ABSTRACT

Title of Dissertation: OPTIMIZING STRATIFIED SAMPLING ALLOCATIONS TO ACCOUNT FOR HETEROSCEDASTICITY AND NONRESPONSE

Jonathan Mendelson
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Michael R. Elliott
Professor Partha Lahiri
Joint Program in Survey Methodology
Michigan Program in Survey and Data Science

Neyman's seminal paper in 1934 and subsequent developments of the next two decades transformed the practice of survey sampling and continue to provide the underpinnings of today's probability samples, including at the design stage. Although hugely useful, the assumptions underlying classic theory on optimal allocation, such as complete response and exact knowledge of strata variances, are not always met, nor is the design-based approach the only way to identify good sample allocations. This thesis develops new ways to allocate samples for stratified random sampling (STSRS) designs.

In Papers 1 and 2, I provide a Bayesian approach for optimal STSRS allocation for estimating the finite population mean via a univariate regression model with heteroscedastic errors. I use Bayesian decision theory on optimal experimental design, which accommodates uncertainty

in design parameters. By allowing for heteroscedasticity, I aim for improved realism in some establishment contexts, compared with some earlier Bayesian sample design work. Paper 1 assumes that the level of heteroscedasticity is known, which facilitates analytical results. Paper 2 relaxes this assumption, which results in an analytically intractable problem. Thus, I develop a computational approach that uses Monte Carlo sampling to estimate the loss for a given allocation, in conjunction with a stochastic optimization algorithm that accommodates noisy loss functions. In simulation, the proposed approaches performed as well or better than the design-based and model-assisted strategies considered, while having clearer theoretical justification.

Paper 3 changes focus toward addressing how to account for nonresponse when designing samples. Existing theory on optimal STSRS allocation generally assumes complete response. A common practice is to allocate sample under complete response, then to inflate the sample sizes by the inverse of the anticipated response rates. I show that this practice overcorrects for nonresponse, leading to excessive costs per effective interview. I extend the existing design-based framework for STSRS allocation to accommodate scenarios with incomplete response. I provide theoretical comparisons between my allocation and common alternatives, which illustrate how response rates, population characteristics, and cost structure can affect the methods' relative efficiency. In an application to a self-administered survey of military personnel, the proposed allocation resulted in a 25% increase in effective sample size compared with common alternatives.

OPTIMIZING STRATIFIED SAMPLING ALLOCATIONS TO ACCOUNT FOR HETEROSCEDASTICITY AND NONRESPONSE

by

Jonathan Mendelson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Michael R. Elliott, Co-Chair
Professor Partha Lahiri, Co-Chair
Professor Roderick J. A. Little
Professor Eric V. Slud
Professor Richard L. Valliant

Dedication

To Amanda, David, and Anna

To my sister, Rebecca, my mom, Michele, and in memory of my dad, David

Acknowledgments

I could not have asked for a better advisor than Dr. Michael Elliott. I am immensely grateful for Mike's guidance and mentorship throughout this process as well as his time—the main drawback of completing my dissertation is that I'll miss our Friday meetings. I would also like to express my deepest appreciation to the rest of my committee—especially my committee co-chair, Dr. Partha Lahiri, as well as Dr. Richard Valliant, Dr. Eric Slud, and Dr. Roderick Little. I am grateful for my committee's many thought-provoking questions, ideas, and useful feedback, many of which led to direct improvements in my work, and others of which will shape how I think about research in the future.

I appreciate the help that Dr. Joseph Sedransk provided during the early stage of my thesis. Joe introduced me to Bayesian decision theory and provided detailed technical feedback and suggestions regarding the theory underlying Chapter 2 and parts of Chapter 3.

I extend my sincerest thanks to the rest of the JPSM/MPSDS community, especially my instructors for Ph.D. Seminar (Dr. James Wagner, Dr. Stanley Presser, and Dr. Frauke Kreuter), from whom I learned a great deal on how to conduct research. Likewise, I thank the numerous outstanding faculty whose courses I took, my classmates for many interesting discussions, and the administrative staff.

I am deeply grateful to my colleagues at Bureau of Labor Statistics for their incredible support while I have been pursuing my degree. I cannot imagine a more supportive employer.

I also acknowledge the University of Maryland supercomputing resources (http://hpcc.umd.edu) made available for conducting the research reported in Chapter 3.

Finally, I am extremely grateful to my family and friends for their support and for believing in me. This especially includes my wife, Dr. Amanda Ginter, without whom I would not have reached this milestone.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction: revisiting the foundations of stratified sample allocation

Random sampling allows for estimates based on sample data of a randomly drawn subset of the population, rather than from a census. This allows for survey designers to achieve accurate population estimates at substantially reduced cost. However, an inefficiently designed sample can lead to excessive costs or inadequate precision. Thus, sampling theory has long focused on how to best design samples. This thesis looks at such issues, with a focus on sample allocation for stratified random samples.

Design-based survey sampling is largely rooted in Neyman's (1934) seminal paper (Brick, 2011; Frankel & Frankel, 1987; Hansen, 1987; Rao, 2011; T. Smith, 1976) and subsequent developments that were consolidated in fundamental sampling texts (e.g., Cochran, 1977; Hansen, Hurwitz, & Madow, 1953; Kish, 1965). Aside from "effectively demolish[ing] the idea of purposive sampling" (Fienberg & Tanur, 1996) in favor of probability sampling, Neyman dispelled the notion that stratified sampling required a proportional allocation. Further, Neyman provided theory for how to optimally allocate sample to strata for single-stage designs and proposed the notions of stratified multi-stage sampling and confidence intervals.

Under the design-based paradigm for finite population sampling, population characteristics are treated as fixed, and inferences are randomization-based, that is, based on the probability distribution for selecting a particular sample according to the given design. This is in contrast to a model-based approach, wherein inferences are based on the probability distribution for a model, as is standard in other areas of statistics (e.g., Little, 2004).

The differences between design-based and model-based approaches to inferences, as well as the foundational underpinnings to sampling, were heavily debated in the 1970s and 1980s (e.g., Basu, 1971; Brewer, 2013; Hansen, Madow, & Tepping, 1983; Royall, 1970; T. Smith, 1976; Valliant, 2023). This controversy focused on inferential issues, such as descriptive inferences for finite populations (e.g., Hansen et al., 1983; T. Smith, 1976) and the role of weights in regression (e.g., DuMouchel & Duncan, 1983). In contrast to this controversy on the use of models in inferences, Kalton (1983) noted that the use of models to inform probability-based sample designs has been relatively uncontroversial, considering that they do not affect the resulting validity of inferences, which are still based on a randomization approach. Likewise, the model-assisted approach (Särndal, Swensson, & Wretman, 1992), as cited by Rao (2011), entails using "a working population model...to find efficient design-consistent estimators."

Although recent Bayesian sampling research has largely focused on inferential topics (e.g., Ghosh, 2009; Rao, 2011; Rao & Molina, 2015), there is an apparently overlooked body of work on Bayesian optimal design, published in the mid-1960s through the early 1980s, which we will review in §2.2.2. This early Bayesian design literature, best exemplified by Draper and Guttman (1968b) for continuous data and Rao and Ghangurde (1972) for categorical

data, entailed development of optimal STSRS designs for estimating population totals and proportions, with extensions toward multivariate sample allocation problems, multi-stage sampling, double sampling for response, multi-phase sampling, and allocation to multiple modes. The Bayesian approach to sample design appears to have been overlooked, as there is little recent work in this area and are few references to this approach in key sampling texts (e.g., Cochran, 1977; Särndal et al., 1992; Valliant, Dever, & Kreuter, 2018), other relevant survey statistics texts (Ghosh & Meeden, 1997; Rao & Molina, 2015), or relevant review articles (Brick, 2011; Frankel & Frankel, 1987; Rao, 2011; T. Smith, 1976). The early Bayesian sample design work generally entailed use of *preposterior* analysis, in which the posterior distribution (conditional on the future data) is averaged over the predictive distribution for the future data, after which the expected loss (e.g., expected posterior variance) could be minimized.

A unified Bayesian framework that can accommodate such design problems is Bayesian decision theory for optimal experimental design (e.g., Lindley, 1972), which has received substantial attention in non-survey contexts (Chaloner & Verdinelli, 1995; Ryan, Drovandi, McGree, & Pettitt, 2016). We will subsequently employ this approach in Chapters 2 and 3, in conjunction with the Bayesian framework for finite population inference, following Ericson (1969a). In doing so, we will assume use of noninformative priors.

Considering that survey sampling theory has long been driven by design-based thinking, especially at the design stage, why should we now revisit the foundations of stratified sampling allocation? A few reasons are:

- **Increasing nonresponse**. Survey response rates have declined dramatically in recent years (Brick, 2013; Luiten, Hox, & de Leeuw, 2020; National Academies of Sciences, Engineering, and Medicine, 2018; Williams & Brick, 2018), posing many consequences (Peytchev, 2013). The use of models is needed to account for the effects of nonresponse (e.g., Brick, 2013), and results in estimators that are no longer purely design-based. However, Neyman allocation assumes complete response, ignoring the effects of nonresponse on precision and on cost efficiency, while leading toward ambiguity for how sample designers should account for nonresponse when determining allocations.

- **Increasing acceptance of Bayesian statistics**. Fienberg (2006) chronicled Bayesian inferences as "com[ing] of age" in the 1960s, with substantial increases in publications and adoption in the decades that followed. The increasing use of Bayesian statistics has extended to certain sample survey contexts, such as small area estimation (e.g., Rao & Molina, 2015) and responsive survey design (Coffey & Elliott, 2023; Coffey, West, Wagner, & Elliott, 2020; Schouten et al., 2018; Wagner, West, Elliott, & Coffey, 2020; West, Wagner, Coffey, & Elliott, 2021). Design-based theory was largely developed before key advancements in Bayesian methodology. Further, Bayesian consideration of sample allocation has received fairly little attention. The Bayesian approach offers several potential uses for surveys (Little, 2006, 2022), including in a design context (Sedransk, 2008). Bayesian methods also can accommodate uncertainty in the population parameters needed to design the sample (e.g., strata variances), which design-based theory assumes are known without error, an issue that was studied by Clark (2013). Further, it is possible to incorporate key aspects of design-based thinking within a

Bayesian framework (Little, 2006, 2022).

- **Computing advances**. Digital computers did not exist when Neyman allocation was first proposed. Computing advances have facilitated the use of more complicated estimators and models in survey sampling, while also being central to advancements in Bayesian methodology. On the latter, as recounted by Hitchcock (2003), an important turning point came in the late 1980s and 1990s with the rise of computing power and popularization of Gibbs sampling (e.g., Gelfand & Smith, 1990), which is a Markov chain Monte Carlo (MCMC) method, and which mitigated an earlier reliance on conjugate priors. However, existing Bayesian sample design literature largely predated modern computing, besides predating recent advancements in Bayesian optimal experimental design (e.g., Ryan et al., 2016).

Chapter 2 uses a Bayesian formulation to identify the optimal allocation for stratified random sampling for estimating the finite population mean under a univariate regression model with heteroscedastic errors. On the latter, the model incorporates a *coefficient of heteroscedasticity*, as may apply in some establishment contexts (e.g., Henry & Valliant, 2009), and which is assumed in Chapter 2 to be known. In contrast with design-based sample allocation theory, the Bayesian formulation does not assume that the strata variances are known. By allowing for heteroscedastic errors, our assumptions aim for improved realism in establishment contexts compared with some earlier Bayesian design work, which did not. We aim to minimize the expected posterior variance for the finite population mean. We derive the approximate expected posterior variance, which allows for computing the optimal allocation through mathematical programming. We assess performance empirically through simulation

studies using numerous artificial populations. The simulations are then repeated in an application to IRS Form 990 data on the revenues of public charities. We find that the proposed methods provide substantial gains compared with a design-based approach, and performance is similar or better than the main model-assisted alternative considered.

The Bayesian optimal design, as per Lindley (1972), minimizes the expected loss (e.g., expected posterior variance) over the space of possible designs (e.g., possible sample allocations). Unless the prior and likelihood are chosen in a manner that facilitates analytical results, Bayesian design can quickly become computationally challenging, by requiring Bayesian analysis on numerous future data sets to evaluate the expected loss for a single design, a process that must be repeated many times during optimization (e.g., Ryan et al., 2016). Chapter 2 assumes the coefficient of heteroscedasticity to be known, which facilitates analytical results. Chapter 3 relaxes this assumption, which results in an analytically intractable problem, leading us to develop a computational approach.

Thus, Chapter 3 develops a fully Bayesian computational approach for stratified sample allocation in heteroscedastic populations with unknown coefficient of heteroscedasticity. We specify a prior on the coefficient of heteroscedasticity and derive the resulting posterior distribution for superpopulation parameters. As a result of this model extension, the objective function (i.e., expected posterior loss) lacks a closed form. We use Monte Carlo sampling to estimate the loss for a given allocation in conjunction with stochastic optimization methods for identifying the optimum. For the latter, we primarily consider simultaneous perturbation stochastic approximation (SPSA), which is a stochastic optimization algorithm that accommodates noisy loss functions, and which we tailor in our application to handle constraints

and guarantee full use of the sample. We compare the proposed methods numerically with design-based, model-assisted, and pseudo-Bayesian alternatives. In simulation and application, the proposed fully Bayesian methods perform as well or better than the alternatives, while providing clearer theoretical justification. As an alternative to SPSA, we also consider Nelder-Mead, which turns out to perform well in simulation, but does not scale well to a higher dimensional application. Beyond providing methodological developments for our specific problem, Chapter 3 illustrates potential opportunities and challenges associated with developing fully Bayesian computational methods for optimal sample design.

Chapter 4 switches gears, returning to a design-based/model-assisted approach for optimal allocation, but now considering the problem of nonresponse. Survey response rates have declined dramatically in recent years, and the types of efforts needed to seriously raise them, such as face-to-face interviewing, are not always feasible. Existing mathematical theory on allocation for single-stage stratified sample designs generally assumes complete response, resulting in ambiguity as to how sample designers should incorporate the effects of nonresponse into the sample allocation process. It is well known that ignoring (anticipated) differential nonresponse when allocating sample can lead to excessive design effects. One commonly used approach aims to account for this issue by allocating a sample under complete response, and then inflating the sample sizes by the inverse of the anticipated response rates. However, we show in Chapter 4 that this method can fail to improve upon an unadjusted allocation, in terms of cost per effective interview, due to ignoring the effects of differential nonresponse on costs.

Therefore, in Chapter 4, we provide mathematical theory on how to allocate single-stage de-

signs in a manner that incorporates the effects of nonresponse on cost efficiency. We derive the optimal allocation for a poststratified estimator under nonresponse, which minimizes either the unconditional variance of our estimator or the expected costs, holding the other constant, and taking into account uncertainty in the number of respondents. We assume a cost model that incorporates effects of nonresponse. We provide theoretical comparisons between our allocation and common alternatives, which illustrate how response rates, population characteristics, and cost structure can affect the methods' relative efficiency. In an application to a self-administered survey of U.S. military personnel, the proposed allocation increases the effective sample size by 25%, compared with common practice. Our results highlight the importance of incorporating the effects of nonresponse on the assumed cost structure when computing allocations.

Chapter 5 summarizes and discusses our contributions and outlines some potential future directions for research.

Chapter 2: Bayesian optimal sample design for establishment surveys with heteroscedasticity: an analytical approach

Abstract

For many survey sample designs, there are analytical results for the optimal sample allocations, but which assume knowledge of certain survey design parameters (e.g., strata variances). Practitioners commonly use these results by substituting survey-based estimates of design parameters in place of their true (unknown) values. Typically, little attention is given to the potential effects of using imperfect information at the sample design stage, even though this could harm sample efficiency.

We consider sample allocation for a univariate regression model with heteroscedastic errors, using Bayesian decision theory to accommodate uncertain design parameters. By allowing for heteroscedasticity, our assumptions may be more realistic in an establishment context than some previous Bayesian design work, which did not. We 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.

2.1 Introduction

Random sampling allows for estimates of finite population parameters based on sample data from a randomly drawn subset of the population, rather than from a census. This allows for survey designers to achieve population estimates of sufficient precision at a greatly reduced cost. However, a poorly designed sample can result in excessive costs or inadequate precision. Thus, 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).

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. Typically, little attention is given to the potential effect of imperfect information on the resulting design, even though this choice could result in an inefficient sample.

In practical terms, given the scale of spending on surveys in the United States, even small improvements in efficiency could lead to meaningful cost savings. Between 1984 and 2004, there were dramatic increases in the budgets for the 10 principal statistical agencies of the United States and in the number of survey respondents approved for their survey efforts

via the Office of Management and Budget (Presser & McCulloch, 2011). Although the exact annual spending on survey research is unclear, federal funding on statistical agencies exceeded $6 billion in Fiscal Year 2015 (Office of Management and Budget, 2017), and the Insights Association reported that domestic private sector marketing research revenues were approximately $18 billion in 2020 (Brereton & Bowers, 2021).

Despite this vast spending on sample surveys, there is little recent literature on designing samples in the presence of uncertain design parameters. From a frequentist perspective, the main exception is Clark's (2013) proposed statistical learning approach for stratified random sampling, which Clark demonstrates can lead to improvements, while noting 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 & Guttman, 1968b; Rao & 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 establishments, wherein distributions are often heavily skewed and where a continuous specification may be more appropriate.

Summing up, there are clear gaps in the existing literature regarding how to best account for uncertain sample design parameters. There is little frequentist work on this topic, and despite a sizable volume of early Bayesian work, there are important practical situations where the assumptions may be unrealistic, including for establishment surveys. Given the numerous sample survey contexts in which design parameters may be estimated inaccurately, we see many research opportunities. For example, different contexts may vary with respect to the type of inaccuracy in the design parameters (e.g., systematic versus variable errors), type of unit surveyed (e.g., establishment versus household), sample design, and/or population model.

Since there are many opportunities, we choose a simple set-up that aims to provide more realistic assumptions for some practical applications compared with previous work, with the notion that other contexts may provide avenues for future work. We examine stratified sampling for establishment surveys wherein there is a population level auxiliary variable, $\{X_{hi}\}$, that is correlated with the survey outcome, $\{Y_{hi}\}$. For each stratum $h = 1, ..., H$, we assume a population model of $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$ where $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$, $X_{hi} > 0$ is known for the population and is treated as fixed, $\{\alpha_h, v_h\}$ are unknown, and $b \geq 0$ is for now assumed to be known (Chapter 3 considers unknown b). The term b accommodates varying levels of heteroscedasticity, as is frequently observed in establishment contexts (e.g., Henry & Valliant, 2009; P. Smith, Pont, & Jones, 2003), and the strata-specific terms allow for group differences. We assume that some pilot data set is available for estimating parameters needed

for designing a main study, which we assume employs stratified random sampling. Given that establishment surveys often stratify based on multiple variables (e.g., size, geography, and industry), we 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., Dalenius & Hodges, 1959; Rivest, 2002; Wright, 1983).

Our contribution entails the derivation of analytical results that are more relevant to establishment surveys than previous work on Bayesian optimal sample design, by virtue of employing a more flexible model. 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. We assess performance through simulation studies using both artificial populations and real data. The proposed Bayesian allocation and associated prediction estimator provide substantial gains compared with Neyman allocation and Horvitz-Thompson estimation. In addition, performance is similar or better than a model-assisted allocation for ratio estimation outlined by Cochran (1977, §6.14).

Background is provided in Section 2.2, and includes attention to stratified sample allocation under our assumed model, sampling with uncertain design parameters, and Bayesian optimal experiment design. Section 2.3 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 2.4 outlines simulation methods for comparing the proposed allocations with alternatives across 90 artificial populations. Section 2.5 presents simulation findings. Section 2.6 repeats the simulation in an application to the log-revenues of public charities, using data from the National Center for Charitable Statistics. Section 2.7 summarizes results, discusses limitations, and suggests future directions.

2.2 Background

Section 2.2 is organized as follows:

- In §2.2.1, we review several topics that are relevant to stratified sampling for establishments, largely from the design-based, model-assisted literature, as is commonly employed in practice. First, we introduce Neyman allocation, after which we briefly describe the common practice in the design stage of substituting estimates in place of unknown quantities. Next, we discuss the motivation behind the separate ratio estimator, as is commonly employed in establishment surveys, as well as associated sample allocations. This leads into discussion of the impact of heteroscedastic errors on sampling.

- In §2.2.2, we review literature on sampling with uncertain design parameters, with a primarily Bayesian focus. We summarize early work on Bayesian optimal sample design, while providing evidence for the apparent lack of more recent work. We briefly summa-

rize recent developments in Bayesian responsive survey design, which help illustrate the promise of Bayesian methods for accommodating uncertain design information, albeit with a different focus and in a different context than our research.

- In §2.2.3, we describe Bayesian decision theory for optimal experimental design, following Lindley (1972). This entails identifying the optimal experiment (i.e., sample design) and estimator for minimizing the expected posterior loss, the latter of which entails averaging the posterior loss (conditional on the future data) over the prior predictive distribution.

2.2.1 Design-based sampling and related topics

2.2.1.1 Optimal allocation for STSRS-π strategy

From a design-based perspective, one must consider a *strategy*, which denotes a sample design in conjunction with an estimator. A basic class of strategies for estimating a population total involves stratified simple random sampling (STSRS) in conjunction with the Horvitz-Thompson (π) estimator. Neyman (1934) showed that the optimal sample allocation for this type of strategy is $n_h \propto N_h S_h$, wherein $S_h = \sqrt{V_h} = \sqrt{\frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2}$ is the stratum standard deviation and $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$ is the stratum population mean. Neyman assumed equal per-unit costs across strata, but this allocation can be modified under variable costs as $n_h \propto N_h S_h / \sqrt{c_h}$, where c_h is the cost of obtaining an interview in stratum h (e.g., Cochran, 1977, §5.5).

2.2.1.2 Plug-in method

Optimal sample allocation calculations typically assume that some population characteristics are known. In practice, some of these design parameters are typically unknown, but estimates may be available, and so a *plug-in* method is commonly used, in which survey estimates from a previous time interval are substituted in place of the true quantities in estimating the optimal allocation (e.g., Clark, 2013; Mahalanobis, 1952). In the case of Neyman allocation, and allowing for costs to vary across strata, this entails an allocation of $n_h \propto N_h \hat{S}_h / \sqrt{c_h}$, where $\hat{S}_h^2$ could be computed as the sample variance from some previous sample.

2.2.1.3 Optimal allocation for the separate ratio estimator

Next, consider a situation where our focus is still on estimating the population total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$, in a context where we now know the values for a related variable, $\{X_{hi}\}$, for the population, but only know the $\{Y_{hi}\}$ for our sample, which is assumed to have been chosen via an STSRS design. Assume that within each stratum, the $\{Y_{hi}\}$ and $\{X_{hi}\}$ are correlated, but that the correlations 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 $X_h = \sum_{i=1}^{N_h} X_{hi}$ and $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ denote the stratum population totals of our outcome and auxiliary variables, respectively. In addition, let $R_h = \frac{X_h}{Y_h}$ be the ratio of stratum population totals. This scenario commonly

employs the separate ratio estimator,

$$\hat{Y}_{Rs} = \sum_{h=1}^{H} \hat{R}_h X_h \tag{2.1}$$

$$= \sum_{h=1}^{H} \frac{\bar{y}_h}{\bar{x}_h} X_h, \tag{2.2}$$

termed *separate* (versus *combined*) by virtue of allowing the strata slopes to vary. Cochran (1977, §6.14) indicated that the optimal allocation for the separate ratio estimator is

$$n_h \propto \frac{N_h S_{dh}}{\sqrt{c_h}} \tag{2.3}$$

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 (2.3) 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 other 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

$$\hat{S}_{dh}^2 = \frac{1}{m_h - 1} \sum_{i \in s_{1h}} \left(y_{hi} - \hat{R}_h x_{hi} \right)^2, \tag{2.4}$$

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, p. 32), suggesting worsened performance when using smaller pilot samples.

2.2.1.4 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 et al. (1992) do not directly provide a solution, Särndal et al. address optimal design for a related model under somewhat different problem formulations. For instance, Särndal et al. 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

$$Y_i = \beta X_i + \varepsilon_i \tag{2.5}$$

$$\mathrm{E_M}(\varepsilon_i) = 0 \tag{2.6}$$

$$\mathrm{Var_M}(\varepsilon_i) = \sigma^2 X_i^b \tag{2.7}$$

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, & Royall, 2000, p. 177). For instance, Henry and Valliant describe several establishment or accounting applications and provide $\{X_i, Y_i\}$ plots for five real populations with a similar structure, and where estimates of b range from 0.62 to 2.03. For another example, P. Smith et al. (2003) indicated that b was commonly between 1 and 1.5 in government establishment surveys in Britain for a separate-slopes model.

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 et al., 1992, Section 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 et al. 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 (2.7) with the weaker assumption that $\text{Var}_\text{M}(Y_i) = v_i$ 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, p. 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 (see Appendix A.1.2 for detail).

Although the model expressed by Equations (2.5–2.7) 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 $\text{Var}_\text{M}(\varepsilon_i) = \sigma^2 X_i^b$ accommodates heteroscedasticity, but it 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.

2.2.2 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 & Guttman, 1968b; Ericson, 1965, 1967a, 1969b, 1972; Khan, 1976a; Rao & Ghangurde, 1972) and proportions (Grosh, 1972; Zacks, 1970). 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 & Guttman, 1968b; Rao & 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 sample allocation problems (Draper & Guttman, 1968a; Khan, 1976b; 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 & Ghangurde, 1972), double sampling for nonresponse (Ericson, 1967b; Rao & Ghangurde, 1972; Singh & Sedransk, 1978, 1984), and multi-phase sampling (Jinn, Sedransk, & Smith, 1987; P. J. Smith & Sedransk, 1982). Ericson (1983a, 1983b) also 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).

More recently, there has been a lack of attention to Bayesian optimal sample designs for finite population sampling. 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., Gelman et al., 2014; Ghosh, 2009; Rao, 2011, pp. 229–230, 259), and in conducting literature searches, I did not find mentions of Bayesian methods for optimal sample design beyond those cited above, with one exception by O'Malley and Zaslavsky (2016) in the context of SAE. However, Sedransk (2008) pointed out that Bayesian methods may be helpful at the design stage. More generally, models are often employed in designing samples (e.g., Hansen et al., 1983; Kalton, 1983; Särndal et al., 1992), regardless of whether the ultimate inferences are design-based.

Recent developments in adaptive or responsive survey design (RSD) further illustrate the potential promise for Bayesian methods at handling scenarios with imperfect information about survey design parameters, albeit in a different setting than our research. RSD proposes to

improve the efficiency of survey data collections by using data collected during earlier phases to inform changes to protocols for later phases (Groves & Heeringa, 2006). Bayesian modeling has recently been applied in RSD to improve estimates of uncertain design parameters, namely, costs (Schouten et al., 2018; Wagner et al., 2020) and response propensities (Coffey et al., 2020; Schouten et al., 2018; West et al., 2021). This has entailed development of a general Bayesian analysis framework for modeling survey design parameters (Schouten et al., 2018), as well as three different applications using data from the National Survey of Family Growth (NSFG; Coffey et al., 2020; Wagner et al., 2020; West et al., 2021), which is a face-to-face survey that uses RSD in the context of a multi-stage area probability sampling design. Methods have been developed to elicit Bayesian priors to improve daily estimated response propensities (Coffey et al., 2020; West et al., 2021), through historical data (West et al., 2021), via literature review (West et al., 2021), or by collecting and pooling expert knowledge obtained through a survey questionnaire (Coffey et al., 2020). In addition, Wagner et al. (2020) developed methods for modeling interviewer costs in NSFG, some of which entailed Bayesian modeling (via use of Bayesian Additive Regression Trees). This recent literature illustrates that Bayesian methods can help to address uncertainty in design parameters through improving inference on these parameters. Several of the authors suggested the potential utility of further work on how to best use the resulting predictions in making data collection or other design decisions. Further, as per Wagner et al., the resulting predictions still have uncertainty, and it may be important to reflect this uncertainty in making design decisions.

In subsequent sections, we will show in detail how Bayesian principles can be applied to

sample optimization problems, while accounting for uncertainty in design parameters. §2.2.3 summarizes a Bayesian decision theoretic framework for optimal experimental design, following Lindley (1972), which is central to our work. We will also draw upon selected results from §5 of Ericson (1969a), which provides Bayesian analysis of a univariate regression model in a manner accommodating heteroscedastic errors; we will describe and cite specific results at the relevant places in §2.3, with further detail provided in Appendix A.2. Finally, although our subsequent analysis does not directly draw upon the key early Bayesian sampling design literature by Draper and Guttman (1968b) for continuous data and Rao and Ghangurde (1972) for categorical data, Appendix A.1.1 provides detailed reviews, further illustrating the Bayesian approach.

From a frequentist perspective, there has also been a lack of recent literature on accounting for uncertainty in determining optimal designs, an exception being Clark's (2013) proposed statistical learning approach for an STSRS design. The proposed allocation, which incorporates smoothing of imprecisely estimated strata variances, is tailored toward a scenario wherein two design data sets are available and where the strata follow some hierarchy. Clark showed via simulation that the proposed approach can improve over a naive plug-in allocation, especially for establishments, while noting that a Bayesian framework may allow for a more efficient solution.

2.2.3 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, pp. 19–21) provided the Bayesian decision theoretic framework on optimal experimental design. This framework can be used to optimize sample allocations even when relevant population quantities for designing the survey are not known upfront. We assume use of the Bayesian formulation, as summarized by Lindley.

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. Under this framework, we are interested in the optimal experiment and estimate that maximize the expected utility, or equivalently, minimize the expected loss.

For an example, which we will subsequently assume use of (in §2.3), consider a loss function of the form $L\left(\hat{\theta}, \theta, e, x\right) = \left(\hat{\theta} - \theta\right)^2$, wherein the parameter is $\theta \in \Theta$; e is the experiment; and where, for purposes of this subsection, $x \in X$ is a sample (i.e., data set) from the sample space X (i.e., space of possible data sets). For brevity, we write $e = \{n_h : h = 1, ..., H\}$, where n_h is the sample size in stratum h, and where we assume use of stratified random sampling. After data collection, assume use of a decision function of the form $\eta : X \to \Theta$; in our context, η is the estimator and $\eta(x) = \hat{\theta}$ is the estimate. In a continuous context, the optimal solution that minimizes the expected value of the loss is:

$$\min_{e} \int_X \min_{\hat{\theta}} \int_\Theta L\left(\hat{\theta}, \theta, e, x\right) p(\theta|x, e) p(x|e)\, d\theta\, dx \tag{2.8}$$

where $p(x|e)$ is a conditional probability density function (pdf) for obtaining a particular sample given the experiment and $p(\theta|x,e)$ is a posterior conditional pdf for the population parameter given the sample and experiment (see Lindley, 1972, for detail). The inner minimization step, which Lindley referred to as the *terminal analysis*, entails averaging the loss function over Θ (while treating the sample as fixed), then deciding on $\hat{\theta}$. Next, the *preposterior analysis* entails averaging the resulting quantity over the predictive distribution for the data, then selecting e to minimize this expected posterior loss.

In Equation (2.8), note that the quantity

$$\int_{\Theta} L\left(\hat{\theta}, \theta, e, x\right) p(\theta|x,e) d\theta = \underset{\theta|x,e}{\mathrm{E}} (\hat{\theta} - \theta)^2 \tag{2.9}$$

is minimized when $\hat{\theta} = \underset{\theta|x,e}{\mathrm{E}}(\theta)$. That is, under squared error loss (SEL), the terminal analysis leads us to use the posterior mean for estimation. Then, (2.8) becomes

$$\min_{e} \int_{X} \mathrm{Var}(\theta|x,e) p(x|e) dx = \min_{e} \left\{ \underset{x|e}{\mathrm{E}} \mathrm{Var}(\theta|x,e) \right\} \tag{2.10}$$

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

Lindley's framework is not specific to survey design, and can accommodate a wide variety of potential objectives, not just SEL. See Chaloner and Verdinelli (1995) for an earlier review of Bayesian optimal experimental design and Ryan et al. (2016) for a recent review of associated computational challenges that can arise, the latter of which we will examine in Chapter 3.

2.3 Analytical results

2.3.1 Overview

In this section, we provide analytical results for stratified random sampling in the establishment context. 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. Specifically, for stratum h with population size N_h, our model of interest is $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, where $\varepsilon_{hi}|(v_h, b, X_{hi}) \overset{ind}{\sim} N(0, v_h X_{hi}^b)$ and $X_{hi} > 0$ for $h = 1, ..., H$ and $i = 1, ..., N_h$. Here, we treat the $\{\alpha_h, v_h\}$ as unknown and assume that $b \geq 0$ is known. Our auxiliary variable $\{X_{hi}\}$ is known for the population, and our interest is in sample-based inferences regarding the finite population mean, $\bar{Y} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$, where $N = \sum_{h=1}^{H} N_h$. We consider optimal study design in a context wherein pilot data are available for design purposes.

After setting up our problem, we apply Bayesian decision theory to find the optimal design in a manner that accounts for uncertainty in the parameters needed to design the sample. This entails four steps:

1. Specify a loss function as a function of the parameter (i.e., finite population mean), the estimate, experiment (i.e., sample allocation), and sample, then find the estimate that minimizes the expected loss (with respect to the parameter space). We specify a squared error loss function, which leads to our loss function being minimized when estimation uses the posterior mean for the finite population mean given the main

sample.

2. Find the posterior loss, which for our problem is the posterior variance of the finite population mean given the main sample.

3. Conduct a preposterior analysis, which entails averaging over the predictive distribution for the main study data given the pilot data, resulting in the expected posterior loss for a given sample size. This step is necessary since the posterior loss is conditional upon the main study data, which have not yet been collected. Here, we use mathematical analysis to derive an approximate closed form for the expected posterior loss for a given allocation via a first-order Taylor series approximation for a sample-based quantity.

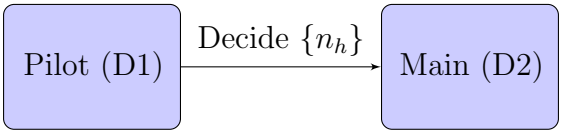
4. Find the optimal allocation for minimizing the expected posterior loss. Given that a closed form is available for the expected posterior loss, we can easily optimize this quantity within the space of possible sample allocations using standard numerical methods.

2.3.2 Problem set-up

Consider a two-phase design involving a pilot study and a main study, wherein the pilot, which has already been conducted, is used to inform the main study's design. Let m_h and n_h refer to the sample sizes in stratum h for the pilot and main study, respectively. Likewise, let $D1$ and $D2$ refer to the pilot data and main study data, respectively. Assume that there are H strata that have been determined upfront, and that each phase involves stratified random sampling (STSRS), wherein simple random sampling is applied independently for each

stratum. We aim to choose the optimal main study sample sizes, $\{n_h\}$, conditional on the pilot data, $D1$, and in a manner that accounts for uncertainty in population characteristics.

Figure 2.1: Study design.



Suppose that our stratum population sizes are $\{N_h : h = 1, ..., H\}$, with a total population size of $N = \sum_h^H N_h$. Assume a population model of $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, with $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$, where $\{X_{hi} > 0 : h = 1, ..., H; i = 1, ..., N_h\}$, and where all values of X_{hi} are known in the population (and are treated as fixed), but α_h and v_h are unknown. We assume for now that the coefficient of heteroscedasticity, b, is known.

Following a special case of a regression model considered by Ericson (1969a, §5.1), and letting $g_h = \frac{1}{v_h}$, we assume a non-informative (and improper) prior on variables (α_h, g_h) with density $\pi(\alpha_h, g_h) \propto \frac{1}{g_h}$, which we write as $\pi(\alpha_h, \frac{1}{v_h}) \propto v_h$. This prior leads to a normal-gamma posterior for variables $(\alpha_h, \frac{1}{v_h})$. Parallel to a choice considered in Draper and Guttman (1968b), we have a decision of whether the pilot data are to be used in subsequent stages of inference, or only for planning purposes. 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, if a research designer wished to make a different decision about the use of such quantities, it would entail only a basic extension of our results.

2.3.2.1 Further notation

See below for further notation for the set-up of the problem:

- Let $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ be the finite population sum for our outcome in stratum h, and let $\bar{Y}_h = \frac{Y_h}{N_h}$ be the finite population mean. Similarly, let X_h and $\bar{X}_h$ be the finite population sum and mean for our auxiliary variable.

- Let $e_1 = \{n_h\}$ denote the experiment, that is, the sample allocations for the main study.

- Let s_{1h} and s_{2h} denote the sets of units sampled in stratum h in the pilot and main study, respectively, and let s_{1h}^C and s_{2h}^C refer to the complements of these sets over the set of stratum population units U_h.

- Let x_{hi} and y_{hi} refer to sample values of the auxiliary variable and outcome variable for $h = 1, ..., H$, and for either $i \in s_{1h}$ (pilot) or $i \in s_{2h}$ (main study).

- We use $\bar{x}_h = \frac{1}{n_h} \sum_{i \in s_{2h}} x_{hi}$ and $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_{2h}} y_{hi}$ to denote main study sample means in stratum h for x and y, respectively.

2.3.3 Fixed, known b

2.3.3.1 Specify loss function, find 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 (as opposed to the typical design-based context). 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, which means that our expected loss is $\underset{\bar{Y}|D2}{\mathrm{E}}\left(\bar{Y} - \hat{\bar{Y}}\right)^2 = \mathrm{Var}\left(\bar{Y}|D2\right)$, and 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.

2.3.3.2 Find 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) 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.2.2, 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:

$$\mathrm{E}\left(\bar{Y} | D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N} \left\{ n_h \bar{y}_h + \left(N_h \bar{X}_h - n_h \bar{x}_h\right) \left(\frac{\sum_{i \in s_{2h}} \frac{x_{hi} y_{hi}}{z_{hi}}}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right) \right\} \tag{2.11}$$

$$\mathrm{Var}\left(\bar{Y} | D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left\{ \frac{1}{n_h - 3} \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{x_{hi} y_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \right\} \tag{2.12}$$

where we define $z_{hi} = x_{hi}^b$ and where, as defined earlier, $\bar{x}_h = n_h^{-1} \sum_{i \in s_{2h}} x_{hi}$ and $\bar{y}_h = n_h^{-1} \sum_{i \in s_{2h}} y_{hi}$ are main study sample means. (See §A.2.3 for detail.) Note that the choice of prior above reflects our intent that the ultimate inferences be based solely on the main study data (i.e., as opposed to also incorporating pilot data).

2.3.3.3 Conduct preposterior analysis

The posterior variance from Equation (2.12) 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 A.3 that

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) \approx \sum_{h=1}^{H} \frac{\mathrm{E}_P(v_h)}{N^2}\left(\frac{n_h-1}{n_h-3}\right)(N_h-n_h)\left[\bar{Z}_h + \frac{\frac{n_h}{N_h}S_{xh}^2 + (N_h-n_h)\bar{X}_h^2}{n_h\bar{Q}_h^2(1+CV_{qh}^2)}\right] \tag{2.13}$$

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

A summary of the derivation of the above result is as follows. First, rewrite Equation (2.12) as $\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H}\frac{1}{N^2(n_h-3)}q_{bh}k(s_{2h})$, where $q_{bh} := \left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]$ and $k(s_{2h}) := \left[\sum_{i\in(U_h\cap s_{2h}^C)} z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]$. Note that q_{bh} is a quadratic form in y whereas $k(s_{2h})$ is a fixed quantity given the set of sample indicators. Further, it turns out that q_{bh} and $k(s_{2h})$ are independent, given that q_{bh} can be re-expressed in a manner that does not depend on the sample indicators. The next step is to average the posterior variance over the distribution of $D2|D1$. Given independence of q_{bh} and $k(s_{2h})$, we analyze terms separately via $\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H}\frac{1}{N^2(n_h-3)}\underset{D2|D1}{\mathrm{E}}(q_{bh})\underset{D2|D1}{\mathrm{E}}(k(s_{2h}))$.

