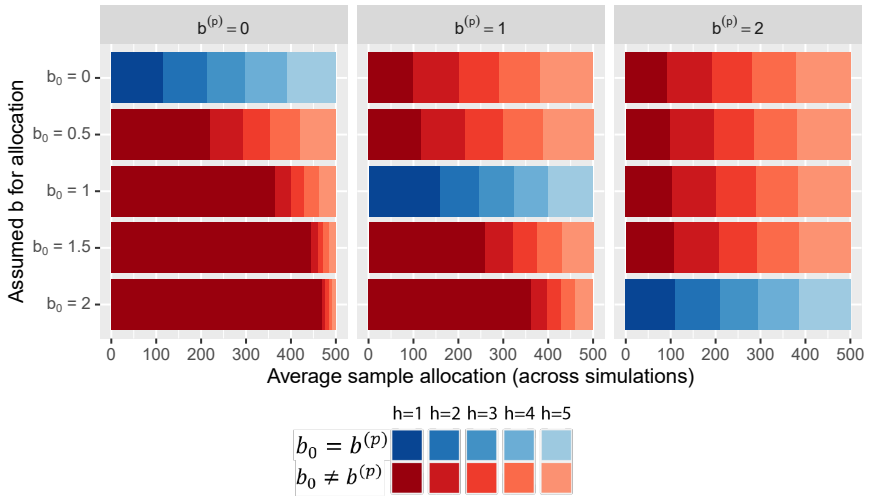
	MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
b = 0	45	34	43	35	38	29	4	10	11	12	7	11	6	10	0	3	-1	12
b = 1	5	10	4	18	10	32	3	3	7	17	11	37	-4	1	5	18	9	40
b = 2	91	84	82	148	133	204	25	31	23	72	55	105	21	17	23	55	39	85

Otherwise, the RT-SR strategies typically had little bias and near-nominal coverage of 95% CIs (excepting the MOS1, $b = 2$ populations, where coverage was closer to 90%; Appendix C.2, Table 4).

5.3 Effects of misspecified b on B-P strategy

Next, we evaluated the effects of misspecifying b at the sample allocation stage when using the B-P strategy—that is, allocation via $b_0 \in \{0, 0.5, 1, 1.5, 2\}$, but where $b_0 \neq b^{(p)}$. The results varied widely across populations, in terms of RMSE and average allocations. In the MOS2 and MOS3 populations, the B-P strategy was fairly insensitive to misspecified b; in contrast, some MOS1 populations were very sensitive to this, especially at the lowest levels of b. To illustrate, Figure 3 shows the average allocations to sampling strata (via Equation 15) for three MOS1 populations, under the baseline $\{\alpha_h, \rho_h\}$ scenario, where $b^{(p)}$ equaled 0, 1, or 2, and where correctly specified allocations are colored in blue (with shade denoting stratum, and bar width denoting average allocation size). For the $b^{(p)} = 0$ population (leftmost panel), the correct specification ($b_0 = 0$; top row, in blue) led to a roughly equal allocation, whereas overestimates of b_0 led to substantial overallocations for stratum 1 (e.g., leftmost panel, bottom row, in red). In contrast, for the corresponding $b^{(p)} = 2$ population (rightmost panel), underestimates of b led to only slight changes in allocation, on average.

Figure 3: Average sample allocations by population and assumed b, selected populations (MOS1, baseline $\{\alpha_h, \rho_h\}$ scenario)



As a result, among MOS1 populations, the RMSEs of the B-P strategy were very susceptible to misspecification of b when $b^{(p)} = 0$, with this effect diminishing at higher levels of $b^{(p)}$. This is illustrated in Table 6, which provides the relative increase in RMSE from misspecified b for selected MOS1 populations.

Beyond the main results above, misspecification under these MOS1 populations still allowed for approximately unbiased estimates, with 95% CIs that usually had reasonable coverage (albeit with overly conservative CIs when $b^{(p)} = 0$ and $b_0 \in \{1.5, 2\}$).

Table 6: Relative increase in RMSE from misspecified b versus correctly specified b among selected MOS1 populations (%)

	$b^{(p)} = 0$						$b^{(p)} = 1$						$b^{(p)} = 2$					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
$b_0 = 0$							7	16	10	12	5	13	3	7	3	5	8	3
$b_0 = 0.5$	11	2	16	17	13	20	7	8	1	6	5	5	5	7	2	5	2	-2
$b_0 = 1$	72	65	87	76	75	86							4	2	0	2	3	-1
$b_0 = 1.5$	182	177	204	193	200	187	8	13	7	8	8	10	3	-1	-5	1	-3	-3
$b_0 = 2$	271	267	281	265	268	240	53	56	48	46	53	46						

6 Application using NCCS data

6.1 Analysis of log revenues

Next, we applied our methods to analyzing tax returns of public charities, using data from the National Center for Charitable Statistics (NCCS), and which were largely based on IRS Form 990 data. Using NCCS Core 1989–2015 Public Charities Fiscal Year Trend data, we analyzed total revenues in fiscal years 2008 and 2013, using Employer Identification Number (EIN) as an identifier, and focusing on organizations with at least $100,000 in 2008 revenue. We focused on the U.S. nonprofit sector, that is, excluding records that NCCS flags as out-of-scope (e.g., due to being overseas or in U.S. territories). We focused on operating public charities (i.e., as opposed to supporting or mutual benefit public charities), to avoid double counting. Further, we focused on organizations with at least $50,000 in gross receipts in each year, to reflect filing thresholds (which were raised to that level in 2010). We also required positive revenues, so that we could subsequently consider log revenue. This resulted in a population of 140,858 domestic operating public charities with at least $100,000 in 2008 revenue, and positive revenue in 2013, and who met filing thresholds. We subsequently categorized cases by nonprofit sector, based on National Taxonomy of Exempt Entities (NTEE) codes; therefore, we dropped an additional 65 cases with missing NTEE category.

Using the resulting data, we aimed to estimate the total log revenue in fiscal year 2013, using log revenue in 2008 as an auxiliary variable. Figure 4 displays the contours of 2D kernel density estimates of the data, by sector.

Figure 4: Contour plot for kernel density estimates of log revenue in 2008 versus 2013, by sector

Next, we estimated b via MCMC in JAGS (Plummer 2017), via R. In doing so, we assumed

an unstratified model of $Y_i = \alpha X_i + \varepsilon_i$, where $\varepsilon_i \overset{ind}{\sim} N(0, \tau/X_i^b)$ (the second parameter being variance), using a normal prior on α, inverse-gamma prior on τ, and uniform prior on b (with support from -3 to 3). After 2,000 burn-in iterations and simulating an additional 20,000 iterations for each of five chains, we retained every tenth observation for a total of 10,000 retained simulations. MCMC diagnostics were plotted using `ggmcmc` (Fernández-i-Marín 2016), including the potential scale reduction factor ($\hat{R}$, Gelman et al. 2003), which suggested approximate convergence. This analysis yielded a posterior mean of $\hat{b} = 0.55$ and 95% equal-tail CI for b of $(0.25, 0.66)$

Next, we used methods that paralleled those of our earlier simulation, as applied to the NCCS data. We formed 24 strata based on the cross-classification of nonprofit sector (8 categories) and 2008 revenue (3 categories), using revenue boundaries that roughly equalized the aggregate unit standard deviations for the three size categories, as applied to log revenue in 2008, and assuming that $b = 0.55$. Population counts for the resulting strata are displayed in Table 7. We then drew $R = 10,000$ equally allocated pilot samples of $n = 360$ units (15 per stratum), each of which was used in applying the strategies of Table 1 for drawing main studies with total sample sizes of $n = 1800$. Strategies were again compared in terms of RMSE, relative bias, and CI properties (see §4.5).

Table 7: NCCS data: population counts by sector and 2008 revenue

	Total Revenue in 2008			
Nonprofit Sector	Under \$300k	\$300k–\$1.2m	\$1.2m+	Total
Arts, culture, and humanities	6,533	4,883	2,870	14,286
Education	5,652	5,897	7,882	19,431
Environment	2,633	2,219	1,308	6,160
Health	4,542	5,668	11,706	21,916
Human services	19,929	19,896	17,795	57,620
International	1,179	959	904	3,042
Public and societal benefit	4,125	3,862	2,876	10,863
Religion	3,822	2,498	1,155	7,475
Total	48,415	45,882	46,496	140,793

Simulation results are summarized in Table 8. The C-SR and B-P strategies offered substantially lower RMSE than the N-HT strategy. All three methods were approximately unbiased and had near-nominal CI coverage, the B-P strategy producing slightly conservative CIs.

Table 8: NCCS simulation results

Strategy	Relative RMSE	1000*RelBias	CI Coverage (%)	1000*CI RelWidth
N-HT	1.427	-0.00	94.7	6.48
C-SR	1.014	-0.01	94.8	4.57
B-P	1.000	-0.00	95.5	4.63

Note: RMSE is displayed relative to that of the B-P strategy.

6.2 Analysis on original scale

In practice, samples are sometimes allocated under models that do not exactly fit the data. Therefore, an important question concerns the performance of B-P methods for scenarios where the model does not fit well. We analyzed such a scenario by repeating the simulation of §6.1, but using the NCCS data on their original scale (i.e., omitting the log transformation), which resulted in a poor fit for our assumed model.

Among the 140,793 charities analyzed in §6.1, 1.3% of units had 2008 revenues exceeding $100 million, which totaled 64% of total 2008 revenues. Considering that our focus is not on extremely large units, which would presumably be selected under any reasonable design, we omitted these units from our analysis, and analyzed the remaining $N = 138,960$ charities. We assumed that the inferential objective is to estimate total revenue in fiscal year 2013, using 2008 revenue as the auxiliary variable.

We again estimated b via MCMC, as in §6.1, but where the X and Y reflect the original scale, which resulted in a posterior mean of $\hat{b} = 1.19$, which we assumed to be the true value of b. Even accounting for heteroscedasticity, we note that the normal distribution is a poor approximation for the underlying data. This is evident from a Q-Q plot for a simple (no-intercept) linear regression on 2013 revenue on 2008 revenue, with analytical weights of $1/X_i^{1.19}$ (see Appendix C.2, Figure 1, righthand panel), which has a far more extreme righthand tail than an analogous plot for the log-transformed data (lefthand panel).