First, we average q_{bh} in three steps as $\underset{D2|D1}{\mathrm{E}}(q_{bh}) = \underset{\alpha_h, v_h|D1}{\mathrm{E}}\left(\underset{s_{2h}}{\mathrm{E}}\left(\underset{D2|\alpha_h, v_h, D1, s_{2h}}{\mathrm{E}}(q_{bh})\right)\right)$. The

innermost expectation, which reflects uncertainty with respect to the model (i.e., on account of the error term), can be evaluated easily via properties of the chi-square distribution, based on results from Appendix A.2.4, which uses some algebraic manipulation and results relating to quadratic forms to show that $v_h^{-1} q_{bh} | (v_h, \alpha_h, b, s_{2h}) \sim \chi^2_{n_h - 1}$ for stratum h. Given that this distribution is free of the main study sample indicators, the middle expectation over the distribution of s_{2h} (reflecting design-based uncertainty) has no impact, and we also see that $q_{bh} | D1$ and $k(s_{2h})$ are independent. The collective result is that $\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y} | D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left(\frac{n_h - 1}{n_h - 3}\right) \left[\mathrm{E}_P(v_h)\right] \mathrm{E}\left(k(s_{2h})\right)$. Next, we compute $\mathrm{E}_P(v_h)$ as a function of pilot data using properties of the inverse-gamma distribution, from Ericson's posterior distribution. We then use a first-order Taylor series approximation for $\mathrm{E}\left(k(s_{2h})\right)$, which assumes large $\{n_h\}$, leading to Equation (2.13).

In the above, the Taylor series (TS) approximation was used on account of the rightmost term in $k(s_{2h})$, which is nonlinear and where the only source of uncertainty is with respect to the sample indicators. 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. This MC estimator could be used to assess the accuracy of the TS approximation or, if needed, could be used in its place. 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 A.3.4.

2.3.3.4 Obtain optimal allocation

Equation (2.13) 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 et al., 2013, Chapter 5). For instance, a common formulation would entail aiming to minimize Equation (2.13) 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 (2.12). Survey designers could add any other relevant constraints, as well (e.g., precision constraints for domain estimation).

After computing an allocation, if Equation (2.13) 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 §2.3.3.3.

2.4 Simulation design

2.4.1 Overview

Simulation methods were used to compare performance of the proposed sampling strategy with alternative stratified random sampling (STSRS) designs. Comparisons were conducted across a series of artificially generated populations that follow our model but which varied

with respect to the population-generating parameters. 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. Then, we compared results primarily in terms of RMSE, relative bias, and coverage and relative width of 95% CIs.

The remainder of §2.4 is organized as follows. In §2.4.2, we describe the population-generating process for creating $P = 90$ populations, which vary with respect to the size measures (3x), the coefficients of heteroscedasticity (5x), and the stratum slopes and correlations (6x). In §2.4.3, we describe pilot sampling. In §2.4.4, we summarize the sampling strategies compared, which included design-based and model-assisted alternatives to the proposed Bayesian methods. In §2.4.5, we describe the main metrics for comparison.

2.4.2 Population data for simulation

Due to confidentiality-related issues, there is a lack of publicly available microdata on establishment surveys. 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, as to evaluate performance across a variety of conditions that may exist in practice. This section describes procedures used for generating a set of bivariate data, $\{X_{hi}, Y_{hi}; h = 1, ..., H; i = 1, ..., N_h\}$, where the $\{X_{hi}\}$ are an auxiliary variable reflecting size (e.g., previous year's revenue for a given firm), the $\{Y_{hi}\}$ reflect a quantity of interest in the current time period (e.g., current year's revenue), and the strat-

ification $\{h : h = 1, ..., H\}$ are defined upfront. This entailed three steps: (1) we drew the measures of size (MOS; i.e., the $\{X_i\}$) from a skewed, continuous distribution; (2) we applied stratification based on the size measures (resulting in $\{X_{hi}\}$); (3) we drew the outcome measure from the model $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, where $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$, for some fixed set of model parameters, $\{\alpha_h, v_h, b\}$.

Our simulation, described next, entailed creating $P = 90$ bivariate populations, which reflected three shapes for the auxiliary variable, five coefficients of heteroscedasticity, and six structures for the slope and correlation for the two variables. We considered a single mechanism for stratification. In total, we generated three sets of X's (with varying distributions), each of which was used to generate thirty sets of Y's (i.e., with varying coefficient of heteroscedasticity, slope, and correlation), resulting in ninety bivariate populations.

2.4.2.1 Generation of size measures

The gamma distribution is relevant to establishment data, which are continuous and heavily skewed (e.g., Andridge & Thompson, 2015). Therefore, the three populations of sizes were generated as truncated gamma distributions, and which varied with respect to their shape, k, and scale, θ, before truncation. The first step entailed drawing 12,500 observations from a gamma distribution with parameters of $(k = 0.2, \theta = 5)$, $(k = 0.6, \theta = 5/3)$, or $(k = 1, \theta = 1)$, depending on the population. Then, following Zangeneh and Little (2015), we dropped the first and last deciles, resulting in truncated gamma distributions of $N = 10,000$ units each; we did this to reduce the numbers of near-zero and extremely large units, the

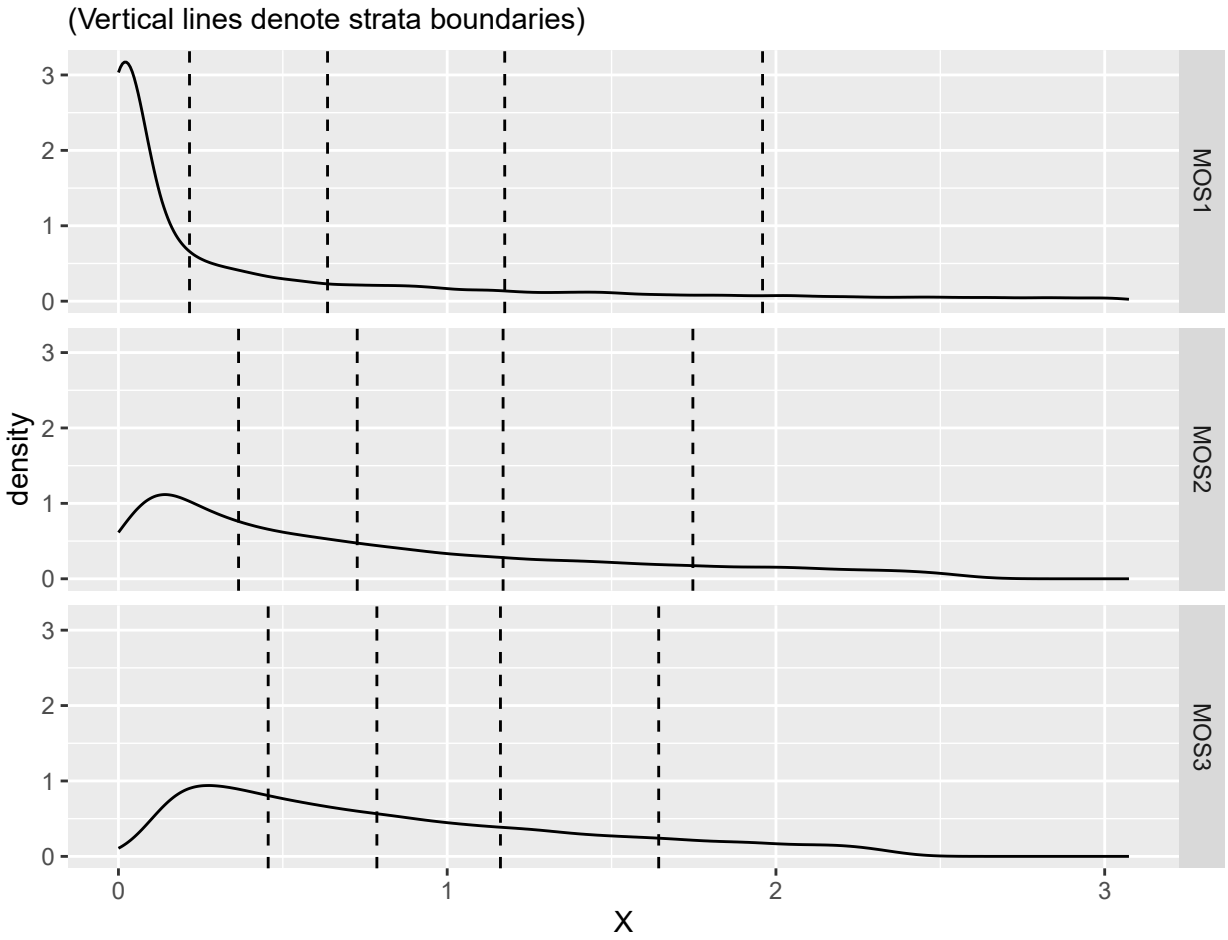
latter of which would presumably be in a take-all stratum under any design and wherein practical considerations would commonly hinge upon reducing non-sampling errors. The shape parameters were varied across populations to reflect that different industries have different size distributions, including different levels of skewness, and the particular choices (i.e., 0.2, 0.6, and 1) follow partly from those used in other related simulations; for instance, Andridge and Thompson (2015) assumed a shape parameter of 1, based on a method-of-moments analysis of three industries, whereas Zangeneh and Little (2015) used a shape parameter of 0.5.

2.4.2.2 Stratification

Next, each set of size measures was grouped into $H = 5$ strata, based on an equal aggregate square root of size rule for X's (see Appendix A.1.2 for a detailed review of equal aggregate stratification rules). 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. In a simulation by Dorfman and Valliant (2000), stratification by aggregate square root of size performed reasonably well across the three populations tested, in conjunction with ratio estimation and (unrestricted) stratified random sampling. The resulting univariate distributions we generated are displayed below in Figure 2.2.

Relating our size measures and stratification to what is done in practice, establishment sur-

Figure 2.2: Distributions of simulated stratified size measures, by MOS



veys commonly stratify by industry and/or unit size; for instance, several of the Census Bureau's economic surveys stratify by size within industry (National Academies of Sciences, Engineering, and Medicine, 2018). Each set of stratified size measures is very roughly analogous to a single industry, stratified by size.

2.4.2.3 Model parameters

Next, we generated the outcome variable for a given pseudo-population by drawing from our model given some fixed set of model parameters. For given $\{X_{hi}\}$ and $\{\alpha_h, v_h, b\}$, we drew the $\{Y_{hi}\}$ from the model $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, where $\varepsilon_{hi} \overset{ind}{\sim} N(0, v_h X_{hi}^b)$ for $h = 1, ..., 5$ and $i = 1, ..., N_h$. For each set of stratified size measures, this process was applied 30 times, using different sets of model parameters.

Given that the coefficient of heteroscedasticity is commonly within the interval $[0, 2]$ in real applications, we considered values of $b = 0$, $b = 0.5$, $b = 1$, $b = 1.5$, and $b = 2$.

We considered six structures for the remaining parameters, $\{\alpha_h, v_h\}$, although instead of choosing the $\{v_h\}$ directly, we selected the $\{v_h\}$ terms in a manner aimed at producing approximately the desired correlation between the size measures and outcome variable for each stratum. More specifically, 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. Given $\{X_{hi}\}$, b, α_h, and ρ_h, the target correlations were approximately achieved by specifying v_h via

$$v_h = \frac{\alpha_h^2 S_{xh}^2 \left(\frac{1}{\rho_h^2} - 1\right)}{\bar{Z}_h}, \tag{2.14}$$

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 A.4.1 for detail).

We purposively chose six sets of slopes and target correlations that allowed for examination of different relationships between the size measure and outcome variable, while recognizing that these parameter sets are far from exhaustive. 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, although the specific parameters observed in real settings depend on the particular application, 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). Specifically, we considered the following scenarios:

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 $\rho_5 = 0.9$ (i.e., strata with smaller units are assumed to have smaller correlations between the X's and Y's) 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 $\alpha_5 = 0.6$.

For illustration, Figure A.1 displays six bivariate populations that used MOS1 and where $b = 1$, for a randomly selected ten-percent sample of each population (to reduce overplotting); each panel reflects a different slope/correlation scenario.

2.4.3 Pilot design

For the pth population, $U_{(p)}$, and the rth simulation, for $p = 1, ..., 90$ and $r = 1, ..., 1000$, we needed to draw a pilot data set, $d_1^{(p,r)} = \left(s_1^{(p,r)}, y_{s_1^{(p,r)}}\right)$, for use in designing the main study, where $s_1^{(p,r)}$ are the sample indicators and $y_{s_1^{(p,r)}}$ are the associated survey data. Each simulation entailed an equally allocated sample of 75 units, which aimed to be large enough

to be realistic for some applications, while not being so large as to undercut our focus on uncertain design parameters. This resulted in pilot sample sizes of $\{m_h = 15 : h = 1, ..., 5\}$, with use of simple random sampling within each stratum. Although, in practice, the existence, size, and sample allocation for a pilot sample will depend on the specific study, equal allocation can be justified in some practical situations, for instance, in situations where a sample designer needs pilot data for estimating stratum-specific design parameters (e.g., the $\{\alpha_h, v_h\}$) but has no reason, a priori, to prioritize certain strata.

2.4.4 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), then drew a single design-based sample for each allocation, and computed the estimated population total and associated 95% equal-tail CI (confidence interval or credible interval) for each sample. We fixed the total sample size at $n = 500$ units. For population p, simulation r, and allocation a, this resulted in the population estimate $\hat{Y}_{(p,a)}^{(r)}$, for $p = 1, ..., P$, $r = 1, ..., R$, and $a = 1, ..., A$. In cases where there were multiple competing methods of producing CIs (e.g., multiple viable variance estimators), we either selected a single method upfront or computed multiple CIs and reported the one that performed better and/or had other practical advantages (e.g., simpler to implement).

We compared our strategy to several key alternative stratified random sampling (STSRS)

designs. Although probability-proportional-to-size (PPS) designs have been suggested to improve the efficiency of designs, using an adequately large number of strata for an appropriately chosen STSRS design will likely limit the potential efficiency loss; further, establishment surveys in practice often use STSRS designs (e.g., National Academies of Sciences, Engineering, and Medicine, 2018, Table 5-2). We assumed constant per-unit costs.

We considered STSRS strategies from various inferential approaches:

- **Design-based**: We considered stratified random sampling (STSRS) via Neyman allocation, in conjunction with the HT-estimator (i.e., N-HT strategy).

- **Design-based/model-assisted**: We considered four approaches outlined or implied by Cochran for the separate ratio (SR) estimator (see §2.2.1.3 and §2.2.1.4). Cochran provides a theoretically optimal allocation of $n_h \propto N_h S_{dh}$, where S_{dh} is the standard deviation of a residual term (see Equation [2.3]). Instead of providing a direct estimator for S_{dh}, Cochran provides rule-of-thumb (RT) solutions that may accommodate some situations in practice. Therefore, our first strategy considered, denoted by C-SR, entailed a plug-in estimate for S_{dh} via Equation (2.4), in conjunction with the SR estimator. The other three strategies, denoted by RT0-SR, RT1-SR, and RT2-SR, entailed the RT allocations of $n_h \propto N_h$, $n_h \propto N_h \sqrt{\bar{X}_h}$, and $n_h \propto N_h \bar{X}_h$, respectively, in conjunction with the SR estimator. These strategies may sometimes be appropriate when we have that $b = 0$, $b = 1$, or $b = 2$, respectively, and when the v_h terms are constant across strata.

- **Bayesian**: We considered our proposed methods under the assumption that the

population-generating parameter, $b = b^{(p)}$, is known, and obtaining inferences via the prediction estimator from Equation (2.11); we refer to this strategy as B-P (i.e., Bayesian allocation/prediction estimator).

The above strategies are summarized in Table 2.1, with additional detail following on the subsequent pages.

Table 2.1: Overview of main estimation strategies for simulation

	Strategy	**Inferential approach**	**Sample allocation**	**Estimator**	**Comments**
N-HT	Neyman/HT	Design-based	Neyman allocation with plug-in estimate of S_h	HT estimator	
C-SR	Cochran plug-in/ separate ratio estimator	Design-based/ model-assisted	Cochran (1977), §6.14, with plug-in estimate $\hat{S}_{dh}$	Separate ratio estimator	$\hat{S}_{dh}$ not directly provided by Cochran
RT0-SR	Prop. allocation/ separate ratio estimator	Design-based/ model-assisted	$n_h \propto N_h$	Separate ratio estimator	Implied by Cochran if S_{dh} is roughly constant
RT1-SR	$b = 1$ rule of thumb/ separate ratio estimator	Design-based/ model-assisted	$n_h \propto N_h\sqrt{\bar{X}_h}$	Separate ratio estimator	Suggested by Cochran if S_{dh} is roughly proportional to $\sqrt{\bar{X}_h}$
RT2-SR	$b = 2$ rule of thumb/ separate ratio estimator	Design-based/ model-assisted	$n_h \propto N_h\bar{X}_h$	Separate ratio estimator	Suggested by Cochran if S_{dh} is roughly proportional to $\bar{X}_h^2$
B-P	Bayesian allocation/ prediction estimator	Bayesian	Minimize Eqn. (2.13) assuming that $b = b^{(p)}$	Compute $\mathrm{E}(Y\|D2, e_1, b)$ via Eqn. (2.11) and $b = b^{(p)}$	

For population p, allocation a, and simulation r, we calculated a CI by computing the variance (estimated variance or posterior variance, as applicable) and assuming normality of the sampling distribution, resulting in bounds of approximately $\hat{Y}^{(r)}_{(p,a)} \pm 1.96\sqrt{\text{var}\left(\hat{Y}^{(r)}_{(p,a)}\right)}$. We used classical variance estimators for the HT and SR estimators. HT variances were estimated via $\text{var}(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$. For SR estimates, we estimated variances in two ways, with results reported for the variance estimator $v_2(\hat{Y}_{Rs}) = \sum_{h=1}^{H}\left(\frac{\bar{X}_h}{\bar{x}_h}\right)^2 \frac{N_h^2(1-f_h)}{n_h}\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \hat{R}_h x_{hi})^2$, where $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$ is the ratio of sample means in stratum h. This variance estimator is obtained by computing the sample residual $e_{hi} = y_{hi} - \hat{R}_h x_{hi}$, applying the usual STSRS variance estimator to the residuals, and then adjusting the strata variances by a factor of $\left(\frac{\bar{X}_h}{\bar{x}_h}\right)^2$. SR variances were also estimated via $v_0(\hat{Y}_{Rs})$, which omits the adjustment of $\left(\frac{\bar{X}_h}{\bar{x}_h}\right)^2$, but the resulting CI properties were nearly identical or slightly worse than those of v_2, and so we only report the latter. For B-P, the posterior variance was computed via Equation (2.12).

For B-P, sample sizes were based on optimizing Equation (2.13), which incorporates a TS approximation for $\text{E}(k_{s_{2h}})$, subject to constraints of $4 \leq n_h \leq N_h$. We also evaluated an alternative strategy, which involved using Monte Carlo (MC) estimates of $\text{E}(k_{s_{2h}})$, using the methods described in Appendix A.3.4, and based on 10,000 random permutations of the population. For each strategy, optimization was conducted using a COBYLA-based algorithm (Powell, 1994), as implemented in NLopt (Johnson, 2014). 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 $\{n_h\}$. As applied to 24-stratum population of approximately 141,000 units (see §2.6.1), each ran

reasonably quickly (i.e., less than a second per optimization run) on a personal computer (Dell OptiPlex 7090), with the MC method requiring some initial set-up time (i.e., 23 minutes on one core for 10,000 random permutations for the given pilot).

Although the B-P strategy assumes that the sample designer knows 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, given the potential inefficiencies that this could introduce. 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 (omitting the true value, $b^{(p)}$), but where the true value was used for estimation.

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 $\left(1+\epsilon_{(p,a)}^{(r)}\right)$, before rounding, for some $\epsilon_{(a)}^{(p,r)}$ close to zero that was chosen to ensure the intended total sample size.

2.4.5 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 is 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, & Haneuse, 2009). For a given population and strategy, three of the simulation-based estimators above (i.e., relative bias, CI coverage, and CI relative width), as well as MSE, can be written as $\hat{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$, where $\hat{\theta}_r$ is a single-draw estimate associated with simulation r, and $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_R$ are i.i.d. random variables from a common distribution with pdf $f(\theta_1)$. For each such estimator, Monte Carlo error was estimated via $sd(\hat{\theta}) = \sqrt{\frac{1}{R}\text{var}(\hat{\theta}_1)}$, where $\text{var}(\hat{\theta}_1) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2$ was the estimated variance for a single draw from $f(\theta_1)$. We then obtained CIs for these metrics via normal approximation.

Given that the allocation methods considered depend on the pilot samples (excepting the rule-of-thumb methods), we also considered the distribution of sample allocations with respect

to this source of variability. For population p and strategy a, we computed the average allocation for stratum h as

$$\bar{n}_{h(p,a)} = \frac{1}{R}\sum_{r=1}^{R} n_{h(p,a)}^{(r)}, \tag{2.15}$$

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

$$sd_{h(p,a)} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(n_{h(p,a)}^{(r)} - \bar{n}_{h(p,a)}\right)^2}. \tag{2.16}$$

2.5 Simulation results

2.5.1 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 (Table A.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 (Table A.3), with any deviations appearing to be attributable to Monte Carlo error. Given the strategies' approximate unbiasedness and near-nominal CI coverage, we can 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 2.2 and 2.3 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. Each cell reflects results from 1,000 simulations of each strategy for the given population structure.

Table 2.2: 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 2.3: 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

Across all populations, N-HT consistently underperformed B-P, increasing the RMSE by 11%–175% for individual populations (Table 2.2). 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 2.3). 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 2.4, 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 2.4: 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

2.5.2 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 2.5, 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 2.5: 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

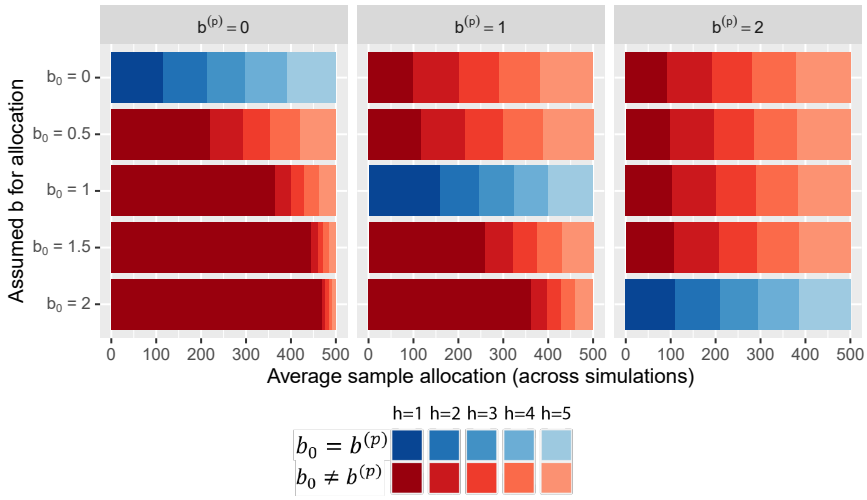
Otherwise, the RT-SR strategies typically had little bias and near-nominal coverage of 95%

CIs (excepting the MOS1, $b = 2$ populations, where coverage was closer to 90%; Table A.4).

2.5.3 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 2.3 shows the average allocations to sampling strata (via Equation 2.15) 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 2.3: 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 2.6, 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 2.6: 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						

Note that the above results on misspecified b apply only when using the B-P strategy, and not to misspecification via RT-SR strategies (e.g., applying RT-SR1 when $b = 2$). Clearly, the latter would have major implications in many practical situations.

2.6 Simulation using NCCS data

2.6.1 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 2.4 displays the contours of 2D kernel density estimates of the data, by sector.

Figure 2.4: 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, Carlin, Stern, & Rubin,

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 2.7. 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 2.1 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 §2.4.5).

Table 2.7: 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 2.8. The C-SR and B-P strategies offered substan-

tially 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 2.8: 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.

2.6.2 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 §2.6.1, 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 §2.6.1, 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 §2.6.1, 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 the level of estimated 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 Figure A.2, 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 a manner analogous to those described in §2.6, except that the size class boundaries were modified to reflect the current section's estimate of b (see Table A.6 for population counts that reflect the current classifications). Given that some of the resulting strata population sizes were fairly small, implementing the classic allocations directly would have led to some allocations of $n_h < 4$ or $n_h > N_h$. In order to avoid such situations, 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 §2.4.4, so that the total sample size was always $n = 1800$.

Simulation results are displayed in Table 2.9. Once again, the C-SR and B-P strategies outperformed N-HT, but by a narrower margin. All three methods produced CIs with slightly less than nominal coverage.

Table 2.9: 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.

2.7 Discussion

2.7.1 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. Under our problem structure, we derived the approximate expected posterior variance, which allowed for the optimal allocation to be identified through mathematical programming. We formulated the optimization problem as minimizing the approximate expected posterior variance subject to constraints of $4 \leq n_h \leq N_h$; additional constraints on sample sizes or other sample-related quantities can be straightforwardly handled via standard optimization software; see Valliant et al. (2013, Chapter 5) for some examples. Although the expected posterior loss was approximate, by virtue of incorporating a first-order Taylor series approximation for a sample-based quantity (and which assumed large $\{n_h\}$), empirical performance for the resulting strategy was nearly identical to an alternative that used Monte Carlo sampling.

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.

2.7.2 Limitations and future directions

We assumed heteroscedastic errors, and that the mean of the outcome follows a linear relationship through the origin with the auxiliary variable. Although our allowing for heteroscedasticity improves realism in comparison to some previous Bayesian optimal design work, our model assumptions will not always be realistic. The proposed allocation is theoretically optimal for the assumed model, loss function, and problem structure, but violations of the assumptions could lead to a biased estimator and/or an inefficient allocation. Therefore, an important avenue of future work entails consideration of alternative population structures, including those that do not fit our model. This could entail considering 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 et al. (2000, pp. 53–54) and set of associations considered in simulation by Zangeneh and Little (2015). It may also be important to consider potential implications of outliers or other influential values, in the context of sample allocation. 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 (2.13), 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.

Some conditions of our simulation had to be fixed, so as to maintain a manageable scope; alternative choices could be employed to extend our applied results. For instance, we did not provide comparisons to PPS designs, as to simplify simulation design. Although adequate stratification can mitigate potential inefficiencies of STSRS designs in comparison to PPS, future work could consider PPS designs. Another alternative we did not consider is Clark's (2013) machine learning frequentist method, which would have required a substantially different simulation set-up.

Although the B-P approach's empirical performance was as good or better than the alternatives under the conditions considered, 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 application to real data on public charities, the B-P strategy performed well, but so did the C-SR strategy, which had roughly similar RMSE. The simulations to artificial data show that the B-P strategy has the potential to improve on C-SR under a correctly specified model, and the application to public charities data showed the B-P strategy performing reasonably in applications to real data, which were not generated following our model. Additional empirical or theoretical work could be conducted to better understand the conditions under which the methods' performance varies and the

reasons why.

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 §2.5.3; 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. Further, Henry and Valliant (2009) provided several examples where existing procedures for estimating b (which we summarize in Appendix A.1.3) did not converge to a stationary point in the interval $[0, 2]$, especially for small samples. Therefore, another area of extension, addressed in Chapter 3, 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. We assume use of a pilot, but there may not always be a literal pilot study available. For instance, for a repeated cross-sectional survey, the prior data may have been collected at a previous time period. 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 [2.13]), which is the mechanism by which the pilot data affect the allocation.

Chapter 3: Computational methods toward fully Bayesian optimal sample design for establishment surveys

Abstract

In sample allocation, imprecise estimates of survey design parameters (e.g., strata variances) can harm efficiency. This can be addressed through a Bayesian decision theoretic formulation. The Bayesian optimal design maximizes the expected utility over the space of possible designs with respect to the future data and model parameters. This can pose computational challenges, by requiring Bayesian analysis on numerous future data sets to evaluate the expected loss for a single design, a process that must be repeated many times during optimization.

We provide a fully Bayesian computational framework for stratified sample optimization. We consider design for the finite population mean under a univariate regression model with heteroscedasticity. We use Monte Carlo sampling to estimate the loss for a given allocation, in conjunction with SPSA, a stochastic optimization method that accommodates noisy loss functions. We compare the proposed methods numerically with design-based, model-assisted, and pseudo-Bayesian alternatives. The proposed methods performed as well or better than

the alternatives, while having clearer theoretical justification. This builds upon our analytical work from Chapter 2.

3.1 Introduction

Chapters 2–3 of this thesis consider stratified sample allocation for establishment surveys, using Bayesian decision theory on optimal experimental design (e.g., Lindley, 1972) to accommodate uncertain design parameters. The Bayesian optimal design maximizes the expected utility over the design space (e.g., space of possible sample designs) with respect to the future data and the model parameters (e.g., strata variances). Unless the prior and likelihood are chosen in a manner that facilitates analytical results, Bayesian design can quickly become computationally challenging due to: (1) the need to average over the space of possible posterior distributions for a single design; and (2) the need to optimize over the space of possible designs (e.g., Ryan et al., 2016).

In Chapter 2, we considered univariate regression, incorporating a coefficient of heteroscedasticity (b) (e.g., Henry & Valliant, 2009), and aiming to minimize the expected posterior variance of the finite population mean. We assumed known b, facilitating analytical results while accommodating uncertainty in other parameters. The resulting objective function was in closed form, which made sample optimization straightforward.

In Chapter 3, we will use a similar problem formulation, but will allow for unknown b. We will first specify a prior for b and derive the resulting posterior distribution for superpopu-

lation parameters. The resulting distribution is analytically intractable, preventing a closed form for the objective function (i.e., expected posterior variance, which is computed by averaging the posterior variance over the predictive distribution for the data). We propose computational methods for sample optimization, which entail: (1) Monte Carlo sampling for estimating the objective function for a given allocation; and (2) an optimization algorithm that requires repeated objective function evaluations in searching for an optimum. Given that our objective function evaluations are expensive and inexact, we propose optimization via simultaneous perturbation stochastic approximation (SPSA; Spall, 1987, 1992), which accommodates noisy functions, does not require an exact analytical expression for the gradient, and scales well to higher dimensionality. Following Sadegh (1997), we accommodate constraints by projecting infeasible points to nearby feasible points. We further modify SPSA to allow for an equality constraint on total sample size. As an alternative to SPSA, we also consider a constrained variant of Nelder–Mead (NM), which is a widely used optimization algorithm (e.g., Conn, Scheinberg, & Vicente, 2009) that outperformed SPSA in a low-dimensional stochastic optimization setting (Huan & Marzouk, 2013).

Although our focus is on fully Bayesian designs (as defined by Ryan et al., 2016), we also consider some pseudo-Bayesian alternatives, which are simpler to implement. One such alternative is a $\hat{b}_{median}$ plug-in allocation, wherein we assume that b is equal to its posterior median given the pilot data and draw upon Chapter 2 results for known b. Such alternatives may assist the computational methods by providing a good starting point for optimization. On the other hand, they have unclear performance for imprecisely estimated b, and so computational methods are needed either to validate or to improve upon them.

After providing pseudo-Bayesian and the proposed fully Bayesian methods, we will compare them empirically to classic approaches through simulation and demonstrate use through application. We will find that the proposed fully-Bayesian methods have empirical performance that matches or exceeds that of the alternatives, while having the clearest theoretical justification. The pseudo-Bayesian methods are also approximately optimal, and are simpler to implement, but this performance is not guaranteed to extend to other contexts. Compared with the best-performing classic strategy (i.e., C-SR strategy, as described in Chapter 2), the proposed methods did slightly better on average, while occasionally achieving substantial improvements in the estimated loss for specific pilot studies. Both SPSA and NM perform well in simulation, but SPSA did better in a higher-dimensional application, raising questions about the scalability of NM to higher dimensions.

Our research contributions are: (1) providing a framework for a computational approach to fully Bayesian optimal sample design in establishment surveys with an unknown coefficient of heteroscedasticity; and (2) providing the first application in survey sampling of a computational approach to finding fully Bayesian optimal experimental design.

The rest of our paper is organized as follows. Section 3.2 reviews computational challenges with fully Bayesian designs and summarizes pertinent results from Chapter 2. Section 3.3 extends our Chapter 2 model to unknown b, which includes a Bayesian analysis of superpopulation parameters and an outline of some pseudo-Bayesian methods. Section 3.4 proposes a computational approach to finding a fully Bayesian optimal design for unknown b. This includes a Monte Carlo estimator for the expected loss for a given allocation, which we use in conjunction with a stochastic optimization algorithm (i.e., SPSA or NM). Section

3.5 outlines simulation methods for comparing performance of the proposed methods with alternatives. Section 3.6 presents simulation findings. Section 3.7 applies the proposed sample allocation methods toward sample design for analyzing tax returns of public charities. Section 3.8 summarizes results, discusses limitations, and suggests future directions.

3.2 Background

Section 3.2 is organized as follows:

- In §3.2.1, we review Bayesian optimal experimental design, primarily in non-survey contexts since computational methods have not yet been applied to surveys. This entails further review and discussion of the computational challenges of fully Bayesian design. We also briefly outline our use of SPSA for stochastic optimization, while deferring a detailed review until §3.4.4.
- In §3.2.2, we summarize key results from Chapter 2 as will be needed in Chapter 3.

3.2.1 Computational challenges of fully Bayesian design

Bayesian decision theory on optimal experimental design (e.g., Lindley, 1972) has been used for a broad array of design problems; see Chaloner and Verdinelli (1995) for a broad review and Ryan et al. (2016) for a recent review of associated computational challenges. As described by Chaloner and Verdinelli, experimental design entails making good choices of

control variables as to improve the eventual statistical inferences. Bayesian decision theory provides a general approach for such choices. An important recent focus of Bayesian optimal experimental design has been developing strategies for reducing computation (e.g., Amzal, Bois, Parent, & Robert, 2006; Bielza, Müller, & Insua, 1999; Huan & Marzouk, 2014, 2013; Long, Scavino, Tempone, & Wang, 2013; Müller & Parmigiani, 1995; Müller, Sansó, & De Iorio, 2004; Overstall & Woods, 2017; Overstall, Woods, & Parker, 2020; Ryan, Drovandi, Thompson, & Pettitt, 2014), although not yet for surveys.

We again assume use of Lindley's framework, as summarized in §2.2.3. To place emphasis on the study design, and noting that Lindley's terminal analysis results in the use of a Bayes estimator of θ (given $p(\theta, x)|e$), we re-write the loss function from §2.2.3 as $l(\theta, e, x)$ where, once again, $\theta \in \Theta$ is the parameter, e is the experiment (e.g., sample allocation), and $x \in X$ is the sample data. Let E refer to the space of possible designs, such that $e \in E$. Letting $L(e)$ denote the expected loss for design e, the optimal experiment e^* can be written as:

$$e^* = \underset{e \in E}{argmin}\, L(e) \tag{3.1}$$

where

$$L(e) = \int_X \int_\Theta l(\theta, e, x)\, p(\theta|x, e) p(x|e)\, d\theta\, dx. \tag{3.2}$$

Hence, our optimal design minimizes the expected loss. Since the true parameter θ is unknown and the future data x have not yet been realized, this entails averaging over the parameter space Θ and the sample space X. Challenges commonly arise due to difficulties

in evaluating Equation (3.2) and the need to optimize over the design space E. For instance, this could require considering hundreds of posterior distributions to estimate $L(e)$ for a single design point, and then repeating this process at numerous design points. These issues are often addressed through strategies to reduce the number of designs that need to be directly evaluated and/or reducing the computational effort associated with evaluating each point.

Given these issues, much of the earlier work, as summarized by Chaloner and Verdinelli (1995), was analytical. In some cases, this involved selection of a prior and likelihood that would allow for analytical results; if this was not possible, alternatives commonly entailed the use of approximations (e.g., normal approximations). Likewise, the previous applications to surveys we described in §2.2.2 focused exclusively on mathematical analysis. More recently, as summarized by Ryan et al. (2016), advances in computing have led to an increasing use of computational methods to solve Bayesian design problems (in non-survey contexts). Ryan et al. (2016) classify computational algorithms for solving such problems based on their method for approximating the posterior distribution and their search framework (i.e., method of searching the design space).

In the absence of a closed form for the expected loss, Equation (3.2) can be evaluated straightforwardly with minimal assumptions through the use of Monte Carlo (MC) integration. This entails approximating the expected loss for a given experimental design via

$$\hat{L}(e) = \frac{1}{R_1} \sum_{i=1}^{R_1} l(\theta, e, x_i), \tag{3.3}$$

where R_1 refers to the number of draws from the predictive distribution for the future data.

We obtain the ith draw by sampling from the prior distribution for the parameters (which in our application is the posterior distribution given the pilot data), selecting a data set x_i from the model, and then evaluating the loss function $l(\theta, e, x_i)$ for that data set, where the loss function measures the performance of the Bayes estimate of θ as follows from Lindley. Since $\theta|x_i$ is not known exactly, but follows some posterior distribution, we can estimate $l(\theta, e, x_i)$ through any appropriate posterior analysis method. For example, if computation were not a concern, this could entail use of Markov chain Monte Carlo (MCMC) to obtain R_2 draws from $f(\theta|x_i)$ for a given future data set x_i; in practice, MCMC is commonly prohibitively expensive for Bayesian design. Some alternatives that have been employed, and which are described by Ryan et al. (2016), include importance sampling, deterministic approximations (e.g., Laplace approximations, numerical quadrature), and approximate Bayesian computation.

Beyond the use of methods to speed up evaluation of an expected loss function, another important focus in computationally-based Bayesian experimental design has been the development of strategies for reducing the number of points that need to be evaluated. We refer to Ryan et al. (2016) for a detailed review, but provide a few examples here to illustrate the breadth of approaches, which reflect the highly varied nature of applications. One approach is use of an emulator or surrogate function to obtain an approximate loss function evaluation at lower computational expense (e.g., Huan & Marzouk, 2014, 2013; Müller & Parmigiani, 1995; Overstall & McGree, 2020; Overstall & Woods, 2017). Given that several optimization algorithms may be ill-suited for optimizing noisy and expensive loss functions, Huan and Marzouk (2014, 2013) focus on efficient methods for stochastic optimization; they further reduce direct sampling from expensive physical models through use of polynomial

chaos surrogate functions. For another example, in a low-dimensional context, Müller and Parmigiani (1995) simulated the utility at many design points spread across the design space, fit a smoothed curve through these observed utilities to estimate the expected utility, and then used deterministic methods to optimize this surrogate function; Overstall and Woods (2017) employ a similar notion in a high dimensional context through a series of conditional optimization steps (also see Overstall et al., 2020). Another manner of dealing with issues of dimensionality is to reparameterize the problem with fewer parameters, which essentially results in searching within a restricted design space (Ryan et al., 2014). A different approach to computation entails defining an artificial probability density function such that the mode reflects areas of the highest utility, allowing MCMC to yield an approximately optimal design (e.g., Amzal et al., 2006; Bielza et al., 1999; Müller et al., 2004).