We formed 24 strata in the same manner as before, except that the size class boundaries were modified to reflect the current section's estimate of b (see Appendix C.2, Table 6 for population counts that reflect the current boundaries). Given that some of the resulting strata population sizes were small, implementing the classic allocations directly would have led to some allocations of $n_h < 4$ or $n_h > N_h$. Therefore, allocations below 4 or above N_h were moved to these bounds; this step was applied in conjunction with the rounding step described in the last paragraph of §4.4, so that the total sample size was always $n = 1800$.

Simulation results are shown in Table 9. The C-SR and B-P strategies again outperformed N-HT, but more narrowly. All strategies produced CIs with slightly below nominal coverage.

Table 9: NCCS simulation results: original scale

Strategy	Relative RMSE	1000*RelBias	CI Coverage (%)	1000*CI RelWidth
N-HT	1.123	0.27	93.2	11.46
C-SR	0.994	-0.24	90.2	9.55
B-P	1.000	2.48	91.7	9.65

Note: RMSE is displayed relative to that of the B-P strategy.

7 Discussion

7.1 Summary of results

We considered sample allocation for stratified random sampling under a univariate regression model with heteroscedastic errors, using Bayesian decision theory to accommodate uncertain design parameters. We focused on optimal allocation for estimating the finite population mean, under squared error loss. We assumed a non-informative prior in conjunction with use of a pilot survey, so that the allocation would be driven by the pilot data rather than via a subjective prior. We derived the approximate expected posterior variance, which allowed for identifying the optimal allocation via mathematical programming. This entailed selecting the $\{n_h\}$ to minimize the approximate expected posterior variance, subject to constraints of $4 \leq n_h \leq N_h$, although additional constraints on sample sizes or other sample-related quantities could have been straightforwardly handled via standard optimization software (see, e.g., Valliant, Dever, and Kreuter 2013, chap. 5). Although the expected posterior loss assumed large $\{n_h\}$ in applying a Taylor series approximation, the resulting empirical performance was nearly identical to that of an alternative that avoided this assumption.

Performance was compared with key design-based and model-assisted strategies across a series of artificial populations with a skewed, continuous size measure, and outcome variables that followed from various applications of our model. Among the main strategies considered, the model-assisted method (C-SR) and proposed Bayesian method (B-P), which incorporate auxiliary information in estimation, consistently outperformed the purely design-based strategy (N-HT), which did not. B-P performed as well or better than C-SR across the different populations, achieving similar or moderately lower RMSEs; for certain types of populations, B-P also led to more stable sample allocations with respect to repeated pilot sampling.

Although the simulations above assumed use of our model, we repeated the simulations using real data. In an application to the log-revenues of public charities, the B-P strategy achieved substantially reduced RMSE compared with N-HT, as did the C-SR strategy. Likewise, B-P and C-SR turned out to outperform N-HT when the public charities application was repeated using the original scale, in spite of a worsened model fit.

We considered rule-of-thumb (RT) strategies in populations where they were potentially appropriate. The C-SR and B-P strategies, which incorporate pilot data, consistently performed as well as or better than the RT strategies, whose allocations are solely based on the auxiliary variable; in some cases, the RT strategies performed substantially worse, particularly for $b = 2$ populations.

We also evaluated the effects of misspecified coefficient of heteroscedasticity at the sample allocation stage under the proposed strategy. We found that under a highly skewed size measure, and when the true b was close to 0, the sample allocations were very sensitive to misspecified b. On the other hand, for many populations considered, misspecification of b did not meaningfully change the resulting allocations or RMSE.

7.2 Limitations and future directions

Although our allowing for heteroscedasticity improves realism in comparison to some previous Bayesian optimal design work, our model assumptions will not always be realistic. Therefore, an important avenue of future work entails development of alternative models, such as multivariate models, different error terms (e.g., lognormal), and/or different conditional mean functions. Some examples of the latter are from the class of general polynomial models considered by Valliant, Dorfman, and Royall (2000, 53–54) and from the set of associations considered in simulation by Zangeneh and Little (2015). It may also be important to consider the implications of outliers or other influential values on allocations. Of course, changes to the population structure may require new derivations for the posterior expected loss, or alternatively, the development of a computational approach (e.g., Ryan et al. 2016).

An important aspect of the problem structure is that separate superpopulation parameters are used for different strata. If there are too many strata relative to the total sample size, then the resulting estimator will be inefficient, since the stratum-specific predictions will be unstable. As an extreme example, considering the $(n_h - 1)/(n_h - 3)$ term in (13), and assuming a target sample size of n, using $H = n/4$ strata would result in four units per stratum regardless of how the strata were formed. Even if the strata are large enough such that most of the population is in strata where $(n_h - 1)/(n_h - 3) \approx 1$, if there are sets of strata with similar model coefficients, then it may be more efficient to pool groups of strata. On the other extreme, having far too few strata may lead to the use of a model that does not adequately capture group differences. Further work could be done to explore these issues.

Although the B-P approach performed as well or better than the alternatives, most of the populations used in simulation were generated according to our model, as was necessary due to the paucity of publicly available establishment data. When repeating the simulation to public charities data, the B-P strategy performed well, but so did the C-SR strategy, which had roughly similar RMSE. The simulations on artificial data show that the B-P strategy can sometimes improve on C-SR under a correctly specified model, whereas the application to public charities data showed the B-P strategy performing reasonably in an application to real data. Additional empirical or theoretical work could be conducted to better understand the conditions under which the methods' performance varies and the reasons why.

Additionally, given that design-based inference is predominant at most statistical agencies—notwithstanding an increasing reliance on models in recent years to combat nonsampling errors (e.g. Valliant 2023)—the reader may wonder about how our proposed optimal Bayes methods relate to design-based inference. While the proposed Bayes estimator relies on Bayesian methodology to develop the sample design, the Bayesian piece—the prior distribution—is based on a simple non-informative prior that often can be viewed as an inferential equivalent to frequentist estimators. Further, the evaluations in this manuscript follow the "calibrated Bayes" paradigm (Little 2006) of using Bayesian methods to develop estimators but to evaluate these estimators using frequentist methods (bias, mean square error).

We assumed that b is known upfront, which facilitated analytical results, but this assumption may sometimes be unrealistic. When b is unknown, it may be reasonable to assume use of a sample-based estimate for sample allocation purposes. Practitioners can assess sensitivity of allocations to different assumed levels of b, following §5.3; if the allocations are fairly

invariant to misspecified b, as occurred in several simulated populations, then assuming an estimated b should work well. In contrast, if allocations are sensitive to misspecified b, then performance of our proposed allocation methods will be unclear. Therefore, another area of extension entails allocating sample when b is unknown or is estimated using a small sample.

Another area of possible extension could entail consideration of alternative loss functions to accommodate other survey goals or even to compromise across multiple objectives. Although we focused on design for a single objective, our methods could straightforwardly be adapted to multi-purpose surveys by constructing a multi-objective loss function as a importance-weighted combination of various single-objective loss functions, in a manner analogous to Valliant and Gentle (1997); an alternative approach would be to use a fuzzy programming technique to compromise between different objectives, following Swain (2016). Although we assumed squared error loss, Chaloner and Verdinelli (1995) and Ryan et al. (2016) provide examples of other objective functions that have been used in Bayesian experimental design.

An additional opportunity for extension is in identifying other ways to express prior knowledge, particularly when a pilot is unavailable. Other approaches to obtaining a prior, which have been employed in the context of responsive survey design for household surveys, entail soliciting and pooling expert opinion (Coffey et al. 2020) and using an intensive literature review (West et al. 2021). Substituting one of these methods in place of pilot data would presumably entail redefining $\underset{\alpha_h, v_h | D1}{\mathrm{E}}(v_h)$ (as pertains to Equation [13]), which is the mechanism by which the pilot data affect the allocation.

References

Brewer, K. R. W. 2002. *Combined Survey Sampling Inference: Weighing Basu's Elephants.* Arnold.

Chaloner, K., and I. Verdinelli. 1995. "Bayesian Experimental Design: A Review." *Statistical Science* 10 (3): 273–304. DOI: https://doi.org/10.1214/ss/1177009939.

Clark, R. G. 2013. "Sample Design Using Imperfect Design Data." *Journal of Survey Statistics and Methodology* 1 (1): 6–23. DOI: https://doi.org/10.1093/jssam/smt002.

Cochran, W. G. 1977. *Sampling Techniques.* John Wiley & Sons, Inc.

Coffey, S., B. T. West, J. Wagner, and M. R. Elliott. 2020. "What Do You Think? Using Expert Opinion to Improve Predictions of Response Propensity under a Bayesian Framework." *Methods, Data, Analyses* 14 (2): 159–194. DOI: https://doi.org/10.12758/mda.2020.05.

Dalenius, T., and J. L. Hodges. 1959. "Minimum Variance Stratification." *Journal of the American Statistical Association* 54 (285): 88–101. DOI: https://doi.org/10.1080/01621459.1959.10501501.

Draper, N. R., and I. Guttman. 1968a. "Bayesian Stratified Two-Phase Sampling Results: k Characteristics." *Biometrika* 55 (3): 587–589. DOI: https://doi.org/10.1093/biomet/55.3.587.

Draper, N. R., and I. Guttman. 1968b. "Some Bayesian Stratified Two-Phase Sampling Results." *Biometrika* 55 (1): 131–139. DOI: https://doi.org/10.1093/biomet/55.1.131.

Ericson, W. A. 1965. "Optimum Stratified Sampling Using Prior Information." *Journal of the American Statistical Association* 60 (311): 750–771. DOI: https://doi.org/10.1080/01621459.1965.10480825.

———. 1967a. "On the Economic Choice of Experiment Sizes for Decision Regarding Certain Linear Combinations." *Journal of the Royal Statistical Society, Series B (Methodological)* 29 (3): 503–512. DOI: https://doi.org/10.1111/j.2517-6161.1967.tb00712.x.

———. 1967b. "Optimal Sample Design with Nonresponse." *Journal of the American Statistical Association* 62 (317): 63–78. DOI: https://doi.org/10.1080/01621459.1967.10482888.

———. 1969a. "Subjective Bayesian Models in Sampling Finite Populations." *Journal of the Royal Statistical Society. Series B (Methodological),* 195–233. DOI: https://doi.org/10.1111/j.2517-6161.1969.tb00782.x.