Although there are many potential computational approaches, we focus on developing methods that can be employed with minimal assumptions. To that end, our starting point will be the MC estimator from Equation (3.3), where $l(\theta, e, x_i)$ is approximated via direct sampling. This results in a nested MC estimator, wherein the outer loop iterates over different future data sets, and the inner loop averages over draws from a given posterior distribution. We will subsequently develop strategies to reduce computation for each evaluation, such as the use of a surrogate function in conjunction with known distributional properties for estimating each posterior loss in a manner that avoids substantial computation for nonsampled units. In optimizing the resulting expected loss, we primarily suggest use of simultaneous perturbation stochastic approximation (SPSA; Spall, 1987, 1992), which accommodates noisy functions and does not require an exact analytical expression for the gradient. Each SPSA iteration

entails estimating the gradient at the current point and then taking a step in the estimated direction of steepest gradient descent. In comparison with Kiefer and Wolfowitz's (1952) classic stochastic approximation algorithm, SPSA uses fewer objective function evaluations for each gradient estimate. We tailor SPSA toward our problem by accommodating constraints and in our choice of gain sequences. Additionally, we consider Nelder–Mead (NM) as an alternative to SPSA, following Huan and Marzouk (2013), who found that it performed well with a limited number of noisy objective functions in a low-dimensional Bayesian optimization setting. Further background on SPSA and NM will be provided in §3.4.4–§3.4.6.

3.2.2 Bayesian optimal design for known b

Here, we briefly summarize key results from Chapter 2, which will be used in Chapter 3.

In Chapter 2, we assumed a model of $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, where $\varepsilon_{hi}|(v_h, b, X_{hi}) \overset{ind}{\sim} N(0, v_h X_{hi}^b)$, with known auxiliary variable $X_{hi} > 0$, for $h = 1, ..., H$ and $i = 1, ..., N_h$, with independence across strata, and where the $\{\alpha_h, v_h\}$ are unknown and $b \geq 0$ is known. We considered sample-based inferences regarding the finite population mean, $\bar{Y} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} Y_{hi}$, where $N = \sum_{h=1}^{H} N_h$. We used Bayesian decision theory to identify the optimal allocation for a main study with sample sizes $e_1 = \{n_h\}$ given a pilot with sample sizes $\{m_h\}$, with the pilot data ($D1$) used only for planning the main study.

In stratum h, we assumed that variables $\left(\alpha_h, \frac{1}{v_h}\right)$ had prior density $\pi\left(\alpha_h, \frac{1}{v_h}\right) \propto v_h$, which led to a normal-gamma posterior for $(\alpha_h, \frac{1}{v_h})$; we assumed independent prior parameters across strata. In applying Bayesian decision theory, we assumed a loss function of

$L(\bar{Y}, \hat{\bar{Y}}, e_1, D2) = \left(\hat{\bar{Y}} - \bar{Y}\right)^2$, where $D2$ is the main study data. This expression is minimized when our estimator is $\hat{\bar{Y}} = \mathrm{E}(\bar{Y}|D2)$. As a result, the posterior loss (i.e., expected loss given the data) is $\mathrm{Var}(\bar{Y}|D2)$. Through application of results from Ericson (1969a), and assuming independence across strata and that $n_h > 3$ for all h, we saw that:

$$\mathrm{E}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N} \left\{ n_h \bar{y}_h + \left(N_h \bar{X}_h - n_h \bar{x}_h\right) \left(\frac{\sum_{i \in s_{2h}} \frac{x_{hi} y_{hi}}{z_{hi}}}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right) \right\} \tag{3.4}$$

$$\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left\{ \frac{1}{n_h - 3} \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{x_{hi} y_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \right\} \tag{3.5}$$

where $z_{hi} := x_{hi}^b$, $\bar{x}_h := n_h^{-1} \sum_{i \in s_{2h}} x_{hi}$ and $\bar{y}_h := n_h^{-1} \sum_{i \in s_{2h}} y_{hi}$ are main study sample means, and s_{2h} refers to the stratum h main study sample indicators.

Next, we re-expressed Equation (3.5) as $\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2(n_h - 3)} q_{bh} k(s_{2h})$, where $q_{bh} = \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{y_{hi} x_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right]$ 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 posterior variance expression is conditional on the main study data, which have not yet been obtained. Therefore, the next step is to average over the predictive distribution for the main study data given the pilot data $(D2|D1)$, which entails considering uncertainty with respect to the posterior distribution for the superpopulation parameters given the pilot data $(\{\alpha_h, v_h\}|D1)$, the sample indicators $(\{s_{2h}\})$, and model uncertainty given fixed parameters and sample indicators $(D2|\alpha_h, v_h, D1, s_{2h})$. We observed that q_{bh} and $k(s_{2h})$ are independent, since $k(s_{2h})$ only depends on the sample indicators, whereas q_{bh} does not depend on the sample indicators. Given this independence, we analyzed terms separately via $\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2(n_h - 3)} \underset{D2|D1}{\mathrm{E}}(q_{bh}) \underset{D2|D1}{\mathrm{E}}(k(s_{2h}))$. Through algebraic manipulation, results relating to quadratic forms, and properties of some known distributions, we

showed that this reduces to

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2,e_1,b\right)=\sum_{h=1}^{H}\frac{1}{N^2}\left(\frac{n_h-1}{n_h-3}\right)\mathrm{E}(v_h|D1)\mathrm{E}\left(k(s_{2h})\right). \tag{3.6}$$

A first-order Taylor series (TS) approximation for $\mathrm{E}(k(s_{2h}))$, which assumed large $\{n_h\}$, then yielded

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2,e_1,b\right)\approx\sum_{h=1}^{H}\frac{\mathrm{E}(v_h|D1)}{N^2}\left(\frac{n_h-1}{n_h-3}\right)(N_h-n_h)\left[\bar{Z}_h+\frac{\frac{n_h}{N_h}S_{xh}^2+(N_h-n_h)\bar{X}_h^2}{n_h\bar{Q}_h^2(1+CV_{qh}^2)}\right], \tag{3.7}$$

where $S_{xh}^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)^2$, $Q_{hi}=\frac{X_{hi}}{\sqrt{Z_{hi}}}$, $\bar{Q}_h=\frac{1}{N_h}\sum_{i=1}^{N_h}Q_{hi}$, $CV_{qh}=\frac{S_{qh}}{\bar{Q}_h}$, $S_{qh}^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(Q_{hi}-\bar{Q}_h\right)^2$, and $\mathrm{E}(v_h|D1)=\frac{1}{m_h-3}\left(\sum_{i\in s_{1h}}\frac{y_{hi}^2}{z_{hi}}-\frac{\left(\sum_{i\in s_{1h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{1h}}\frac{x_{hi}^2}{z_{hi}}}\right)$ (where s_{1h} refers to the stratum h pilot study sample indicators).

The optimal allocation, $\underset{e_1}{argmin}\,\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2,e_1,b\right)$, could then be obtained via standard mathematical programming techniques (see, e.g., Valliant et al., 2013, Chapter 5). For instance, minimize Equation (3.7) subject to a constraint on total costs, $\sum_{h=1}^{H}n_hc_h\leq C$, and sample size constraints of the form $4\leq n_h\leq N_h$. As an alternative to a TS approximation of $\mathrm{E}(k(s_{2h}))$, we also showed how to obtain results via Monte Carlo (MC) approximations, although doing so did not discernibly affect the empirical performance (possibly a consequence of using few strata and a reasonably large n). In simulation, the proposed Bayesian strategy provided substantial empirical gains compared with a design-based approach and similar or better performance than the main model-assisted approach considered.

3.3 Model extension to unknown b

The preceding results assumed known b, but in practice, b is often unknown. Therefore, §3.3 discusses challenges associated with extending the previous results to unknown b. In §3.3.1, we reintroduce our problem formulation for Chapter 3. In §3.3.2, we conduct a Bayesian analysis of superpopulation parameters, which entails specifying a prior for b and determining the posterior distribution. In §3.3.3, we briefly outline challenges with unknown b in the context of applying Lindley's paradigm. In §3.3.4, we briefly describe some pseudo-Bayesian approaches to our problem, before turning our attention to a fully Bayesian computational approach in §3.4.

3.3.1 Problem formulation

In Chapter 3, we assume the same problem specification as in Chapter 2 (which was summarized in §3.2.2), the only difference being that we now allow b to be unknown (and will assign it a prior). That is, as before, our goal is to identify the Bayesian optimal design for the main study for estimating the finite population mean, using a squared error loss function. As before, our decision variables are the main study sample sizes, $e_1 = \{n_h\}$, conditional on the pilot data ($D1$). Likewise, the above specification results in the posterior loss again being the posterior variance of the finite population mean given the main sample.

Under this formulation, we aim to identify the optimal design,

$$\underset{e_1}{argmin} \underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1\right), \tag{3.8}$$

and where this expression no longer conditions on b.

3.3.2 Bayesian analysis of superpopulation parameters

To allow for an unknown b, we assume a prior of $\pi\left(b, \left\{\alpha_h, \frac{1}{v_h}\right\}\right) \propto \prod_{h=1}^{H} v_h I_{[b_0,b_1]}(b)$ where $\prod_{n=1}^{H} v_h$ is the same as Ericson's prior (and is a Jeffreys prior on $\{\alpha_h, v_h\}$) and $I_{[b_0,b_1]}(b)$ is an indicator variable that is equal to 1 if $b_0 \leq b \leq b_1$ and 0 otherwise. For convenience, assume that the sample auxiliary variable $\{x_{hi}\}$ and outcome variable $\{y_{hi}\}$ are indexed by the sample. The likelihood is $y_{hi}|b, \boldsymbol{\alpha}, \boldsymbol{v} \sim N\left(\alpha_h x_{hi}, v_h x_{hi}^b\right)$, wherein $\boldsymbol{\alpha} = \{\alpha_h : h = 1, ..., H\}$, $\boldsymbol{v} = \{v_h : h = 1, ..., H\}$, and the conditioning on the $\{x_{hi}\}$ is implicit. The joint likelihood is

$$L(\mathbf{y}|b, \boldsymbol{\alpha}, \boldsymbol{v}) = \left(\prod_{h=1}^{H}\prod_{i=1}^{n_h} x_{hi}\right)^{-b/2} \left(\prod_{h=1}^{H} v_h^{n_h}\right)^{-1/2} exp\left(-\sum_{h=1}^{H}\sum_{i=1}^{n_h} \frac{(y_{hi} - \alpha_h x_{hi})^2}{2v_h x_{hi}^b}\right), \tag{3.9}$$

where $\mathbf{y} = \{y_{hi} : h = 1, ..., H; i = 1, ..., n_h\}$, and the posterior distribution is

$$\pi\left(b, \boldsymbol{\alpha}, \boldsymbol{v}|\boldsymbol{y}\right) \propto I_{[b_0,b_1]}(b) \left(\prod_{h=1}^{H}\prod_{i=1}^{n_h} x_{hi}\right)^{-b/2} \left(\prod_{h=1}^{H} v_h^{1-\frac{n_h}{2}}\right) exp\left(-\sum_{h=1}^{H} \frac{1}{2v_h} \sum_{i=1}^{n_h} \frac{(y_{hi} - \alpha_h x_{hi})^2}{x_{hi}^b}\right). \tag{3.10}$$

Conditional on b, the above is the joint pdf of H independent normal-gamma random variables. Using this fact, we find that the marginal posterior of b is

$$\pi(b|\boldsymbol{y}) \propto \frac{I_{[b_0,b_1]}(b)\left(\prod_{h=1}^{H}\prod_{i=1}^{n_h} x_{hi}\right)^{-b/2}}{\prod_{h=1}^{H}\left\{\left[\sum_{i=1}^{n_h} y_{hi}^2 x_{hi}^{-b} - \frac{\left(\sum_{i=1}^{n_h} y_{hi}x_{hi}^{1-b}\right)^2}{\sum_{i=1}^{n_h} x_{hi}^{2-b}}\right]^{\frac{n_h-1}{2}}\sqrt{\sum_{i=1}^{n_h} x_{hi}^{2-b}}\right\}} \tag{3.11}$$

(see Appendix B.1 for detail).

3.3.3 Sample design challenges with unknown b

Unfortunately, the distribution expressed by Equation (3.11) does not appear to lend itself to analytical results for Equation (3.8). In order to use methods analogous to those employed in Chapter 2, we would need a formula for the expected posterior loss, $\underset{D2|D1}{\mathrm{E}}\mathrm{Var}\left(\bar{Y}|D2, e_1\right)$, but with b unknown, we lack even a formula for the inner term, $\mathrm{Var}(\bar{Y}|D2, e_1)$. To further illustrate challenges, one possible avenue of inquiry is to consider whether we could make use of Equations (3.4) and (3.5) in some way. For instance, consider these formulas in conjunction with the decomposition

$$\mathrm{Var}\left(\bar{Y}|D2, e_1\right) = \underset{b|D2}{\mathrm{E}}\mathrm{Var}\left(\bar{Y}|b, D2, e_1\right) + \underset{b|D2}{\mathrm{VarE}}\left(\bar{Y}|b, D2, e_1\right). \tag{3.12}$$

Even if some method could be identified for evaluating $\underset{D2|D1}{\mathrm{E}}\left[\underset{b|D2}{\mathrm{E}}\mathrm{Var}\left(\bar{Y}|b, D2\right)\right]$, evaluating $\underset{D2|D1}{\mathrm{E}}\left[\underset{b|D2}{\mathrm{VarE}}\left(\bar{Y}|b, D2\right)\right]$ appears to be particularly challenging. Given that this problem appears to be analytically intractable, we see utility for alternative approaches.

3.3.4 Pseudo-Bayesian approaches

Before turning our attention toward developing computational methods for fully Bayesian sample optimization, we briefly outline some pseudo-Bayesian alternatives, which are simple to implement but do not follow directly from Bayesian decision theory and have unclear performance for imprecisely estimated b.

$\hat{\mathbf{b}}_{\mathbf{MAP}}$ plug-in allocation. Let $\hat{b}_{MAP} = \underset{b}{argmax}\, \pi(b|\mathbf{y})$ denote a maximum a posteriori (MAP) estimate of b computed using the pilot data (D1), where $\mathbf{y}$ is a vector reflecting the pilot data (and where we assume the argmax to be a single point). This can be computed by specifying bounds for b (e.g., $b_0 = 0$ and $b_1 = 2.5$) and using numerical optimization to maximize Equation (3.11) for the pilot data. Then, obtain a plug-in allocation by assuming that $b = \hat{b}_{MAP}$ and obtaining the allocation via methods for known b (i.e., plug this value for b into Equation (3.7) and select the sample allocation that minimizes the resulting expression). Related alternatives, which seem advantageous for asymmetric distributions, entail plugging in the posterior median or mean of $b|D1$ in place of the MAP estimate.

Obtain inferences on the optimal sample allocation for fixed b. Let $\mathbf{n}^{opt}$ be an H-dimensional vector reflecting the optimal allocation for some fixed b. Assuming that each value of b can be mapped onto a single optimal allocation (for fixed b), and given that b is unknown, we could treat $\mathbf{n}^{opt}$ as a vector of random variables and consider the posterior distribution $\pi(\mathbf{n}^{opt}|D1)$. To obtain a single allocation for unknown b, we could then compute the posterior mean, $\mathrm{E}(\mathbf{n}^{opt}|D1)$. To illustrate how this method could be implemented, draw

$R = 2500$ values of b from its posterior distribution given the pilot data, as reflected by the pdf in Equation (3.11) (e.g., assuming bounds of $b_0 = 0$ and $b_1 = 2.5$), obtain the optimal allocation for each, and then average over the R allocations.

These pseudo-Bayesian allocations may potentially perform reasonably under some scenarios, but it is unclear how well they will do if b is estimated imprecisely (e.g., using a small or modestly sized pilot study), especially if the plug-in allocation is sensitive to choice of b (as Chapter 2 showed to be the case for some simulated populations). Further, they do not follow from Lindley's framework, since the loss function does not reflect uncertainty in b. Fully Bayesian methods are needed, either to improve upon these pseudo-Bayesian approaches (if possible) or to validate their use (if it turns out that these pseudo-Bayesian approaches are already approximately optimal).

3.4 Computational methods for optimal design given unknown b

Next, we propose a computational approach for obtaining an optimal allocation for unknown b, which aims to make minimal assumptions while being computationally feasible. Our approach proposes: (1) Monte Carlo (MC) sampling for averaging the posterior loss over the predictive distribution given some sample allocation; and (2) efficient stochastic optimization methods that can accommodate the noisy and expensive loss function of (1). For our problem, (1) entails averaging over the predictive distribution $D2|D1, e_1$; in doing so, we take advantage of analytical results for fixed b. For (2), we suggest SPSA, while considering Nelder–Mead as an alternative.

The remainder of this section is as follows. §3.4.1 provides an MC sampling approach to finding the finite population mean, and which incorporates distributional information in conjunction with an emulation function to substantially reduce computation. §3.4.2 uses two-stage sampling theory to obtain an approximate variance estimator for the MC estimator, which has two nested steps. This variance estimator is helpful for managing quality-runtime tradeoffs when implementing the MC estimator. §3.4.3 parameterizes our decision problem using $H-1$ decision variables, with an additional stratum defined implicitly (e.g., $n_H = n - \sum_{h-1}^{H-1} n_h$, assuming fixed n). §3.4.4 reviews SPSA and describes our modifications to allow for constrained optimization. §3.4.5 extends SPSA as to better handle bounds for the implicitly defined stratum. §3.4.6 briefly reviews Nelder–Mead as an alternative and summarizes our implementation. §3.4.7 outlines our handling of the mismatch between the Monte Carlo estimator, which has a discrete domain, and the continuous optimization algorithms used.

3.4.1 Nested Monte Carlo (MC) estimator for the expected posterior loss

In this section, we provide an approach using MC sampling to average the posterior loss over the predictive distribution for a given sample allocation, while using minimal assumptions. The approach avoids the approximations used in §2.3.3.3 (i.e., Taylor series approximation for $\mathrm{E}(k(s_{2h}))$ and treatment of b as fixed if it is unknown). The basic approach is described in §3.4.1.1. The approach requires sampling from various posterior densities for b, which we describe how to do efficiently in §3.4.1.2. In §3.4.1.3, we describe use of an emulation

function to speed up certain intermediate calculations.

3.4.1.1 Summary of basic procedure

Our interest is in simulating

$$\underset{D2|D1}{\mathrm{E}} \mathrm{Var}(\bar{Y}|D2) = \int \mathrm{Var}\left(\bar{Y}|D2\right) f(D2|D1)\, \mathrm{d}D2 \tag{3.13}$$

for a given design vector $e_1 = \{n_h\}$, and where the conditioning on e_1 is implicit. Let $\phi = (b, \boldsymbol{\alpha}, \boldsymbol{v})$ refer to the list of superpopulation parameters, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_H)$ and $\boldsymbol{v} = (v_1, ..., v_H)$. Equation (3.13) can be simulated via two nested steps, wherein the outer step entails drawing $R1$ main study data sets from the distribution $f(D2|D1)$ and the inner step entails estimating $\mathrm{Var}\left(\bar{Y}|D2\right)$ via R2 draws from $f(b|D2)$ and associated applications of (3.4) and (3.5), the latter of which are applied toward estimating (3.12). More specifically, the nested steps are:

1. **Outer step:** Draw R1 main study data sets via $f(D2|D1) = \int f_1\left(D2|\phi\right) f_2\left(\phi|D1\right)\, \mathrm{d}\phi$, where the (r_1)st data set, $d_2^{(r_1)} = \left(s_2^{(r_1)}, y_{s_2^{(r_1)}}\right)$, comprises the population indices of sampled units, $s_2^{(r_1)}$, and the associated survey data, $y_{s_2^{(r_1)}} = \left\{y_{hi}^{(r_1)} : h = 1, ..., H; i \in s_2^{(r_1)}\right\}$, for $r_1 = 1, ..., R1$. To do so, we first use the pilot data, $d_1 = (s_1, y_{s_1})$, to draw a set of superpopulation parameters, $\phi^{(r_1)} = \left\{b^{(r_1)}, \boldsymbol{\alpha}^{(r_1)}, \boldsymbol{v}^{(r_1)}\right\}$, via

   $f_2\left(\phi|D1\right) = f_3(b|D1) f_4\left(\{\alpha_h, v_h\}\,|D1, b\right)$, where f_3 is the univariate distribution from Equation (3.11) and f_4 is the joint pdf of H independent normal-gamma distributions

implied by Equation (3.10). Considering that f_3 lacks an exact inverse cumulative distribution function, we obtain approximate draws from f_3 via an adaptive grid-based sampling method, which we describe in §3.4.1.2. After obtaining $\phi^{(r_1)}$, we sample from f_1 by drawing a design-based stratified random sample (i.e., set of sample indicators and associated auxiliary data) and then using the sample indicators, auxiliary data, and model to draw $\{y_{hi}^{(r_1)} : h = 1, ..., H; i \in s_2^{(r_1)}\}$ via $y_{hi}^{(r_1)} \sim N\left(\alpha_h^{(r_1)} x_{hi}^{(r_1)}, v_h^{(r_1)} z_{hi}^{(r_1)}\right)$, where $z_{hi}^{(r_1)} = \left(x_{hi}^{(r_1)}\right)^{b^{(r_1)}}$. See Appendix B.2.1 for detail.

2. **Inner step:** Observe that Equation (3.12) decomposes $\text{Var}(\bar{Y}|D2)$ into components corresponding with Equations (3.4) and (3.5), each of which can be simulated by drawing from $f(b|D2)$. Therefore, for given r_1 (and associated data, $d_2^{(r_1)}$), we can approximately draw $\{b^{(r_1,r_2)} : r_2 = 1, ..., R2\}$ from $f(b|D2)$ via the adaptive grid-based procedure. Let $E^{(r_1,r_2)}$ and $V^{(r_1,r_2)}$ denote the quantities computed via Equations (3.4) and (3.5), and where the latter will subsequently incorporate a slight approximation (in §3.4.1.3) to speed up computation. We can estimate $\text{Var}\left(\bar{Y}|D2 = d_2^{(r_1)}\right)$ via

$$V^{(r_1)} = EV^{(r_1)} + VE^{(r_1)}, \tag{3.14}$$

where $EV^{(r_1)} = \frac{1}{R2}\sum_{r_2=1}^{R2} V^{(r_1,r_2)}$, $VE^{(r_1)} = \frac{1}{R2-1}\sum_{r_2=1}^{R2}\left(E^{(r_1,r_2)} - \bar{E}^{(r_1)}\right)^2$, and $\bar{E}^{(r_1)} = \frac{1}{R2}\sum_{r_2=1}^{R2} E^{(r_1,r_2)}$.

After doing so, the MC estimated loss is then

$$\hat{\underset{D2|D1}{\text{E}}}\,\text{Var}(\bar{Y}|D2) = \frac{1}{R1}\sum_{r_1=1}^{R1} V^{(r_1)}, \tag{3.15}$$

the conditioning on the $\{e_1\}$ again being implicit.

Note that an alternative to the inner step above would have been to simulate Var $\left(\bar{Y}|D2\right)$ by obtaining $R2$ predictions of $\bar{Y}$ from $f\left(\bar{Y}|D2\right) = \int f_1\left(\bar{Y}|\phi\right) f_2\left(\phi|D2\right) \mathrm{d}\phi$ (see Appendix B.3 for detail). The proposed approach is more intuitively appealing than this alternative, considering that Equation (3.15) has fewer sources of MC error. Further, the proposed approach enables scalability to large populations via the modification described in §3.4.1.3.

3.4.1.2 Use of adaptive grid sampling for marginal posterior of b

The procedure of §3.4.1.1 necessitates sampling from $R1+1$ complicated univariate posterior densities for b, each of which has the form of Equation (3.11). To sample from a given $f(b)$ efficiently, in the absence of an exact inverse cumulative distribution function, we employ an adaptive grid-based sampling method based on Ritter and Tanner (1991; also see Ritter & Tanner, 1992), but omitting the Gibbs sampling aspect, which is unnecessary for our application. This sampling procedure entails forming a sequence of approximate inverse-cdf functions, $\{\hat{F}_{b,k}^{-1}(q) : k = 0, 1, ..., K\}$, where k denotes the iteration. Each iteration entails evaluating $f(b)$ for a set of points in the interval $[b_0, b_1]$ and using these evaluations, in conjunction with previous evaluations, to approximate $F_b^{-1}(q)$ via a piecewise linear function. At initialization, $\hat{F}_{b,0}^{-1}(q)$ is constructed by evaluating equally spaced points on the interval $[b_0, b_1]$; subsequent iterations evaluate points at equally spaced approximate quantiles, based on the previous iteration's approximation, so that the resulting inverse-cdf is more accurate at areas of higher density. See Appendix B.2.2.2 for detail. Ultimately, we can sample

R points from the approximate distribution by selecting $U_1, ..., U_R \overset{iid}{\sim} Unif(0,1)$ and then computing $\hat{F}_b^{-1}(U_1), ..., \hat{F}_b^{-1}(U_R)$, where $\hat{F}_b^{-1} := \hat{F}_{b,K}^{-1}$.

3.4.1.3 Emulation function for $Z_h(b)$

The algorithm in §3.4.1.1 is of computational complexity $O(R1 \cdot R2 \cdot N)$, since it involves $R1 \cdot R2$ applications of Equations (3.4) and (3.5), the latter of which uses all N units. Since this will not scale well to large populations, we suggest a modification that effectively reduces the complexity of Equation (3.15) to $O(R1 \cdot R2 \cdot n)$, by approximating Equation (3.5). The overarching idea entails pre-processing the population data to allow for fast and accurate approximations of population-based quantities, for purposes of the repeated applications of (3.5) that are needed to compute (3.15). Specifically, rewrite Equation (3.5) as

$$\text{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left\{ \frac{1}{n_h - 3} \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{x_{hi} y_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \left[Z_h(b) - \sum_{i \in s_{2h}} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \right\} \quad (3.16)$$

where $Z_h(b) := \sum_{i \in U_h} Z_{hi} = \sum_{i \in U_h} X_{hi}^b$ is a function of b (given the auxiliary variable) and is the only quantity above that is population-based (rather than sample-based). In pre-processing, compute an emulation function $\hat{Z}_h(b)$. To do so, exactly evaluate $Z_h(b)$ at a large number of equally spaced points from $[b_0, b_1]$ (e.g., we will subsequently use $b \in \{0, 0.01, ..., 2.49, 2.50\}$, for this purpose). For arbitrary point $b \in [b_0, b_1]$, let b_l and b_u be the nearest points below and above b, respectively, where exact evaluations have been obtained, and where $b_l \leq b \leq b_u$. Estimate $Z_h(b)$ as a weighted geometric average of $Z_h(b_l)$ and $Z_h(b_u)$

via

$$\hat{Z}_h(b) = \left[Z_h\left(b_l\right)\right]^w \left[Z_h\left(b_u\right)\right]^{1-w}, \tag{3.17}$$

where the weight, $w = \frac{b_u - b}{b_u - b_l}$, reflects b's position relative to the nearby exactly evaluated points. This is evaluated quickly, since the exact evaluations are pre-computed and stored in an array. For given b, we substitute $\hat{Z}_h(b)$ in place of $Z_h(b)$ to reduce computational time by a factor of roughly $1 - \frac{n}{N}$. For the remainder of this paper, we assume that applications of Equation (3.5) are made by substituting $\hat{Z}_h(b)$ in place of $Z_h(b)$, and noting the equivalency between Equations (3.5) and (3.16). These applications of Equation (3.5) are made when computing the quantities $\left\{V^{(r_1,r_2)}\right\}$, which enter into Equation (3.15) by way of the $EV^{(r_1)}$ term in Equation (3.14).

3.4.2 Choice of R1 and R2

Identifying an optimal allocation requires searching the design space, which will entail numerous evaluations of the MC estimator from Equation (3.15). In doing so, $R1$ and $R2$ must be chosen to provide adequately accurate objective function evaluations within the available runtime. We manage such accuracy-runtime tradeoffs by treating the choice of $R1$ and $R2$ as an intermediate optimization problem, noting parallels to two-stage sampling. In doing so, we also provide a variance estimator for the MC estimator, which will be useful in comparing allocations. Details are provided in Appendix B.2.3 and are briefly summarized here.

To simplify notation, let $f_{ij} = E^{(r_1,r_2)}$ and $g_{ij} = V^{(r_1,r_2)}$ (with the righthand terms defined as in §3.4.1), and temporarily use z_{ij} to denote a PSU-like quantity. Then, we can rewrite Equation (3.15) as:

$$\bar{\bar{z}} := \underset{D2|D1}{\hat{\mathrm{E}}} \mathrm{Var}(\bar{Y}|D2) = \frac{1}{R1 \cdot R2} \sum_{i=1}^{R1} \sum_{j=1}^{R2} z_{ij} \tag{3.18}$$

where $z_{ij} := g_{ij} + \frac{R2}{R2-1} \left(f_{ij} - \bar{f}_i\right)^2$ and $\bar{f}_i := \frac{1}{R2} \sum_{j=1}^{R2} f_{ij}$. Assuming large $R2$, we can estimate the variance of our MC estimator for some arbitrary $(R1, R2)$ via

$$\mathrm{var}\left(\bar{\bar{z}}|R1, R2\right) \doteq \frac{\hat{S}_1^2}{R1} + \frac{\hat{S}_2^2}{R1 \cdot R2}, \tag{3.19}$$

where $\hat{S}_1^2 = s_1^2 - \frac{s_2^2}{R2'}$, $s_1^2 = \frac{1}{R1'-1} \sum_{i=1}^{R1'} \left(\bar{z}_i - \bar{\bar{z}}\right)^2$, and $\hat{S}_2^2 = s_2^2 = \frac{1}{R1'(R2'-1)} \sum_{i=1}^{R1'} \sum_{j=1}^{R2'} \left(z_{ij} - \bar{z}_i\right)^2$ are based on some high-quality MC estimates (e.g., $R1' = R2' = 10000$), and where we have implicitly conditioned on some sample allocation and grid sampling criteria. This result arises by noting that our MC sampling process approximately resembles an epsem two-stage sample from a certain type of infinitely large superpopulation, allowing application of corresponding results for a finite population (e.g., Cochran, 1977, §10.3–§10.4), but ignoring the fpc terms. Then, by assuming that the procedures in §3.4.1 have a runtime of roughly $t(R1, R2) \doteq \beta_1 \cdot R1 + \beta_2 \cdot R1 \cdot R2$, we can select some $R1, R2$ to minimize $\hat{t}(R1, R2)$ subject to constraints on precision (e.g., $cv(\bar{\bar{z}}) = 0.005$), and assuming adequate $R2$.

For example, in an early test application, these methods and precision constraints suggested use of $R1 = 3895$ and $R2 = 2$. These quantities were then modified to $R1 = 3000$ and $R2 = 200$, as to meet the assumption of large $R2$, at modest cost to runtime and precision.

In comparison with a naive selection of $R1 = R2 = 1000$, the resulting choice was anticipated to yield a two-thirds reduction in variance for similar estimated runtime, while allowing for an estimated 588 objective function evaluations per day with estimated CV of 0.00565.

Note that Equation (3.19) can be used in two different ways, whether for planning purposes (as described above), or for post-hoc analysis of an MC estimated loss. For planning purposes, in §3.5.4 and §3.7.1, we will estimate S_1^2 and S_2^2 a single time (with precision controlled by choice of $R1'$ and $R2'$), enabling us to roughly but quickly anticipate the CVs for various choices of $(R1, R2)$. We will also use Equation (3.19) in a post-hoc manner to estimate the variances associated with individual applications of Equation (3.15), by using the underlying data to also estimate $\hat{S}_1^2$ and $\hat{S}_2^2$, so that the resulting variance estimates reflect the actual allocations used.

3.4.3 Parameterization of decision problem for stochastic optimization

Although optimization would ideally use an exact objective function, we only have the MC estimated loss, $\hat{\underset{D2|D1}{\mathrm{E}}} \mathrm{Var}\left(\bar{Y}|D2, e_1\right)$, as per Equation (3.15). Standard optimization methods, which typically assume an exact loss function, might not perform well. In a noisy optimization context, two potentially viable algorithms are SPSA and NM, as will be discussed in §3.4.4–§3.4.6.

Our formulation in §3.3.1 entails minimizing the expected posterior variance, subject to lower and upper bounds on the strata sample sizes (e.g., $4 \leq n_h \leq N_h$), and assuming a

constraint on the overall sample (e.g., limit on total sample size or costs). For purposes of applying NM or SPSA, we assume use of an (H-1)–dimensional parameterization, where the strata are reindexed such that the allocation to stratum H is implicitly defined. The decision variables, $(n_1, n_2, ..., n_{H-1})$, indicate allocations to the explicitly defined strata, whereas the allocation to the implicit stratum is a function of the other allocations and the overall sample constraint, for example, $n_H = n - \sum_{h=1}^{H-1} n_h$ (assuming fixed total sample size of n). We note this, given that different methods are required for handling explicit and implicit constraints, in an optimization context.

3.4.4 Simultaneous perturbation stochastic approximation (SPSA)

Stochastic approximation (SA) denotes certain iterative stochastic algorithms for optimizing functions that are measured noisily (e.g., Kushner, 2010). We consider simultaneous perturbation SA (SPSA; Spall, 1987, 1992), which accommodates noisy functions and does not require an exact analytical expression for the gradient. In comparison with the earlier finite difference SA (FDSA; Kiefer & Wolfowitz, 1952), SPSA uses fewer objective function evaluations for each gradient estimate.

Suppose we aim to identify the parameter $\theta \in \Theta$ that minimizes the value of some exact but unknown objective function, $f(\theta)$, which is only evaluated approximately via $\hat{f}(\theta)$, and where $\mathrm{E}\left(\hat{f}(\theta)\right) = f(\theta)$. After specifying an initial guess, θ_0, each iteration k entails computing an updated guess, θ_{k+1}, for $k = 0, 1, ..., K-1$, by computing an approximation to the gradient

and moving in the estimated direction of gradient descent, $\hat{g}_k(\theta)$, via the updating rule

$$\theta_{k+1} = \theta_k - a_k \hat{g}_k(\theta_k), \tag{3.20}$$

for $k = 0, 1, ..., K-1$, where $\{a_k\}$ is a positive sequence of numbers controlling the step size, with $\sum_k a_k = \infty$. Kiefer and Wolfowitz considered a one-dimensional decision space and estimated a finite difference gradient approximation via

$$\hat{g}_k(\theta) = \frac{\hat{f}(\theta_k + c_k) - \hat{f}(\theta_k - c_k)}{2c_k} \tag{3.21}$$

where $\{c_k\}$ is a positive sequence of numbers, such that $c_k \to 0$ as $k \to \infty$, $\sum_k a_k c_k < \infty$, and $\sum_k a_k^2 c_k^{-2} < \infty$. For example, Kiefer and Wolfowitz suggested $c_k = k^{-1/3}$ (in conjunction with $a_k = \frac{1}{k}$). As applied in D dimensions, FDSA requires $2 \cdot D$ function evaluations per iteration, with gradient estimation via $\hat{g}_k(\theta) = \sum_{d=1}^{D} \frac{\hat{f}(\theta_k + c_k e_d) - \hat{f}(\theta_k - c_k e_d)}{2c_k} e_d$, where e_d refers to the dth unit vector (e.g., see Ruppert, 1983, for detail).

In contrast, SPSA (Spall, 1987, 1992) only requires two objective function evaluations per gradient estimate, by virtue of varying all dimensions simultaneously. On iteration k, this entails generating a D-dimensional random perturbation vector, $\Delta_k = (\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kD})^t$, where the components are i.i.d. and follow some probability distribution with mean 0 and satisfying conditions from Spall (1992). The typical choice is to assume that each Δ_{kd} follows a symmetric Bernoulli ± 1 distribution, that is, $\Pr(\Delta_{kd} = 1) = \Pr(\Delta_{kd} = -1) = 0.5$, for $d = 1, 2, ..., D$ and $k = 0, 1, ..., K-1$. After evaluating $\hat{f}(\theta_k \pm c_k \Delta_k)$, the gradient is

estimated via

$$\hat{g}_k(\theta_k) = \frac{\hat{f}\left(\theta_k + c_k\Delta_k\right) - \hat{f}\left(\theta_k - c_k\Delta_k\right)}{2c_k} \begin{bmatrix} \Delta_{k1}^{-1} \\ \vdots \\ \Delta_{kD}^{-1} \end{bmatrix} \tag{3.22}$$

and a step is taken via Equation (3.20), given the SPSA gain sequences $a_k = \frac{a}{(A+k+1)^\alpha}$ and $c_k = \frac{c}{(k+1)^\gamma}$, which assume nonnegative SPSA coefficients a, c, A, α, and γ, and which we assume are chosen such that the gain sequences $\{a_k\}$ and $\{c_k\}$ tend toward 0 as $k \to \infty$. Under assumptions from Spall (1992), Equation (3.22) is an approximately unbiased gradient estimate, with bias of order $O(c_k^2)$; Spall uses this to show SPSA's convergence under certain conditions and analyze asymptotic efficiency. Spall (1998) provides practical recommendations on the gain sequences and other implementation details, many of which we incorporate in our application. Spall also notes that it is sometimes useful to average multiple independent applications of Equation (3.22) for a given θ_k to improve the stability of SPSA in early iterations.

Following Sadegh (1997), we modify SPSA to accommodate constraints by projecting infeasible points to nearby feasible points, in the context of Equations (3.20) and (3.22). We modify Equation (3.20) to

$$\theta_{k+1} = P\left(\theta_k - a_k\hat{g}_k(\theta_k)\right), \tag{3.23}$$

where $P(\theta)$ is a function for projecting (i.e., moving) θ to the nearest feasible point of the same

dimensionality. For example, if vector $\theta = (n_1, n_2, ..., n_D)^t$ has bounds on its components of $l_d \leq n_d \leq u_d$, for $d = 1, 2, ..., D$, and where the $\{n_d : d = 1, 2, ..., D\}$ reflect explicitly defined decision variables, define a projection function by setting the dth component of the projected point to l_d if $n_d < l_d$, to u_d if $n_d > u_d$, and to n_d, otherwise. For Equation (3.22), the objective function evaluations are changed to be at $P_k(\theta) \pm c_k\Delta_k$, where $P_k(\theta)$ is a projection function in which the bounds on the decision variables are moved inward by $c_k|\Delta_{kd}|$, and where any implicitly defined bounds (if applicable) may also need to be adjusted inward (e.g., by $c_k|\sum_{d=1}^{D}\Delta_{kd}|$). In our sample optimization context, we propose using projection functions of the form $P(\{n_h\}) = P_e(\{n_h\} + \varepsilon)$, where $P_e(\{n_h\})$ projects points violating explicit bounds onto the bounds (i.e., for the $H-1$ decision variables), and where ε, computed iteratively, is the scalar closest to 0 such that the constraints for the implicitly defined stratum are also satisfied. For example, if a candidate allocation, $\{n_h\}$, results in too many units allocated to the implicit stratum (i.e., $n_H > u_H$), then projection may entail adding $\varepsilon = (n_H - u_H)/(H-1)$ units to each of the remaining strata, assuming that doing so would not violate upper bounds for any decision variables; if it would, then a larger ε is needed. Ultimately, although estimating gradients via projected points may lead to noisier estimates near bounds, the potential effects are mitigated in later iterations since $\{a_k\}$ and $\{c_k\}$ tend to zero as $k \to \infty$.

3.4.5 Extending SPSA for improved handling of implicitly defined bounds

Although our $(H-1)$–dimensional parameterization ensures full use of the sample, doing so under certain conditions (e.g., large H) could cause unstable or infeasible allocations to the implicitly defined stratum, supposing that the components of the perturbation vector, $\Delta_k = (\Delta_{k\,1}, \Delta_{k\,2}, ..., \Delta_{k\,H-1})$, are i.i.d., as occurs under the typical choice (e.g., i.i.d. Bernoulli ± 1 random variables). The current section addresses this issue through a different choice of distribution for Δ_k, essentially by imposing a restriction on the perturbation around the implicit stratum; since the resulting components of Δ_k are no longer i.i.d., we apply a corresponding modification to SPSA's gradient estimator.

More specifically, under our assumed parameterization from §3.4.3, the perturbation around stratum H will be the negative sum of the perturbation around the other $(H-1)$ strata. As a result, if the components of the perturbation vector are i.i.d., and if H is large, c_k is large, and/or the optimum is close to the implicit stratum boundary, then one or both of $P_k(\theta) \pm c_k\Delta_k$ may be infeasible, which would be apparent in practice since the algorithm would fail to run. Note also that the variance of the perturbation around the implicit stratum would be $H-1$ times that around the other strata, which suggests the potential for erratic movements along the implicit stratum if H is large.

We suggest, as a workaround, imposing a restriction on the joint distribution of the D elements of Δ_k, and then applying a corresponding modification to Equation (3.22) to ensure valid gradient estimates. For an example (which we will subsequently employ in §3.7), on

iteration k, and assuming $D > 2$, draw scalar B_k as symmetric Bernoulli(± 1), then randomly assign $\lfloor \frac{D-1}{2} \rfloor$ components of Δ_k to each have a value of B_k and give each of the other components the complementary value of $-B_k$. For purposes of computing $P_k(\theta)$, the implicit bounds would need to be moved inward by c_k if D is odd or $2 \cdot c_k$ if D is even. Given that the $\{\Delta_{kd}\}$ are no longer independent, Equation (3.22) no longer yields approximately unbiased gradient estimates, and must be revisited.

Let $\hat{g}_{kd}(\theta_k)$ denote the dth term of $\hat{g}_k(\theta_k)$. As per Spall (1992), a Taylor series expansion, after some algebra, leads to $\mathrm{E}\left(\hat{g}_{kd}(\theta_k)|\theta_k\right) = L'_d(\theta_k) + \sum_{i \neq d} L'_i(\theta_k)\mathrm{E}\left(\frac{\Delta_{ki}}{\Delta_{kd}}\right) + O(c_k^2)$, where $L'_i(.)$ denotes the ith component of the noise-free gradient, $g(.)$, at the given point. Ignoring the $O(c_k^2)$ term, the approximate bias of $\hat{g}_{kd}(\theta_k)$ is $b_{kd} \doteq \sum_{i \neq d} L'_i(\theta_k)\mathrm{E}\left(\frac{\Delta_{ki}}{\Delta_{kd}}\right)$. Spall only considered cases where the components of Δ_k were independent, which, in conjunction with a bounded inverse moment condition, led to $E\left(\frac{\Delta_{ki}}{\Delta_{kd}}\right) = 0$.

We assume that the $\{\Delta_{ki}\}$ are marginally distributed as symmetric Bernoulli (± 1) random variables, and which are exchangeable, but, in contrast to Spall's assumption, are not necessarily independent. For practical purposes, we must also define the joint distribution in a way that allows for movement along the implicitly defined dimension. Let ρ denote the correlation between the ith and jth component of Δ_k. We show in Appendix B.4.1 that an approximately unbiased estimate for $L'_d(\theta_k)$ is

$$\hat{\tilde{g}}_{kd}(\theta_k) = \frac{\hat{g}_{kd}(\theta_k) - D\rho\hat{\bar{g}}_k(\theta_k)}{1-\rho}, \tag{3.24}$$

where $\hat{\bar{g}}_k(\theta_k) := \frac{\sum_{d=1}^{D} \hat{g}_{kd}(\theta_k)}{D(1-\rho+D\rho)}$, $\hat{g}_{kd}(\theta_k)$ is an uncorrected estimate computed via the original

formula, and ρ can be readily computed upfront based on the specific joint distribution used. On the latter, consider the two-step procedure described earlier (i.e., draw scalar $B_k \sim$ symmetric Bernoulli(± 1), then randomly assign $\lfloor \frac{D-1}{2} \rfloor$ elements of Δ_k to receive the shared value B_k, with the complementary set receiving the shared value $-B_k$). Through basic probability theory, these assumptions lead to $\rho = \frac{-1}{D}$, if D is odd, and $\rho = \frac{-1}{D} + \frac{3}{D(D-1)}$, if D is even, and where we assume $D > 2$ (see Appendix B.4.2 for detail).

3.4.6 Nelder–Mead

As an alternative to SPSA, we also considered Nelder–Mead (NM; Nelder and Mead, 1965), which can work well in practice with a limited number of objective function evaluations (e.g., Conn et al., 2009; Lagarias, Reeds, Wright, & Wright, 1998). NM is an unconstrained direct search method, insofar as its decision rules rely solely on a series of comparisons of function evaluations. Huan and Marzouk (2013) suggested that this could be an advantage in noisy optimization settings over methods that involve estimating the gradient. Huan and Marzouk also found NM to outperform SPSA in a low-dimensional Bayesian optimization setting.

We based our NM implementation on Conn et al.'s interpretation (2009, pp. 141–145), which resolves some ambiguities. In adapting it to our setting, we: handled constraints as in Box (1965); incorporated Barton and Ivey's suggested modifications to the shrink step (1996) to improve performance for stochastic loss functions; incorporated oriented restarts, as per Kelley (1999a), to remedy stagnation; and handled collapses of simplex geometry by reinitializing the simplex in the neighborhood of the previous best point. Although we

found NM to perform well in simulation, Han and Neumann (2006) suggested that NM's performance may deteriorate in higher dimensions, which appeared to be the case in our application in §3.7. For a more detailed review of NM, see Appendix B.5.

3.4.7 Rounding of non-integer sample sizes

SPSA and NM require the ability to evaluate the loss function at continuous inputs, but the MC-estimated loss from §3.4.1 is only defined for integer sample allocations. We handle this discrepancy by evaluating the estimated loss function for continuous $\{n_h\}$ at nearby integer allocation $\{\lfloor n_h \rfloor + I_h\}$, where the $\{I_h\}$ are $(0, 1)$ random variables that are jointly assigned via a PPS systematic sampling step, and where $\Pr(I_h = 1) = n_h - \lfloor n_h \rfloor$, for $h = 1, 2, ..., H$.

3.5 Simulation design

Simulation methods were used to evaluate performance of the proposed computational methods toward fully Bayesian design. We aimed to see whether the proposed methods could identify an approximately optimal allocation and to compare performance with alternative strategies. For each of two artificial populations generated under our model, this entailed drawing $R = 1000$ pilot studies, then applying A allocation methods to each. For pilot r and allocation a, we then drew $S = 30$ main study samples and obtained the corresponding inferences.

§3.5 is organized as follows. In §3.5.1, we describe the two populations considered, which

had been generated in the Chapter 2 simulation. The main population considered was one in which the fully Bayesian methods had some potential for improving upon the pseudo-Bayesian allocations. We also considered a second population, which had a different level of heteroscedasticity. In §3.5.2, we outline the sampling strategies that were evaluated, with focus on Bayesian and pseudo-Bayesian approaches. We also considered design-based and model-assisted approaches. In §3.5.3, we outline evaluation methods and metrics. For each pilot, we estimated the expected posterior loss for each final allocation. We evaluated performance across pilots relative to the true population value (e.g., RMSE). We also considered reliability of the proposed stochastic optimization methods with respect to repeated applications. §3.5.4 provides additional detail on computing.

3.5.1 Population structure

The Chapter 3 simulation used two of the populations that had been generated in the Chapter 2 simulation, and which followed our model, but which varied with respect to the assumed level of heteroscedasticity (i.e., $b = 0.5$ versus $b = 1$). These two populations had been generated as follows:

- **Auxiliary variable**: The gamma distribution is relevant to establishment data, which are continuous and heavily skewed (e.g., Andridge & Thompson, 2015). Therefore, as a first step, we drew 12,500 units as gamma with shape $k = 0.2$ and scale $\theta = 5$. Following Zangeneh and Little (2015), we dropped the first and last deciles; we did so as to reduce the numbers of near-zero and extremely large units, the latter of which

presumably be in a take-all stratum under any design. In sum, this step resulted in selecting 10,000 units from a truncated gamma distribution, which were used as X's.

- **Stratification**: Units were assigned to $H = 5$ strata via an equal aggregate square root of size rule (see Appendix A.1.2 for a detailed review of equal aggregate stratification rules). In a simulation by Dorfman and Valliant (2000), stratification by aggregate square root of size performed reasonably well across the three populations tested, in conjunction with ratio estimation and (unrestricted) stratified random sampling. We operationalized this step 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. This step resulted in $\{X_{hi}\}$, with strata ordered via ascending size measures (i.e., $\bar{X}_1 \leq \bar{X}_2 \leq ... \leq \bar{X}_5$). Relating our $\{X_{hi}\}$ to practice, we note that establishment surveys commonly stratify by unit size, often within industry; for instance, this is done in the Census Bureau's economic surveys (National Academies of Sciences, Engineering, and Medicine, 2018).

- **Outcome variable**: Next, for each of two sets of $(b, \{\alpha_h, v_h\})$, we drew $\{Y_{hi}\}$ as independent draws from conditional distributions specified by $Y_{hi}|X_{hi} \sim N(\alpha_h X_{hi}, v_h X_{hi}^b)$. In doing so, instead of choosing the $\{v_h\}$ directly, we specified a set of target correlations, $\{\rho_h\}$, then solved for the $v_h|\alpha_h, \rho_h, b$ that would approximately result in the target correlations. Specifically, after specifying $(b, \{\alpha_h, \rho_h\})$, we computed $v_h = \alpha_h^2 S_{xh}^2 \left(\frac{1}{\rho_h^2} - 1\right) / \bar{Z}_h$, 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 $\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} X_{hi}$. We had showed in Appendix A.4.1 that this choice approxi-

mately yields the desired correlations, on expectation, under some simplifying assumptions. For the current simulation, we specified constant slopes of $\alpha_1 = \alpha_2 = ... = \alpha_5 = 1$ and target correlations of $\rho_1 = \rho_2 = ... = \rho_5 = 0.7$, and then considered two populations with different choices of b, namely, $b = 0.5$ and $b = 1$.

These two populations were from a larger set of populations considered in the Chapter 2 simulation, the latter of which considered several additional sets of population-generating parameters. For instance, Chapter 2 also considered two alternative choices for truncated gamma distribution parameters, which had resulted in less highly skewed populations (versus the size measure described above) and three additional choices for b (i.e., 0, 1.5, and 2). The particular population choices for Chapter 3 were based primarily on the results of §2.5.3, which examined the sensitivity of Bayesian designs for known b to misspecification of b. Relative to a pseudo-Bayesian design, we would expect a fully Bayesian design to hold more promise for populations where the Bayesian allocation for known b is sensitive to misspecification of b. In contrast, for populations where the assumed value of b at the design stage makes little difference to the pseudo-Bayesian allocation, it is unclear why fully Bayesian design would add much. In Chapter 2, we had found that allocations for the two other size measures considered were fairly insensitive to misspecification of b; we had also found that lower true values of b led to greater susceptibility to misspecified b. Hence, in Chapter 3, we opted to primarily analyze one of the more highly skewed populations, and where $b = 0.5$; we supplemented this choice by also considering an analogous $b = 1$ population.

3.5.2 Strategies considered

In Chapter 3, we again assumed that an equally allocated pilot study of size $n = 75$ was used to allocate a main study sample of size $n = 500$, according to each allocation method, and subject to $4 \leq n_h \leq N_h$ for all h. We reused the $R = 1000$ pilot samples drawn in Chapter 2, to facilitate comparisons with previously evaluated methods. We considered the main design-based and model-assisted strategies from Chapter 2 (i.e., N-HT, C-SR, RT0-SR, RT1-SR, and RT2-SR strategies, as described in §2.4.4). We also considered the following pseudo-Bayesian and Bayesian allocations:

- **Pseudo-Bayesian (PB) allocations**. We considered the four PB allocations described in §3.3.4. This included the $\hat{b}_{MAP}$ plug-in allocation, where we assumed $b = \hat{b}_{MAP}$, and obtained the allocation using methods for known b, as well as alternative allocations that assumed b equal to its posterior mean or median, given the pilot data. We refer to these as the PB-MAP, PB-mean, and PB-median allocations. We also considered the $\mathrm{E}(\mathbf{n}^{opt}|D1)$ allocation described earlier (i.e., PB-nopt allocation). All four PB approaches required a method of identifying an optimal design for given b; this was achieved using a COBYLA-based algorithm (Powell, 1994), as implemented in NLopt (Johnson, 2014).

- **Bayesian allocations**. We considered SPSA and NM variants of a fully Bayesian, computational approach, following §3.4 (i.e., B-SPSA and B-NM allocations). This entailed computing the MC estimated loss via Equation (3.15) and conducting con-

strained optimization via SPSA or NM, as in §3.4.4 and §3.4.6, respectively (the SPSA modifications of §3.4.5 not being necessary until §3.7). The first stratum (i.e., containing the smallest units) was designated as the implicit stratum (see §3.4.3). Both algorithms were initialized at an equal allocation; although a well-informed initial guess can speed up some optimization algorithms, doing so here would have clouded comparisons to PB approaches. For NM, the remaining points of the simplex were identified via the axis-by-axis method (e.g., Baudin, 2010, pp. 18–19), with step length of 80.

For all PB and Bayesian approaches, we assumed bounds of $0 \leq b \leq 2.5$. In addition to the main approaches above, we considered an idealized Bayesian allocation that made use of the knowledge of the true value of b for population p, $b^{(p)}$ (i.e., 0.5 or 1, as applicable; B-oracle allocation), for comparison purposes. As in Chapter 2, for all strategies, the final sample allocations were rounded, incorporating adjustments when necessary to ensure total sample sizes of 500. For simulation r and allocation a, this entailed multiplying the unrounded sample sizes by $\left(1 + \epsilon_{(a)}^{(r)}\right)$, before rounding, for some $\epsilon_{(a)}^{(r)}$ close to zero that was chosen to result in the intended total sample size.

For PB and Bayesian allocations, we considered three inferential approaches, focusing on the third:

- **Conditional posterior mean and variance, normal-based CIs**: Use the formulas for $\mathrm{E}(Y|D2, e_1, b)$ and $\mathrm{Var}(Y|D2, e_1, b)$, where Y is the finite population total, and condition on the assumption that b is equal to its MAP estimate given the main study. Use normal-based CIs (i.e., bounds of $\hat{Y} \pm 1.96 \cdot sd\sqrt{\hat{Y}}$). This is simple to implement,

but does not accommodate uncertainty in b.

- **Unconditional posterior mean and variance, normal-based CIs**: Compute $\mathrm{E}(Y|D2, e_1)$ and $\mathrm{Var}(Y|D2, e_1)$ by approximately sampling from $f(b|D2)$ (via adaptive grid-based methods), and then estimating the posterior variance as $\mathrm{Var}\left(Y|D2, e_1\right) = \underset{b|D2}{\mathrm{E}} \mathrm{Var}\left(Y|b, D2, e_1\right) + \underset{b|D2}{\mathrm{Var}} \mathrm{E}\left(Y|b, D2, e_1\right)$ through repeated applications of Equations (3.4) and (3.5). Use normal-based CIs.

- **Unconditional approach, posterior sampling**: Sample from the posterior distribution for the finite population total via $f(Y|D2) = \int f_1(Y|\phi) f_2(\phi|D2) d\phi$, where $\phi = \{b, \alpha_h, v_h\}$ are the superpopulation parameters, and then form a 95% equal-tail CI based on the .025 and .975 quantiles. In doing so, draw R' sets of superpopulation parameters via $f_2(\phi|D2) = f_3(b|D2) f_4\left(\{\alpha_h, v_h\}|D2, b\right)$, using adaptive grid-based methods for f_3 and drawing each $\left(\alpha_h, \frac{1}{v_h}\right)|b, D2$ from the relevant normal-gamma distribution (see A.2.2.2). Use each set of superpopulation parameters to obtain a complete set of predictions of nonsampled units, in the course of predicting the finite population total.

Reporting in §3.6 assumes use of the third inferential approach, unless otherwise noted. The third approach has the fewest assumptions, but we considered the others since they might sometimes be used in practice. For the second and third methods, we used $R' = 10,000$ draws for a given data set. For the B-oracle allocation, which used knowledge of the true value of b, inferences followed a posterior sampling approach.

Most strategies considered assumed use of pilot data for estimating sample-based quantities

at the design stage; these *pilot-based strategies* are summarized in Table 3.1. The remaining strategies are summarized in Table 3.2. As in Chapter 2, classic variance estimators were used for HT and SR estimation.

Table 3.1: Summary of pilot-based strategies considered

Strategy	**Allocation**	**Estimation method**
Neyman-HT	$n_h \propto N_h \hat{S}_h$, where $\hat{S}_h^2 = \text{var}\left(y_{hi}\right)$	Horvitz-Thompson
Cochran-SR	$n_h \propto N_h \hat{S}_{dh}$, where $\hat{S}_{dh}^2 = \text{var}\left(y_{hi} - \frac{\hat{Y}_h}{\hat{X}_h} x_{hi}\right)$	Separate ratio estimator
PB-MAP	Optimize Equation (3.7) given $b = \hat{b}_{MAP}$	Unconditional posterior sampling
PB-mean	Optimize Equation (3.7) given $b = \hat{b}_{mean}$	Unconditional posterior sampling
PB-median	Optimize Equation (3.7) given $b = \hat{b}_{median}$	Unconditional posterior sampling
PB-nopt	Compute $\text{E}\left(\mathbf{n}^{opt} \middle| D1\right)$ as defined in §3.3.4	Unconditional posterior sampling
B-NM	Optimize Equation (3.15) via NM	Unconditional posterior sampling
B-SPSA	Optimize Equation (3.15) via SPSA	Unconditional posterior sampling
B-oracle	Optimize Equation (3.7) given $b = b^{(p)}$	Posterior sampling (given $b = b^{(p)}$)

Table 3.2: Summary of pilot-invariant strategies considered

Strategy	**Allocation**	**Estimation method**
RT0-SR	$n_h \propto N_h$	Separate ratio estimator
RT1-SR	$n_h \propto N_h \sqrt{\bar{X}_h}$	Separate ratio estimator
RT2-SR	$n_h \propto N_h \bar{X}_h$	Separate ratio estimator

3.5.3 Performance comparison

This section outlines how we compared the strategies' performance. §3.5.3.1 describes our comparisons of various allocations with respect to the loss function. §3.5.3.2 describes how we considered performance with respect to the true population value, across repeated applications. The methods were analogous to those from the Chapter 2 simulation, except that the Monte Carlo error took into account the clustered simulation design. §3.5.3.3 describes how we examined reliability of the proposed stochastic optimization methods.

3.5.3.1 Evaluation relative to loss function

First, we evaluated how different allocations performed with respect to the expected posterior loss, $\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1\right)$. Minimizing this quantity gives the best possible performance for a particular pilot, assuming that the model assumptions hold and that the associated Bayesian inferential methods are used. For each final allocation, the expected posterior loss was estimated via a high-quality MC estimate, via Equation (3.15), and specifying $R1 = 5000$ and $R2 = 500$, which were anticipated in the simulation planning stage to yield CVs of very roughly 0.6%. These MC estimated losses were computed independently of any sample optimization process. MC error was estimated via the variance estimator from Equation (3.19).

3.5.3.2 Evaluation relative to true population value

Next, we examined performance relative to the true underlying population value. The procedures described above entailed drawing $R = 1,000$ pilot data sets, and obtaining A allocations according to each method. For the rth pilot and ath allocation, we drew $S = 30$ main study data sets, each of which was used to estimate the finite population total and associated confidence interval or credible interval (CI) via one or more estimation methods. We had chosen S substantially greater than 1 since doing so meaningfully increased the number of effective simulations, without having much effect on total runtime. In aggregate, these procedures resulted in a set of estimates, $\{\hat{Y}_{(a,e)}^{(r,s)}\}$, and associated CIs, for pilot $r = 1, 2, ..., 1000$, main study sample $s = 1, 2, ..., 30$, allocation $a = 1, 2, ..., A$, and estimation method $e = 1, ..., a_e$, where a_e denotes the number of estimation methods used for allocation a, and where the true population total, $Y = \sum_{i \in U} Y_i$, was known.

For each strategy, we then estimated the root mean square error (RMSE), relative bias, and coverage and relative width of 95% CIs, and the associated MC error of each, using formulas analogous to those described in Chapter 2 (see §2.4.5), the only difference being that each metric in Chapter 3 was estimated as the average of R i.i.d. cluster means instead of the average of R i.i.d. single-draw estimates. Specifically, the Monte Carlo–estimated relative bias, CI coverage, CI relative width, and MSE for strategy (a, e) each have the form $\hat{\theta}_{(a,e)} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_{(a,e)}^{(r)}$, where $\hat{\theta}_{(a,e)}^{(r)} := \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_{(a,e)}^{(r,s)}$ can be thought of as a cluster mean, and where $\hat{\theta}_{(a,e)}^{(r,s)}$ is a single-draw estimated metric associated with simulation (r, s). For example, to estimate the MSE for the finite population total under strategy (a, e), define

$\hat{\theta}^{(r,s)}_{(a,e)} := \left(\hat{Y}^{(r,s)}_{(a,e)} - Y\right)^2$, and apply the preceding formula. Monte Carlo error is estimated via $sd\left(\hat{\theta}_{(a,e)}\right) = \sqrt{\frac{1}{R}\text{var}(\hat{\theta}^{(1)}_{(a,e)})}$, where $\text{var}\left(\hat{\theta}^{(1)}_{(a,e)}\right) := \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)}_{(a,e)} - \hat{\theta}_{(a,e)}\right)^2$ is the estimated variance for a cluster mean.

3.5.3.3 Reliability of stochastic optimization

In our simulation, for a given population, two sources of variability are: (1) variability from the choice of pilot data set; and (2) variability from the stochastic optimization methods. Regarding the latter, for a fixed pilot data set, repeated runs of the stochastic optimization methods could be expected to result in at least slightly different allocations, on account of randomness in the MC estimated losses and/or due to the algorithms themselves (mainly in terms of SPSA's method for gradient estimation). Therefore, as an initial simulation, we examined (2) by drawing 20 pilot studies from the $b = 0.5$ population, and obtaining the B-NM and B-SPSA allocations five times per pilot study, using different random seeds to initialize the optimization routines. We examined results, in terms of the MC estimated loss (see §3.5.3.1) and the resulting allocations themselves. The results of this analysis (see §3.6.1) informed the decision to compute one allocation per method for each of 1000 pilot studies (e.g., as opposed to 250 pilots with 4 allocations each). Results were considered separately from the main simulation, to simplify analysis.

3.5.4 Computational considerations

The simulations were carried out using the Deepthought2 and Zaratan high performance computing (HPC) clusters at the University of Maryland (UMD). Deepthought2 had 488 compute nodes, nearly each of which had 20 cores, 128 GB, and dual Intel Ivy Bridge E5-2680v2 processors running at 2.80 GHz. Mid-study, Deepthought2 was replaced by the Zaratan cluster, which has 360 main compute nodes, each of which has 128 cores, 512 GB, and dual AMD Zen3 7763 64-core CPUs. The simulation was designed so that individual NM or SPSA runs would take roughly 7.5 hours using a single core, assuming use of the maximum number of function evaluations, with a hard cap of 12 hours per run, and assuming use of Deepthought2 (assumed hereafter, unless stated otherwise). Computing was primarily in R (R Core Team, 2022), with Equations (3.11) and (3.16) implemented in C++ (via Rcpp; Eddelbuettel and François, 2011) using a log-scale. Benchmarking was conducted via the R package microbenchmark (Mersmann, 2021) and the RStudio profiler.

For a given run, NM was stopped when the simplex approximately converged to a point, or when the maximum of 550 function evaluations was reached, whichever came first, with up to five oriented restarts and five restarts for collapsed simplices. Approximate convergence was defined as having a simplex diameter (i.e., maximum distance between any two vertices) of less than 0.1. Each objective function evaluation used $R1 = 1400$ outer loops and $R2 = 200$ inner loops, which were anticipated at the simulation design stage to have a runtime of 49 seconds/each with estimated CVs of roughly 1.2%.

Each SPSA run was for 5,000 iterations, and each iteration's gradient estimate was the average of two independent applications of Equation (3.22) (i.e., using different perturbation vectors), and drawing the Δ_{kd} as i.i.d. and following a symmetric Bernoulli ± 1 distribution. Each objective function evaluation used $R1 = 30$ and $R2 = 200$ and had an anticipated runtime of 1.3 seconds and CV of roughly 8.2%. SPSA coefficients largely followed Spall's rules of thumb (1998), with some modifications. We set $\alpha = 0.602$ and $\gamma = 0.101$, following Spall. We set $A = 100$, to avoid excessive early movements. Although Spall suggested setting c approximately equal to the standard deviation of the measurement noise, doing so in trial runs led to excessively noisy gradient estimates and occasionally erratic early movements; after some trial and error, we set $c = 3$. Following Spall, we selected a in a manner aimed at controlling early movements; we did so programmatically, by computing 200 SP gradient estimates at the initial allocation (using two perturbation vectors each), and then choosing a so that the change to each stratum's allocation would typically not exceed 2 units. As a further failsafe against erratic movements, we also capped the maximum step size (i.e., Euclidian distance between θ_{k+1} and θ_k) at 3; steps exceeding this were rescaled.

The choices of NM and SPSA coefficients were informed by a small-scale simulation that examined how different levels of noise and parameter choices affected the algorithms' performance for noisy versions of an exact loss function that was known and could be quickly computed. This involved examining SPSA's performance across repeated runs under various sets of algorithm coefficients and levels of induced noise (including no noise). The exact test function was defined as the Chapter 2 approximate expected posterior loss given known b (Equation [2.13]), for some fixed pilot and assumed level of b.

To facilitate post-hoc evaluation of the algorithms' progress over time, after conducting each optimization run, we re-estimated the loss function at every NM iteration (at the current best point) and every tenth SPSA iteration using a medium-quality MC estimate via Equation (3.15). For NM, the re-estimated losses used $R1 = 1000$ and $R2 = 200$ (anticipated CV of 1.4%); for SPSA, these choices were $R1 = R2 = 200$ (anticipated CV of 3.4%). For each algorithm, R1 was chosen so that the expected runtime of a set of MC re-estimated losses for one optimization run was about an hour on a single core.

Simulations for the $b = 0.5$ and $b = 1$ populations were conducted under roughly equivalent conditions, a minor exception being that the latter population used the Zaratan cluster (in place of Deepthought2), which was faster (i.e., average reduction in actual NM and SPSA runtimes by 38% and 39%, respectively). Simulations for the $b = 1$ population also incorporated a minor change to how the upper bounds were specified, but which appeared to have no consequence. In particular, bounds for the $b = 1$ population were specified via $4 \leq n_h \leq min\left(N_h, n - 4(H-1)\right)$, as opposed to the earlier $4 \leq n_h \leq N_h$, considering that $n_h > n - 4(H-1)$ implies an infeasible allocation. This change only affects the handling of infeasible points, but there were no such points considered by SPSA, nor did there appear to be any such points considered by NM, in which case there would be no impact.

3.6 Simulation results

Next, we report on the simulation results. Simulation results for the two populations were extremely similar. Therefore, reporting assumes use of the $b = 0.5$ population, unless oth-

erwise stated, supplemented by results for the $b = 1$ population in Appendix B.7. Further, we primarily focus on the strategies that incorporated pilot data in the allocations, namely, the two main classic strategies (N-HT and C-SR), the four PB methods, and the two main Bayesian methods (B-NM and B-SPSA). For each of the PB and Bayesian allocations, we assume inferences follow an unconditional approach, via sampling from the posterior distribution, since results were nearly identical to the other two inferential approaches considered (see Tables B.1 and B.2 for the $b = 0.5$ and $b = 1$ populations, respectively).

3.6.1 Reliability of stochastic optimization

First, we report on our preliminary simulation, as described in §3.5.3.3. This analysis informed the decision to compute one allocation per method for each of 1000 pilot studies (versus applying each allocation method multiple times to a smaller number of pilots).

For each fully Bayesian allocation method, although repeated applications for a fixed pilot resulted in variable sample allocations, the variability in the corresponding estimated losses was roughly in line with what would be expected on account of MC error (see Figure B.1 for an illustrative example). To help quantify this, we can consider the estimated loss for a given method, pilot, and optimization run relative to the average losses for the remaining four runs for the same method and pilot. For NM and SPSA, these relative estimated losses were usually within the interval [0.99, 1.01], as would be expected if the methods produced stable results; further, the numbers of NM and SPSA relative estimated losses outside of this interval were roughly on par with analogous quantities for the non-stochastic pilot-based

allocations considered. Thus, to the extent that the B-NM or B-SPSA allocations themselves varied, this appeared likely to be primarily attributable to the MC estimated loss being insensitive to moderate changes in sample allocations. The results were consistent with the response surface being fairly flat for a large area near the optimum, and in which repeated NM or SPSA runs yielded slightly different allocations but which were all approximately optimal.

3.6.2 Performance relative to loss function

Next, we considered performance relative to the expected posterior loss, as per §3.5.3.1. Given that the MC-estimated losses varied considerably for each method across pilots, we considered the estimated losses relative to that produced by the B-SPSA for the given pilot. Box plots summarizing results across pilots by allocation strategy are provided in Figures 3.1 and 3.2, for the pilot-based strategies and RT-based strategies, respectively, with the latter displayed separately to allow the X-axis to have a wider range. The PB and B-NM allocations generally achieved estimated losses that were very similar to those of B-SPSA. The main classic allocations typically performed well, also, but had much longer tails, indicating that for some pilots, their use resulted in much higher estimated losses. The RT0 and RT2 allocations tended to lead to substantially higher estimated losses than the other methods. The preceding findings above also applied to the $b = 1$ population (see Figures B.2 and B.3), the main difference being that in this latter population, the tails for the Neyman and Cochran allocations were even longer (compared with the equivalent allocations for the

$b = 0.5$ population).

Figure 3.1: Estimated loss box plots by pilot-based allocation method

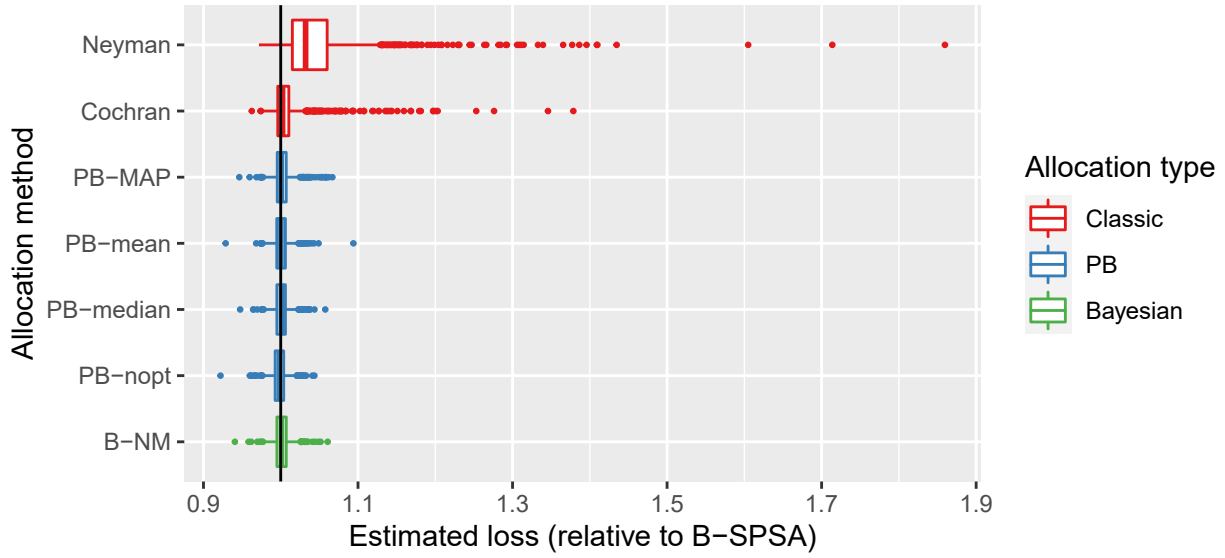
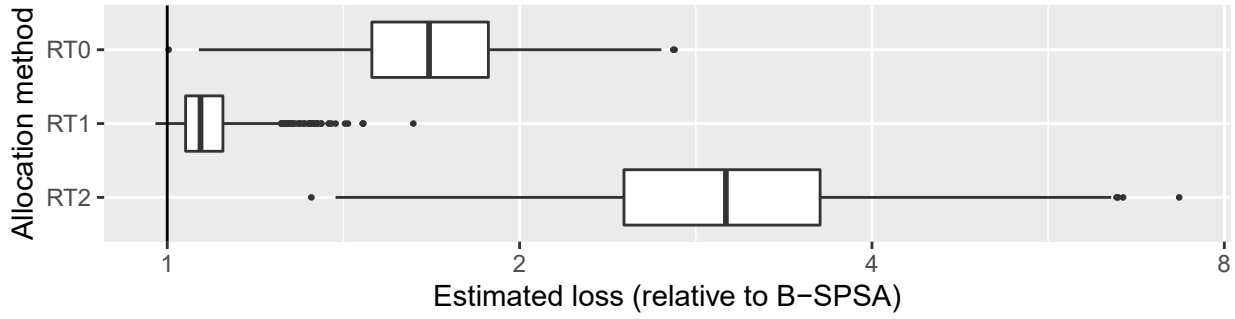


Figure 3.2: Estimated loss box plots by rule of thumb-based allocation method (relative to B-SPSA)



3.6.3 Performance relative to true population value

Table 3.3 displays results relative to the true population value, as per §3.5.3.2. Compared with B-SPSA, N-HT had a 45% higher RMSE, and C-SR had 3% higher RMSE; the PB and Bayesian strategies had nearly identical RMSEs. Considering the three rule-of-thumb

strategies, RT1-SR obtained similar performance to C-SR, whereas RT0-SR and RT2-SR had considerably higher RMSE. All main methods were approximately unbiased and nearly all had approximately-nominal coverage for 95% CIs (with a minor exception for RT2-SR, which had 94% coverage). We also considered the B-oracle strategy (not displayed), which used the knowledge that $b = 0.5$. The B-oracle strategy had a relative RMSE of 0.995, indicating that in the absence of precise estimates of other design parameters, simply knowing the true value of b did not meaningfully reduce RMSE; similar results were obtained for the $b = 1$ population (i.e., relative RMSE of 0.999). The findings for the $b = 1$ population (Table B.3) were quite similar to those described above, a minor difference being that the RMSE of the RT1-SR strategy was no longer practically distinguishable from that of the B-SPSA strategy (i.e., relative RMSE of 1.007).

Table 3.3: Simulation results

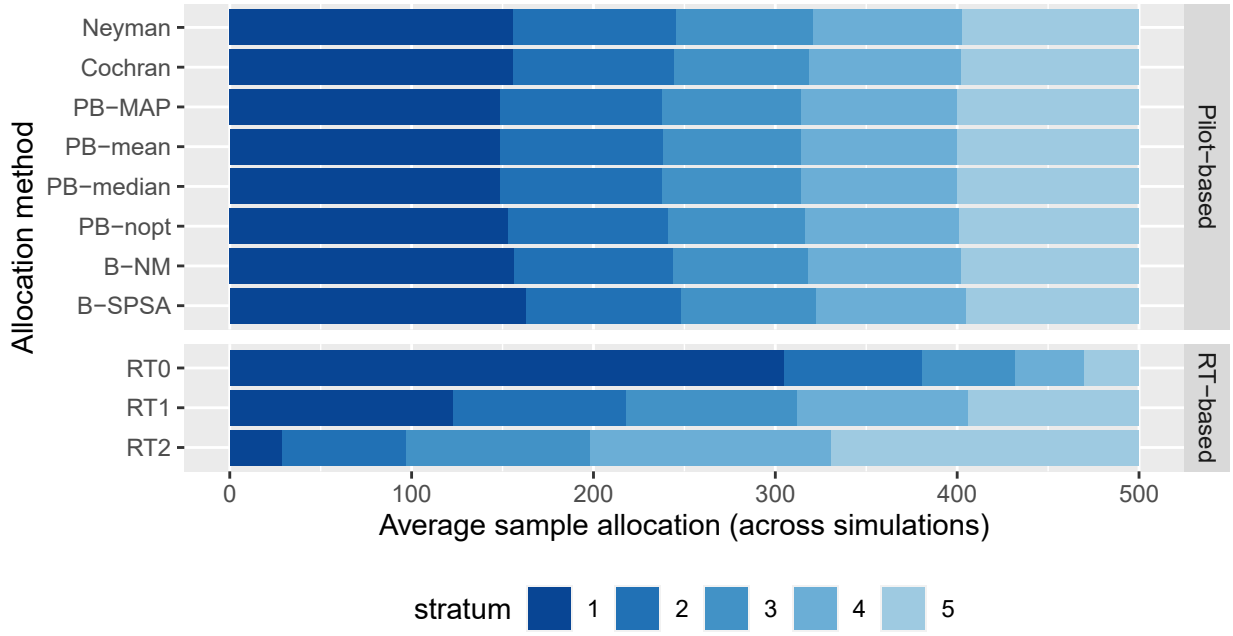
Strategy	Relative RMSE	1000 * RelBias	CI Coverage (%)	1000 * CI RelWidth
N-HT	1.446	0.03	94.9	58.6
C-SR	1.031	0.08	95.0	41.9
PB-MAP	1.006	0.54	95.1	41.1
PB-mean	1.003	0.60	95.1	41.1
PB-median	1.000	0.61	95.2	41.1
PB-nopt	1.004	0.80	95.1	41.0
B-NM	0.999	0.47	95.2	41.1
B-SPSA	1.000	0.65	95.2	41.0
RT0-SR	1.283	0.02	94.7	51.8
RT1-SR	1.029	-0.12	95.0	41.9
RT2-SR	1.658	-0.12	94.0	65.2

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

The relatively worse performance of the N-HT strategy compared with the other pilot-based strategies was apparently attributable to not incorporating the auxiliary variable in estima-

tion. For example, the main strategies resulted in fairly similar average allocations (Figure 3.3, top group), as was also the case for the $b = 1$ population (Figure B.4). Likewise, there were no meaningful differences across the pilot-based methods in the stability of the allocations (i.e., when considering the standard deviation of the allocations to an individual stratum for a method, across repeated pilots).