———. 1969b. "Subjective Bayesian Models in Sampling Finite Populations: Stratification." In *New Developments in Survey Sampling,* edited by H. Smith and N. L. Johnson, 326–357. Wiley.

———. 1972. *Subjective Bayesian Models in Sampling Finite Populations: Stratification.* Technical report 15. Ann Arbor, Michigan: University of Michigan, Department of Statistics.

———. 1973. *A Bayesian Approach to Two-Stage Sampling.* Technical report 26. Ann Arbor, Michigan: University of Michigan, Department of Statistics.

———. 1975. *A Bayesian Approach to Two-Stage Sampling.* Technical report AFFDL-TR-75-145. Wright-Patterson AFB, Ohio: Air Force Dynamics Laboratory.

———. 1983a. *Response Bias and Error in Sampling Finite Populations.* Technical report 122. Ann Arbor, Michigan: University of Michigan, Department of Statistics.

———. 1983b. *Survey Sampling: Modeling, Randomization, and Robustness: A Subjective Bayesian Approach.* Technical report 121. Ann Arbor, Michigan: University of Michigan, Department of Statistics.

———. 1985. *Bayesian Inference in Finite Populations.* Technical report 129. Ann Arbor, Michigan: University of Michigan, Department of Statistics.

———. 1988. "Bayesian Inference in Finite Populations." In *Handbook of Statistics 6: Sampling,* 213–246. Elsevier. DOI: https://doi.org/http://dx.doi.org/10.1016/S0169-7161(88)06011-0.

Fernández-i-Marín, X. 2016. "ggmcmc: Analysis of MCMC Samples and Bayesian Inference." *Journal of Statistical Software* 70 (9): 1–20. DOI: https://doi.org/10.18637/jss.v070.i09.

Foreman, E. K. 1991. *Survey Sampling Principles.* New York: Marcel Dekker.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis.* 3rd ed. Chapman & Hall/CRC Boca Raton, FL, USA. DOI: https://doi.org/10.1201/b16018.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis.* 2nd ed. Chapman & Hall/CRC Boca Raton, FL, USA. DOI: https://doi.org/10.1201/9780429258480.

Ghosh, M. 2009. "Bayesian Developments in Survey Sampling." Chap. 29 in *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis,* edited by D. Pfeffermann and C. R. Rao, 155–189. Elsevier. DOI: https://doi.org/10.1016/s0169-7161(09)00229-6.

Godambe, V. P., and V. M. Joshi. 1965. "Admissibility and Bayes Estimation in Sampling Finite Populations - I." *The Annals of Mathematical Statistics* 36 (6): 1707–1722. DOI: https://doi.org/10.1214/aoms/1177699799.

Grosh, D. L. 1972. "A Bayes Sampling Allocation Scheme for Stratified Finite Populations with Hyperbinomial Prior Distributions." *Technometrics* 14 (3): 599–612. DOI: https://doi.org/10.1080/00401706.1972.10488949.

Haltiwanger, J., R. S. Jarmin, and J. Miranda. 2013. "Who Creates Jobs? Small Versus Large Versus Young." *Review of Economics and Statistics* 95 (2): 347–361. DOI: https://doi.org/10.1162/rest_a_00288.

Henry, K. A., and R. Valliant. 2009. "Comparing Sampling and Estimation Strategies in Establishment Populations." *Survey Research Methods* 3 (1): 27–44. DOI: https://doi.org/10.18148/srm/2009.v3i1.72.

Isaki, C. T., and W. A. Fuller. 1982. "Survey Design under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77 (377): 89–96. DOI: https://doi.org/10.1080/01621459.1982.10477770.

Jinn, J. H., J. Sedransk, and P. Smith. 1987. "Optimal Two-Phase Stratified Sampling for Estimation of the Age Composition of a Fish Population." *Biometrics* 43 (2): 343–353. DOI: https://doi.org/10.2307/2531817.

Johnson, S. G. 2014. *The NLopt Nonlinear-Optimization Package.*

Khan, M. Z. 1976. "Optimum Allocation in Bayesian Stratified Two-Phase Sampling When There Are m-Attributes." *Metrika* 23 (1): 211–219. DOI: https://doi.org/10.1007/bf01902867.

———. 1976. "Optimum Allocation in Bayesian Stratified Two-Phase Sampling (a General Case)." *Journal of the Indian Statistical Association* 14:65–74.

Koehler, E., E. Brown, and S. J.-P. Haneuse. 2009. "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses." *The American Statistician* 63 (2): 155–162. DOI: https://doi.org/10.1198/tast.2009.0030.

Lindley, D. V. 1972. *Bayesian Statistics: A Review.* Society for Industrial / Applied Mathematics. DOI: https://doi.org/10.1137/1.9781611970654.

Little, R. J. 2006. "Calibrated Bayes: a Bayes/frequentist roadmap." *The American Statistician* 60 (3): 213–223. DOI: https://doi.org/10.1198/000313006x117837.

National Academies of Sciences, Engineering, and Medicine. 2018. *Reengineering the Census Bureau's Annual Economic Surveys.* National Academies Press. DOI: https://doi.org/10.17226/25098.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. DOI: https://doi.org/10.2307/2342192.

O'Malley, A. J., and A. M. Zaslavsky. 2016. "Optimal Small-Area Estimation and Design When Nonrespondents Are Subsampled for Follow-Up." *Journal of Survey Statistics and Methodology* 4 (1): 2–21. DOI: https://doi.org/10.1093/jssam/smv036.

Office of Management and Budget. 2017. *Statistical Programs of the United States Government, Fiscal Year 2017.* Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/information_and_regulatory_affairs/statistical-programs-2017.pdf (accessed April 2017).

Plummer, M. 2017. *JAGS Version 4.3.0 User Manual [computer Software Manual].* Available at: https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x (accessed February 2022).

Powell, M. J. D. 1994. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation." In *Advances in Optimization and Numerical Analysis,* edited by S. Gomez and J. Hennart, 51–67. Springer. DOI: https://doi.org/10.1007/978-94-015-8330-5_4.

Raiffa, H., and R. Schlaifer. 1961. *Applied Statistical Decision Theory.* Boston: Division of Research, Harvard Business School.

Rao, J. N. K. 2011. "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal." *Statistical Science* 26 (2): 240–256. DOI: https://doi.org/10.1214/10-sts346.

Rao, J. N. K., and P. D. Ghangurde. 1972. "Bayesian Optimization in Sampling Finite Populations." *Journal of the American Statistical Association* 67 (338): 439–443. DOI: https://doi.org/10.1080/01621459.1972.10482406.

Rivest, L.-P. 2002. "A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys." *Survey Methodology* 28 (2): 191–198.

Ryan, E. G., C. C. Drovandi, J. M. McGree, and A. N. Pettitt. 2016. "A Review of Modern Computational Algorithms for Bayesian Optimal Design." *International Statistical Review* 84 (1): 128–154. DOI: https://doi.org/10.1111/insr.12107.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling.* New York, NY: Springer.

Schouten, B., N. Mushkudiani, N. Shlomo, G. Durrant, P. Lundquist, and J. Wagner. 2018. "A Bayesian Analysis of Design Parameters in Survey Data Collection." *Journal of Survey Statistics and Methodology* 6 (4): 431–464. DOI: https://doi.org/10.1093/jssam/smy012.

Sekkappan, R. M. 1981a. "Generalization of Ericson's Subjective Bayesian Models in Sampling Finite Populations for k Characteristics." *Journal of the Indian Statistical Association* 19:159–164.

———. 1981b. "Subjective Bayesian Multivariate Stratified Sampling from Finite Populations." *Metrika* 28 (1): 123–132. DOI: https://doi.org/10.1007/bf01902884.

Singh, B., and J. Sedransk. 1978. "Sample Size Selection in Regression Analysis When There Is Nonresponse." *Journal of the American Statistical Association* 73 (362): 362–365. DOI: https://doi.org/10.1080/01621459.1978.10481583.

———. 1984. "Bayesian Inference and Sample Design for Regression Analysis When There Is Nonresponse." *Biometrika* 71 (1): 161–170. DOI: https://doi.org/10.1093/biomet/71.1.161.

Smith, P., M. Pont, and T. Jones. 2003. "Developments in Business Survey Methodology in the Office for National Statistics, 1994–2000." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52 (3): 257–286. DOI: https://doi.org/10.1111/1467-9884.03571.

Smith, P. J., and J. Sedransk. 1982. "Bayesian Optimization of the Estimation of the Age Composition of a Fish Population." *Journal of the American Statistical Association* 77 (380): 707–713. DOI: https://doi.org/10.1080/01621459.1982.10477875.

Soland, R. M. 1967. "Optimal Multivariate Stratified Sampling with Prior Information." *Scandinavian Actuarial Journal* 1967 (1-2): 31–39. DOI: https://doi.org/10.1080/03461238.1967.10406205.

Solomon, H., and S. Zacks. 1970. "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas." *Journal of the American Statistical Association* 65 (330): 653–677. DOI: https://doi.org/10.1080/01621459.1970.10481115.

Swain, A. K. P. C. 2016. "A Note on Optimum Allocation for Multiobjective Sample Surveys Using Fuzzy Programming." *Journal of Survey Statistics and Methodology* 4 (2): 171–187. DOI: https://doi.org/10.1093/jssam/smv038.

Valliant, R. 2023. "Hansen Lecture 2022: The Evolution of the Use of Models in Survey Sampling." *Journal of Survey Statistics and Methodology* 12 (2): 275–304. DOI: https://doi.org/10.1093/jssam/smad021.

Valliant, R., J. A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples.* Springer. DOI: https://doi.org/10.1007/978-1-4614-6449-5.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach.* John Wiley.

Valliant, R., and J. E. Gentle. 1997. "An Application of Mathematical Programming to Sample Allocation." *Computational Statistics & Data Analysis* 25 (3): 337–360. DOI: https://doi.org/10.1016/s0167-9473(97)00007-8.

Wagner, J., B. T. West, M. R. Elliott, and S. Coffey. 2020. "Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context." *Journal of Official Statistics* 36 (4): 907. DOI: https://doi.org/10.2478/jos-2020-0043.

Wald, A. 1950. *Statistical Decision Functions.* Wiley.