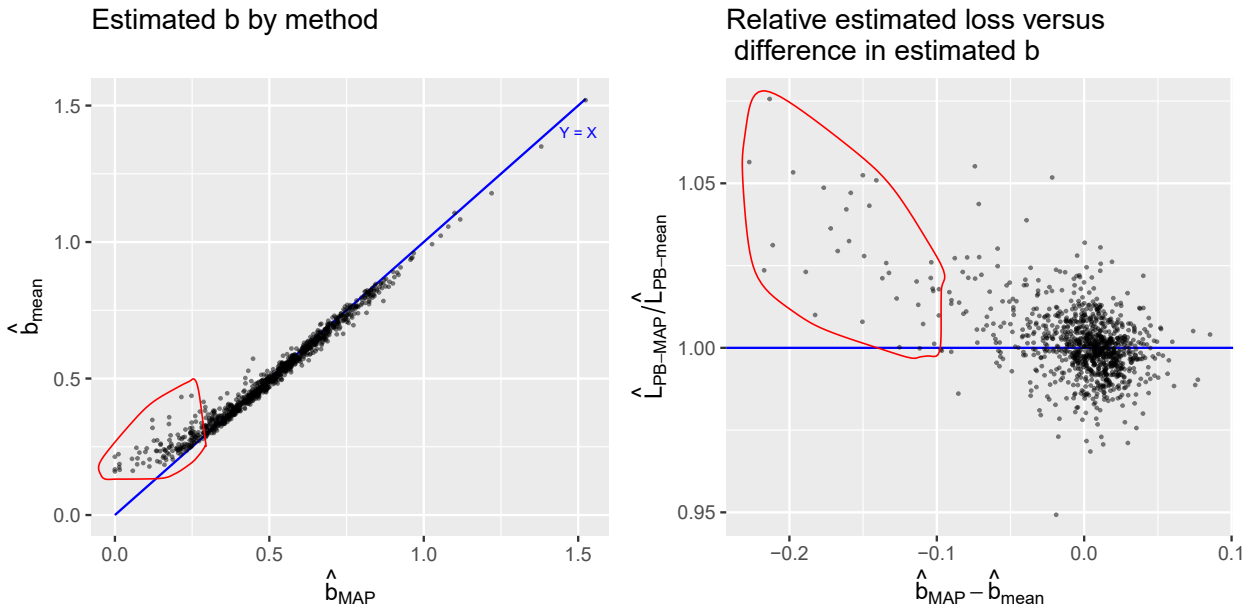
Figure 3.3: Average allocations to strata by method, across simulations



The three plug-in PB allocations (PB-MAP, PB-mean, PB-median) achieved extremely similar average performance, presumably since the MAP estimate was usually quite similar to the posterior mean and median given an individual pilot. For some pilots, the posterior distributions for b were asymmetric; in those cases, the posterior mean and median appeared to provide better estimates of b than the MAP. This is illustrated in Figure 3.4: the lefthand panel compares $\hat{b}_{mean} = \mathrm{E}(b|D1)$ with $\hat{b}_{MAP} = \underset{b}{argmax}\, \pi(b|D1)$, with each point represent-

ing a pilot; the righthand panel examines how the difference between these quantities (X axis) relates to the ratio of estimated losses of the corresponding PB allocations (Y axis). The figure shows that for pilots in which $\hat{b}_{MAP}$ was particularly close to 0 (left panel, points circled in red), the posterior mean was often substantially higher; in those cases, the PB-MAP allocation tended to lead to slightly higher estimated losses relative to the PB-mean allocation. Although not displayed, similar results were obtained when comparing $\hat{b}_{MAP}$ with the posterior median (and the associated allocations). Note that for the $b = 1$ population, there was better agreement between the different estimates of b, as very few such estimates were near 0 (see Figure B.5).

Figure 3.4: Estimated b and resulting loss by method: MAP versus posterior mean.



Note. Estimates are based on pilot data; each point denotes results for one pilot. On the lefthand panel, the red polygon encloses points where $\hat{b}_{MAP} \leq 0.25$, in which case $\hat{b}_{mean}$ was often substantially higher and closer to the true value of $b = 0.5$. On the righthand panel, the red polygon encloses points where $\hat{b}_{MAP} - \hat{b}_{mean} \leq -0.1$, in which case the associated MAP-based allocation tended to result in a higher estimated loss.

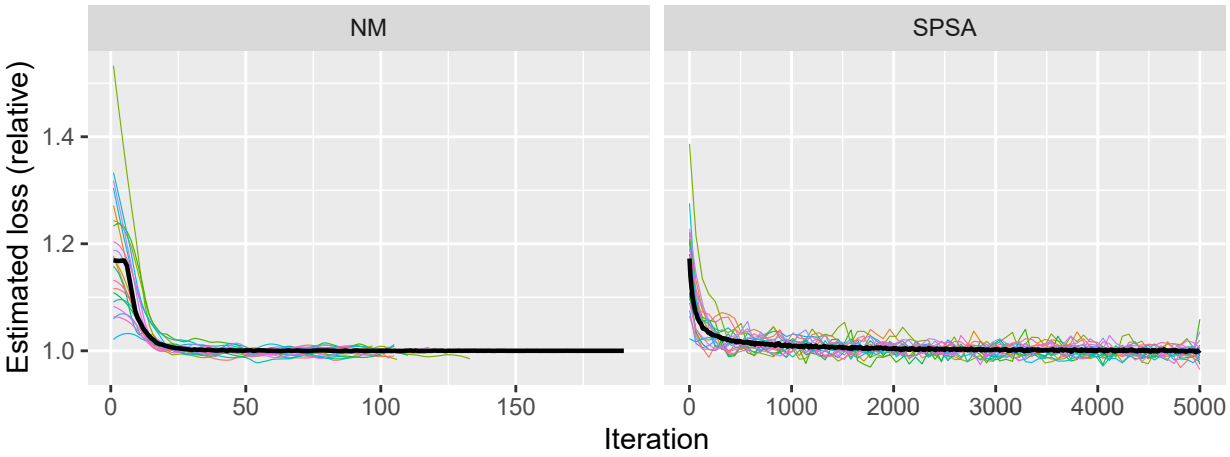
3.6.4 Computational aspects

The NM runs had a mean of runtime 4.0 hours (min = 2.1; max = 6.9), with averages of 106.5 iterations and 298.1 total objective function evaluations. These runs were all below the target runtime, since each run achieved approximate convergence before reaching the maximum number of function evaluations. In no cases did the simplex geometry collapse. The NM runs averaged 19.4 regular shrinks and 4.8 oriented restarts. On average, the NM-estimated loss at the current best point had a relative bias of -1.2% (i.e., on account of NM's tendency to retain unusually good measurements).

For SPSA, runtime averaged 7.5 hours, with all but eight runs taking 7.2—7.7 hours (the remaining eight runs shared an apparently slower computing node, and averaged 8.6 hours). In no cases did any points need to be projected; on average, 0.2% of steps had their movement size capped (min = 0.0%; max = 1.1%).

For both NM and SPSA, progress tended to be fastest in the earlier iterations, with diminishing gains in later iterations, and where the magnitude of gains from the starting point varied considerably across pilots. To illustrate, Figure 3.5 displays the average estimated loss for each method by iteration (black line) and the LOESS-smoothed estimated loss for 20 randomly chosen pilots (thin colored lines). Note that for NM, improvement at the best vertex does not yet appear for the first five or so iterations, as the simplex adapts to the shape of the response surface. Similar patterns of progress were observed for the $b = 1$ population (Figure B.6).

Figure 3.5: Estimated loss at candidate allocation by iteration and method, overall and for randomly selected pilots



Note. Estimated losses were computed following §3.5.4 (second-to-last paragraph), with estimated CVs of 3.4% for SPSA and 1.4% for NM, and were considered relative to the high quality estimated loss at the final allocation (via §3.5.3.1). Overall results (black line) denote averages across all 1000 pilots; pilot-specific results (thin colored lines) are LOESS-smoothed curves, shown for 20 randomly selected pilots. Given that the number of iterations varied across NM runs, for purposes of averaging results across pilots, the estimated loss for iterations after convergence was assumed to be equal to the high quality estimated loss at the final allocation.

3.7 NCCS application

3.7.1 Methods

Next, we applied our methods to analyzing tax returns of public charities, using the same data set described in §2.6, and briefly summarized here. The data were primarily based on IRS Form 990 data, as cleaned and aggregated by the National Center for Charitable Statistics (NCCS). The population reflects 140,793 domestic operating public charities with

at least \$100,000 in 2008 revenue, positive revenue in 2013, meeting filing thresholds, and classifiable by nonprofit sector. Among this population, X_{hi} and Y_{hi} reflect log revenue in fiscal years 2008 and 2013, respectively, for unit i within stratum h, the latter of which reflected nonprofit sector by revenue class (see Table 2.7).

We drew one equally allocated pilot sample of $m = 360$ units (15 units per stratum), which was used to allocate a main study with total sample size of $n = 1800$ units, with bounds of $4 \leq n_h \leq min\left(N_h, n - 4(H-1)\right)$ (considering that $n_h > n-4(H-1)$ implies an infeasible allocation). We considered the same strategies described in §3.5.2, except that SPSA employed the modifications described in §3.4.5 given the large number of strata used. Performance was considered relative to the loss function, as in §3.5.3.1, except that the high-quality MC estimates used $R1 = 100,000$ and $R2 = 500$. We examined reliability of stochastic optimization by computing the B-SPSA and B-NM allocations ten times each. To facilitate post-hoc analysis, after each optimization run, we reevaluated the loss function at the current best point for every NM iteration and at every fifth SPSA iteration via Equation (3.15), and using $R1 = 1,000$ and $R2 = 200$.

The application was conducted on the Zaratan HPC cluster, as described in §3.5.4. Individual NM or SPSA runs were designed to run within three days using a single core. The stratum with the most units (i.e., Human Services, Under \$300k) was implicitly defined (see §3.4.3).

SPSA used a similar set-up as described in §3.5.4, except with some changes to account for the larger number of strata, use the larger budgeted runtime, and otherwise improve performance. Each iteration's gradient estimate was the average of five independent applications

of Equation (3.24), with $D = 23$ and $\rho = -1/23$, as per §3.4.5. Each objective function evaluation used $R1 = 60$ and $R2 = 200$ and had an anticipated runtime of 4.0 seconds and CV of 4.1% for the initial allocation and 1.9% near the anticipated optimum. We chose a programmatically, similar to before, but aiming for early movements not to exceed 10 units (with maximum step size of 15). We chose $c = 10$, as to speed up progress in comparison to the earlier choice of 3. Other SPSA variables were as before (i.e., 5000 iterations; $A = 100$).

NM used a similar set-up as in §3.5.4, except that the maximum number of evaluations was increased to 3600 (mainly on account of the increased dimensionality) and the objective function evaluations used $R1 = 1000$ and $R2 = 200$, with anticipated runtimes of 64 seconds/each and CVs of 1.0% and 0.5% for the initial allocation and anticipated optimum, respectively.

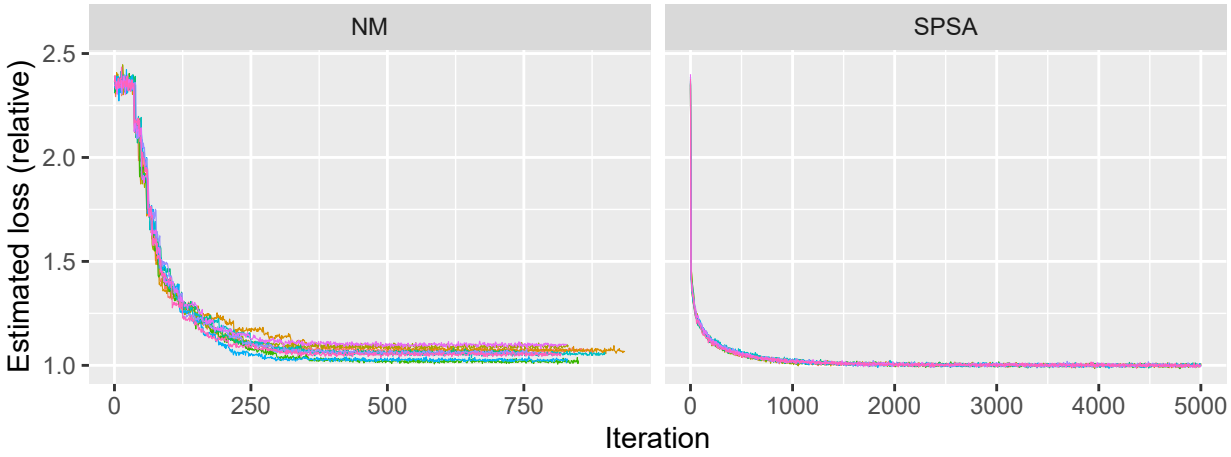
3.7.2 Results

First, we examined reliability of the stochastic optimization methods. Figure 3.6 displays the estimated loss for B-NM and B-SPSA runs by iteration, relative to the average final B-SPSA estimated loss, and colored by run. For SPSA (righthand panel), the runs exhibited extremely similar behavior, to the point that it is difficult to visually distinguish between the ten different colored curves, which largely overlap. The SPSA runs achieved rapid progress in the first few hundred iterations, with diminishing returns in later iterations, and essentially a flat slope by the end. The flat slopes for the final iterations could be explained by having reached an optimum, but could alternatively be explained by the choice of SPSA gain

sequences, which affect step sizes. To rule out the latter, we plotted the LOESS-smoothed SPSA-estimated gradients by iteration (Figure B.7), observing estimated gradients of around zero by the end of each run, providing additional evidence of having reached an approximate optimum.

NM also showed steady improvements in early iterations (Figure 3.6, lefthand panel), but progress became essentially flat after a few hundred iterations, with the NM runs' results varying modestly and achieving slightly worse performance than the SPSA runs.

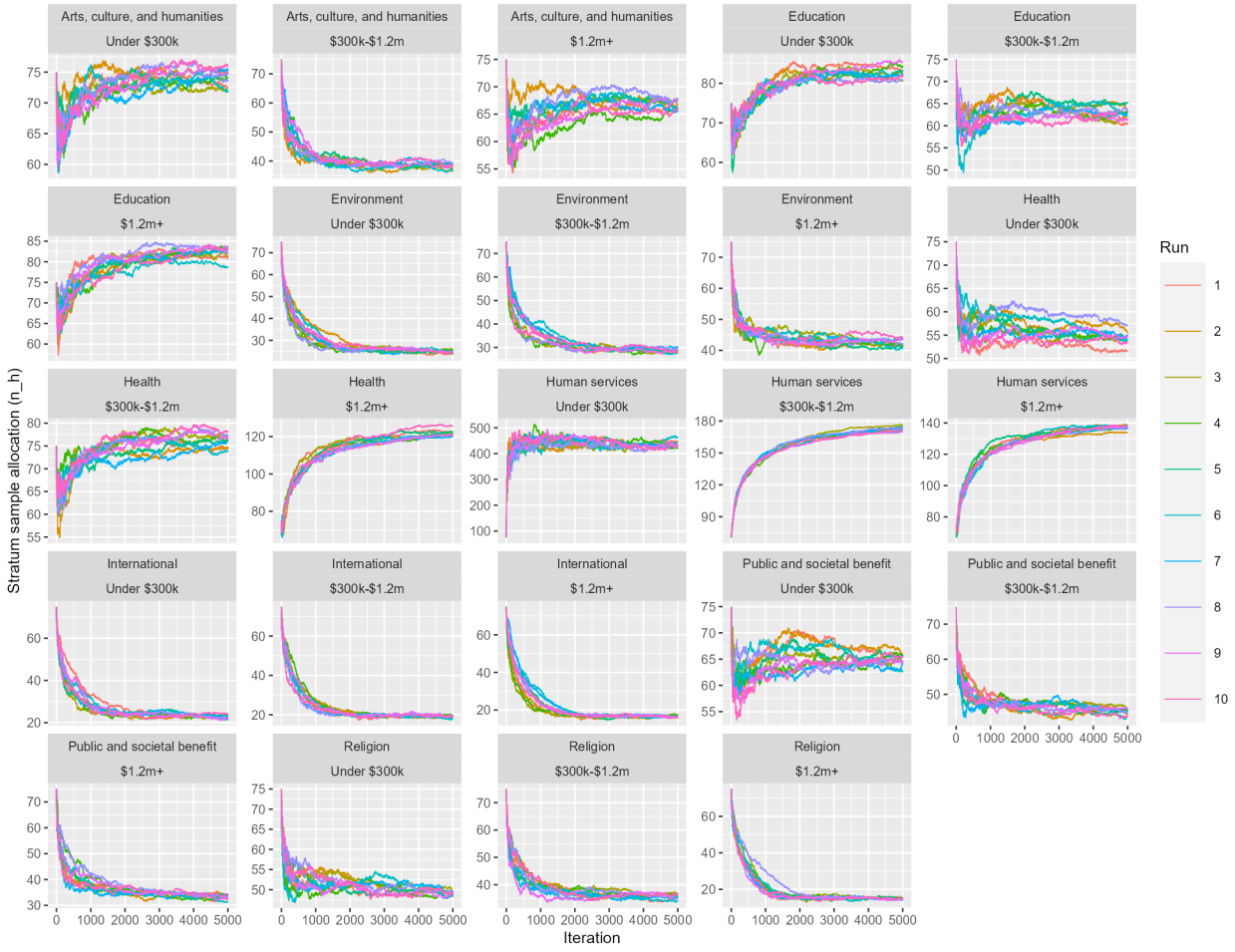
Figure 3.6: Estimated loss at candidate allocation by iteration and method



Note. Each color denotes one of ten optimization runs; each plot displays all ten curves (sometimes overlapping). Losses were estimated directly, independent of optimization, and were considered relative to the average final B-SPSA estimated loss.

Next, we examined how the strata allocations themselves vary across repeated stochastic optimization runs, starting with B-SPSA. Figure 3.7 shows the B-SPSA sample allocations by iteration—each panel denotes a stratum, with curves colored by run. The final allocations were reasonably tightly clustered, and did not change too much in the later iterations.

Figure 3.7: B-SPSA sample allocations by stratum, run, and iteration

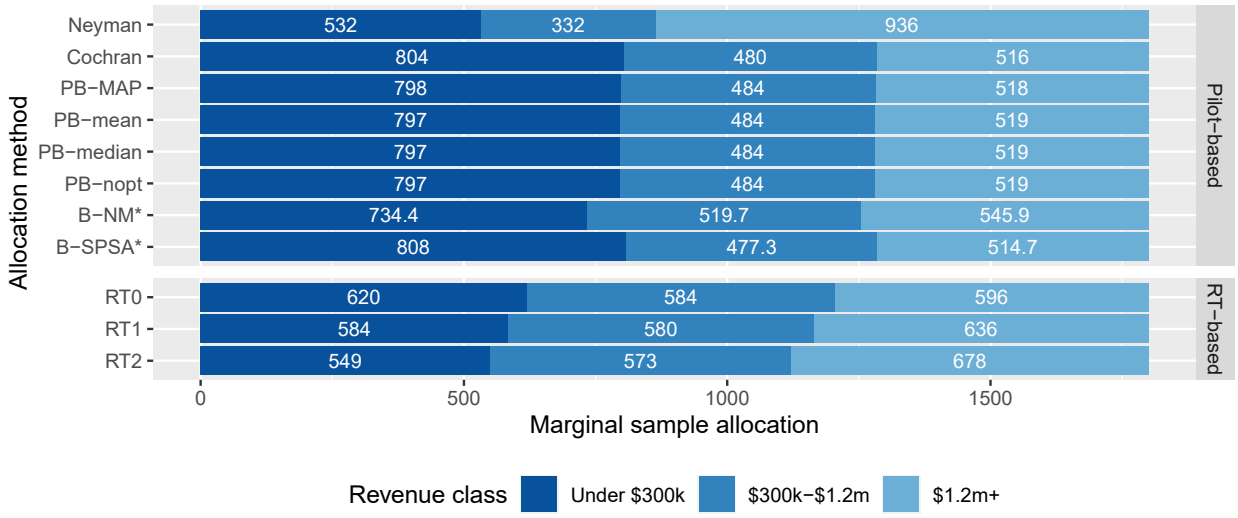


In contrast with B-SPSA, the B-NM allocations varied considerably across runs. This is seen in Figures B.8 and B.9, which display the B-NM allocations by iteration at the simplex centroid and current best point, respectively. Each figure has separate panels by strata that plot that stratum's allocation (Y axis) by iteration (X axis), colored by run. We observed that in some strata, allocations varied across runs by factors of at least two.

Next, we examined the sample allocations, provided in Table B.4 (and averaged across runs for B-NM and B-SPSA), and for which we present marginal distributions in Figures 3.8 and

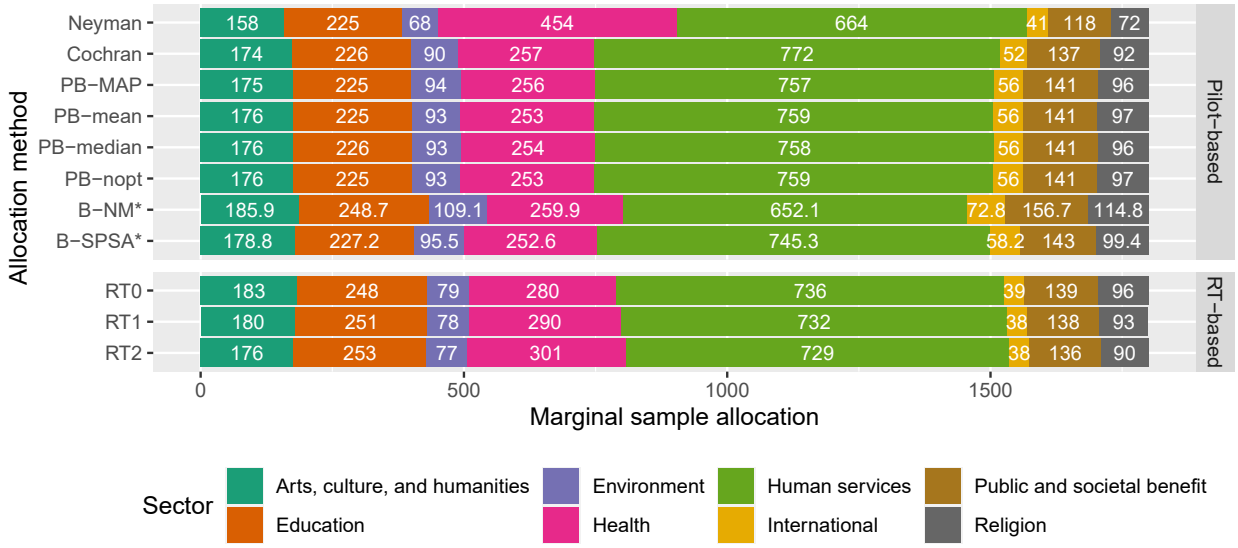
3.9, for ease of readability. The Cochan, PB, and B-SPSA allocations were quite similar. The PB allocations were nearly identical to each other, in spite of $\hat{b}_{MAP} = 0.00$ being substantially less than the posterior median (0.92) and mean (1.02) for the given pilot. B-NM tended to result in lower allocations to Stratum 13 (Human Services, under \$300k) relative to others; the apparent optimum for this stratum was farther away from the initial simplex than for other strata. Neyman resulted in by far the largest allocation toward the larger revenue class; we observed that the associated strata consistently had the most variability with respect to the pilot $\{y_{hi}\}$, but not necessarily with respect to model residuals (e.g., as used in Cochran's allocation).

Figure 3.8: Marginal allocations to revenue classes by allocation method



** B-NM and B-SPSA allocations are averaged across the ten runs.*

Figure 3.9: Marginal allocations to nonprofit sector by allocation method



** B-NM and B-SPSA allocations are averaged across the ten runs.*

Finally, we examined performance of the allocations. Table 3.4 shows the estimated loss of the allocations, averaged across runs for B-NM and B-SPSA, and considered relative to B-SPSA; each entry has a CV of 0.06% or less. The Cochran, PB, and B-SPSA allocations performed best (with individual SPSA runs having values of 0.999–1.001). The average B-NM run did 6% worse than B-SPSA (with individual runs having values of 1.02–1.10) The estimated losses of the RT allocations exceeded those from most other allocations.

In this application, we observe that a single run of B-SPSA would have been adequate, each allocation being approximately optimal. Given that we have access to ten such allocations, which are clustered tightly together, we would in practice average strata allocations across the ten runs and use that as the final B-SPSA allocation. In contrast, B-NM per-

Table 3.4: Estimated loss by allocation

Allocation	Est. loss (relative to B-SPSA)
Neyman	1.28
Cochran	1.00
PB-MAP	1.00
PB-mean	1.00
PB-median	1.00
PB-nopt	1.00
B-NM*	1.06
B-SPSA*	1.00
RT0	1.15
RT1	1.18
RT2	1.21

* B-NM and B-SPSA allocations are averaged across the ten runs.

formance varied modestly across runs and the allocations themselves varied considerably. Therefore, in practice, we would designate the best performing B-NM allocation as the final allocation—namely, Run 4, which was nearly optimal (est. loss of 1.02), albeit which slightly underperformed in comparison to several other methods.

3.8 Discussion

3.8.1 Summary of results

We provided a fully Bayesian computational framework for stratified sample allocation in establishment surveys. We focused on design for estimating the finite population mean under a univariate regression model with unknown coefficient of heteroscedasticity. We used Monte Carlo sampling to estimate the loss function for a given allocation, in conjunction with SPSA, a stochastic optimization method that accommodates noisy loss functions. In

doing so, we tailored SPSA toward our application, as to handle constraints and through a parameterization that guarantees full use of the sample. Empirically, the proposed methods (B-SPSA) appeared to be approximately optimal, performing as well or better than the other methods, while having clearer theoretical justification. The PB allocations also turned out to be approximately optimal, but this is not guaranteed for other populations and pilots.

Our work sheds light on the relative advantages and disadvantages of SPSA and NM in a sample optimization setting. We generally prefer SPSA since it seems more scalable toward harder problems (e.g., higher dimensions; noisier loss functions) and has clearer theoretical justification. Although SPSA was a little more complicated to set up (in terms of choosing SPSA variables), it provided more options for remediating poor performance. NM seems a reasonable alternative in a low-dimensional setting, as it performed well in the simulation, was simpler to customize toward a particular application, and is widely known. However, NM did not scale as well to the higher dimensionality of the application, where the estimated loss was slightly sub-optimal and the allocations themselves varied greatly across runs, the latter of which suggests risks of NM in high dimensional settings. Our preference of SPSA over NM is contrary to that of Huan and Marzouk (2013), who indicated that NM was "relatively less sensitive to the noise level" in their physics application. One possible explanation for the disconnect is that the best optimization algorithm is often application-specific. A different possibility could relate to the choice of SPSA variables—we found in a test application that certain SPSA variable choices could lead to poor performance.

Given the potential sensitivity of SPSA's performance to the choice of SPSA coefficients, we next discuss this choice. Spall's rules of thumb (1998) provided us a good starting point,

but Spall noted that the choices of gain sequences are critical to performance and vary by application. Indeed, in our preliminary test cases, the rules of thumb led to excessively small c (e.g., small fractional quantity), leading to occasional sharp movements away from the optimum. We remediated this issue primarily by increasing c to a small positive integer; as a further safeguard, we capped the maximum step size. Given that we planned for full runs to take hours (in the simulation) or days (in the application), we found it helpful to develop the coefficients using noisy versions of a quickly-computed test function that would follow from a known b scenario. Although our algorithm's runtime scales with target sample size, it would have been possible to accommodate larger sample sizes by using multiple CPUs, reducing quality of the MC estimator (e.g., reducing R1), and/or reducing the number of objective function evaluations (e.g., reduce number of iterations and/or only compute Equation (3.22) once per iteration).

We initialized SPSA at an equal allocation, to facilitate comparison with alternatives, but in other applications, a better starting allocation (e.g., PB-median or Cochran plug-in) may reduce the number of iterations needed and/or allow for smaller early step sizes. As a caveat, starting from an approximate optimum would lead to flat progress, potentially obscuring any algorithmic implementation issues (e.g., poor choice of gain sequences). In that case, it would be particularly important for practitioners to check that the gradient in later iterations is approximately flat. Although doing so would not preclude the possibility of having identified a local optimum, rather than global optimum, this risk is mitigated to some extent by noise in the MC estimates, and could be further reduced by running the algorithm multiple times from different starting points and seeing whether performance varies.

Returning to consider the other allocations, several performed well in simulation or application, although this is not guaranteed in other populations. The PB allocations were approximately optimal and nearly identical, with slightly worsened performance of PB-MAP for pilots where $f(b|D1)$ was asymmetric. Among classic methods, C-SR performed best, with good simulation performance overall, but occasionally with a substantially higher estimated loss for particular pilots. C-SR's overall good performance is not inconsistent with the Chapter 2 results for fixed b, where it was approximately optimal for some populations but sub-optimal for others, compared with a Bayesian approach. In the current chapter, N-HT did substantially worse in simulation due to not using the auxiliary data in estimation, and with a much less efficient allocation in the application. The RT1-SR method did as well as C-SR in the simulation, but not in the application. The RT0-SR and RT2-SR strategies substantially underperformed B-SPSA in both simulation and application.

3.8.2 Limitations and future directions

Future work could consider different ways to parameterize constrained sample allocation problems for use with SPSA. We ensured full use of the sample through an $(H-1)$–dimensional parameterization, with one stratum defined implicitly as having the remaining allowable sample, coupled with a restriction on the joint distribution of the stochastic perturbation vector's components. This required using a modified gradient estimator, which we showed to be approximately unbiased, and which worked well in practice. A different approach would have been to define all H strata explicitly; this would avoid need for some

SPSA modifications, while possibly introducing other issues (e.g., due to the optimum being on a boundary). Future work could formally consider the convergence properties of our proposed approach, and/or could consider alternative parameterizations. In the absence of such work, practitioners may find it helpful to re-run SPSA multiple times with different parameterizations (e.g., different choices of implicit stratum) to see if results vary.

The proposed methods implicitly assume constant per-unit costs, but could straightforwardly be adapted to allow costs to vary by strata. Ignoring fixed costs, and supposing a maximum cost of C, we would change §3.4.3 so that the decision variables reflect costs, say, $\{C_1, C_2, ..., C_{H-1}\}$, where $C_h = c_h \cdot n_h$ denotes total costs of stratum h, assumed to have known per-unit costs of c_h, for $h = 1, 2, ..., H-1$, and where the total costs for the last stratum are defined implicitly as $C_H = C - \sum_{h=1}^{H-1} C_h$ and have per-unit costs of c_H. We can translate any bounds on sample sizes to costs straightforwardly, by multiplying all sides of the sample constraint by c_h and substituting C_h in place of $c_h \cdot n_h$. This would also require an analogous modification to §3.4.7 to ensure that the sample sizes are integers and that $\mathrm{E}(\widetilde{C}_h) = C_h$ (although no longer requiring that $\sum_h^H \widetilde{C}_h = \sum_h^H C_h$, since this might not be possible). For example, evaluate C_h at $\widetilde{C}_h = c_h \left(\lfloor \frac{C_h}{c_h} \rfloor + I_h \right)$, where the I_h's are binary variables drawn via Poisson sampling with $\Pr(I_h = 1) = C_h - c_h \lfloor \frac{C_h}{c_h} \rfloor$.

Hence, our methods can straightforwardly accommodate lower and upper bounds on sample sizes and/or costs. More work may be necessary to identify ways to accommodate more complicated types of constraints. In such scenarios, one option may be to use different types of projection functions. A different approach would be to incorporate the constraints into the objective function via penalty functions (e.g., Wang & Spall, 2008).

We assumed that the coefficient of heteroscedasticity is constant across strata. If this assumption is violated, the proposed sample allocation might not work well. In this case, one option would be to separately analyze the coefficient of heteroscedasticity for different subgroups of the data. Further work could be done to explore this issue.

Our Bayesian analysis of b required specifying a prior support, which we set to be $(0, 2.5)$. If the true value of b is at the edge or outside of the prior support, posterior inferences about b could be compromised, which might negatively affect the resulting allocations. Practitioners may find it helpful to assess the sensitivity of $f(b|D1)$ to their choice of prior, and to widen the support if needed. For example, if it is plausible that the population is homoscedastic (i.e., $b = 0$), it may help to extend the lower bound to $b_0 = -0.5$.

Another line of future work could entail developing scalable methods of Bayesian optimization for other types of inferential objectives, beyond optimization for the finite population mean. Our methods are easily adapted for other inferential quantities (e.g., quantiles) and other types of loss functions beyond squared error loss, but at a potentially considerable computational cost. For example, suppose that our interest hinged on some other finite population quantity, $\theta = h\left(\{Y_{hi}\}\right)$, in place of $\bar{Y}$. This change could be accommodated by modifying the inner step of the MC estimator to estimate $\text{Var}(\theta|D2)$ via $R2$ draws from $f\left(\theta|D2\right) = \int f_1\left(\theta|\phi\right) f_2\left(\phi|D2\right) \mathrm{d}\phi$ (e.g., as in Appendix B.3.1). Although this alternative approach is more general, each MC estimate has computational complexity of order $O(R1 \cdot R2 \cdot N)$, and where each draw of θ requires a set of predictions for all nonsampled units. In §3.4.1, we obviated this need and reduced computation by a factor of $1 - \frac{n}{N}$ through use of expressions for the conditional mean and variance in conjunction with a fast approx-

imation of a population-based quantity. A different approach would have been to modify the alternative procedure of Appendix B.3.1 for very large populations by generating predictions only for some large random subset of the $N - n$ nonsampled units, and extrapolating their responses to the other nonsampled units, assuming that doing so did not distort the objective function. Clearly, design for other types of inferential objectives would be facilitated through additional analytical work to develop accurate and fast surrogate functions for approximating the loss.

Although we used a Bayesian formulation, future work could adapt the proposed simulation-based optimization methods to frequentist contexts. We assumed squared error loss for estimating the finite population mean, $\bar{Y}$, given the data, $D2$; following the Bayesian approach, we chose the estimator to minimize the posterior loss, resulting in estimation via $\mathrm{E}(\bar{Y}|D2)$ and a posterior loss of $\mathrm{Var}(\bar{Y}|D2)$, the latter of which averages the loss over the posterior distribution of the data. We then chose an allocation to minimize the expected posterior loss with respect to the predictive distribution for the future data. The proposed methods could be adapted to a frequentist context by changing the estimator and loss function to reflect frequentist design criteria, and where the loss function could still be averaged over some assumed statistical distributions for $\bar{Y}|D2$ and for $D2$ (given the allocation), not necessarily Bayesian.

Chapter 4: Stratified sample allocation under anticipated nonresponse

Abstract

Survey response rates have declined dramatically in recent years, increasing the costs of data collection. Despite this, there is little existing research on how to most efficiently allocate samples in a manner that incorporates response rate information. Existing mathematical theory on allocation for single-stage stratified sample designs generally assumes complete response. A common practice is to allocate sample under complete response, then to inflate the sample sizes by the inverse of the anticipated response rates. However, we show that this method can fail to improve upon an unadjusted allocation, due to ignoring the associated increase in the cost-per-interview.

We provide mathematical theory on how to allocate single-stage designs in a manner that incorporates the effects of nonresponse on cost efficiency. We derive the optimal allocation for the poststratified estimator under nonresponse, which minimizes either the unconditional variance of our estimator or the expected costs, holding the other constant, and taking into account uncertainty in the number of respondents. We assume a cost model that incorporates effects of nonresponse. We provide theoretical comparisons between our allocation and com-

mon alternatives, which illustrate how response rates, population characteristics, and cost structure can affect the methods' relative efficiency. In an application to a self-administered survey of U.S. military personnel, the proposed allocation increases the effective sample size by 25%, compared with common practice.

4.1 Introduction

In recent years, survey response rates have continued to decline (Brick & Williams, 2013; Luiten et al., 2020; National Research Council, 2013; Williams & Brick, 2018). This poses many consequences (Peytchev, 2013), including increased costs of data collection. Although nonresponse has received a great deal of attention, there appears to be no generally accepted theoretical framework for how to incorporate its impact on the efficiency of data collection when designing samples. Lohr and Brick (2014) did so in the case of dual frame random digit dialing (DFRDD) surveys, linking nonresponse to the allocation through their cost model, but there appears to be no analogous and general treatment for (single-frame) single-stage stratified random sampling (STSRS) designs. Szeitl and Rudas (2022) considered a sample allocation that adjusts for anticipated response rates, but did not consider costs. The lack of theory on this issue is especially problematic considering that nonresponse can vary meaningfully by subgroup and is often a main driver of per-unit costs in surveys where mail is a main contact mode.

The lack of attention toward these issues is evidenced by reviewing a few key textbooks in sampling (Lohr, 2021; Särndal et al., 1992; Valliant et al., 2018), where nonresponse is

generally treated separately from sample allocation for STSRS designs, and without comprehensive consideration for how it may affect sample efficiency. These textbooks are generally comprehensive, so we attribute the lack of coverage on the topic to the lack of relevant published literature. Of the three books, the most direct coverage is in Valliant et al., who, after providing the basic theory of optimal STSRS under full response (§3.1.2), provided an applied example of accounting for anticipated sample loss in the planning stage (§6.5.1). The numerical example, which reflects common practice, entailed computing an allocation assuming full response, then inflating the intended allocation of respondents to account for various forms of sample loss (e.g., due to housing vacancies, noncontact, screener nonresponse, ineligibility, and noncooperation); however, no theoretical treatment was provided, presumably because it had not yet been developed. Särndal et al. addressed nonresponse as a special topic (Chapter 15), focusing primarily on inferential issues rather than on implications for sample design; to the extent that design was considered, the authors focused on topics such as design features to reduce nonresponse (e.g., via an increased number of contacts and follow-up attempts), as well as double sampling for nonresponse. However, low response rates can arise in spite of best efforts to facilitate survey participation within available budget, and the types of efforts needed to seriously raise response rates (e.g., via face-to-face interviews) may not always be feasible. Recent texts by Lohr (2021) and Kalton (2020) likewise treated nonresponse separately from STSRS allocation.

This issue can be muddled by ambiguous terminology for "sample size," which is not always clearly defined, and is sometimes taken to refer to the original sample, before nonresponse, but sometimes used to connote the number of completed interviews that have been obtained.

At the design stage, sample designers need to consider both the original sample size, which they have direct control over, and the (expected) realized sample size after accounting for various sample losses (e.g., due to nonresponse). On the other side of data collection, at the inferential stage, "sample size" commonly refers to the set of eligible respondents, since the nonrespondents are not observed.

Sample designers' interpretations of these ambiguities can meaningfully affect allocations. When applying Neyman allocation or proportional allocation, does this refer to an allocation of expected respondents, as described in Valliant et al. (2018, §6.5.1), or does it refer to the original sample, as Szeitl and Rudas (2022) claim to be general practice for implementing proportional allocation? This choice can have a large effect on the efficiency of the design. However, current survey sampling theory does not provide an unambiguous framework for how to address it. Further, under a simple cost model for self-administered surveys that we will subsequently consider, and assuming known response rates that vary by strata, neither reflects an optimal allocation of resources. Neyman allocation of expected respondents may perform well if the per-respondent costs do not vary across strata, but this condition is clearly violated when response rates meaningfully affect costs and vary by strata. In the opposite direction, completely ignoring the strata-specific response rates at the design stage can lead to substantial underrepresentation of low response rate groups, hampering precision.

In considering this problem, some key assumptions are the use of a list-based sampling frame that allows identification of groups with different response rates, and where we assume use of a model for costs per respondent based on response rate information. Although we do not require a specific mode, it is important to consider mail as a contact mode. In recent years,

there have been some transitions from telephone toward self-administered and mixed mode surveys (Olson et al., 2021), driven by numerous factors, such as declining cost efficiency of telephone surveys (Lavrakas et al., 2017), development of address-based sampling (ABS) via the US Postal Service's Computerized Delivery Sequence File (Battaglia et al., 2016; Harter et al., 2016; Iannacchione, 2011; Link, Battaglia, Frankel, Osborn, & Mokdad, 2008), ability of mail to facilitate delivery of survey incentives, and development of push-web methodologies (e.g., Dillman, 2017), among others.