West, B. T., J. Wagner, S. Coffey, and M. R. Elliott. 2021. "Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis Versus Literature Review." *Journal of Survey Statistics and Methodology* 11 (2): 367–392. DOI: https://doi.org/10.1093/jssam/smab036.

Wright, R. L. 1983. "Finite Population Sampling with Multivariate Auxiliary Information." *Journal of the American Statistical Association* 78 (384): 879–884. DOI: https://doi.org/10.1080/01621459.1983.10477035.

Zacks, S. 1970. "Bayesian Design of Single and Double Stratified Sampling for Estimating Proportion in Finite Population." *Technometrics* 12 (1): 119–130. DOI: https://doi.org/10.1080/00401706.1970.10488639.

Zangeneh, S. Z., and R. J. Little. 2015. "Bayesian Inference for the Finite Population Total from a Heteroscedastic Probability Proportional to Size Sample." *Journal of Survey Statistics and Methodology* 3 (2): 162–192. DOI: https://doi.org/10.1093/jssam/smv002.

A Application and extension of Ericson's results

Under our problem formulation in Section 3 of the main paper, the model for stratum h can be viewed as a special case of a regression model considered by Ericson (1969a, sec. 5.1), insofar as we assume use of a diffuse prior and incorporate the coefficient of heteroscedasticity directly in the variance structure, whereas Ericson's model allowed for other priors and only assumed that the variance function was some known function of the auxiliary variable. This section applies Ericson's results to our problem, and we also obtain an important distributional result that is used in our preposterior analysis.

Specific sections are as follows:

- A.1 provides a definition of the normal-gamma distribution.
- A.2 applies a diffuse prior to Ericson's posterior distribution of the superpopulation parameters given the data and provides key results. This section primarily uses Ericson's original notation, but the results are translated back to our own notation in A.2.2.
- A.3 provides the posterior mean and variance of $\bar{Y}|D2, e_1, b$.
- A.4 uses some algebraic manipulation and results relating to quadratic forms to obtain the distribution of a term that appears in the expression for $\text{Var}\left(\bar{Y}_h|D2, e_1, b\right)$.

A.1 Normal-gamma definition

- Suppose $X|T \sim N\left(\mu, \frac{1}{\lambda T}\right)$, where the parameters are mean and variance. We have a conditional pdf of $f(x|\tau) \propto \sqrt{\lambda\tau} exp\left(\frac{-(x-\mu)^2}{2}\lambda\tau\right)$.
- Suppose further that $T|\alpha, \beta \sim \Gamma(\alpha, \beta)$, where the parameters are shape and rate; that is, our pdf is $f(\tau) = \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}$
- Then $(X, T) \sim NormalGamma(\mu, \lambda, \alpha, \beta)$, with pdf $f(x, \tau) = \frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-\frac{1}{2}}e^{-\beta\tau}e^{\frac{-\tau\lambda(x-\mu)^2}{2}}$

Note that this means that $\frac{1}{T}|(\alpha, \beta) \sim IG(\alpha, \beta)$. Via properties of the IG distribution, we have that $\text{E}\left(\frac{1}{T}\right) = \frac{\beta}{\alpha-1}$ (assuming $\alpha > 1$) and $\text{Var}\left(\frac{1}{T}\right) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (assuming $\alpha > 2$). We also have that $\text{E}(X) = \mu$ and $\text{Var}(X) = \frac{\beta}{\lambda(\alpha-1)}$.

A.2 Ericson's posterior distribution under a diffuse prior

Ericson (1969a) provides relevant results for a univariate regression model, which we will summarize using his notation, with subsequent application to our problem. To assist the reader, a key for selected notation is provided below:

Table 1: Notational key

Description	Ericson's Notation	Our Notation
Model	$X_i\|(y_i,\alpha,h,z_i) \overset{ind}{\sim} N\left(\alpha y_i, \frac{z_i}{h}\right)$	$Y_{hi}\|(X_{hi},\alpha_h,v_h,b) \overset{ind}{\sim} N(\alpha_h X_{hi}, v_h X_{hi}^b)$
Outcome variable	X_i	Y_{hi}
Auxiliary data	y_i	X_{hi}
Slope	α	α_h
Variance component 1	$v = \frac{1}{h}$	v_h
Variance component 2	$z_i = g(y_i)$	$z_{hi} = x_{hi}^b$

Ericson considered a population of N distinguishable elements, labeled by the integers from the label set $\mathcal{N} = \{1, 2, ..., N\}$, where X_i is the unknown value of an outcome variable and y_i is some known and positively-valued auxiliary variable for the ith member of the population, defined for $i \in \mathcal{N}$. The sample, $(s, \mathbf{x})$, comprises some set of indices of distinct population units, $s = \{i_1, ..., i_n\} \subseteq \mathcal{N}$, in conjunction with their observed values x_j of X_j, $j \in s$. The X_i's are viewed as exchangeable and independent random variables coming from a common distribution, conditional on the superpopulation parameters. In §5.1, Ericson considers an unstratified model of $X_i|(y_i,\alpha,h,z_i) \overset{ind}{\sim} N\left(\alpha y_i, \frac{z_i}{h}\right)$, wherein the second parameter in the normal distribution is variance, and where $z_i = g(y_i)$, for some pre-specified and positively valued function g. Given this model, he writes the joint prior for the population vector $\mathbf{X} = (X_1, X_2, ..., X_N)$ in Equation (63) as:

$$p(\mathbf{X}|\alpha, h) \propto \prod_{i=1}^{N} \left(\frac{h}{z_i}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{h}{z_i}(X_i - \alpha y_i)^2\right\} \tag{17}$$

Further, Ericson assumed a normal-gamma prior on (α, h), written in Equation (64) as having density

$$f'(\alpha, h) \propto exp\left\{-\frac{1}{2}hn'(\alpha - \bar{\alpha}')^2\right\} h^{\frac{1}{2}\delta(n')} exp\left\{-\frac{1}{2}h\nu'\upsilon'\right\} h^{\frac{1}{2}\nu'-1} \tag{18}$$

where $\delta(n') = 0$ if $n' = 0$ and $\delta(n') = 1$ otherwise. From this pdf, we can see that Ericson has assigned the parameters a distribution of $(\alpha, h) \sim NG\left(\bar{\alpha}', n', \frac{1}{2}\left[\delta(n') + \nu' - 1\right], \frac{1}{2}\nu'\upsilon'\right)$, as per the normal-gamma definition in Appendix A.1.

This leads to a posterior distribution that is normal-gamma, which he writes in Equations (67–71) as

$$f\{\alpha, h|(s, \mathbf{x})\} \propto exp\left\{-\frac{1}{2}hn''(\alpha - \bar{\alpha}'')^2\right\} h^{\frac{1}{2}\delta(n'')} exp\left\{-\frac{1}{2}h\nu''\upsilon''\right\} h^{\frac{1}{2}\nu''-1}, \tag{19}$$

where

$$n'' = \sum_{i \in s} y_i^2 / z_i + n', \tag{20}$$

$$\bar{\alpha}'' = \frac{1}{n''} \left(\sum_{i \in s} \frac{x_i y_i}{z_i} + n' \bar{\alpha}' \right), \tag{21}$$

$$\nu'' \upsilon'' = \left\{ \nu' \upsilon' + \frac{n'}{n''} \sum_{i \in s} \left(\frac{x_i - \bar{\alpha}' y_i}{\sqrt{z_i}} \right)^2 + \frac{1}{n''} \sum_{i \in s} \frac{y_i^2}{z_i} \sum_{i \in s} \frac{x_i^2}{z_i} - \frac{1}{n''} \left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right)^2 \right\}, \tag{22}$$

$$\nu'' = n + \nu' + \delta(n') - \delta(n''), \tag{23}$$

and where $\delta(n'') = 0$ if $n'' = 0$ and $\delta(n'') = 1$ otherwise. Equivalently, we have that $\alpha, h|(s, \mathbf{x}) \sim NG\left(\bar{\alpha}'', n'', \frac{1}{2}\left[\delta(n'') + \nu'' - 1\right], \frac{1}{2}\nu''\upsilon''\right)$. From this, we see that $h|(s, \mathbf{x}) \sim \Gamma\left(\frac{1}{2}\left[\delta(n'') + \nu'' - 1\right], \frac{1}{2}\nu''\upsilon''\right)$, and equivalently, $\frac{1}{h}|(s, \mathbf{x}) \sim IG\left(\frac{1}{2}\left[\delta(n'') + \nu'' - 1\right], \frac{1}{2}\nu''\upsilon''\right)$, where IG refers to the inverse-gamma distribution.

Ericson noted that a diffuse prior of $f'(\alpha, h) \propto h^{-1}$ could be considered as a special case of the prior in (18) by putting $\nu' = n' = 0$. We do so here. Thus, Equations (20–23) can be simplified as:

$$n'' = \sum_{i \in s} y_i^2 / z_i \tag{24}$$

$$\bar{\alpha}'' = \frac{1}{n''} \left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right) \tag{25}$$

$$\nu'' \upsilon'' = \sum_{i \in s} \frac{x_i^2}{z_i} - \frac{\left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right)^2}{\sum_{i \in s} \frac{y_i^2}{z_i}} \tag{26}$$

$$\nu'' = n - 1 \tag{27}$$

$$\delta(n'') = 1 \tag{28}$$

where $z_i = g(y_i)$.

Summing up, the posterior distribution under the diffuse prior of $f'(\alpha, h) \propto h^{-1}$ can be written as $\alpha, h|(s, \mathbf{x}) \sim NG\left(\bar{\alpha}'', n'', \frac{n-1}{2}, \frac{1}{2}\nu''\upsilon''\right)$, with terms as defined in (24–27) above.

A.2.1 Posterior distribution of 1/h

From the above, and continuing to use Ericson's notation, we have that $\frac{1}{h} \sim IG\left(\frac{n-1}{2}, \frac{1}{2}\nu''\upsilon''\right)$. If we define $q = \frac{1}{n-3}\left(\sum_{i \in s} \frac{x_i^2}{z_i} - \frac{\left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right)^2}{\sum_{i \in s} \frac{y_i^2}{z_i}} \right)$, then we have that $\frac{1}{h} \sim IG\left(\frac{n-1}{2}, \frac{n-3}{2} q\right)$.

Via properties of the IG distribution, we have that if $X \sim IG(\alpha, \beta)$, then $\mathrm{E}(X) = \frac{\beta}{\alpha - 1}$ (assuming $\alpha > 1$) and $\mathrm{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (assuming $\alpha > 2$).

Hence, assuming that $n > 3$, we have:

$$\mathrm{E}\left(\frac{1}{h}\right) = q \tag{29}$$

Assuming that $n > 5$, we have:

$$\mathrm{Var}\left(\frac{1}{h}\right) = \frac{2q^2}{n-5} \tag{30}$$

A.2.2 Ericson's posterior under our notation

Ericson's posterior distribution for the superpopulation parameters, which were provided earlier using Ericson's notation, can be summarized in our notation as follows. Letting $g_h = \frac{1}{v_h}$, assume a prior on (α_h, g_h) with density $\pi(\alpha_h, g_h) \propto \frac{1}{g_h}$; write this as $\pi(\alpha_h, \frac{1}{v_h}) \propto v_h$. This prior, in conjunction with our model and associated assumptions, leads to a posterior distribution of $(\alpha_h, \frac{1}{v_h})|b, D2 \sim NG(\mu_h, c_h, \gamma_h, \delta_h)$ where $z_{hi} = x_{hi}^b$, $a_h = \sum_{i=1}^{n_h} \frac{y_{hi}^2}{z_{hi}}$, $d_h = \sum_{i=1}^{n_h} \frac{x_{hi}y_{hi}}{z_{hi}}$, $c_h = \sum_{i=1}^{n_h} \frac{x_{hi}^2}{z_{hi}}$, $\mu_h = \frac{d_h}{c_h} = \frac{\sum_{i=1}^{n_h} x_{hi}y_{hi}/z_{hi}}{\sum_{i=1}^{n_h} x_{hi}^2/z_{hi}}$, $\gamma_h = \frac{n_h - 1}{2}$, and $\delta_h = \frac{1}{2}\left(a_h - \frac{d_h^2}{c_h}\right) = \frac{1}{2}\left\{\sum_{i=1}^{n_h} y_{hi}^2/z_{hi} - \frac{\left[\sum_{i=1}^{n_h} x_{hi}y_{hi}/z_{hi}\right]^2}{\sum_{i=1}^{n_h} x_{hi}^2/z_{hi}}\right\}$. Hence, we have that $v_h|b, D2 \sim IG(\gamma_h, \delta_h)$ and $\alpha_h|v_h, b, D2 \sim N\left(\mu_h, \frac{v_h}{c_h}\right)$.

A.3 Posterior of finite population mean under diffuse prior

For the finite population mean μ, then Ericson's Equations (73) and (74) with his notation are:

$$\mathrm{E}\left\{\mu|(s, \mathbf{x})\right\} = \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{\alpha}''\bar{y}_{\bar{s}} \tag{31}$$

$$\mathrm{Var}\left\{\mu|(s, \mathbf{x})\right\} = \frac{\nu''\upsilon''}{\nu'' - 2}\frac{(n''z_{\bar{s}} + y_{\bar{s}}^2)}{n''N^2} \tag{32}$$

where the subscript s denotes a sample quantity, subscript $\bar{s}$ denotes a quantity for elements not sampled, and his model is $X_i|(y_i, \alpha, h) \overset{ind}{\sim} N\left(\alpha y_i, \frac{z_i}{h}\right)$, and other terms are as described in Appendix A.2.

Under a diffuse prior of $\pi(\alpha, h) \propto h^{-1}$, via Ericson's notation, we have:

$$\begin{aligned}\mathrm{E}\left\{\mu|(s, \mathbf{x})\right\} &= \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{\alpha}''\bar{y}_{\bar{s}} \\ &= \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{y}_{\bar{s}}\frac{\left(\sum_{i\in s}\frac{x_i y_i}{z_i}\right)}{\sum_{i\in s} y_i^2/z_i}\end{aligned} \tag{33}$$

Translating this back to our own notation yields:

$$\mathrm{E}\left(\bar{Y}_h|D2,e_1,b\right)=\frac{n_h}{N_h}\bar{y}_h+\left(\frac{N_h\bar{X}_h-n_h\bar{x}_h}{N_h}\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right) \tag{34}$$

We can then combine results across strata to get:

$$\begin{aligned}\mathrm{E}\left(\bar{Y}|D2,e_1,b\right)&=\sum_{h=1}^{H}\frac{N_h}{N}\left\{\frac{n_h}{N_h}\bar{y}_h+\left(\frac{N_h\bar{X}_h-n_h\bar{x}_h}{N_h}\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right)\right\}\\&=\sum_{h=1}^{H}\frac{1}{N}\left\{n_h\bar{y}_h+\left(N_h\bar{X}_h-n_h\bar{x}_h\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right)\right\}\end{aligned} \tag{35}$$

For the variance, under a diffuse prior, and via Ericson's notation, we have

$$\begin{aligned}\mathrm{Var}\left\{\mu|(s,\mathbf{x})\right\}&=\frac{\nu''\upsilon''}{\nu''-2}\frac{(n''z_{\bar{s}}+y_{\bar{s}}^2)}{n''N^2}\\&=(n-3)^{-1}\left[\sum_{i\in s}\frac{x_i^2}{z_i}-\frac{\left(\sum_{i\in s}\frac{x_iy_i}{z_i}\right)^2}{\sum_{i\in s}\frac{y_i^2}{z_i}}\right]\frac{1}{N^2}\left[z_{\bar{s}}+\frac{y_{\bar{s}}^2}{n''}\right]\\&=(n-3)^{-1}\left[\sum_{i\in s}\frac{x_i^2}{z_i}-\frac{\left(\sum_{i\in s}\frac{x_iy_i}{z_i}\right)^2}{\sum_{i\in s}\frac{y_i^2}{z_i}}\right]\frac{1}{N^2}\left[z_{\bar{s}}+\frac{y_{\bar{s}}^2}{\sum_{i\in s}y_i^2/z_i}\right],\end{aligned} \tag{36}$$

assuming that $n>3$. Translating this back to our own notation yields

$$\mathrm{Var}\left(\bar{Y}_h|D2,e_1,b\right)=(n_h-3)^{-1}\left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}}-\frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\frac{1}{N_h^2}\left[\sum_{i\in(U_h\cap s_{2h}^C)}z_{hi}+\frac{\left(N_h\bar{X}_h-n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right], \tag{37}$$

assuming that $n_h>3$ for all h, and where s_{2h}^C refers to the complement of s_{2h} and U_h refers to the population within stratum h. We then combine results across strata to get:

$$\mathrm{Var}\left(\bar{Y}|D2,e_1,b\right)=\sum_{h=1}^{H}\frac{1}{N^2}\left\{\frac{1}{n_h-3}\left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}}-\frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\left[\sum_{i\in(U_h\cap s_{2h}^C)}z_{hi}+\frac{\left(N_h\bar{X}_h-n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\right\} \tag{38}$$

A.4 Quadratic form distribution

This subsection uses results relating to quadratic forms to obtain the distribution of $q_{bh}=\left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}}-\frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]$, which is a quadratic form in y. Specifically, we will

show that $v_h^{-1} q_{bh} | (v_h, \alpha_h, b, s_{2h}) \sim \chi^2_{n_h - 1}$. As per previous notation and set-up, (v_h, α_h, b) are model parameters, and s_{2h} is the set of sample indicators in stratum h.

To simplify notation, temporarily focus on a single stratum, ignore the stratum subscripts, and write the set of sample indicators as s. For the remainder of this subsection, assume that the model parameters, (b, v, α), are known. Also suppose that s is fixed; given that the population auxiliary data, $\{X_i : i = 1, ..., N\}$ are known, it follows that the sample auxiliary data, $\{x_i : i = 1, ..., n\}$, are known. We have that $y_i \sim N(\alpha x_i, v z_i)$, where $z_i = x_i^b$. Let $a_i = \frac{y_i}{\sqrt{z_i}}$ and $d_i = \frac{x_i}{\sqrt{z_i}}$. Also define the vectors $a = (a_1, ..., a_n)^t$ and $d = (d_1, ..., d_n)^t$. Note that a is random with respect to the model, whereas d is fixed. We have:

$$\begin{aligned}
\sum_i y_i^2/z_i - \frac{\left(\sum_i y_i x_i/z_i\right)^2}{\sum_i x_i^2/z_i} &= \sum_i a_i^2 - \frac{\left(\sum_i a_i d_i\right)^2}{\sum_k d_k^2} \\
&= \sum_i a_i^2 - \frac{\sum_i a_i^2 d_i^2 + \sum_{i \neq j} a_i d_i a_j d_j}{\sum_k d_k^2} \\
&= \sum_i a_i^2 \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n a_i a_j \left(\frac{-d_i d_j}{\sum_k d_k^2}\right) \\
&= a^t D a \qquad (39)
\end{aligned}$$

where D is an $n \times n$ matrix, in which cell D_{ij} in row i and column j is defined as:

$$D_{ii} = 1 - \frac{d_i^2}{\sum_k d_k^2}, \qquad \text{i= } 1, ..., n \qquad (40)$$

$$D_{ij} = \frac{-d_i d_j}{\sum_k d_k^2}, \qquad \text{i= } 1, ..., n; j = 1, ..., n; i \neq j \qquad (41)$$

Let $D' = D^2$. (Hence, we have $d'_{rc} = \sum_{i=1}^n d_{ri} d_{ic}$.) Then examining the diagonals, we have:

$$\begin{aligned}
D'_{ii} &= D_{ii}^2 + \sum_{j=1, j \neq i}^n D_{ij} D_{ji} \\
&= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\left(\sum_k d_k^2\right)^2} \sum_{j=1, j \neq i}^n d_j^2 \\
&= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\left(\sum_k d_k^2\right)^2} \left(\left[\sum_{j=1}^n d_j^2\right] - d_i^2\right) \\
&= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\sum_k d_k^2} \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right) \\
&= 1 - \frac{d_i^2}{\sum_k d_k^2} = D_{ii} \qquad (42)
\end{aligned}$$

For cell ij where $i \neq j$, we have:

$$\begin{aligned} D'_{ij} &= \sum_{k=1}^{n} D_{ik} D_{kj} \\ &= \frac{1}{\left(\sum_k d_k^2\right)^2} \sum_{k=1}^{n} d_k^2 d_i d_j \\ &= \frac{d_i d_j}{\sum_k d_k^2} = D_{ij} \end{aligned} \tag{43}$$

Next, note that a general result regarding quadratic forms is that if $Y_1, ..., Y_n$ are independent normal random variables where Y_i has mean μ_i and known variance σ^2, and matrix A is symmetric and idempotent and has dimension $n \times n$, then random variable $\frac{Y^t A Y}{\sigma^2}$ is distributed as non-central chi-square with $tr(A)$ degrees of freedom and non-centrality parameter of $\lambda = \frac{\mu^t A \mu}{\sigma^2}$, where $\mu = (\mu_1, ..., \mu_n)^t$.