Here, we provide a few examples of the types of surveys with mail as a contact mode where useful stratifiers may exist. In some cases, list-based frames of special populations are available with rich auxiliary information, covering populations such as registered voters (Green & Gerber, 2006), physicians (DesRoches et al., 2015), military personnel (Mason et al., 1995; Office of People Analytics, 2022), and taxpayers (Brick, Contos, Masken, & Nord, 2010). For ABS surveys, Valliant, Hubbard, Lee, and Chang (2014) showed that commercially-appended variables can improve the efficiency of sample design; a different option for address frames is to use geographically appended neighborhood characteristics (e.g., Chen & Kalton, 2015), or even to match records to a secondary sampling frame to identify likely members of a rare population (e.g., Brick, Andrews, & Mathiowetz, 2016), to provide auxiliary information as part of a single-frame sampling approach.

The rest of our paper is organized as follows. In §4.2, we develop basic theory for optimal sample design under nonresponse. Nonresponse is linked to the allocation through a cost model that decomposes marginal costs based on whether they are associated with respondents or nonrespondents. In §4.3, we compare the theoretical efficiency of our proposed allocation

with alternatives, namely, Neyman allocations of invitees or of expected respondents, where the former is unadjusted and the latter adjusts by the inverse of the anticipated response rates. Although the proposed method improves upon both alternatives, the magnitude of efficiency gains depends on the cost structure, population characteristics, and alternative considered. In §4.4, we apply our allocation to a self-administered survey of military personnel, in which the response rates varied greatly by group. We consider how to best allocate a sample of 50,000 invitees, finding that the proposed allocation increases the effective sample size by 25%, compared with either of two commonly used alternatives. In §4.5, we summarize key findings, discuss limitations, and suggest future directions.

4.2 Optimal allocation for the finite population mean under nonresponse

§4.2 develops theory for optimal sample design under nonresponse, where the focus is on stratified sample allocation for estimating the finite population mean through a poststratified estimator. §4.2 is organized as follows. §4.2.1 provides key definitions and notation. §4.2.2 briefly reviews existing theory under complete response. §4.2.3 states our assumptions about the response propensities and provides the resulting conditional variance and mean of our estimator given the numbers of respondents (within strata). §4.2.4 describes our distributional assumptions about the numbers of respondents, which we use to compute an unconditional variance of our estimator in §4.2.5. §4.2.6 describes our assumed cost structure, which decomposes the costs within each stratum based on whether they are attributable to respondents or nonrespondents. §4.2.7 provides the optimal allocation.

4.2.1 Setup and notation

We assume that there is some population of $N = \sum_{h=1}^{H} N_h$ units, divided into H mutually exclusive strata, where N_h refers to the number of elements in stratum h, and where the elements are distinguishable by their index (hi) and exchangeable within stratum. We consider sample design for estimating the finite population mean, defined as

$$\bar{Y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h, \tag{4.1}$$

where $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$ denotes the stratum h mean and Y_{hi} denotes the value for unit (hi). In stratum h, let U_h denote the set of population units, $s_h \subset U_h$ refer to the original sample (before nonresponse), and $s_{Rh} \subset s_h$ refer to the realized sample (i.e., set of completed interviews), the latter of which is not known at the time of sample design. Assume that in stratum h, s_h is drawn as a simple random sample of n_h units (from U_h), resulting in r_h completed interviews. For purposes of simplicity, we assume that all members of U_h are known to be eligible units of the population (where eligibility is defined as in AAPOR, 2016).

We assume use of the poststratified estimator, defined under nonresponse as

$$\hat{\bar{Y}} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i \in s_{Rh}} \frac{y_{hi}}{r_h}, \tag{4.2}$$

where $\{y_{hi}\}$ refers to responding sample members' values of $\{Y_{hi}\}$, indexed by the responding

sample, and where we assume that poststratification follows the original strata. Under the special case of complete response (i.e., $s_{Rh} = s_h$), this estimator reduces to the HT estimator (Horvitz & Thompson, 1952).

4.2.2 Existing theory for complete response

Under existing sampling theory for STSRS designs, which assumes complete response, the optimal design minimizes either cost or variance, subject to holding the other constant, and where the variance refers to the design-based variance for the HT estimator for the finite population mean. Under this formulation, a well-known result is that the optimal allocation is $n_h \propto N_h S_h / \sqrt{c_h}$, where $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi} - \bar{Y}_h\right)^2$ is the stratum h variance, and where we temporarily define c_h under complete response as the per-unit cost in stratum h (e.g., Cochran, 1977, pp. 96–98). Under the special case where unit costs are constant across strata, this reduces to Neyman allocation, or $n_h \propto N_h S_h$ (Neyman, 1934). Another common allocation is proportional allocation, or $n_h \propto N_h$, which is theoretically optimal when the costs and variances are both constant across strata.

4.2.3 Conditional mean and variance of poststratified estimator

We assume that the survey design features, other than the sample allocation, are fixed in advance, and that population member i within stratum h has some response propensity of ϕ_{hi}, which indicates that sample member's probability of replying to the survey. We assume

use of a quasi-randomization model (Oh & Scheuren, 1983), wherein nonresponse is treated as an additional sampling phase, and where the response status of distinct sample units is assumed to be independent. Following Little (2022), we assume that the response propensity is conditioned on all survey and design variables. Let $\bar{\phi}_h = \frac{1}{N_h}\sum_{i=1}^{N_h}\phi_{hi}$ denote the average response propensity in stratum h and $\bar{\phi} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\phi_{hi}$ denote the population average.

To simplify analysis, for the remainder of this paper, we further assume constant response propensities within strata, that is, $\bar{\phi}_h = \phi_{hi}$ for all $i \in U_h$, and where we assume the $\{\bar{\phi}_h\}$ to be known. Although somewhat restrictive, these assumptions allow a natural extension of existing sampling theory under complete response, while relaxing the latter's untenable assumption that $\bar{\phi}_h = 1$. Under our assumptions, and conditional on realized strata sample sizes of $\{r_h\}$, each of the $\prod_{h=1}^{H}\frac{1}{\binom{N_h}{r_h}}$ possible realized samples has identical probability, meaning that the set of respondents is conditionally STSRS. Therefore, we have

$$\mathrm{E}\left(\hat{\bar{Y}}|\{r_h\}\right) = \bar{Y} \tag{4.3}$$

$$\begin{aligned}\mathrm{Var}\left(\hat{\bar{Y}}|\{r_h\}\right) &= \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(1-\frac{r_h}{N_h}\right)\frac{S_h^2}{r_h} \\ &= \sum_{h=1}^{H}\left(\frac{N_h^2 S_h^2}{N^2 r_h} - \frac{N_h S_h^2}{N^2}\right),\end{aligned} \tag{4.4}$$

assuming that $r_h \geq 1$ for all h (as is necessary to compute $\hat{\bar{Y}}$).

4.2.4 Use of (zero-) truncated binomial distribution

It would be natural to assume that $r_h \sim Binom(n_h, \bar{\phi}_h)$, if not for the fact that r_h could be zero. Therefore, for purposes of modeling the numbers of respondents in a given stratum, we assume that each r_h follows a (zero-) *truncated binomial distribution* (e.g., Rider, 1955; Stephan, 1945), which has mass proportional to that of the binomial distribution, after conditioning on $r_h \geq 1$.

Formally, let X denote a random variable with probability mass function (pmf) of

$$p(k; n, p) = \Pr(X = k; n, p) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{1 - (1-p)^n} \propto \binom{n}{k} p^k (1-p)^{n-k}, \tag{4.5}$$

for $k = 1, 2, ..., n$, and with zero mass, otherwise. Call this a *truncated binomial distribution* with parameters (n, p) and write the above as $X \sim TBinom(n, p)$. In Appendix C.1, we show that

$$\mathrm{E}(X) = \frac{np}{1 - q^n} \text{ and} \tag{4.6}$$

$$\mathrm{Var}(X) = \frac{np}{1 - q^n} \left(1 - p + np - \frac{np}{1 - q^n} \right), \tag{4.7}$$

where $q = 1 - p$. The first inverse moment can be computed via $\mathrm{E}\left(\frac{1}{X}\right) = \sum_{k=1}^{n} \frac{1}{k} p(k; n, p)$, although approximations are also available (e.g., Mendenhall & Lehman, 1960; Walker, 1984). As a consequence of the assumed quasi-randomization model, conditioning on $r_h \geq 1$ leads to $r_h | (n_h, \phi_h) \sim TBinom\,(n_h, \phi_h)$, for $h = 1, 2, ..., H$, and where the $\{r_h\}$ are mutually

independent.

Assuming adequately large sample sizes and expected numbers of respondents, the binomial distribution can be reasonably approximated by the truncated binomial distribution. For example, assuming minimum response propensities of $\bar{\phi}_h \geq .01$ and also assuming $n_h\bar{\phi}_h \geq 7$, we find that $0 \leq \frac{\mathrm{E}\left(TBinom(n_h,\bar{\phi}_h)\right)}{\mathrm{E}\left(Binom(n_h,\bar{\phi}_h)\right)} - 1 < .001$. We interpret this to mean that the relative absolute bias from misspecifying a binomial random variable as truncated binomial, and vice versa, is less than 0.1%, under these conditions.

4.2.5 Unconditional variance of poststratified estimator

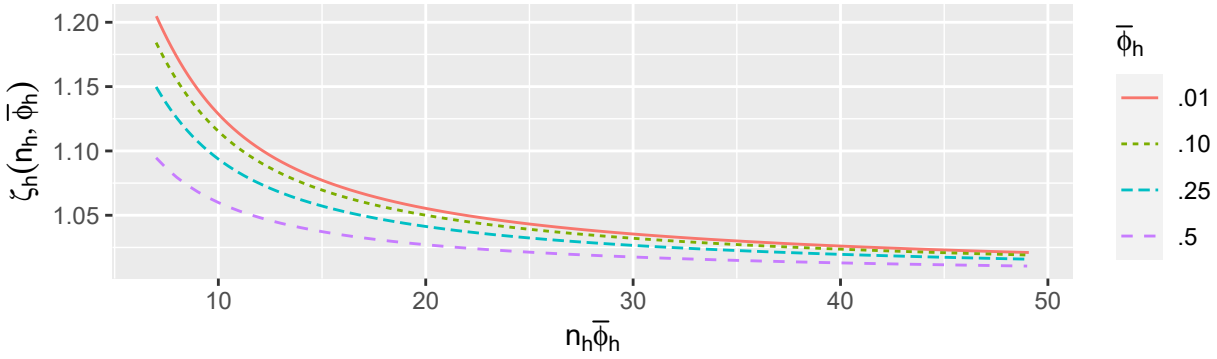
The variance of the poststratified estimator can be decomposed as $\mathrm{Var}(\hat{\bar{Y}}) = \mathrm{EVar}\left(\hat{\bar{Y}}|\{r_h\}\right) + \mathrm{VarE}\left(\hat{\bar{Y}}|\{r_h\}\right)$. Since $\mathrm{E}(\hat{\bar{Y}}|\{r_h\})$ does not depend on the $\{r_h\}$, the VarE term is zero, resulting in

$$\begin{aligned}\mathrm{Var}\left(\hat{\bar{Y}}\right) &= \mathrm{EVar}\left(\hat{\bar{Y}}|\{r_h\}\right) \\ &= \sum_{h=1}^{H}\left(\frac{N_h^2 S_h^2 \zeta_h(n_h,\bar{\phi}_h)}{N^2 n_h \bar{\phi}_h} - \frac{N_h S_h^2}{N^2}\right), \end{aligned} \tag{4.8}$$

where $\zeta_h(n_h,\bar{\phi}_h) := \mathrm{E}\left(\frac{1}{r_h}\right)\mathrm{E}(r_h)$ is a variance inflation term that captures the effect of the variability in the number of respondents for a given allocation. Although not employed here, a first-order Taylor series approximation for $\mathrm{E}\left(\frac{1}{r_h}\right)$ would lead to $\zeta_h(n_h,\bar{\phi}_h) \doteq 1$; this approximate inequality holds when both n_h and $n_h\bar{\phi}_h$ are adequately large. On the other hand, if the $\{n_h\bar{\phi}_h\}$ and/or $\{\bar{\phi}_h\}$ vary across strata, ignoring the $\{\zeta_h\}$ terms at the sample

allocation stage may lead to a modest loss in efficiency. To illustrate, Figure 4.1 plots $\zeta_h(n_h, \bar{\phi}_h)$ versus the expected number of respondents, $n_h\bar{\phi}_h$ (assumed to be at least 7), and holding the response rate constant at 1%, 10%, 25%, or 50%.

Figure 4.1: Zeta versus expected respondents, by response propensities



4.2.6 Cost model

Suppose that total survey costs can be expressed as

$$C = C_0 + \sum_{h=1}^{H} C_h, \tag{4.9}$$

where C_0 denotes fixed costs and C_h denotes total variable costs within stratum h, for $h = 1, 2, ..., H$. Suppose further that within each stratum, we can decompose marginal costs based on whether they are responding or nonresponding members of the original sample. For example, in a paper survey by mail, costs associated with all nonrespondents could entail prepaid incentives, postage, and printing; costs associated with respondents could include these as well as contingent (postpaid) incentives, return postage, and labor associated with

scanning paper questionnaires and manual coding of open-ended responses. We write the total variable costs in stratum h as

$$\begin{aligned} C_h &= r_h c_{R_h} + (n_h - r_h) c_{NR_h} \\ &= c_{NR_h} \left(r_h(\tau_h - 1) + n_h \right), \end{aligned} \tag{4.10}$$

where c_{R_h} and c_{NR_h} denote the marginal costs in stratum h for a single respondent or nonrespondent, respectively, and are assumed to be known, and where $\tau_h := \frac{c_{R_h}}{c_{NR_h}}$ is the ratio of costs for respondents to that of nonrespondents.

In (4.10), if we ignore the strata subscript, then $(\tau - 1)$ can be thought of as the marginal per-unit costs associated with respondents (e.g., use of postpaid incentives, perhaps modestly offset by savings from not needing to send as many mail pieces to those responding early in the field period) relative to nonrespondents. Although τ is specific to particular survey administration choices, self-administered surveys commonly entail $\tau \geq 1$. For example, in a survey with mail as the primary contact mode, web as the primary response mode, prepaid incentives but no postpaid incentives, and minimal use of open-ended responses, we might see $c_R \approx c_{NR}$, leading to $\tau = 1$. On the other hand, in a mail survey with use of postpaid incentives and where there are marginal labor costs associated with responses (e.g., scanning questionnaires and coding open-ended responses), we would expect to see higher values of τ (e.g., 2 or 3).

A common special case of our cost model, and which is typical of single-mode surveys, entails *common cost structures* for respondents and nonrespondents across strata, that is, $c_{R_h} = c_R$

and $c_{NR_h} = c_{NR}$ for known (c_R, c_{NR}) and for all $h \in \{1, 2, ..., H\}$. This scenario implies constant $\tau_h = \tau$ for known $\tau = \frac{c_R}{c_{NR}}$. However, we do not assume that this special case will always apply, considering that multi-mode surveys sometimes tailor mode based on strata characteristics, for example, based on modeled probability of responding via paper only (U.S. Census Bureau, 2022). Other situations where strata cost structures may vary include tailored incentives (Jackson, McPhee, & Lavrakas, 2020; Link & Burks, 2013), tailored response mode (Jackson, Medway, & Megra, 2023), or the use of matched telephone numbers for nonresponse followup in a mixed-mode survey (e.g., Olson et al., 2021, §5).

Through our cost structure (via [4.9] and [4.10]), and assuming that $\mathrm{E}(r_h) \doteq n_h \bar{\phi}_h$ (the result being only approximate due to the denominator of [4.6]), the survey's expected total costs are

$$\mathrm{E}(C) = C_0 + \sum_{h=1}^{H} n_h c_{NR_h} \left(\bar{\phi}_h (\tau_h - 1) + 1 \right). \tag{4.11}$$

Let $c_h = \frac{\mathrm{E}(C_h)}{n_h}$ denote the expected cost per invitee. The per-invitee costs under the general cost structure considered and two important special cases are summarized in Table 4.1. We will subsequently provide an allocation under the general cost structure, which can be specialized toward the other cases when appropriate.

Table 4.1: Summary of cost structures considered

Cost structure scenario	Assumptions	Expected cost per invitee
General (strata-specific)	$\{\tau_h\}$ and $\{c_{NR_h}\}$ known	$c_h = c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)$
Common cost structure (across strata)	$\{\tau_h = \tau\}$ and $\{c_{NR_h} = c_{NR}\}$ for known τ and c_{NR}	$c_h = c_{NR}\left(\bar{\phi}_h(\tau - 1) + 1\right)$
Constant cost per invitee	$\{\tau_h = 1\}$ and $\{c_{NR_h} = c_{NR}\}$ for known c_{NR}	$c_h = c_{NR}$

4.2.7 Optimization step

We assume that our objective is to minimize the (unconditional) variance, $\text{Var}\left(\hat{\bar{Y}}\right)$, subject to fixed expected costs, or to minimize the expected costs, subject to fixed variance. This objective is equivalent to minimizing $V'C'$, holding one of the two terms constant, where $V' := \sum_{h=1}^{H} \frac{N_h^2 S_h^2 \zeta_h(n_h, \bar{\phi}_h)}{N^2 n_h \bar{\phi}_h}$ and $C' := \sum_{h=1}^{H} n_h c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)$ are the only components of the unconditional variance and expected cost that depend on the $\{n_h\}$. Define $a_h = \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{n_h \bar{\phi}_h}}$ and $b_h = \sqrt{n_h c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)}$. Ignoring the constant $\frac{1}{N^2}$ term, minimizing $V'C'$ is equivalent to minimizing $\left(\sum_h a_h^2\right)\left(\sum_h b_h^2\right)$. The Cauchy–Schwartz inequality indicates that $\left(\sum_h a_h b_h\right) \leq \left(\sum_h a_h^2\right)\left(\sum_h b_h^2\right)$. The righthand side is minimized only if $\frac{a_h}{b_h} = k$ for some constant k. Therefore, the optimum is at

$$n_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{\bar{\phi}_h c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)}}, \tag{4.12}$$

assuming that $n_h \leq N_h$ for all h; if this assumption is not met, or if the sample designer wishes to impose additional constraints (e.g., stratum sample size requirements; domain

precision requirements), then the optimum can be computed by minimizing $V'C'$ through mathematical programming (see, e.g., Valliant et al., 2018, Chapter 5). Note that (4.12) is recursively defined, since $\zeta_h(n_h, \bar{\phi}_h)$ incorporates n_h. Calculations can be simplified by assuming that $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$. Alternatively, the optimum can be computed iteratively, as follows:

1. For iteration $k = 1$, calculate n_h^1 (i.e., via [4.12] or through a mathematical programming approach) under the assumption that $\zeta_h(n_h, \bar{\phi}_h) = 1$.

2. For each subsequent iteration k, for $k = 2, 3, ...$, do the following:

   (a) Compute $\zeta_h(n_h^{k-1}, \bar{\phi}_h)$ via use of the truncated binomial distribution.

   (b) Compute n_h^k under the assumption that $\zeta_h(n_h, \bar{\phi}_h) = \zeta_h(n_h^{k-1}, \bar{\phi}_h)$.

   (c) If the largest component of $n_h^k - n_h^{k-1}$ has magnitude below some tolerance (e.g., $1e-8$), or if the maximum number of iterations has been reached, accept n_h^k as the optimum; otherwise, continue.

When applying step 2(a) above, two implementation suggestions are as follows. First, if there are any strata where the allocation may lead to $\mathrm{E}(r_h) < 3.5$, we suggest computing $\zeta_h^k(.)$ at $\zeta_h^k\left(max\left(n_h^{k-1}, \left\lceil \frac{3.5}{\bar{\phi}_h} \right\rceil\right), \bar{\phi}_h\right)$, to protect from the fact that the truncated binomial distribution may no longer be a good approximation for the binomial distribution below these levels, and given that we observed numerically that $\zeta_h(n_h, \bar{\phi}_h)$ is roughly maximized for given $\bar{\phi}_h$ (fixed at various levels between .01 and 1) when $n_h \approx \frac{3.5}{\bar{\phi}_h}$. The second suggestion addresses the mismatch between (4.12), which generally yields continuous n_h, and the definition of

$\zeta_h(n_h, \bar{\phi}_h) = \mathrm{E}(r_h)\mathrm{E}\left(\frac{1}{r_h}\right)$, which assumes that n_h is an integer (given that we assumed r_h to follow a discrete distribution). As a workaround, for purposes of step 2(a), we suggest defining $\zeta_h(n_h, \bar{\phi}_h)$ for continuous n_h as a weighted average of its evaluations at $\lfloor n_h \rfloor$ and $\lfloor n_h \rfloor + 1$, via

$$\zeta_h'(n_h, \bar{\phi}_h) = w_h \cdot \zeta_h(\lfloor n_h \rfloor, \bar{\phi}_h) + (1 - w_h) \cdot \zeta_h(\lfloor n_h \rfloor + 1, \bar{\phi}_h), \tag{4.13}$$

where $w_h = (\lfloor n_h \rfloor + 1) - n_h$. In comparison to simply rounding continuous n_h, the use of a weighted average mitigates convergence issues that could occur when the fractional component of n_h is close to 0.5.

Equation (4.12) can alternatively be written as

$$n_h\bar{\phi}_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{c_{NR_h}\left((\tau_h - 1) + \frac{1}{\bar{\phi}_h}\right)}}, \tag{4.14}$$

where $n_h\bar{\phi}_h$ reflects expected respondents. The denominator now reflects the expected cost per respondent, placing emphasis on the effect of nonresponse on the cost efficiency of data collection. Considering that self-administered surveys commonly have very low response rates, in several practical circumstances, $(\tau_h - 1)$ will be small relative to $\frac{1}{\bar{\phi}_h}$, meaning that the low response rates are driving the bulk of per-respondent costs (on expectation), in which case it may be reasonable to approximate (4.14) by assuming that $\tau_h \approx 1$. Under this situation, and further assuming that c_{NR_h} is constant across strata (as is commonly the case for single-mode surveys), (4.14) reduces to $n_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{\bar{\phi}_h}}$. This allocation is also

appropriate if 'cost' is defined as the number of invitees (as opposed to a monetary cost).

4.3 Comparison to existing methods

Section 4.3 compares the theoretical efficiency of our proposed allocation with Neyman allocations of invitees and expected respondents, which differ based on whether they adjust for the inverse of the anticipated response propensities. These comparisons involve considering approximate variance ratios. We provide plots and discussion that illustrate how the relative efficiency of the allocations varies based on cost structure (e.g., choice of τ) and population characteristics (e.g., response propensities). §4.3.1 provides the approximate variance ratio for the variance of an alternative design to that of our proposed design, under some simplifying assumptions. §4.3.2 examines the variance ratios under a constant cost per invitee scenario. §4.3.3 examines the variance ratios under a common cost structure scenario, which allows for $\tau > 1$.

4.3.1 Ratio of approximate variances

Here, we examine the efficiency of our proposed allocation, compared with alternatives. From (4.12), by assuming that the strata are adequately sized such that $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$, the approximately optimal allocation can be written as

$$n_h^{opt} \propto \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}, \tag{4.15}$$

where $c_h = \frac{\mathrm{E}(C_h)}{n_h} = c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)$ is the expected cost per invitee in stratum h, as in Table 4.1. As an alternative, consider some arbitrary allocation $\{n_h^{alt}\}$, where we assume $n_h^{alt} \geq 1$ for all h, and where we assume equivalent total expected costs of $\sum_h n_h^{opt} c_h = \sum_h n_h^{alt} c_h$. Let $AV\left(\hat{\bar{Y}} | \{n_h^0\}\right) := \frac{1}{N^2}\sum_h \frac{N_h^2 S_h^2}{n_h^0 \bar{\phi}_h}$ denote the approximate variance from (4.8), assuming a negligible finite population correction (i.e., $\frac{r_h}{N_h}$ close to zero), and again assuming $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$. The ratio of approximate variances for the two designs turns out to be

$$\frac{AV\left(\hat{\bar{Y}} | \{n_h^{alt}\}\right)}{AV\left(\hat{\bar{Y}} | \{n_h^{opt}\}\right)} = \frac{\left(\sum_h T_h \alpha_h\right)\left(\sum_h \frac{T_h}{\alpha_h}\right)}{\left(\sum_h T_h\right)^2}, \tag{4.16}$$

where we define $T_h = \frac{N_h S_h \sqrt{c_h}}{\sqrt{\bar{\phi}_h}}$ and $\alpha_h = \frac{n_h^{alt}}{n_h^{opt}}$ for $h = 1, 2, ..., H$ (see Appendix C.2 for detail).

4.3.2 Constant cost per invitee scenario

Consider again the important special case where $\tau_h = 1$ and c_{NR_h} is constant across strata (i.e., *constant cost per invitee* scenario from Table 4.1), as would apply if the objective was to minimize the variance subject to fixed overall sample size. Under this scenario, the approximately optimal allocation becomes $n_h^{opt} \propto \frac{N_h S_h}{\sqrt{\bar{\phi}_h}}$, again assuming that $\zeta_h(n_h, \bar{\phi}_h) \approx 1$. As alternatives, consider allocations of $n_h^{Ninv} \propto N_h S_h$ and $n_h^{Nresp} \propto \frac{N_h S_h}{\bar{\phi}_h}$, where the superscript conveys that these refer to Neyman allocations of invitees and respondents, respectively. Note that use of $\{n_h^{Nresp}\}$ is a common practice (as implied by Valliant et al., 2018, §6.5.1) and specializes to Szeitl and Rudas's (2022) proposed allocation in the case where S_h is constant across strata. These two alternative allocations turn out to have equivalent variances (under the given special case), which, in comparison to the proposed allocation, yield

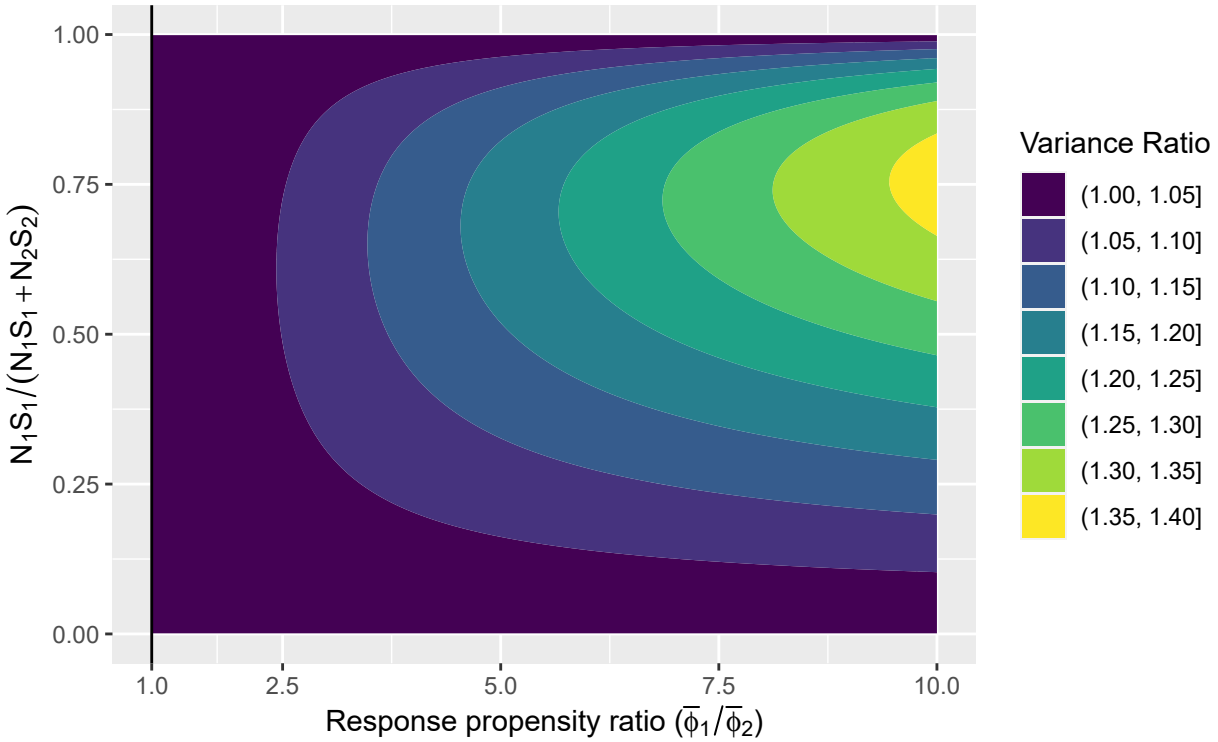
variance ratios of

$$VR := \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{\left(\sum_h N_h S_h\right)\left(\sum_h \frac{N_h S_h}{\bar{\phi}_h}\right)}{\left(\sum_h \frac{N_h S_h}{\sqrt{\bar{\phi}_h}}\right)^2} \tag{4.17}$$

(see Appendix C.2 for detail). This illustrates that when allocating a fixed number of invitees, adjusting by the inverse of the expected response rates (i.e., response propensities) does not actually reduce the resulting (approximate) variance, as both are equally sub-optimal.

To illustrate how sample efficiency can vary as a function of the $\{N_h S_h\}$ and $\{\bar{\phi}_h\}$, consider a 2-stratum situation. Under $H = 2$, (4.17) reduces to $VR = \frac{(\beta+1)\left(\frac{\beta}{\gamma}+1\right)}{\left(\frac{\beta}{\sqrt{\gamma}}+1\right)^2}$, where $\beta = \frac{N_1 S_1}{N_2 S_2}$ and $\gamma = \frac{\bar{\phi}_1}{\bar{\phi}_2}$ are the strata ratios of aggregate standard deviations and response propensities, respectively. Figure 4.2 provides a contour map of the resulting VR as a function of the stratum 1 share of the aggregate standard deviations (i.e., $\frac{N_1 S_1}{N_1 S_1 + N_2 S_2} = \frac{\beta}{1+\beta}$) and the ratio of expected response rates (i.e., γ), and where we assume without loss of generality that $\bar{\phi}_1 \geq \bar{\phi}_2$. The figure illustrates that under the given scenario, our proposed allocation produces the largest efficiency gains when the stratum with the larger aggregate standard deviation (i.e., larger $N_h S_h$) has a substantially larger response rate than the other stratum. For a given γ, with $H = 2$, VR obtains its highest value when $\beta = \sqrt{\gamma}$, as is straightforward to verify analytically by considering $\frac{\partial VR}{\partial \beta}$.

Figure 4.2: Variance ratio: common practice versus proposed allocation under H = 2, by response propensity ratio and relative aggregate standard deviation

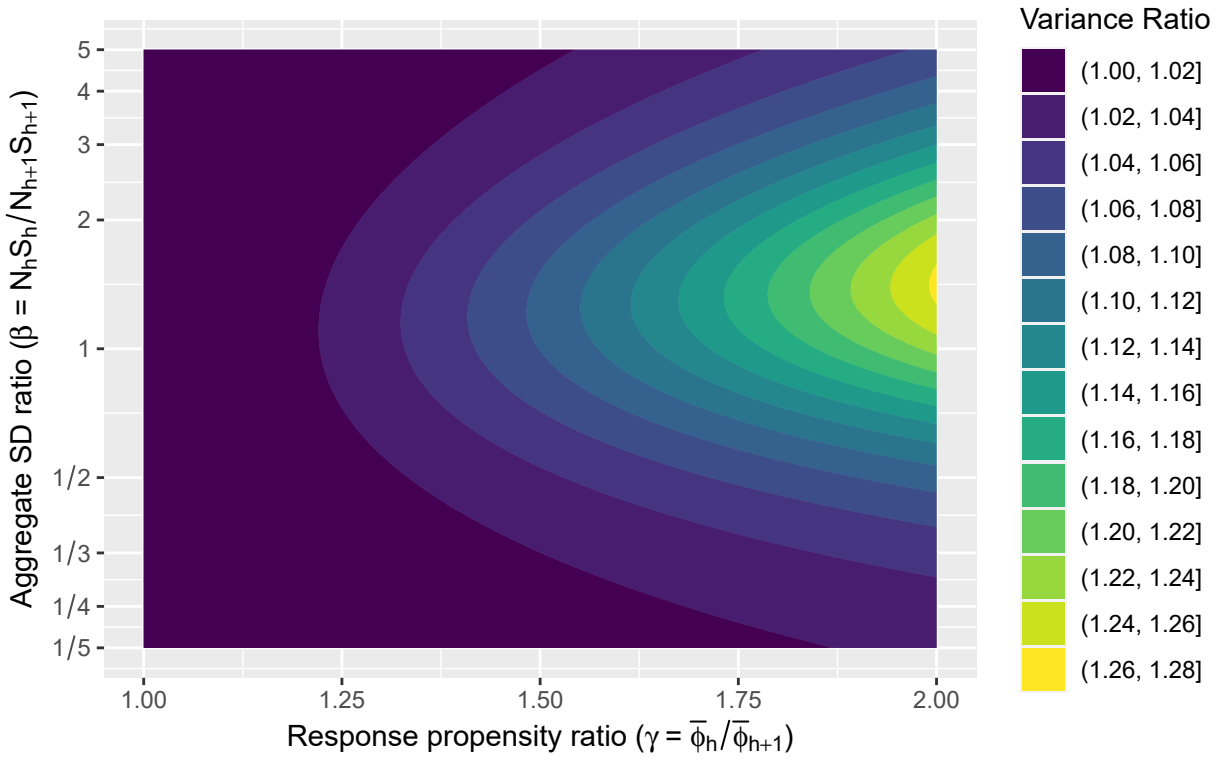


Note. Formula assumes $\tau_h = \tau$, $c_{NR_h} = c_{NR}$ for constant c_{NR_h}, $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$, and $1 - \frac{r_h}{N_h} \doteq 1$.

Although results for H-dimensional scenarios depend on the type of population, we can get a rough sense of what happens with larger dimensionality by considering a certain subset of populations that could be parameterized in two dimensions. Specifically, assume that the strata aggregate standard deviations and response rates are geometric sequences with ratios of $\beta = \frac{N_j S_j}{N_{j+1} S_{j+1}}$ and $\gamma = \frac{\bar{\phi}_j}{\phi_{j+1}}$, respectively, for $j = 1, 2, ..., H-1$. Under these assumptions, the variance ratio is $VR = \frac{\left(\sum_{j=0}^{H-1} \beta^j\right)\left(\sum_{j=0}^{H-1} \left(\frac{\beta}{\gamma}\right)^j\right)}{\left(\left(\sum_{j=0}^{H-1} \left(\frac{\beta}{\sqrt{\gamma}}\right)^j\right)\right)^2}$. Figure 4.3 displays the VR in this scenario by β (Y-axis; log-scale) and γ (X-axis), and assuming that $\gamma \geq 1$ (considering that $(\beta = \beta_0, \gamma = \gamma_0)$ and $(\beta = \frac{1}{\beta_0}, \gamma = \frac{1}{\gamma_0})$ yield equivalent results). As before, for fixed

β, larger values of γ lead to larger VRs (and using the assumption of $\gamma \geq 1$). Although not displayed, for fixed γ, numerical results appear to suggest that VR is maximized when $\beta = \sqrt{\gamma}$.

Figure 4.3: Variance ratio: common practice versus proposed allocation under H = 5, by response propensity ratio and aggregate standard deviation ratio



Note. Formula assumes $\tau_h = \tau$, $c_{NR_h} = c_{NR}$ for constant c_{NR_h}, $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$, and $1 - \frac{r_h}{N_h} \doteq 1$. Y axis is displayed on logarithm scale.

4.3.3 Common cost structure scenario

Although Figure 4.2 assumes that $\tau_h = 1$, a somewhat more general situation is the *common cost structure* scenario from Table 4.1, which entails an assumption of $\tau_h = \tau$ for known constant τ (while still assuming $c_{NR_h} = c_{NR}$ for constant c_{NR}); further assume that $\tau \geq 1$.

Then, (4.16), as applied to the Neyman allocations of invitees and respondents, after some simplification becomes

$$VR_{Ninv} := \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{\left(\sum_h N_h S_h c_h\right)\left(\sum_h \frac{N_h S_h}{\bar{\phi}_h}\right)}{\left(\sum_h N_h S_h \sqrt{\frac{c_h}{\bar{\phi}_h}}\right)^2} \tag{4.18}$$

$$VR_{Nresp} := \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{\left(\sum_h N_h S_h \frac{c_h}{\bar{\phi}_h}\right)\left(\sum_h N_h S_h\right)}{\left(\sum_h N_h S_h \sqrt{\frac{c_h}{\bar{\phi}_h}}\right)^2}, \tag{4.19}$$

and where the expected per-invitee strata costs are $c_h = c_{NR}\left(\bar{\phi}_h\left(\tau - 1\right) + 1\right)$ under the current assumptions. Under this scenario, it turns out that $AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right) \leq AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right)$, the inequality being strict as long as $\tau > 1$ and the response propensities vary by strata. This result can be shown by considering $h(\tau) := \frac{AV(\hat{\bar{Y}}|\{n_h^{Ninv}\})}{AV(\hat{\bar{Y}}|\{n_h^{Nresp}\})}$ in conjunction with $h'(\tau) := \frac{\partial h(\tau)}{\partial \tau}$, and then showing $\tau \geq 1 \implies h'(\tau) \geq 0$, to be used with the fact that $h(1) = 1$ (see Appendix C.3 for detail). Noting that $\{n_h^{opt}\}$ is optimal, this further implies that

$$AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right) \leq AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right) \leq AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right). \tag{4.20}$$

The relative efficiency of these allocations, as quantified by (4.18) and (4.19), can be further examined by considering the effects that τ and the response propensities have on the allocations themselves. The Neyman allocations of invitees and respondents have different sources of inefficiency. Compared with $\{n_h^{opt}\}$, the inefficiency of $\{n_h^{Ninv}\}$ arises from sampling more units than is optimal in the strata with large response propensities, leading to larger design effects, but which are mitigated to some extent by the larger number of expected respondents. In contrast, $\{n_h^{Nresp}\}$ allocates more than is optimal to strata with lower response

propensities, resulting, on average, in a higher cost per (nominal) respondent than $\{n_h^{opt}\}$ and fewer respondents, although these negative effects are partially mitigated since the $\{n_h^{Nresp}\}$ allocation results in lower design effects.