In our case, we have that $a_i \sim N\left(\frac{\alpha x_i}{\sqrt{z_i}}, v\right)$, for $i = 1, ..., n$, with $a_1, ..., a_n$ independent. Likewise, we have that $\frac{a_i}{\alpha} \sim N(d_i, v\alpha^{-2})$, for $i = 1, ..., n$. Given that D is symmetric and idempotent with rank $n - 1$, it follows that $\frac{\left(\frac{a}{\alpha}\right)^t D \left(\frac{a}{\alpha}\right)}{v\alpha^{-2}} = \frac{a^t D a}{v} \sim \chi^2$ with $df = n - 1$ and non-centrality parameter of $\lambda = \frac{\left(\mathrm{E}\left[\frac{a}{\alpha}\right]\right)^t D \left(\mathrm{E}\left[\frac{a}{\alpha}\right]\right)}{v\alpha^{-2}} = \frac{d^t D d}{v\alpha^{-2}}$.

Note that matrix D can be rewritten as $D = I - \frac{dd^t}{d^t d}$ where $d = (d_1, ..., d_n)^t$ and I is an $n \times n$ identity matrix. Hence, we have that $d^t D d = d^t \left(I - \frac{dd^t}{d^t d}\right) d = d^t I d - \frac{d^t d d^t d}{d^t d} = 0$, and likewise, $\lambda = 0$. Thus, we have that $\frac{a^t D a}{v} \sim \chi^2_{n-1}$.

From here, we see that $\mathrm{E}(a^t D a) = v\mathrm{E}(\chi^2_{n-1}) = v(n-1)$.

Note that the above used the abbreviated notation, the parameters were treated as fixed, and the sample indicators were treated as fixed. Using the more detailed notation, we have shown that $v_h^{-1} q_{bh} | \left(v_h, \alpha_h, b, s_{2h}\right) \sim \chi^2_{n_h - 1}$ and that $\underset{D2 | \alpha_h, v_h, b, s_{2h}}{\mathrm{E}} (q_{bh}) = v_h(n_h - 1)$.

B Analysis for fixed and known b

Recall from Equation (12) of the main paper that:

$$\text{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left\{ \frac{1}{n_h - 3} \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{y_{hi} x_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \right\} \tag{44}$$

Let $q_{bh} = \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{y_{hi} x_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right]$ and $k(s_{2h}) = \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right]$. (This notation is used given that the first quantity is a quadratic form and the second quantity is fixed given the sample indicators, as the population-level auxiliary data are known.) Then we have:

$$\text{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2(n_h - 3)} q_{bh} k(s_{2h}) \tag{45}$$

B.1 Analysis relating to q_{bh}

Conditional on the pilot data (D1) and the main study sample allocation (e_1), Equation (45) has three sources of uncertainty, in relation to: (1) the posterior distribution for the superpopulation parameters conditional on the pilot data; (2) the outcome variable, conditional on the superpopulation parameters and sample indicators; and (3) the set of sample indicators (which we refer to as the 'design-based sample'). We will analyze q_{bh} and $k(s_{2h})$ separately since it turns out that these terms are independent, which will result in $\text{E}(q_{bh}k(s_{2h})) = \text{E}(q_{bh})\text{E}(k(s_{2h}))$. These terms are independent since it turns out that q_{bh} does not depend on the units selected in the design-based sample, s_{2h}, whereas $k(s_{2h})$ only depends on the design-based sample.

Thus, as a first step, we will evaluate $\underset{D2|D1}{\text{E}}(q_{bh})$ through a series of conditional expectations. For now, treating the design-based sample as fixed, we will use subscripts of M and P, respectively, to distinguish whether the expectation is respect to the model or with respect to the posterior distribution of the parameters given the pilot data. Let $\text{E}_M(Z) = \underset{D2|\alpha_h, v_h, D1, s_{2h}}{\text{E}}(Z)$ and $\text{E}_P(Z) = \underset{\alpha_h, v_h|D1}{\text{E}}(Z)$ for any Z. Then assuming that the strata sample sizes, $\{n_h\}$, are fixed, we have:

$$\underset{D2|D1}{\text{E}}(q_{bh}|s_{2h}) = \text{E}_P(\text{E}_M(q_{bh}|s_{2h})) \tag{46}$$

Appendix A.4 shows that $v_h^{-1} q_{bh}|(v_h, \alpha_h, b, s_{2h}) \sim \chi^2_{n_h - 1}$. Note that the resulting distribution is free of s_{2h}, resulting in $v_h^{-1} q_{bh}|(v_h, \alpha_h, b) \sim \chi^2_{n_h - 1}$. Since the distribution $(\alpha_h, v_h)|D1$ also does not depend on s_{2h}, it follows that the distribution of $q_{bh}|D1$ does not depend on the

main sample, and is independent from $k(s_{2h})$. As a result of this independence, we have that

$$\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2(n_h - 3)} \left[\underset{D2|D1}{\mathrm{E}}(q_{bh})\right] \mathrm{E}\left(k(s_{2h})\right) \tag{47}$$

In addition, from the properties of the χ^2 distribution, we have that $\mathrm{E}_M(q_{bh}|s_{2h}) = v_h(n_h - 1)$. Substituting this into (46), and then using the fact that $q_{bh}|D1$ does not depend on s_{2h}, yields:

$$\underset{D2|D1}{\mathrm{E}}(q_{bh}) = (n_h - 1)\mathrm{E}_P(v_h) \tag{48}$$

Substituting this into (47) gives:

$$\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \frac{n_h - 1}{n_h - 3} \left[\mathrm{E}_P(v_h)\right] \mathrm{E}\left(k(s_{2h})\right) \tag{49}$$

B.2 First-order Taylor series approximation in relation to $k(s_{2h})$

In the above equation, $\mathrm{E}_P(v_h) = \underset{\alpha_h, v_h|D1}{\mathrm{E}}(v_h)$ can be computed at the time of sampling for the main study based on the posterior distribution for the parameters given the pilot data. The term $k(s_{2h})$ is a nonlinear function of the main study sample, which has not yet been collected. Thus, we obtain a first-order Taylor series approximation of $\mathrm{E}(k(s_{2h}))$ so that our expected posterior loss is in terms of known population characteristics.

For most of this section, we will temporarily omit the stratum and sampling phase subscripts, in order to simplify notation. Recall that a simple random sample (within stratum) is assumed. Recall also that $k(s) = \sum_{i \in ns} z_i + \frac{(N\bar{X} - n\bar{x})^2}{\sum_{i \in s} x_i^2/z_i}$, where $z_i = x_i^b$ and where ns refers to the nonsampled units. For the lefthand term, we have that $\mathrm{E}\left(\sum_{i \in ns} z_i\right) = (N - n)\bar{Z}$ where $Z_i = X_i^b$ and $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$. For the righthand term, assuming that $n \to \infty$ and $N \to \infty$, we can reasonably use a first-order Taylor series approximation, wherein

$$\mathrm{E}\left[\frac{\left(N\bar{X} - n\bar{x}\right)^2}{\sum_{i \in s} x_i^2/z_i}\right] \approx \frac{\mathrm{E}\left[\left(N\bar{X} - n\bar{x}\right)^2\right]}{\mathrm{E}\left(\sum_{i \in s} x_i^2/z_i\right)}, \tag{50}$$

and where the the only source of uncertainty is with respect to the design (noting our previous assumptions that the $\{X_i\}$ and b are known and are treated as fixed).

For the numerator, we have

$$\begin{aligned}
\mathrm{E}\left[\left(N\bar{X} - n\bar{x}\right)^2\right] &= \mathrm{Var}\left[N\bar{X} - n\bar{x}\right] + \left(\mathrm{E}\left[N\bar{X} - n\bar{x}\right]\right)^2 \\
&= \mathrm{Var}(n\bar{x}) + (N-n)^2\bar{X}^2 \\
&= n^2\left(1 - \frac{n}{N}\right)\frac{S_x^2}{n} + (N-n)^2\bar{X}^2 \\
&= (N-n)\left[\frac{n}{N}S_x^2 + (N-n)\bar{X}^2\right], \qquad (51)
\end{aligned}$$

where $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$. For the denominator, if we let $Q_i = \frac{X_i}{\sqrt{Z_i}}$, $\bar{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i$, $S_q^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Q_i - \bar{Q}\right)^2$, and $CV_q = \frac{S_q}{\bar{Q}}$, then we find

$$\begin{aligned}
\mathrm{E}\left(\sum_{i\in s} x_i^2/z_i\right) &= \frac{n}{N}\sum_{i\in U} Q_i^2 \\
&= \frac{n}{N}\left((N-1)S_q^2 + N\bar{Q}^2\right) \\
&\approx n(S_q^2 + \bar{Q}^2) \qquad (52) \\
&= n\bar{Q}^2(1 + CV_q^2) \qquad (53)
\end{aligned}$$

Hence, we have a first-order Taylor series approximation of

$$\begin{aligned}
\mathrm{E}\left(k(s)\right) &\approx (N-n)\bar{Z} + \frac{(N-n)\left[\frac{n}{N}S_x^2 + (N-n)\bar{X}^2\right]}{n\bar{Q}^2(1+CV_q^2)} \qquad (54) \\
&= (N-n)\left[\bar{Z} + \frac{\frac{n}{N}S_x^2 + (N-n)\bar{X}^2}{n\bar{Q}^2(1+CV_q^2)}\right]. \qquad (55)
\end{aligned}$$

Reintroducing strata subscripts, we substitute this quantity into Equation (49) to yield

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) \approx \sum_{h=1}^{H}\frac{\mathrm{E}_P(v_h)}{N^2}\left(\frac{n_h - 1}{n_h - 3}\right)(N_h - n_h)\left[\bar{Z}_h + \frac{\frac{n_h}{N_h}S_{xh}^2 + (N_h - n_h)\bar{X}_h^2}{n_h\bar{Q}_h^2(1+CV_{qh}^2)}\right], \qquad (56)$$

which assumes large $\{n_h\}$.

B.3 Expected value of v_h with respect to posterior given pilot data

In the above equation, $\mathrm{E}_P(v_h) = \underset{\alpha_h, v_h|D1}{\mathrm{E}}(v_h)$ can be obtained using the posterior distribution from the pilot data. This could be easily estimated via standard Bayesian computation. Alternatively, if we use the pilot data in conjunction with a diffuse prior on the model parameters then we can draw upon results from Appendix A.2, which provides basic results of Ericson's posterior distribution under the use of a diffuse prior. Under such a diffuse prior

of $\pi\left(\alpha_h, \frac{1}{v_h}\right) \propto v_h$, then translating the notation of Equation (29) to the current notation yields:

$$\mathrm{E}_P(v_h) = \frac{1}{m_h - 3}\left(\sum_{i \in s1h} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{1h}} \frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{1h}} \frac{x_{hi}^2}{z_{hi}}}\right) \tag{57}$$

where, as before, m_h is the pilot sample size in stratum h, s_{1h} is the set of units in the pilot sample in stratum h, and $z_{hi} = x_{hi}^b$.

B.4 Sampling-based approximation for $k(s_{2h})$

This section outlines an efficient Monte Carlo (MC) sampling-based method to approximating $\mathrm{E}(k(s_{2h}))$, as an alternative to the TS approximation of Appendix B.2. This MC approach can be incorporated directly into the optimization process, while avoiding a potential computational bottleneck that could arise from a naive implementation.

Recall that $k(s_{2h}) = \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}}\right]$, where $Z_{hi} = X_{hi}^b$ and b is known. For purposes of this section, define $j_h = \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}}$. We have that

$$\mathrm{E}\left(k(s_{2h}|n_h)\right) = (N_h - n_h)\bar{Z}_h + \mathrm{E}(j_h|n_h). \tag{58}$$

For optimization purposes, we will need to approximately evaluate $\mathrm{E}(j_h|n_h)$ at numerous possible allocations. For a given stratum, consider R random permutations of elements for large R (e.g., 10,000). For a given permutation, treat the first n_h units of stratum h as an SRS of size n_h, and compute a sample estimate for every possible value of n_h, using cumulative sums. This will yield the vector $\left\{\hat{\mathrm{E}}(j_h|n_h) : n_h = 1, 2, ..., N_h\right\}$, in advance of optimization.

More specifically, consider a single stratum, and ignore strata subscripts to simplify notation. Let $(i_{r1}, i_{r2}, ..., i_{rN})$ refer to the rth random permutation of the population indices, for $r = 1, 2, ..., R$. For the rth permutation, compute $\mathbf{j}_r = (j_{r1}, j_{r2}, ..., j_{rN})$, where we define

$$j_{rn} = \frac{(N\bar{X} - \sum_{j=1}^n X_{i_{rj}})^2}{\sum_{j=1}^n \frac{X_{i_{rj}}^2}{Z_{i_{rj}}}}, \tag{59}$$

and where the summations of $\{X_i\}$ and $\left\{\frac{X_i^2}{Z_i}\right\}$ are computed at all levels via cumulative sums, as to share computation across different sample sizes. Our MC estimator is thus

$$\hat{\mathrm{E}}(k(s_2)|n) = (N - n)\bar{Z} + \bar{j}_{.n}, \tag{60}$$

where $\bar{j}_{.n} = \frac{1}{R}\sum_{r=1}^R j_{rn}$. The MC estimate of (60) will reflect the sample mean of R i.i.d. random variables, $j_{h1}, ..., j_{hR}$, reflecting a common distribution, $f(j_{hi}|n_h)$; we suggest choosing R

adequately high so the CV of (60), which can be estimated empirically, is small enough across potential sample sizes such that choosing larger R would not meaningfully alter the resulting allocations. Repeating this process separately for each stratum and then reintroducing strata subscripts will yield H vectors of Monte Carlo estimates of $\{k(s_{2h}|n_h) : n_h = 1, ..., N_h\}$. These vectors are precomputed in advance of optimization with large R, subsequently allowing for quick and accurate evaluation for any given integer sample allocation.

The quantity $k(s_{2h})$ is only defined for discrete $\{n_h\}$, but many optimization methods assume that decision variables are continuous. This disconnect can be accommodated straightforwardly through linear interpolation, for example, obtain the precomputed MC estimates for stratum h sample sizes of $\lfloor n_h \rfloor$ and $\lceil n_h \rceil$, then compute a weighted average via

$$\hat{\mathrm{E}}(k(s_{2h})|n_h) = (1 - frac_h)\,\hat{\mathrm{E}}(k(s_{2h})|\lfloor n_h \rfloor) + (frac_h)\,\hat{\mathrm{E}}(k(s_{2h})|\lceil n_h \rceil), \tag{61}$$

where $frac_h = n_h - \lfloor n_h \rfloor$ denotes the fractional allocation component. This step should enable continuous optimization algorithms to yield reasonable approximations of the optimum allocation in scenarios where the fractional component of the allocation does not have much impact (e.g., for large strata sample sizes). If a more exact situation is needed, integer programming could be employed in the neighborhood of the continuous solution.

C Additional simulation detail

C.1 Selection of v_h to achieve target correlation

Our simulation (from Sections 4 and 5 of the main paper) involved generating pseudo-populations that followed our model for a given set of size measures, $\{X_{hi}\}$, and model parameters, $\{b, \alpha_h, v_h\}$. In doing so, after directly specifying b and $\{\alpha_h\}$, we specified target correlations of $\{\rho_h\}$ and then determined $\{v_h\}$ via Equation (14) of the main paper so as to approximately result in the desired correlations between $\{X_{hi}, Y_{hi}\}$ for the H strata.

Equation (14) of the main paper was derived as follows: To simplify notation, ignore strata subscripts. Define $\rho_{xy} = \frac{C_{xy}}{S_x S_y}$, $S_x^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})^2$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y})^2$, and $C_{xy} = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$, wherein $\bar{X} = N^{-1}\sum_{i=1}^N X_i$ and $\bar{Y} = N^{-1}\sum_{i=1}^N Y_i$. We aim to specify v such that $\mathrm{E}(\rho_{xy}) \approx \rho_{xy}^0$ for some specified ρ_{xy}^0, and where variability is considered with respect to the distribution for $\{Y_i\}|(\{X_i\}, \alpha, v, b)$. We find that $\mathrm{E}(C_{xy}) = \alpha S_x^2$ and $\mathrm{E}(S_y^2) = \alpha^2 S_x^2 + v\bar{Z}$, where $\bar{Z} = N^{-1}\sum_{i=1}^N X_i^b$. Assuming that $\mathrm{E}(\rho_{XY}) \approx \frac{\mathrm{E}(C_{xy})}{\mathrm{E}(S_x S_y)}$ and $\mathrm{E}(S_y) \approx \sqrt{\mathrm{E}(S_y^2)}$, we have that $\mathrm{E}(\rho_h) \approx \frac{\alpha S_x}{\sqrt{\alpha^2 S_x^2 + v\bar{Z}}}$. Solving for v and reintroducing strata subscripts leads to an approximate result analogous to Equation (14) of the main paper, meaning that the data-generating mechanism described earlier would approximately yield the desired correlations, on expectation, if the assumptions hold.

C.2 Supplementary figures and tables

Table 2: 1,000*RelBias by main strategy and population-generating characteristics

		MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
		1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
N-HT	b = 0	-0.2	0.3	-0.2	0.6	0.3	0.5	0.1	-0.3	-0.2	-0.3	-0.4	-0.2	0.7	-0.2	-0.1	0.2	0.0	-0.4
	b = 0.5	0.7	0.5	0.1	0.6	-0.9	-0.0	0.1	0.1	0.2	0.1	0.3	0.4	-0.4	0.2	-0.1	0.3	-0.2	0.3
	b = 1	0.0	-0.3	0.6	1.0	-0.8	-0.3	-0.0	0.1	-0.1	-0.5	0.1	-0.6	-0.2	-0.4	-0.0	-0.1	-0.1	-0.4
	b = 1.5	-0.6	0.4	0.4	1.1	0.6	0.7	0.2	0.1	-0.1	-0.3	0.1	0.1	-0.2	-0.0	-0.3	-0.3	0.3	0.1
	b = 2	0.1	0.9	0.1	0.6	-0.9	-0.3	0.4	-0.7	-0.0	0.4	0.3	-0.3	0.2	-0.5	0.2	-0.1	-0.0	0.2
C-SR	b = 0	-1.2	-0.3	-0.0	0.0	-0.6	0.2	-0.5	0.1	-0.0	0.5	-0.1	0.4	-0.2	0.3	-0.1	-0.2	-0.5	-0.2
	b = 0.5	-0.1	-0.8	-0.2	-0.5	-0.3	0.9	0.3	-0.4	0.1	-0.7	-0.1	-0.3	0.0	0.2	0.0	0.1	0.1	-0.2
	b = 1	-0.3	0.4	-0.2	-0.4	-0.2	-0.1	-0.1	0.1	0.2	0.1	0.2	0.3	-0.1	0.2	0.1	0.2	-0.1	0.7
	b = 1.5	0.1	-0.0	0.0	-0.5	-0.1	-0.7	0.1	-0.3	0.0	-0.1	-0.1	0.8	-0.0	0.1	-0.3	-0.0	0.4	0.1
	b = 2	0.1	0.2	0.2	0.3	0.2	0.9	0.0	-1.4	0.1	-0.3	0.2	-0.4	-0.3	0.2	0.1	-0.0	-0.2	0.1
B-P	b = 0	1.0	1.1	0.4	-1.5	0.2	2.4	0.0	0.9	0.1	-0.0	-0.2	1.6	0.3	-0.6	-0.1	0.6	0.3	0.5
	b = 0.5	1.1	0.8	0.0	-0.7	-0.3	1.3	-0.0	0.6	-0.2	-0.8	-0.4	-0.2	0.4	0.0	0.1	-0.1	-0.3	0.9
	b = 1	-0.2	-0.4	-0.1	0.0	0.0	1.1	-0.3	0.2	-0.2	-0.4	-0.2	0.6	0.1	0.4	0.0	-0.3	0.3	-0.1
	b = 1.5	-0.4	-0.3	-0.1	1.0	-0.3	-0.8	-0.3	0.5	0.1	0.3	0.1	0.1	0.2	-0.4	0.1	0.1	-0.3	-0.4
	b = 2	-0.4	2.3	0.6	2.9	1.6	2.4	-0.4	1.5	0.2	-1.1	1.2	0.7	0.5	-0.6	-0.0	-1.1	0.7	0.4