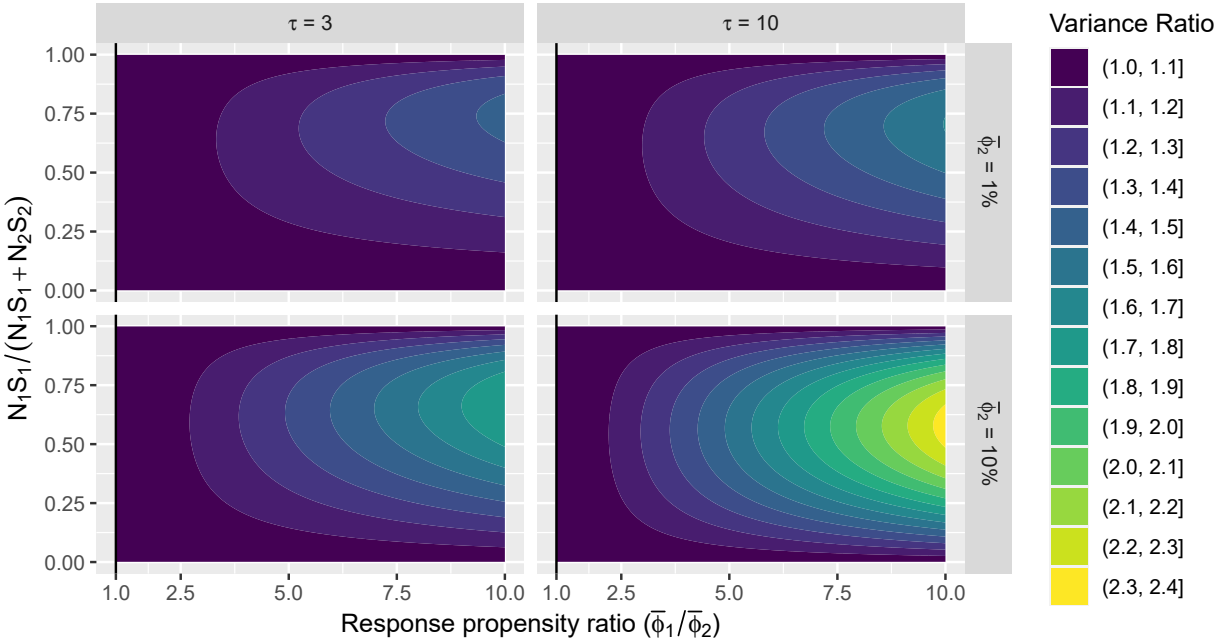
Holding other quantities constant, and assuming that response propensities vary by strata, larger values of τ exacerbate the relative inefficiency of the $\{n_h^{Ninv}\}$ allocation compared to the $\{n_h^{opt}\}$ allocation. Although larger τ increases the expected per-respondent costs for all allocations, the effect is relatively greater on the $\{n_h^{Ninv}\}$ allocation, due to its having the highest nominal expected response rates. Further, under the optimal allocation, the negative effects of larger τ are partially mitigated due to heavier oversampling of the lower response propensity strata, which partially offsets the increase in costs-per-respondent while also reducing the expected design effects. In contrast to the above, increased τ mitigates the relative inefficiency of the $\{n_h^{Nresp}\}$ allocation, relative to that of $\{n_h^{opt}\}$. As τ increases, $\{n_h^{opt}\}$ increasingly resembles $\{n_h^{Nresp}\}$, since costs are increasingly driven by the costs of full interviews. If τ is large enough such that the costs per expected respondent are roughly constant across strata, then $\{n_h^{opt}\}$ and $\{n_h^{Nresp}\}$ will be roughly equivalent, leading to $VR_{Nresp} \approx 1$.

To further illustrate how relative sample efficiency is affected by population characteristics, we again consider a 2-stratum situation, but now allowing for $\tau > 1$ (in contrast to Figure 4.2). Under $H = 2$, (4.18) and (4.19) reduce to $VR_{Ninv} = \frac{(1+\beta\eta)\left(1+\frac{\beta}{\gamma}\right)}{\left(1+\beta\sqrt{\frac{\eta}{\gamma}}\right)^2}$ and $VR_{Nresp} = \frac{\left(1+\frac{\beta\eta}{\gamma}\right)(1+\beta)}{\left(1+\beta\sqrt{\frac{\eta}{\gamma}}\right)^2}$, respectively, where $\eta = \frac{c_1}{c_2}$ is the strata ratio of per-invitee costs and where, as before, $\beta = \frac{N_1 S_1}{N_2 S_2}$ and $\gamma = \frac{\bar{\phi}_1}{\bar{\phi}_2}$. Figures 4.4 and 4.5 display the approximate variance ratios (VRs) of Neyman allocations of invitees and respondents, respectively, to that of the proposed allocation, for various sets of population parameters, and again assuming that $\bar{\phi}_1 \geq \bar{\phi}_2$. As

before, the contour maps denote the resulting VR as a function of the stratum 1 share of aggregate standard deviations (Y axis) and ratio of response propensities (X axis). The plots are analogous to Figure 4.2, except that since $\tau > 1$, we now have that $VR_{Ninv} \neq VR_{Nresp}$, and we also must specify $\bar{\phi}_2$, since the results now depend not only on the relative response propensities but also on their specific levels. We consider four levels for the parameters $(\tau, \bar{\phi}_2)$, with choices of τ varying across columns (3 or 10) and choices of $\bar{\phi}_2$ varying across rows (1% or 10%, which lead to $\bar{\phi}_1 \in [1\%, 10\%]$ or $\bar{\phi}_1 \in [10\%, 100\%]$, depending on the plot).

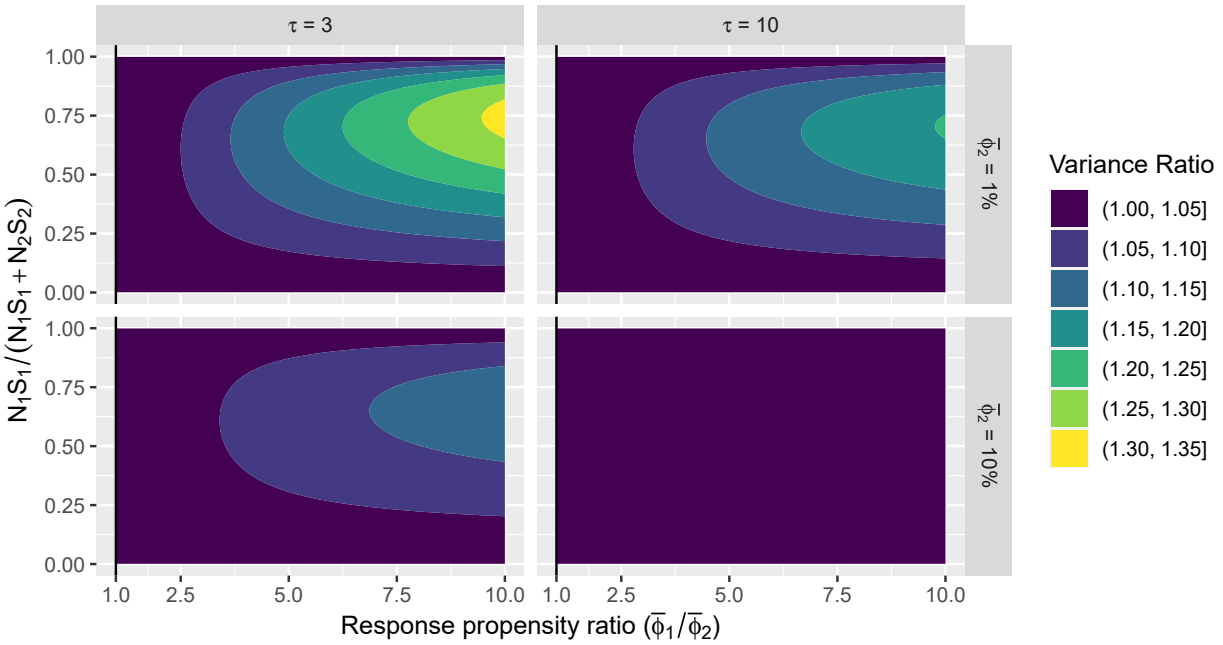
As with Figure 4.2, Figures 4.4 and 4.5 illustrate that the proposed allocation produces larger efficiency gains when the stratum with the larger $N_h S_h$ also has the larger response propensities. Further, we observe that higher levels of τ exacerbate the relative inefficiencies of the $\{n_h^{Ninv}\}$ allocation and mitigate those of the $\{n_h^{Nresp}\}$ allocation, for the reasons discussed two paragraphs above. With respect to the effect of lower response propensities, we observe that lower values for $\bar{\phi}_2$ (in conjunction with constant $\frac{\bar{\phi}_1}{\bar{\phi}_2}$) lead to a greater VR_{Ninv} and reduced VR_{Nresp}, while mitigating differences between the $\{n_h^{Ninv}\}$ and $\{n_h^{Nresp}\}$ allocations. On the latter, decreases in overall response propensities (while holding $\bar{\phi}_1 : \bar{\phi}_2$ constant) result in per-invitee costs that more closely resemble c_{NR}, such that the allocations more closely resemble those of the constant cost-per-invitee scenario, and which leads the VRs to more closely resemble (4.17).

Figure 4.4: Variance ratio: Neyman allocation of invitees versus proposed allocation under H = 2, by assumed population characteristics



Note. Formula assumes $\tau_h = \tau$, $c_{NR_h} = c_{NR}$ for constants τ and c_{NR_h}, $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$, and $1 - \frac{r_h}{N_h} \doteq 1$.

Figure 4.5: Variance ratio: Neyman allocation of respondents versus proposed allocation under H = 2, by assumed population characteristics



Note. Formula assumes $\tau_h = \tau$, $c_{NR_h} = c_{NR}$ for constants τ and c_{NR_h}, $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$, and $1 - \frac{r_h}{N_h} \doteq 1$.

4.4 Example. Post election voting survey of military personnel

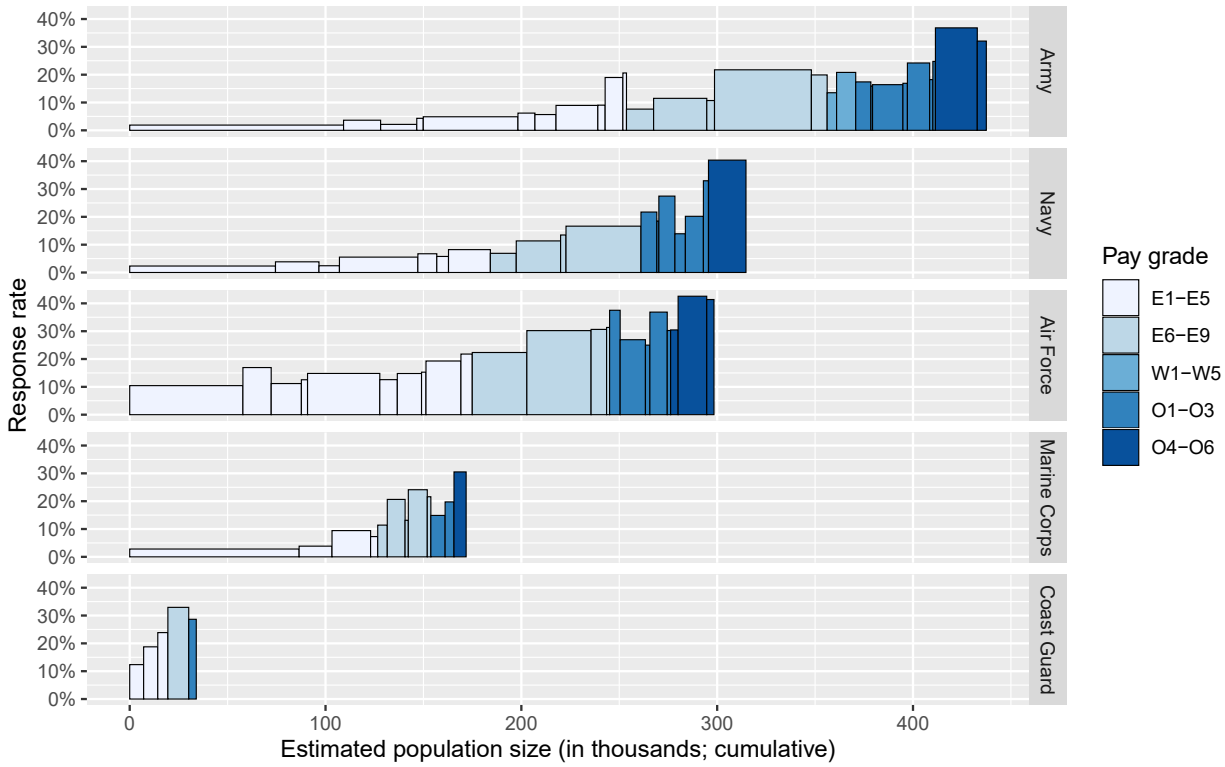
The Federal Voting Assistance Program (FVAP) is an agency within the U.S. Department of Defense (DoD) that assists U.S. citizens in the military and/or who live overseas with voting. As part of this, every two years, it conducts Post-Election Voting Surveys of U.S. Active Duty Military (PEVS-ADM), with a sample that is based on personnel records from the Defense Manpower Data Center (DMDC), and for which public use data are available. FVAP's 2016 survey effort (i.e., 2016 PEVS-ADM) entailed inviting sample members via mail and email to complete a web questionnaire, with up to four mail contacts and eight

email contacts each. The survey had an original sample size of $n = 90,982$, drawn as an STSRS, and incorporated a response rate experiment; to simplify analysis, we exclusively examine the control group ($n = 77,333$) (Federal Voting Assistance Program, 2017a, 2017b). We further omit from our analyses $1,785$ sample members who were identified as ineligible via administrative data prior to fielding; these units were in the July 2016 sampling frame, but were identified via September 2016 personnel records as no longer being on active duty (e.g., due to separations or retirements). Resulting analyses reflect an estimated population size of 1.28 million personnel.

The sample entailed use of 212 strata that were based on a cross-classification of duty location (US or overseas), Service (Army, Navy, Marine Corps, Air Force, or Coast Guard), age group (18–24, 25–29, 30–34, or 35+), pay grade group (E1–E5, E6–E9, W1–W5, O1–O3, or O4–O6), and sex (male or female). The survey does not report strata-specific response rates, but these can be inferred or approximated via the public use data set, which provides design weights (i.e., base weights), disposition codes, and a "poststrata" variable, the latter of which is a collapsed stratum variable that appears to denote original stratum for around 85% of the sample (with coarsened categories for the remaining sample).

Examining the data, we observe an overall response rate of 14% (AAPOR RR2, weighted), but with poststrata-specific rates that mostly range from 1.9% (Army males ages 18–24 with pay grade E1–E5, stationed in the US) to 43% (Air Force males with pay grade O4–O6, stationed in the US). The substantial variability in response rates by subgroup is illustrated in Figure 4.6, which plots the response rates by poststrata (denoted by bordered regions), with widths scaled by population sizes.

Figure 4.6: PEVS-ADM poststrata response rates by population size and selected characteristics



Note. Bordered regions denote distinct poststrata and have widths scaled to estimated population sizes, the latter of which were computed via design weights. Plot does not display poststrata with fewer than 100 invitees, which reflect an estimated 1.8% of the population. Three Navy poststrata do not distinguish between pay grades W1–W5 and O1–O3 and are displayed here as O1–O3, since summary data from FVAP (2017b) imply that they are roughly 92% O1–O3.

To illustrate how the proposed allocation methods differ from alternatives, we consider the constant cost per invitee scenario, as might be applied to the FVAP data. We consider how a sample of $n = 50,000$ invitees might be allocated, under the assumption of constant S_h across strata, as is commonly employed in multi-purpose surveys. We stratify based on the "poststrata" variable since the original stratifier is unavailable. We compare the proposed allocation with Neyman allocations of invitees and respondents. For the proposed

allocation, we consider both the exact version (i.e., incorporating the $\zeta_h(n_h, \bar{\phi}_h)$ term) and the approximate version (i.e., assuming that ζ_h is constant). Screener responses indicate that an estimated 99.3% of the invited population meets screener-based eligibility criteria; therefore, we simplify analysis by treating nonresponse as being a single phase. In doing so, we estimate stratum response propensities as base-weighted proportions of invitees who provide an eligible, complete response to the survey; we include all invited units in the denominator (as opposed to excluding ineligibles, as in AAPOR response rates), since our focus is on capturing the effects of sample loss on costs. Further, we simplify analysis by assuming that the underlying response propensities, $\{\bar{\phi}_h\}$, are equal to their estimates, $\{\hat{\bar{\phi}}_h\}$.

As a first step, and to facilitate interpretation, we temporarily ignore variability in the number of respondents over repeated replications of the survey, by assuming for now that $r_h = \mathrm{E}(r_h) = n_h \bar{\phi}_h$. For each allocation, Table 4.2 provides the unweighted response rate, design effect from weighting ($deff_w$; Kish, 1992), and effective number of respondents, which is approximated as $r_{eff} = \frac{r}{deff_w}$, where $r = \sum_{h=1}^{H} r_h$ (see Table C.1 for the complete allocations). Under these assumptions, the proposed allocation has an effective sample size that is 25% higher than the alternatives. In this example, there is minimal practical difference between the approximate and exact versions of the proposed allocation; this appears to be attributable to the fact that most of the population is in adequately sized strata, such that $\zeta_h(.)$ is very nearly constant.

Table 4.2: Summary of PEVS-ADM allocations under assumption that $r_h = n_h\bar{\phi}_h$

Allocation	Notation	Response rate (%)	Respondents	$deff_w$	Effective respondents
Neyman/invitees	$n_h \propto N_h S_h$	13.6	6,794	2.37	2,865
Neyman/respondents	$n_h \propto \frac{N_h S_h}{\bar{\phi}_h}$	5.7	2,865	1.00	2,865
Proposed/approx.	$n_h \propto \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}$	9.0	4,494	1.26	3,570
Proposed/exact	$n_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{\bar{\phi}_h c_h}}$	9.0	4,496	1.26	3,570

Note: Assumes that S_h is constant across strata and that $n = 50,000$. Assumes constant cost per invitee scenario (i.e., $\tau_h = 1$ and $c_{NR_h} = c_{NR}$).

Table 4.2 above assumes that $r_h = n_h\bar{\phi}_h$, but doing so ignores the effects of the variability in $\{r_h\}$ on the expected variance of our estimator (i.e., [4.2]), as is captured by the $\zeta_h(n_h, \bar{\phi}_h)$ term in (4.8). For a given allocation, the randomness in the $\{r_h\}$ causes the variance to increase by a factor of

$$RelIV = \frac{\text{Var}\left(\hat{\bar{Y}}\right)}{\text{Var}\left(\hat{\bar{Y}} | r_h = n_h\bar{\phi}_h\right)}, \tag{4.21}$$

where the unconditional and conditional variance terms are computed via (4.8) and (4.4), respectively. To quantify this effect, we computed $RelIV - 1$ for the allocations in Table 4.2, which assumed $n = 50,000$; to assess sensitivity to the overall sample size, we also applied the same allocation methods in an analogous fashion toward allocating samples of sizes $n = 25,000$ and $n = 100,000$; in doing so, we assumed constant S_h. Results are displayed in Table 4.3. The middle data column indicates that the randomness in the $\{r_h\}$ could be expected to cut the effective sample size associated with Table 4.2 by roughly 2%–

3%, depending on the allocation, with slightly more severe results for the Neyman allocation of respondents than for the other allocations. The other columns highlight that the specific results are sensitive to the overall sample size when holding the number of strata fixed.

Table 4.3: Percentage increase in variance from randomness in r_h, by PEVS-ADM allocation method and total invitees

Allocation	$n = 25$k	$n = 50$k	$n = 100$k
Neyman/invitees	4.5	2.1	1.0
Neyman/respondents	5.9	3.0	1.4
Proposed/approx.	3.6	1.7	0.8
Proposed/exact	3.4	1.6	0.8

The previous analyses assumed use of a multi-purpose survey, but some surveys optimize the design for a single survey estimator. Therefore, we next considered the performance of the allocations, had they been applied toward a specific variable. For a given dichotomous study variable, we computed each of the allocations considered in Table 4.2, using similar procedures as before, the only difference being that we now incorporated the $\{S_h\}$ into the allocations. We repeated this process for four different survey variables, summarized in Table 4.4, and which are all based on self-reports. For allocation purposes, we assumed that S_h was equal to $\hat{S}_h = \sqrt{\frac{N_h}{N_h - 1}\hat{P}_h(1 - \hat{P}_h)}$, where $\hat{P}_h$ was the base-weighted proportion for some given variable. In computing $\hat{P}_h$, we simplified analysis by excluded item-missing data from the denominator; this amounted to fewer than 0.5% of responses for each variable. After computing the allocations, we then computed unconditional variances via (4.8).

Table 4.4: List of 2016 PEVS-ADM measures considered for variable-specific allocations

Survey Variable	Description
Registered voter	Registered to vote in the United States for the November 2016 election
UOCAVA ADM	Respondent lives 50 or more miles away from their place of voter registration (if registered) or from their legal voting residence (if not registered)
Voted	Respondent indicates having "definitely voted" in the November 2016 election
FVAP use	Respondent used one or more of FVAP's products or services for voting assistance for the November 2016 election

Table 4.5 displays the resulting variances. For each variable considered, the approximate and exact versions of the proposed allocation yielded equivalent variances (after rounding). The Neyman-type allocations had variances that were 21%–26% higher than those of the (exact) proposed allocations.

Table 4.5: 100,000 * Variance of PEVS-ADM estimator, for a given allocation method

Allocation method	Reg. voter	UOCAVA ADM	Voted	FVAP use
Neyman/invitees	1.60	1.50	1.61	1.41
Neyman/respondents	1.63	1.51	1.62	1.42
Proposed/approx.	1.32	1.20	1.29	1.12
Proposed/exact	1.32	1.20	1.29	1.12

Note: Assumes $n = 50,000$, constant cost per invitee scenario (i.e., $\tau_h = 1$ and $c_{NR_h} = c_{NR}$), and that $S_h = \hat{S}_h$.

4.5 Discussion

4.5.1 Summary of results

Existing mathematical theory on allocation for single-stage stratified sample designs typically does not distinguish between the invited sample and the responding sample. Aside from leading to ambiguity about how to best account for nonresponse at the design stage, the existing framework can lead to inefficient allocations, by ignoring the effects of nonresponse on cost efficiency. It is well known that ignoring the response rates at the sample design stage can lead to excessive design effects. However, common practice, which inflates the sample sizes by the inverse of the anticipated stratum-specific response propensities, overcorrects for this issue, leading to excessive costs per interview.

We provided an optimal allocation for the poststratified estimator under nonresponse, and which strikes a better balance between design effects and costs per interview than the existing practices above. Our allocation minimizes either the unconditional variance of our estimator or the expected costs, holding the other constant. Under some simplifying assumptions, the variance considered is analogous to the standard expression under complete response, except that it not only reflects the expected sample loss but also captures the effects of the variability in the responding sample sizes across repeated iterations of the survey. A key feature of our allocation is that it incorporates a simple cost model that assumes that costs within strata can be decomposed into costs associated with respondents and with nonrespondents. This cost need not be a literal monetary cost—for example, a constraint on the total sample size

can be accommodated by specifying the constant cost per invitee scenario.

We examined the allocations' relative efficiency by considering the approximate ratio of the variance under the proposed design to that under the alternatives (i.e., Neyman allocations of respondents or invitees). Our method produces the greatest gains when response propensities vary greatly across strata; gains are greater when the strata with higher response propensities also account for more of the population variability on the outcome measure. Under a constant cost per invitee, the existing alternatives (i.e., Neyman allocations of respondents or invitees) are equally inefficient. In scenarios where survey costs are driven by the costs associated with obtaining full interviews (rather than the costs of inviting a large number of nonrespondents), then this partially mitigates the inefficiency of the Neyman allocation of respondents, while exacerbating that of the Neyman allocation of invitees.

We further illustrated advantages of our proposed method through an application to a survey of U.S. military personnel, in which subgroup response rates generally ranged from 1.9% to 43%. Under an equal cost-per-invitee scenario, the proposed method would have increased the effective sample size by roughly 25%, ignoring the variability in the $\{r_h\}$, the latter of which would not have meaningfully changed results for the given design. This calculation assumed constant strata variances, but similar results were obtained when re-calculating the allocations for four different sets of strata variances, specific to key survey variables.

4.5.2 Limitations and future directions

We assumed that the response propensities within stratum were constant (i.e., $\phi_{hi} = \bar{\phi}_h$ for $i \in U_h$), which implies that we have assumed nonrespondents' data to be *missing at random* (MAR), as defined by Little and Rubin (2019). If the data were *missing not at random* (MNAR, as per Little and Rubin), meaning that there were unobserved confounders of nonresponse even after conditioning on characteristics that are observable for the invited sample, then it would not be clear how to obtain valid inferences, let alone how to best allocate a sample. This limitation is not unique to our methods—standard theory on optimal allocations for STSRS ignores nonresponse entirely, and nonresponse weighting adjustments typically assume a MAR mechanism (Brick, 2013). Note that a MAR assumption does not necessarily imply that $\phi_{hi} = \bar{\phi}_h$, considering there could be additional available auxiliary variables (or levels thereof), beyond those used in stratification, that are associated with residual differences in nonresponse.

We assumed use of a poststratified estimator under nonresponse, and where the poststrata were assumed to be equivalent to the original design strata. If a different estimator is ultimately used, the proposed design will no longer be optimal. Other estimators are commonly used in practice—for example, survey weights may incorporate nonresponse adjustments and/or calibration adjustments, sometimes applied in multiple steps, using a form of calibration other than poststratification (e.g., raking or linear calibration), and/or incorporating adjustment variables that are not limited to the original stratification variables; see Valliant et al. (2018, Chapters 13–14) or Brick (2013) for detailed reviews of weighting. Our

assumed poststratified estimator is but one special case of a weighted estimator that could be constructed via such procedures. This mismatch is not specific to our proposed methods, considering that classic theory on optimal allocation also assumes a poststratified estimator. If survey designers anticipate that weights will vary within design strata, they may consider whether to account for this when computing allocations. However, doing so may not always be straightforward. For example, previous survey data used to estimate design parameters may have employed different strata or weighting procedures. Accurately anticipating design variances may be further complicated when nonparametric methods are used to adjust for nonresponse (e.g., Buskirk & Kolenikov, 2015; Fay & Riddles, 2017; Rizzo, Kalton, & Brick, 1996). Additional research is needed to identify how to best allocate samples in such scenarios.

We assumed that at least one respondent will be obtained in each stratum, that is, $r_h \geq 1$, since otherwise our estimator and its variance would be undefined. Although there may be some real situations where some strata have no or very few respondents, such situations can be mitigated by the sample designer through use of adequately sized strata at the design stage, or by modifying the estimator after the data are collected (e.g., collapsing adjustment cells for applying a poststratification adjustment to the base weights). Considering that incorporating such situations would add some complexity, without necessarily leading to any meaningful practical gain, we instead chose to approximate the distributions for the $\{r_h\}$ as standard binomial where support for 0 has been removed.

We assumed 100% eligibility, meaning that units in the sampling frame were assumed to be members of the target population. This may sometimes be unrealistic. In surveys of low-

incidence populations, individual eligibility might not be known from the frame, although it may be possible to construct sampling strata with substantially different eligibility rates—see, for example, U.S. Census Bureau (2022), Brick et al. (2016), and Chen and Kalton (2015). In such cases, the strata eligibility rates can meaningfully affect the optimal allocations, since it will be more efficient to interview units in strata with higher concentrations of the target group, while still being necessary to obtain some interviews from remaining strata to avoid bias from undercoverage. Even in surveys with high eligibility rates, there may still be some loss of the sample due to screening out ineligibles. Future efforts could address sample allocation under nonresponse and with incomplete eligibility. For example, §4.2 could be adapted by treating eligibility status in a manner analogous to domain estimation (e.g., see Cochran, §2.13), with inferences via the domain mean, to be computed via a combined ratio estimate of the domain total to the domain size. The resulting conditional variance could then be derived in a manner analogous to Cochran's proof of his Theorem 6.5 (p. 166, Cochran), but under incomplete response. Under certain simplifying assumptions about data missingness, the resulting conditional variance would be analogous to (4.4), except that in place of S_h^2, the strata variance would be of a residual term underlying the estimator's implied model. Incomplete eligibility may also necessitate changes to §4.2's cost model and underlying response structure as to account for various forms of sample loss, some of which may be depend on screener-based eligibility.

We simplified analysis by assuming that the response propensities and strata variances are known exactly; in practice, these survey design parameters are commonly estimated from previous data. Clark (2013) provided examples of authors who have commented on how

errors in estimating S_h^2 could lead to inefficiencies (e.g., P. Smith et al., 2003; Lohr, 2009, p. 90) or who had considered its theoretical impact (e.g., Cochran, 1977, §5.1; Evans, 1951). Clark proposed smoothing strata variances that follow a hierarchical structure via a composite of direct estimates computed at various levels of aggregation; in Chapters 2 and 3, we proposed using a Bayesian formulation to accommodate uncertainty. This work focused on sampling variances, in isolation. Szeitl and Rudas (2022) considered the effect of misspecified response rates in their comparison of Neyman allocations of invitees and respondents, but assumed the number of respondents to be fixed, thereby ignoring the often-severe effects of response rates on costs. Further theoretical and empirical research could be conducted to better understand the cumulative effect of such errors, and on how to best account for them at the design stage. Although we did not consider such issues, it appears likely that our proposed allocation may be somewhat less sensitive to misspecified response propensities than Neyman allocation of respondents, especially for $\tau = 1$, considering the allocations' denominators.

We assumed that the cost structure is known and that the marginal unit costs associated with nonrespondents can be distinguished from those of respondents, with a motivating example that assumed self-administration. As with any sample optimization problem that considers costs, errors in the cost model may reduce sample efficiency, on top of potentially leading either to cost overruns or to underutilization of the available budget. In discussing survey costs and errors, Groves (1989, p. 79) lamented that "survey researchers have given much less attention to survey cost models than to survey errors." Over 30 years later, Olson, Wagner, and Anderson (2020) argued that there still remained a lack of systematic information

about costs, partly due to a lack of common language about costs. However, some of the challenges with costs discussed by Groves (e.g., discontinuities and nonlinearities) and/or by Olson et al. (e.g., challenges accurately disaggregating labor costs) are particularly salient in an interviewer-administered context, and may be somewhat mitigated for self-administered surveys. Likewise, Groves and Olson et al. acknowledged that linking costs to errors is more straightforward in the context of cost models for sample optimization than it is for non-sampling errors. For statisticians applying (4.12) to specific self-administered surveys, we suspect that many of the main cost components per-invitee (e.g., postage, printing, prepaid incentives) or per-respondent (e.g., postpaid incentives) may already be available (e.g., from budgeting departments) and with adequate accuracy for sample design purposes. In some instances, it might be possible to further refine these estimates via use of paradata (e.g., microdata on contact attempts and/or incentives delivery) or by considering how marginal labor costs associated with data entry affect τ. Also, different mode assumptions may require changes to the cost model; for example, if a large portion of resources is dedicated toward following up with non-respondents (e.g., as in a sequential mixed-mode design, progressing from less expensive to more expensive modes), then it may be worth dividing respondents into subgroups based on the timing of their responses.

We also assumed that the strata unit costs for respondents (c_{R_h}) and nonrespondents (c_{NR_h}) are constant. In practice, this assumption will not always hold—for example, even controlling for strata and response status, units might vary in terms of the number and/or mode(s) of contact attempts; respondents may further differ in response mode. A different formulation might have treated unit costs within strata, conditional on sample membership and response

status, as being random variables. Considering that the costs entered the optimization step by way of their expected strata totals (i.e., expected value of [4.10]), then using the latter formulation would not have changed the allocations, so long as the strata unit costs by response status assumed for allocation purposes, say, $\{\tilde{c}_{R_h}\}$ and $\{\tilde{c}_{NR_h}\}$, reflect their underlying averages, and assuming that the response propensities are constant within strata (as was used in [4.11]) and that the strata are left unchanged. On the latter, if a more detailed stratification allows for a better representation of cost, that could plausibly lead, in some cases, to improved efficiency (i.e., lower cost per effective interview), although a more detailed stratification also implies more variability in the sampling weights and in the responding strata sample sizes, which could plausibly lead to a net negative effect in other cases.

We examined sample allocation under nonresponse in a single-frame, single-stage design. Although Lohr and Brick (2014) provided an analogous and thorough treatment for telephone surveys with nonresponse, their analysis was specific to a dual-frame context and involved different types of design decisions (e.g., screener vs. overlap design; choice of compositing factor; allocation to frames, rather than to strata). Lohr and Brick assumed cost models that directly incorporated the effects of nonresponse, but noted that "such a direct relationship between costs and response rates is not discussed often." Future work could consider incorporating the effect of nonresponse on sample efficiency in additional contexts, such as multi-stage designs or multi-mode designs, as might be accommodated via changes to the assumed variance structure and/or cost structure.

Nonresponse not only affects the cost efficiency of data collection, but also leads to uncertain

precision and can lead to uncertain costs, especially if response rates are low and there are substantial marginal costs associated with response (e.g., large postpaid incentives). We handled uncertainty in the responding sample size by averaging the costs and conditional variance over an assumed distribution for the number of respondents and then aiming to minimize the expected costs or unconditional variance, holding the other constant. However, if response rates are low, this may lead to unacceptably high risks of cost overruns or inadequate precision, especially for domains. Future work could consider alternative formulations that are aimed at mitigating such risks. This could entail treating costs and realized variances as random quantities and imposing distributional assumptions, such as our assumption that the number of respondents in stratum h is conditionally truncated binomial (given the underlying response propensities), and which could include additional assumptions on other survey design parameters (e.g., analyzing the stratum response propensities via a Bayesian beta-binomial model). Treating costs and realized variances as random allows for computing relevant probabilities, such as the probability that a given allocation yields total costs within the maximum budget, or that some specified precision criteria for the realized sample are met. Under a mathematical programming approach, constraints could be imposed on these probabilities.

Chapter 5: Summary and future directions

Neyman's seminal paper and subsequent developments of the next two decades transformed the practice of survey sampling and continue to provide the underpinnings of today's probability samples. Although still hugely relevant, the underlying assumptions are not always met. Even putting aside issues of measurement and coverage error, a purely design-based approach to inferences is generally not possible due to nonresponse, which has posed increasing challenges. Existing design-based sample optimization theory assumes complete response, resulting in ambiguous terminology and a lack of clarity on how to best account for nonresponse at the design stage. Design-based optimal allocation theory also assumes that the survey design parameters are exactly known, even though they are generally not known. Therefore, this thesis revisited these underlying assumptions, and developed now ways to optimize STSRS designs. In Chapters 2 and 3, we developed Bayesian approaches for optimal allocation for heteroscedastic populations; in Chapter 4, we shifted focus toward addressing nonresponse under a design-based (model-assisted) framework.

The remainder of this chapter is as follows. §5.1 summarizes and discusses Chapters 2 and 3. §5.2 summarizes Chapter 4.

5.1 Bayesian design for heteroscedastic data

5.1.1 The Bayesian approach to design

As we saw in Chapters 2 and 3, design-based theory is also not the only way to identify good allocations. Bayesian optimal experimental design (Lindley, 1972) is a flexible and general approach that accommodates uncertainty in design parameters and can be applied toward sample optimization problems. In a finite population context, Bayesian optimal design can be used in conjunction with the Bayesian framework for finite population inference (Ericson, 1969a). Although there was an early Bayesian sample optimization literature (e.g., Draper & Guttman, 1968b; Rao & Ghangurde, 1972), reviewed in Chapter 2, the Bayesian approach to sample design has been largely overlooked. Although it is unclear exactly why this is, the Bayesian literature on optimal sample allocation, largely published from the mid-1960s through the early 1980s, arose at a time of inferential controversy in survey statistics (model-based versus design-based) and in other areas of statistics (e.g., frequentist versus Bayesian). Considering that the Bayesian literature on sample optimization predates modern computing and leaves some important problems unsolved, we revisited this approach. We further noted that the Bayesian formulation has been used extensively in other areas of statistics, the earlier applications generally being analytical (e.g., Chaloner & Verdinelli, 1995), with active recent attention toward computation (Ryan et al., 2016), as discussed in Chapter 3.

In reviewing the early Bayesian design literature, we observed that existing work does not appear to adequately address optimal design for establishment populations, whose distribu-

tions may be continuous and heavily skewed (e.g., Andridge & Thompson, 2015). Under a model-assisted approach, the level of heteroscedasticity (e.g., Henry & Valliant, 2009) can meaningfully affect the optimal allocation. However, the models employed in the early Bayesian design literature do not appear to adequately address such populations. For example, Draper and Guttman (1968b) explored Bayesian allocation for a continuous outcome variable, but assumed a normal likelihood with fixed strata means and variances, which does not accommodate heteroscedasticity. Rao and Ghangurde (1972) considered optimal allocation for categorical data via a Dirichlet-multinomial model, which may be appropriate in other contexts, but not necessarily for skewed, continuous data. Therefore, in Chapters 2 and 3, we incorporated the coefficient of heteroscedasticity into our assumed population mean.

The Bayesian optimal design, as per Lindley (1972), selects the experiment and estimate for minimizing the expected loss with respect to the parameter space and future data. Under squared error loss (SEL), this is equivalent to minimizing the expected posterior variance for the parameter of interest, as we saw from the *terminal analysis* (following Lindley) in §2.2.3. The earlier Bayesian sample design work did not always follow Lindley's framework, but in some cases implicitly used a squared error loss function. For instance, Draper and Guttman (1968b) and Rao and Ghangurde (1972) did not formally conduct a terminal analysis, but considered the objective of minimizing the expected posterior variance (for the superpopulation mean and finite population mean, respectively).

5.1.2 Bayesian approach for heteroscedastic data

In Chapters 2 and 3, 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 employed Lindley's framework and SEL for the finite population mean. Under this formulation, the optimal allocation minimizes the expected posterior variance for the finite population mean.

In Chapter 2, we assumed known b, facilitating analytic results while accommodating uncertainty in other parameters. We assumed a non-informative prior on the strata slopes and variance proportionality terms 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 under our population model, which allowed for the optimal allocation to be identified through mathematical programming. Performance of the proposed strategy 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; we also repeated our simulation in an application toward analyzing revenues of public charities, using publicly available IRS Form 990 data. The proposed strategy provided substantial empirical gains in comparison to a design-based approach, while also doing as well or better than the main model-assisted approach considered, with the extent of gains depending on the population. For certain types of populations, the proposed allocation also led to more stable sample allocations, with respect to repeated pilot samples. We also evaluated the effects of misspecified coefficient of

heteroscedasticity on the proposed allocation at the sample design stage, finding that some populations were very sensitive to misspecification of this coefficient, whereas others were minimally affected.

In §2.7.2, we discussed limitations of our research and potential future directions. One limitation arose from our assumption that the coefficient of heteroscedasticity was known, which might not always be realistic; we addressed this issue in Chapter 3. Other areas of future work entail considering alternative population structures, sample designs, loss functions, and/or ways of expressing prior knowledge. Although our simulation considered a large number of populations and several alternative allocations, we had to fix some conditions, as to maintain a manageable scope. For example, our comparisons to a model-assisted approach focused on ratio estimation, and the alternative allocations were all STSRS designs; future empirical work could perhaps consider GREG estimation and/or PPS designs. Although our assumed model employed weaker assumptions than some previous Bayesian work, future work might consider different error terms (e.g., lognormal), different conditional mean functions, and/or the potential implications of outliers for sample allocation.

5.1.3 Computational approach for Bayesian allocation

Chapter 3 provided a computational approach to fully Bayesian optimal sample design in establishment surveys with an unknown coefficient of heteroscedasticity. The Bayesian optimal design maximizes the expected utility over the space of possible designs with respect to the future data and model parameters. This can pose computational challenges, by requiring

Bayesian analysis on numerous future data sets to evaluate the expected loss for a single design, a process that must be repeated many times during optimization. Chapter 2 avoided this issue by assuming that the coefficient of heteroscedasticity was known, which facilitated analytical results. In contrast, in Chapter 3, we specified a (non-informative) prior on b and derived the resulting posterior distribution for superpopulation parameters, which resulted in an analytically intractable posterior distribution. Thus, we developed a computational approach for sample optimization.