Table 3: CI coverage rate (%) by main strategy and population-generating characteristics

		MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
		1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
N-HT	b = 0	94.9	94.0	94.9	95.4	94.6	94.0	95.3	93.8	95.6	95.2	95.4	95.1	94.3	95.0	94.8	95.3	94.8	94.0
	b = 0.5	95.4	95.7	95.0	94.7	95.8	95.2	95.5	94.8	95.3	94.4	95.2	94.4	94.6	95.0	93.9	95.7	96.0	95.4
	b = 1	94.8	94.8	95.8	95.8	95.5	95.4	94.8	95.1	95.9	94.7	93.9	95.3	94.4	94.8	94.7	96.4	95.3	93.6
	b = 1.5	95.4	95.1	94.2	94.9	95.4	94.9	95.2	94.9	94.2	95.4	94.1	95.2	95.4	94.8	95.2	95.1	95.4	94.8
	b = 2	96.0	95.3	96.3	94.2	95.2	94.9	95.4	96.3	95.3	94.2	95.0	94.6	94.9	94.6	96.0	95.0	95.4	93.8
C-SR	b = 0	94.6	95.9	96.1	95.2	94.4	93.7	95.3	94.9	96.6	94.5	95.6	95.9	94.4	95.5	95.4	94.9	93.5	94.7
	b = 0.5	93.3	93.6	95.6	95.3	93.5	94.6	95.1	96.1	94.0	94.3	95.3	95.0	94.6	94.6	93.8	95.4	93.7	95.6
	b = 1	94.6	94.9	94.5	95.0	93.5	94.6	95.2	93.5	95.9	94.7	94.9	93.0	95.7	94.7	95.0	95.1	95.1	94.4
	b = 1.5	95.2	93.9	94.8	95.8	96.3	94.8	93.8	94.1	94.3	94.4	95.5	93.7	95.6	96.2	95.1	95.4	94.9	95.1
	b = 2	95.3	96.2	94.5	95.4	95.3	95.2	95.2	95.9	94.7	95.0	94.8	95.2	94.4	93.6	94.8	95.1	94.7	94.9
B-P	b = 0	96.1	94.5	96.4	96.1	95.4	95.4	93.8	95.4	95.3	95.8	95.7	94.0	94.5	95.7	95.2	94.5	94.8	95.7
	b = 0.5	94.8	95.1	95.4	95.1	95.7	95.0	95.3	95.4	96.0	94.7	95.2	94.5	95.2	96.5	94.3	95.5	94.8	94.8
	b = 1	95.2	95.7	95.1	95.6	95.3	95.4	95.8	95.2	96.1	95.0	96.3	96.0	95.6	96.0	95.0	95.1	94.5	96.3
	b = 1.5	95.9	94.8	94.6	95.2	95.2	93.8	95.3	94.2	94.5	95.5	94.8	95.2	95.9	95.6	96.1	95.0	94.7	95.0
	b = 2	95.3	95.1	95.0	94.1	95.1	94.2	94.5	94.8	93.5	94.4	94.7	95.0	95.9	95.3	96.1	95.3	94.9	95.9

Table 4: RT-SR results: 1,000*RelBias and CI coverage (%), by population

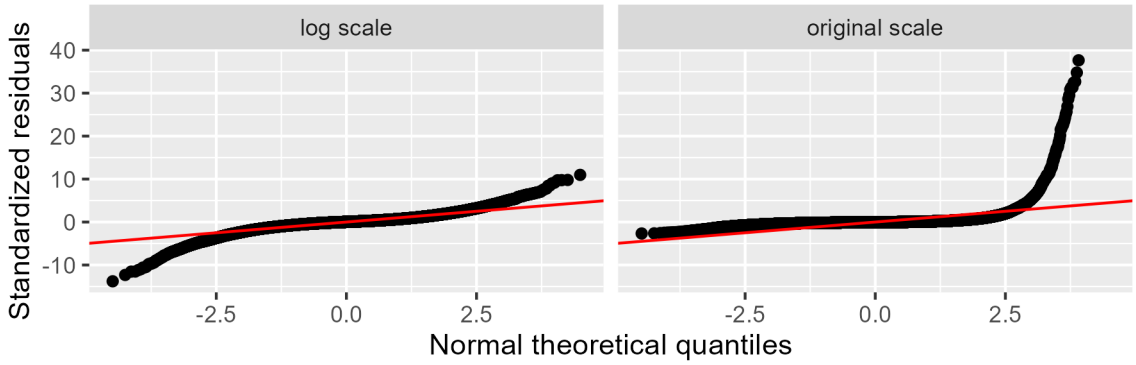
		MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
		1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
RelBias*1000	b = 0	-0.6	-1.1	0.1	-0.0	-0.4	-0.6	-0.0	-0.1	0.0	-0.0	-0.0	0.4	0.2	0.1	-0.1	0.4	0.1	-0.2
	b = 1	0.0	0.2	0.2	-0.2	-0.6	0.1	0.3	0.8	-0.0	-0.1	-0.0	-0.4	0.6	0.1	-0.0	0.4	-0.2	-0.1
	b = 2	-0.4	-0.1	0.2	-0.7	-0.9	1.7	0.0	0.3	0.0	0.5	-0.2	1.2	-0.1	0.2	0.0	0.4	-0.0	0.0
Coverage (%)	b = 0	93.8	95.4	96.1	95.2	94.0	95.3	95.1	95.4	95.0	95.3	95.3	95.0	95.3	94.1	95.9	94.6	94.8	95.9
	b = 1	94.2	94.1	93.7	94.8	94.3	93.9	94.5	93.7	95.5	95.2	94.7	94.2	95.3	95.0	93.9	95.0	94.1	93.9
	b = 2	89.8	92.2	91.3	91.6	89.2	90.7	95.7	93.9	94.3	93.3	94.5	93.9	93.8	95.3	94.2	93.3	94.1	94.7

Table 5: B-P results under model misspecification: 1,000*RelBias and CI coverage (%) by b_0 and $b^{(p)}$ among selected MOS1 populations

		$b^{(p)}=0$						$b^{(p)}=1$						$b^{(p)}=2$					
		1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
RelBias*1000	$b_0=0$	1.0	1.1	0.4	-1.5	0.2	2.4	-0.0	0.6	0.1	0.2	0.0	-0.2	-0.7	0.6	0.5	2.7	1.9	2.3
	$b_0=0.5$	0.9	-0.6	0.5	-1.6	0.5	0.8	-0.6	0.1	0.2	0.4	0.9	-0.8	-0.5	2.3	0.7	3.1	1.7	2.0
	$b_0=1$	1.0	-0.7	0.3	-1.3	-0.3	2.0	-0.2	-0.4	-0.1	0.0	0.0	1.1	-0.4	1.3	0.4	2.9	2.5	2.6
	$b_0=1.5$	2.2	-1.8	0.1	-3.4	0.4	-0.1	0.0	0.2	-0.1	-0.1	-0.0	-0.2	-0.6	1.5	0.4	3.0	1.5	2.3
	$b_0=2$	1.6	-3.8	-0.5	-1.8	0.8	3.1	-0.2	0.2	0.1	-0.4	-0.3	-0.1	-0.4	2.3	0.6	2.9	1.6	2.4
Coverage (%)	$b_0=0$	96.1	94.5	96.4	96.1	95.4	95.4	95.8	94.6	95.0	93.4	95.5	93.8	96.0	95.5	95.8	94.9	93.7	95.6
	$b_0=0.5$	94.3	94.8	95.1	95.0	96.1	94.1	94.3	95.1	95.3	94.4	94.7	95.1	94.9	95.2	95.1	94.6	95.8	96.6
	$b_0=1$	93.7	95.3	94.8	95.9	95.4	95.8	95.2	95.7	95.1	95.6	95.3	95.4	95.0	95.2	94.9	94.6	95.2	94.8
	$b_0=1.5$	97.0	96.5	96.7	97.1	95.7	98.1	95.7	95.5	95.5	94.7	95.6	95.1	94.0	96.2	96.4	94.6	96.3	95.0
	$b_0=2$	98.6	98.1	98.3	98.2	98.8	98.1	96.1	96.0	95.9	95.6	95.4	97.6	95.3	95.1	95.0	94.1	95.1	94.2

Note: Gray background denotes use of the correct specification (i.e., $b_0 = b^{(p)}$).

Figure 1: Q-Q plots for unstratified regressions of IRS data, based on log-transformed and original scales



Note. Lefthand panel is a Q-Q plot for an unstratified, no-intercept simple linear regression of the NCCS data from Section 6.1 of the main paper, where X_i and Y_i denote log revenues of unit i for 2008 and 2013, respectively, with analytical weights of by $1/X_i^{0.55}$. Righthand panel is an analogous Q-Q plot for the NCCS data on the original scale, as in Section 6.2 of the main paper, meaning that X_i and Y_i denote revenues of unit i for 2008 and 2013, respectively, and with analytical weights of $1/X_i^{1.19}$. Righthand panel does not display six observations with standardized residuals greater than 40.

Table 6: NCCS data: population counts by sector and 2008 revenue (with size classes based on original scale)

	Total Revenue in 2008			
Nonprofit Sector	Under \$1.8m	\$1.8m–\$12m	\$12m–\$100m	Total
Arts, culture, and humanities	12,239	1,709	314	14,262
Education	12,996	4,507	1,646	19,149
Environment	5,246	781	118	6,145
Health	11,675	5,755	3,185	20,615
Human services	44,088	10,824	2,609	57,521
International	2,332	482	176	2,990
Public and societal benefit	8,740	1,702	366	10,808
Religion	6,681	701	88	7,470
Total	103,997	26,461	8,502	138,960