The computational approach entailed Monte Carlo (MC) sampling for estimating the objective function for a given allocation, in conjunction with simultaneous perturbation stochastic approximation (SPSA), a stochastic optimization method that accommodates noisy loss functions. The MC estimator incorporated distributional information in conjunction with an emulation function to substantially reduce computation; we then used two-stage sampling theory to obtain an approximate variance estimator to quantify the MC error of the estimated loss. In optimization, we tailored SPSA toward our application, as to handle constraints and through a parameterization that guarantees full use of the sample. As an alternative to SPSA, we also considered Nelder–Mead (NM). We also considered pseudo-Bayesian (PB) alternatives, such as employing the Chapter 2 allocation under the assumption that the coefficient of heteroscedasticity equals some pilot-based estimate.

We compared each variant of the fully Bayesian approach (i.e., B-SPSA and B-NM) to pseudo-Bayesian, design-based, and model-assisted alternatives through simulation to two continuous, skewed populations with differing levels of heteroscedasticity, and in an application to the revenues of nonprofits, the latter of which used IRS Form 990 data. Empirically,

the proposed methods (i.e., B-SPSA) appeared to be approximately optimal, performing as well or better than the other methods, while having clearer theoretical justification. The pseudo-Bayesian (PB) allocations also turned out to be approximately optimal, but they were not derived from first principles, and their performance is not guaranteed for other populations and pilots. The B-NM approach performed well in the simulation, which had limited dimensionality, but did not scale as well to the population, which had larger dimensionality.

We provided some suggestions, discussed limitations, and suggested areas for future work. We provided implementation suggestions for SPSA, considering that its performance is specific to the choice of gain sequences and to the application. Although the proposed methods assumed constant per-unit costs, it is simple to modify the methods to allow for variable costs, and we explained how to do so. Although sample allocation problems typically focus on optimization for the finite population mean, the Bayesian formulation could be applied toward other inferential quantities. However, doing so can lead to potentially considerable computational costs, which we had reduced through use of expressions for the conditional mean and variance in conjunction with a quick and accurate approximation for a needed population-based quantity. Therefore, an area of future work could entail developing scalable methods of Bayesian optimization for other types of inferential objectives.

5.2 Sample allocation under anticipated nonresponse

In Chapter 4, we changed our focus toward addressing the effect of nonresponse on cost efficiency at the sample allocation stage. Response rates have declined dramatically in recent

years. However, existing design-based theory on optimal allocation is formulated under complete response. Ignoring the effects of nonresponse leads to excessive design effects, but the common practice of adjusting the allocations by the inverse of the strata response rates overcompensates for this issue, leading to excessive costs per interview. Although Lohr and Brick (2014) considered the effects of nonresponse on cost efficiency in allocating sample to frames for DFRDD telephone surveys, there has been no analogous and general treatment for allocation to sampling strata in STSRS designs. We argued that this lack of theory is particularly an issue considering recent transitions toward self-administered and mixed mode surveys, for which it can sometimes be possible to partition the population into strata with substantially different response rates.

We expanded the standard formulation for optimal allocation (e.g., Neyman, 1934; Cochran, 1977, §5.5) to account for nonresponse. We distinguished between the original sample of invitees and the responding sample, with strata h having sample sizes of n_h invitees and r_h respondents, respectively. Some key assumptions were that the response propensities were constant within strata (i.e., $\bar{\phi}_h = \phi_{hi}$), that it was possible to determine the marginal costs associated with respondents (relative to those of nonrespondents), and that estimation would employ a poststratified estimator under nonresponse (using the original sampling strata as poststrata). The proposed allocation of invitees is

$$n_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{\bar{\phi}_h c_h}}, \tag{5.1}$$

where $c_h = \frac{\mathrm{E}(C_h)}{n_h} = c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)$ is the expected cost per invitee, c_{NR_h} is the unit

cost of nonrespondents, $\tau_h = \frac{c_{R_h}}{c_{NR_h}}$ is the relative cost of respondents to that of nonrespondents in stratum h, and $\zeta_h(n_h, \bar{\phi}_h) = \mathrm{E}(r_h)\mathrm{E}\left(\frac{1}{r_h}\right)$ captures the effect of variability in the number of respondents for a given allocation. On the latter, we suggested modeling $\zeta_h(.)$ as standard binomial but where support for 0 is removed, so that the poststratified estimator (under nonresponse) and its variance are defined. We outlined how to compute this allocation exactly considering its recursive definition.

We then provided theoretical and empirical comparisons between the proposed allocation, written as $\{n_h^{opt}\}$, and Neyman allocations of invitees or respondents, written as $\{n_h^{Ninv}\}$ and $\{n_h^{Nresp}\}$, respectively. The theoretical comparisons explored how the cost structure and population characteristics affected the allocations' relative efficiency. Under some simplifying assumptions (i.e., negligible finite population correction; negligible effect of the $\zeta_h(.)$ term), our allocation performs best (in terms of cost per effective interview), followed by $\{n_h^{Nresp}\}$, followed by $\{n_h^{Ninv}\}$, and assuming $\tau \geq 1$. The inefficiency of $\{n_h^{Ninv}\}$ arises from sampling more units than is optimal in strata with larger response propensities, whereas the inefficiency of $\{n_h^{Nresp}\}$ arises from excessive costs per interview. The proposed allocation provides a better balance between design effects and cost per completed interview compared with these existing practices. At $\tau = 1$ (i.e., constant cost per invitee), the Neyman allocations of invitees and respondents are equally inefficient. Increasing the value of τ increases the cost per effective interview for all allocations, but exacerbates the relative inefficiency of $\{n_h^{Ninv}\}$ (compared to the optimum), while mitigating that of $\{n_h^{Nresp}\}$, the latter of which approaches the optimum as τ becomes adequately large. In an application to a self-administered survey of military personnel, under the constant cost per invitee scenario, the proposed methods

resulted in a 25% increase in effective sample size compared with the alternatives.

Our proposed allocation can be specialized toward the standard design-based optimal allocation (i.e., $n_h \propto \frac{N_h S_h}{\sqrt{c_h}}$) by assuming 100% response propensities, and can further be specialized toward Neyman allocation by assuming constant costs across strata. Likewise, our assumed strategy (i.e., STSRS plus poststratified estimator under nonresponse) specializes, under complete response, to the STSRS-HT strategy. Thus, some of the underlying limitations (e.g., assumption of 100% eligibility; known strata variances and cost model) also apply to standard design-based theory. Likewise, our assumption that nonrespondents' data are missing at random underlies standard weighting practices, and our assumption that response rates are constant within strata and known upfront underlie the standard practice of inflating the sample by the inverse of the anticipated response rates. In contrast with common practice, we did not ignore the effect of variability in the $\{r_h\}$ on the unconditional variance. Although our approach, at the invitation stage, oversamples units in strata with low response propensities, it does so to a lesser extent than a Neyman allocation of expected respondents, and therefore may also be somewhat less susceptible to misspecification of response propensities.

We highlighted the importance of considering the effect of nonresponse on cost efficiency when allocating single-stage stratified random samples. Future research is needed to address limitations with the current approach and to extend consider the effects of nonresponse on cost efficiency to additional designs or contexts. One avenue for doing so would be to modify the underlying variance and/or cost structures to encompass different scenarios. For example, we suggested how the conditional variance could be modified straightforwardly to accommodate incomplete eligibility by treating eligibility in a manner analogous to domain

estimation. Another important area to consider is how to best accommodate the effects of nonresponse on cost efficiency in a multi-stage context, perhaps especially for sequential mixed-mode designs; this will require additional research on costs, which Olson et al. (2020) have cogently argued for.

Appendix A: Chapter 2 Appendices

A.1 Additional literature review

A.1.1 Detailed summaries of early Bayesian optimal sample design literature

A.1.1.1 Draper & Guttman (1968)

Draper and Guttman (1968b) derive an optimal allocation for two phases of STSRS sampling, with fixed strata $h = 1, ..., H$, in which the data are distributed as $x_{hi}, y_{hj}|(\mu_h, \sigma_h^2) \overset{ind}{\sim} N(\mu_h, \sigma_h^2)$, where $\{x_{hi}, y_{hj} : h = 1, ..., H; i = 1, ..., m_h; j = 1, ..., n_h\}$ refer to the $m = \sum_h m_h$ pilot observations (which have already been collected) and $n = \sum_h n_h$ main study observations, respectively, and with independent, locally uniform prior distributions on (μ_h, σ_h^2). They derive an optimal allocation for the main study, wherein they aim to minimize the expected value of the second-phase posterior variance of the superpopulation quantity $\mu = \sum_{h=1}^{H} W_h \mu_h$, where $\left\{W_h = \frac{N_h}{N} : h = 1, ..., H\right\}$ are the known proportions of the population in the strata. This optimization is conducted subject to fixed total costs, $C = \sum_{h=1}^{H} c_h(m_h + n_h)$, where $\alpha C = \sum_{h=1}^{H} c_h m_h$ is the total first-phase cost and $(1 - \alpha)C = \sum_{h=1}^{H} c_h n_h$ is the total

second-phase cost, with per-unit costs of c_h in stratum h. In doing so, they first compute a first-phase posterior distribution, which they use as the prior distribution for the second-phase analysis, after which they compute the joint posterior of the $\{\mu_h, \sigma_h^2\}$, then integrate out the $\{\sigma_h^2\}$, in order to obtain the marginal posterior of the $\{\mu_h\}$. From this posterior distribution, they obtain the posterior variance of $\text{Var}\left(\mu_h | \underline{\text{x}}_{hi}, \underline{\text{y}}_{hj}\right) = \frac{SS_h}{m_h+n_h-3}\frac{1}{m_h+n_h}$, wherein SS_h is a sum of square–type quantity for stratum h, based on both phases' data, and $\underline{\text{x}}_{hi}$ and $\underline{\text{y}}_{hj}$ are the vectors of pilot and main study observations in stratum h. Since the second phase data have not yet been observed, they conduct a preposterior analysis wherein they substitute $\underset{\underline{\text{y}}_{hi}|\underline{\text{x}}_{hj}}{\text{E}}(SS_h)$ in place of SS_h in the above formula, and solve for the optimal allocation, which gives a working solution that is approximately equivalent to Neyman allocation, but which may need to be modified for some strata in which it would imply a negative second-phase allocation.[1] Given that they plan to use both phases' data in their ultimate inference about μ, this working allocation may imply negative values of n_h, which can be interpreted as indicating that for some strata, more than the optimal number of units across both phases have already been selected in the pilot. They work around this complication by providing an iterative solution, in which the iterative step involves setting the second-phase allocations for such strata to zero, and recomputing the remaining allocations; with this step to be applied repeatedly until there are no negative sample allocations. In §3 of their paper, they conduct a similar analysis, but for a scenario wherein the pilot data are only to be used for determining the main study sample allocation, rather than also being used for the ultimate inference about μ; without the added complication of avoiding negative allocations,

[1] Specifically, they find an allocation across both phases of $(n_h + m_h) \propto \frac{W_h u_h}{\sqrt{c_h}}$, wherein $u_h^2 = (m_h - 3)^{-1}\sum_{i=1}^{m_h}(x_{hi} - \bar{x}_h)^2$, assuming that this allocation is possible (i.e., that there have not already been too many observations for any strata in the pilot).

this latter analysis leads to an allocation that is approximately equal to Neyman allocation. In later sections of their paper, they provide numerical examples and discussions of results, and obtain results for the scenario where the $\{W_h\}$ are unknown.

Thus, in sum, Draper & Guttman's methods of analysis entail specifying a model and prior distribution, obtaining the second-phase posterior distribution, conducting a preposterior analysis that involves averaging the posterior variance over the predictive distributions of the unknown quantities, and then using this expected posterior variance to obtain an optimal sample allocation given the pilot data; they also provide alternative solutions, depending on whether the pilot data will be combined with the main study in obtaining their ultimate inferences. Note that although they do not explicitly refer to exchangeability, they do assume the use of a common model within each stratum. Their methods have parallels to the methods I will use later for determining the main study optimal allocation in the presence of uncertain information, although they use a different model and focus on the superpopulation parameters. Further, I will assume that the pilot data are not used in the ultimate inferences, although this decision does not have major implications for the math involved.

Given that their allocation is an approximately Neyman allocation for the scenario wherein the pilot data are not used for the ultimate inferences, this can be viewed as providing a Bayesian justification for Neyman allocation, but with an explicitly specified model. However, a main drawback of their allocation is that their model makes possibly unrealistic assumptions about the population distribution. Specifically, they assume a normal likelihood for the stratum observations, with constant variance within a given stratum. There are several scenarios under which this may be unrealistic.

A.1.1.2 Rao & Ghangurde (1972)

Rao and Ghangurde (1972) use a similar approach to Draper and Guttman (1968b) in obtaining an optimal sample allocation for inference in the context of categorical, rather than continuous, data, and with a focus on estimating finite population parameters, rather than the superpopulation parameters. For an STSRS design with known strata sizes, they specify a population model, conduct a preposterior analysis that entails averaging the posterior variance over the predictive distribution of the unknown quantities, and then determine the optimal strata sample sizes to minimize this expected posterior variance. They also provide results for unknown strata sizes, double sampling for nonresponse, and two-stage sampling.

For their first analysis, they assume that each stratum has a population of N_h units and a simple random sample drawn without replacement of size n_h, for a total population of size $N = \sum_{h=1}^{H} N_h$ and sample size $n = \sum_{h=1}^{H} n_h$. They further assume that stratum h has a finite set of scale-points $Y'_{ht} = (t = 1, ..., T_h)$, and that inference is about the finite population mean $\bar{Y} = \frac{1}{N} \sum_{h=1}^{H} \sum_{t=1}^{T_h} N_{ht} Y'_{ht}$, where N_{ht} is the number of units in stratum h having measurement t. Assuming exchangeability within strata and independence across strata, the likelihood for the N_{ht} is:

$$L(\mathbf{N}) = \prod_{h=1}^{H} \frac{\prod_{t=1}^{T_h} \binom{N_{ht}}{n_{ht}}}{\binom{N_h}{n_h}} \tag{A.1}$$

where $\mathbf{N}_h = (N_{h1}, ..., N_{hT_h})^T$ and $\mathbf{n}_h = (n_{h1}, ..., n_{hT_h})^T$ are the vectors of N_{ht}'s and n_{ht}'s in stratum h and $\mathbf{N} = (\mathbf{N}_1, ..., \mathbf{N}_H)$ and $\mathbf{n} = (\mathbf{n}_1, ..., \mathbf{n}_H)$ are the corresponding matrices for the

full population and full sample.

Rao and Ghangurde then assume a compound-multinomial prior on the $\mathbf{N}_h$ for stratum h, draw upon results from Hoadley (1969) in determining the posterior distribution of the nonsampled units, $\mathbf{N}-\mathbf{n}$, and then draw upon results from Hartley and Rao (1968) in determining the expected posterior variance of $\bar{Y}$. Specifically, they assume that there are superpopulation parameters of $\{p_{ht} : h = 1, ..., H; t = 1, ..., T_h\}$ such that $p_{ht} = \Pr(Y_{hi} = Y'_{ht})$ is the probability that the ith population unit in stratum h has label Y'_{ht}. They use a Dirichlet prior with parameters $\boldsymbol{\nu}_h = (\nu_{h1}, ..., \nu_{hT_h})$ of $p(p_{h1}, ..., p_{hT_h}|\boldsymbol{\nu}_h) \propto \prod_{t=1}^{T_h} p_{ht}^{\nu_{ht}-1}$ in conjunction with a conditionally multinomial prior of $p(N_{h1}, ..., N_{ht}|p_{h1}, ..., p_{ht}) \propto \prod_{t=1}^{T_h} \left((p_{ht})^{N_{ht}} [N_{ht}!]^{-1}\right)$, which leads to a Dirichlet-multinomial joint prior for the $(\mathbf{N}_h, \mathbf{p}_h)$, where $\mathbf{p}_h = (p_{h1}, ..., p_{hT_h})$. Integrating out the $\mathbf{p}_h$, since we are only interested in the $\mathbf{N}_h$, leads to a compound-multinomial prior for the $\mathbf{N}_h$ of:

$$p(N_{h1}, ..., N_{ht}|\nu_{h1}, ..., \nu_{hT_h}) = \binom{N_h + \nu_h - 1}{\nu_h - 1}^{-1} \prod_{t=1}^{T_h} \binom{N_{ht} + \nu_{ht} - 1}{\nu_{ht} - 1} \tag{A.2}$$

and is denoted by $CMtn(\mathbf{N}_h; \boldsymbol{\nu}_h)$, where $\nu_h = \sum_{t=1}^{T_h} v_{ht}$. For fixed h, this leads to a posterior distribution for $\mathbf{N}_h - \mathbf{n}_h$, denoted by $p(\mathbf{N}_h - \mathbf{n}_h|\mathbf{n}_h)$, that is $CMtn(\mathbf{N}_h - \mathbf{n}_h; \boldsymbol{\nu_h} + \mathbf{n}_h)$. Given independence of strata, this leads to a posterior distribution for the unobserved units of:

$$p(\mathbf{N} - \mathbf{n}|\mathbf{n}) = \prod_{h=1}^{H} p(\mathbf{N}_h - \mathbf{n}_h|\mathbf{n}_h) \tag{A.3}$$

which is the product of the relevant $CMtn$ densities.

Subsequent to data collection, the above posterior distribution can be used to obtain the posterior variance of $\bar{Y}$. However, given that the sample has not been drawn, it is necessary to perform a preposterior analysis, of the sort that was conducted by Draper and Guttman. Noting that Hartley and Rao (1968) had computed this expected posterior variance of $\bar{Y}$, Rao and Ghangurde provide the expected posterior variance as a function of the $\mathbf{N}_h, \boldsymbol{\nu}_h, \mathbf{n}_h$, for fixed $\mathbf{n}_h$. At this point, the particular form of the expected variance allows for solving for the optimal sample allocations via convex programming methods. Rao and Ghangurde also provide an extension of Draper and Guttman's sequential algorithm, which allows for not only lower bounds of 0 for strata sample sizes, but also upper bounds.

Although I do not draw directly from Rao and Ghangurde's results, their work is instructive, and their model will be more appropriate than a normal model in some circumstances. Like them, I use a preposterior analysis, aiming to minimize the expected posterior loss (in their case, the expected posterior variance), with a focus on inference for finite population quantities.

A.1.2 Equal aggregate stratification rules

Although we aim to determine optimal sample sizes for strata that are fixed upfront, establishment surveys sometimes use a different formulation, namely, equal aggregate stratification rules. Equal aggregate rules are longstanding stratification rules that entail sorting based on a size measure (or monotone transformation thereof) as to approximately equalize the strata sums (e.g., Valliant et al., 2000, p. 189); in contrast to our formulation, these rules

aim to optimally construct strata for equal allocation. More specifically, given the complexities associated with probability proportional to size without replacement (PPSWOR) designs of fixed sample size, Särndal et al. note that an approximately optimal solution is to apply model-based stratification (Wright, 1983) in conjunction with an equal aggregate σ rule (Särndal et al., 1992, §12.4–§12.5). As described in detail by Särndal et al. and Wright, this entails sorting the units by their σ_i's, choosing strata boundaries as to approximately equalize the strata sums, and allocating sample equally to strata. Formally, if U_h refers to the population in stratum h, for $h = 1, ..., H$, and where $U := \bigcup_{h=1}^{H} U_i$, this allocation leads to the equation $\sum_{i \in U_h} \sigma_i \doteq \frac{1}{H} \sum_{i \in U} \sigma_i$ holding true for each stratum h; equivalently, $N_h \bar{\sigma}_h$ is roughly constant across strata, where we define $\bar{\sigma}_h = \frac{1}{N_h} \sum_{i \in U_h} \sigma_i$ as the average unit standard deviation within stratum h. As per Wright, this is near-optimal when there are enough strata, by virtue of controlling the within-strata variability of the unit standard deviations; as an extreme example, H could be set to $n/2$, which in many populations would lead to selection probabilities approximating that of the optimal PPS design, although Wright refers to several applications that achieved near-optimality with far fewer strata (e.g., 93%–98% efficiency with 8–10 strata). As noted by Särndal et al. (and several others), as special cases, this leads to stratification rules of equal size ($b = 0$), equal aggregate square root size ($b = 1$), and equal aggregate size ($b = 2$). Specifically, we have that:

- $b = 0$ leads to equally sized strata, wherein $N_h \doteq \frac{N}{H}$ for $h = 1, ..., H$;
- $b = 1$ leads to $\sum_{i \in U_h} X_i^{1/2} \doteq \frac{1}{H} \sum_{i \in U} X_i^{1/2}$ for $h = 1, ..., H$; and
- $b = 2$ leads to $\sum_{i \in U_h} X_i \doteq \frac{1}{H} \sum_{i \in U} X_i$ for $h = 1, ..., H$.

Although these stratification rules assume that sample are equally allocated to strata, this stratification, in conjunction with equal allocation, leads to a sampling rate in stratum h of approximately $f_h \propto \bar{\sigma}_h$. Therefore, under our problem formulation (treating strata as fixed but allowing sample sizes to vary), the corresponding allocation is implied to be $n_h \propto N_h \bar{\sigma}_h$, which is implied to be near-optimal as long as the stratification adequately controls the within-strata variability of the unit standard deviations (using similar logic as in Wright, 1983). Likewise, under our problem formulation, values of $b = 0$, $b = 1$, and $b = 2$ will lead to near-optimal allocations of $n_h \propto N_h$, $n_h \propto \sum_{i=1}^{N_h} \sqrt{X_{hi}}$, and $n_h \propto N_h \bar{X}_h$, respectively, assuming adequate stratification. Assuming equal costs, this allocation for $b = 2$ is exactly the same as Cochran's corresponding rule-of-thumb allocation, and under some circumstances (e.g., equal costs and fairly constant X_{hi}'s within strata), the allocation for $b = 1$ may be similar to Cochran's corresponding rule-of-thumb allocation of $n_h \propto N_h \sqrt{\bar{X}_h}$.

Although equal aggregate stratification rules may be useful and appropriate in some scenarios, stratification may sometimes be designed for multiple purposes, preventing use of optimal size boundaries, in which case equal allocation might not be appropriate to assume. For instance, in addition to size, establishment surveys are often stratified by geography and/or industry, which may be necessary as to accommodate domain precision criteria. Even within other groupings (e.g., industry), stratification might not strictly follow from optimizing the size boundaries, but might be constrained by practical considerations, such as a need to coordinate with other samples (e.g., previous time periods or other surveys) or a desire to employ common boundaries across industry groups.

A.1.3 Estimating the coefficient of heteroscedasticity

Given that the coefficient of heteroscedasticity can vary by population and has meaningful implications for the sample allocation, an important consideration in sample design can be estimation of b in populations where it is unknown. Consider a multivariate generalization of the model expressed earlier by Equations (2.5–2.7), wherein $\mathrm{E_M}(Y_i) = \mathbf{X}_i^T\boldsymbol{\beta}$ and $\mathrm{Var_M}(Y_i) = \sigma^2 X_i^b$ for size measure X_i and vector of covariates $\mathbf{X}_i$. As described by Henry and Valliant (2009, §3.4) and Valliant et al. (2013, p. 53), b can be estimated iteratively (see also Roshwalb, 1987), wherein the two-part iterative step is as follows for step $k = 1, 2, ...K$:

i. Estimate a linear regression model for $\mathrm{E_M}(Y_i) = \mathbf{X}_i^T\boldsymbol{\beta}$ via weighted least squares, weighted by unity weights on the first iteration and weighted by $1/X_i^{\hat{b}_{(k-1)}}$ thereafter, wherein $\hat{b}_{(k-1)}$ is the estimate of b from the previous iteration. Use this model to obtain the kth set of model-estimated residuals, $\{e_{i(k)}\}$.

ii. Use ordinary least squares (OLS) to regress the squared residuals from step [i] on the natural logarithm of the x_i's, using the estimated slope as the kth estimate for b. That is, we estimate the model $\log(\varepsilon_{i(k)}^2) = \beta_0 + b \cdot \log(x_i)$, resulting in $\hat{b}_{(k)}$.

This procedure ends on the Kth iteration after some convergence criterion has been met (e.g., relative change of the estimated b falls below some threshold). Details are available via Henry and Valliant (2009) and Roshwalb (1987), and R implementation is available via the `gammaFit()` routine using the `PracTools` package (Valliant, Dever, & Kreuter, 2020).

A.2 Application and extension of Ericson's results

Under our problem formulation in §2.3, the model for stratum h can be viewed as a special case of a regression model considered by Ericson (1969a, §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.2.1 provides a definition of the normal-gamma distribution.
- A.2.2 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 distribution is provided using our own notation in A.2.2.2.
- A.2.3 provides the posterior mean and variance of $\bar{Y}|D2, e_1, b$.
- A.2.4 uses some algebraic manipulation and results relating to quadratic forms to obtain the distribution of the q_{bh} term defined in §2.3.3.3.

A.2.1 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 $\mathrm{E}\left(\frac{1}{T}\right) = \frac{\beta}{\alpha-1}$ (assuming $\alpha > 1$) and $\mathrm{Var}\left(\frac{1}{T}\right) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (assuming $\alpha > 2$). We also have that $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \frac{\beta}{\lambda(\alpha-1)}$.

A.2.2 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 A.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:

$$p(\mathbf{X}|\alpha, h) \propto \prod_{i=1}^{N} \left(\frac{h}{z_i}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{h}{z_i}(X_i - \alpha y_i)^2\right\} \tag{A.4}$$

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

$$f'(\alpha, h) \propto exp\left\{-\frac{1}{2}hn'(\alpha - \bar{\alpha}')^2\right\} h^{\frac{1}{2}\delta(n')} exp\left\{-\frac{1}{2}h\nu'\upsilon'\right\} h^{\frac{1}{2}\nu'-1} \tag{A.5}$$

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.2.1.

This leads to a posterior distribution that is normal-gamma, which he writes in Equations

(67–71) as

$$f\left\{\alpha, h|(s, \mathbf{x})\right\} \propto exp\left\{-\frac{1}{2}hn''(\alpha - \bar{\alpha}'')^2\right\} h^{\frac{1}{2}\delta(n'')} exp\left\{-\frac{1}{2}h\nu''\upsilon''\right\} h^{\frac{1}{2}\nu''-1}, \tag{A.6}$$

where

$$n'' = \sum_{i \in s} y_i^2/z_i + n', \tag{A.7}$$

$$\bar{\alpha}'' = \frac{1}{n''}\left(\sum_{i \in s} \frac{x_i y_i}{z_i} + n'\bar{\alpha}'\right), \tag{A.8}$$

$$\nu''\upsilon'' = \left\{\nu'\upsilon' + \frac{n'}{n''}\sum_{i \in s}\left(\frac{x_i - \bar{\alpha}' y_i}{\sqrt{z_i}}\right)^2 + \frac{1}{n''}\sum_{i \in s}\frac{y_i^2}{z_i}\sum_{i \in s}\frac{x_i^2}{z_i} - \frac{1}{n''}\left(\sum_{i \in s}\frac{x_i y_i}{z_i}\right)^2\right\}, \tag{A.9}$$

$$\nu'' = n + \nu' + \delta(n') - \delta(n''), \tag{A.10}$$

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 (A.5) by putting $\nu' = n' = 0$. We do so here. Thus, Equations (A.7–A.10) can

be simplified as:

$$n'' = \sum_{i \in s} y_i^2 / z_i \tag{A.11}$$

$$\bar{\alpha}'' = \frac{1}{n''} \left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right) \tag{A.12}$$

$$\nu'' \upsilon'' = \sum_{i \in s} \frac{x_i^2}{z_i} - \frac{\left(\sum_{i \in s} \frac{x_i y_i}{z_i} \right)^2}{\sum_{i \in s} \frac{y_i^2}{z_i}} \tag{A.13}$$

$$\nu'' = n - 1 \tag{A.14}$$

$$\delta(n'') = 1 \tag{A.15}$$

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 (A.11–A.14) above.

A.2.2.1 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:

$$\mathrm{E}\left(\frac{1}{h}\right) = q \tag{A.16}$$

Assuming that $n > 5$, we have:

$$\mathrm{Var}\left(\frac{1}{h}\right) = \frac{2q^2}{n-5} \tag{A.17}$$

A.2.2.2 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 $v_h|b, D2 \sim IG\left(\gamma_h, \delta_h\right)$ and $\alpha_h|v_h, b, D2 \sim N\left(\mu_h, \frac{v_h}{c_h}\right)$.

A.2.3 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:

$$\mathrm{E}\left\{\mu|(s,\mathbf{x})\right\} = \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{\alpha}''\bar{y}_{\bar{s}} \tag{A.18}$$

$$\mathrm{Var}\left\{\mu|(s,\mathbf{x})\right\} = \frac{\nu''\upsilon''}{\nu''-2}\frac{(n''z_{\bar{s}}+y_{\bar{s}}^2)}{n''N^2} \tag{A.19}$$

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.2.2.

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

$$\begin{aligned}\mathrm{E}\left\{\mu|(s,\mathbf{x})\right\} &= \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{\alpha}''\bar{y}_{\bar{s}} \\ &= \frac{n}{N}\bar{x}_s + \frac{N-n}{N}\bar{y}_{\bar{s}}\frac{\left(\sum_{i\in s}\frac{x_i y_i}{z_i}\right)}{\sum_{i\in s} y_i^2/z_i}\end{aligned} \tag{A.20}$$

Translating this back to our own notation yields:

$$\mathrm{E}\left(\bar{Y}_h|D2,e_1,b\right) = \frac{n_h}{N_h}\bar{y}_h + \left(\frac{N_h\bar{X}_h - n_h\bar{x}_h}{N_h}\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right) \tag{A.21}$$

We can then combine results across strata to get:

$$\begin{aligned}\mathrm{E}\left(\bar{Y}|D2,e_1,b\right) &= \sum_{h=1}^{H}\frac{N_h}{N}\left\{\frac{n_h}{N_h}\bar{y}_h + \left(\frac{N_h\bar{X}_h - n_h\bar{x}_h}{N_h}\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right)\right\} \\ &= \sum_{h=1}^{H}\frac{1}{N}\left\{n_h\bar{y}_h + \left(N_h\bar{X}_h - n_h\bar{x}_h\right)\left(\frac{\sum_{i\in s_{2h}}\frac{x_{hi}y_{hi}}{z_{hi}}}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right)\right\} \qquad \text{(A.22)}\end{aligned}$$

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

$$\begin{aligned}\mathrm{Var}\left\{\mu|(s,\mathbf{x})\right\} &= \frac{\nu''\upsilon''}{\nu''-2}\frac{(n''z_{\bar{s}}+y_{\bar{s}}^2)}{n''N^2} \\ &= (n-3)^{-1}\left[\sum_{i\in s}\frac{x_i^2}{z_i} - \frac{\left(\sum_{i\in s}\frac{x_iy_i}{z_i}\right)^2}{\sum_{i\in s}\frac{y_i^2}{z_i}}\right]\frac{1}{N^2}\left[z_{\bar{s}} + \frac{y_{\bar{s}}^2}{n''}\right] \\ &= (n-3)^{-1}\left[\sum_{i\in s}\frac{x_i^2}{z_i} - \frac{\left(\sum_{i\in s}\frac{x_iy_i}{z_i}\right)^2}{\sum_{i\in s}\frac{y_i^2}{z_i}}\right]\frac{1}{N^2}\left[z_{\bar{s}} + \frac{y_{\bar{s}}^2}{\sum_{i\in s}y_i^2/z_i}\right], \qquad \text{(A.23)}\end{aligned}$$

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

$$\mathrm{Var}\left(\bar{Y}_h|D2,e_1,b\right) = (n_h-3)^{-1}\left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\frac{1}{N_h^2}\left[\sum_{i\in(U_h\cap s_{2h}^C)}z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right], \qquad \text{(A.24)}$$

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:

$$\mathrm{Var}\left(\bar{Y}|D2,e_1,b\right) = \sum_{h=1}^{H}\frac{1}{N^2}\left\{\frac{1}{n_h-3}\left[\sum_{i\in s_{2h}}\frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{2h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\left[\sum_{i\in(U_h\cap s_{2h}^C)}z_{hi} + \frac{\left(N_h\bar{X}_h - n_h\bar{x}_h\right)^2}{\sum_{i\in s_{2h}}\frac{x_{hi}^2}{z_{hi}}}\right]\right\} \qquad \text{(A.25)}$$

A.2.4 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:

$$\begin{aligned}
\sum_i y_i^2/z_i - \frac{\left(\sum_i y_i x_i/z_i\right)^2}{\sum_i x_i^2/z_i} &= \sum_i a_i^2 - \frac{\left(\sum_i a_i d_i\right)^2}{\sum_k d_k^2} \\
&= \sum_i a_i^2 - \frac{\sum_i a_i^2 d_i^2 + \sum_{i\neq j} a_i d_i a_j d_j}{\sum_k d_k^2} \\
&= \sum_i a_i^2 \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right) + \sum_{i=1}^n \sum_{j=1, j\neq i}^n a_i a_j \left(\frac{-d_i d_j}{\sum_k d_k^2}\right) \\
&= a^t D a \qquad \text{(A.26)}
\end{aligned}$$

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

$$D_{ii} = 1 - \frac{d_i^2}{\sum_k d_k^2}, \qquad \text{i= } 1, ..., n \tag{A.27}$$

$$D_{ij} = \frac{-d_i d_j}{\sum_k d_k^2}, \qquad \text{i= } 1, ..., n; j = 1, ..., n; i \neq j \tag{A.28}$$

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

$$\begin{aligned} D'_{ii} &= D_{ii}^2 + \sum_{j=1, j\neq i}^{n} D_{ij} D_{ji} \\ &= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\left(\sum_k d_k^2\right)^2} \sum_{j=1, j\neq i}^{n} d_j^2 \\ &= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\left(\sum_k d_k^2\right)^2} \left(\left[\sum_{j=1}^{n} d_j^2\right] - d_i^2\right) \\ &= \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right)^2 + \frac{d_i^2}{\sum_k d_k^2} \left(1 - \frac{d_i^2}{\sum_k d_k^2}\right) \\ &= 1 - \frac{d_i^2}{\sum_k d_k^2} = D_{ii} \end{aligned} \tag{A.29}$$

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

$$\begin{aligned} D'_{ij} &= \sum_{k=1}^{n} D_{ik} D_{kj} \\ &= \frac{1}{\left(\sum_k d_k^2\right)^2} \sum_{k=1}^{n} d_k^2 d_i d_j \\ &= \frac{d_i d_j}{\sum_k d_k^2} = D_{ij} \end{aligned} \tag{A.30}$$

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 $v_h^{-1} q_{bh} | (v_h, \alpha_h, b, s_{2h}) \sim \chi^2_{n_h - 1}$ and that $\underset{D2|\alpha_h, v_h, b, s_{2h}}{\mathrm{E}} (q_{bh}) = v_h(n_h - 1)$.

A.3 Analysis for fixed and known b

Recall from Equation (2.12) that:

$$\text{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \left\{ \frac{1}{n_h - 3} \left[\sum_{i \in s_{2h}} \frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i \in s_{2h}} \frac{y_{hi} x_{hi}}{z_{hi}}\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \left[\sum_{i \in (U_h \cap s_{2h}^C)} z_{hi} + \frac{\left(N_h \bar{X}_h - n_h \bar{x}_h\right)^2}{\sum_{i \in s_{2h}} \frac{x_{hi}^2}{z_{hi}}} \right] \right\} \tag{A.31}$$

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:

$$\text{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2 (n_h - 3)} q_{bh} k(s_{2h}) \tag{A.32}$$

A.3.1 Analysis relating to q_{bh}

Conditional on the pilot data (D1) and the main study sample allocation (e_1), Equation (A.32) 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

$\mathrm{E}(q_{bh}k(s_{2h})) = \mathrm{E}(q_{bh})\mathrm{E}(k(s_{2h}))$. 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 $\underset{D2|D1}{\mathrm{E}}(q_{bh})$ 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 $\mathrm{E}_M(Z) = \underset{D2|\alpha_h,v_h,D1,s_{2h}}{\mathrm{E}}(Z)$ and $\mathrm{E}_P(Z) = \underset{\alpha_h,v_h|D1}{\mathrm{E}}(Z)$ for any Z. Then assuming that the strata sample sizes, $\{n_h\}$, are fixed, we have:

$$\underset{D2|D1}{\mathrm{E}}(q_{bh}|s_{2h}) = \mathrm{E}_P(\mathrm{E}_M(q_{bh}|s_{2h})) \tag{A.33}$$

Appendix A.2.4 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

$$\underset{D2|D1}{\mathrm{E}}\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2(n_h-3)}\left[\underset{D2|D1}{\mathrm{E}}(q_{bh})\right]\mathrm{E}\left(k(s_{2h})\right) \tag{A.34}$$

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 (A.33), and then using the fact that $q_{bh}|D1$ does not depend on s_{2h},

yields:

$$\underset{D2|D1}{\mathrm{E}}(q_{bh}) = (n_h - 1)\mathrm{E}_P(v_h) \tag{A.35}$$

Substituting this into (A.34) gives:

$$\underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) = \sum_{h=1}^{H} \frac{1}{N^2} \frac{n_h - 1}{n_h - 3} \left[\mathrm{E}_P(v_h)\right] \mathrm{E}\left(k(s_{2h})\right) \tag{A.36}$$

A.3.2 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

$$\mathrm{E}\left[\frac{\left(N\bar{X}-n\bar{x}\right)^2}{\sum_{i\in s}x_i^2/z_i}\right] \approx \frac{\mathrm{E}\left[\left(N\bar{X}-n\bar{x}\right)^2\right]}{\mathrm{E}\left(\sum_{i\in s}x_i^2/z_i\right)}, \tag{A.37}$$

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

$$\begin{aligned}
\mathrm{E}\left[\left(N\bar{X}-n\bar{x}\right)^2\right] &= \mathrm{Var}\left[N\bar{X}-n\bar{x}\right] + \left(\mathrm{E}\left[N\bar{X}-n\bar{x}\right]\right)^2 \\
&= \mathrm{Var}(n\bar{x}) + (N-n)^2\bar{X}^2 \\
&= n^2\left(1-\frac{n}{N}\right)\frac{S_x^2}{n} + (N-n)^2\bar{X}^2 \\
&= (N-n)\left[\frac{n}{N}S_x^2 + (N-n)\bar{X}^2\right], \qquad \text{(A.38)}
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{E}\left(\sum_{i\in s}x_i^2/z_i\right) &= \frac{n}{N}\sum_{i\in U}Q_i^2 \\
&= \frac{n}{N}\left((N-1)S_q^2 + N\bar{Q}^2\right) \\
&\approx n(S_q^2+\bar{Q}^2) \qquad \text{(A.39)} \\
&= n\bar{Q}^2(1+CV_q^2) \qquad \text{(A.40)}
\end{aligned}$$

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

$$\mathrm{E}\left(k(s)\right) \approx (N-n)\bar{Z} + \frac{(N-n)\left[\frac{n}{N}S_x^2 + (N-n)\bar{X}^2\right]}{n\bar{Q}^2(1+CV_q^2)} \tag{A.41}$$

$$= (N-n)\left[\bar{Z} + \frac{\frac{n}{N}S_x^2 + (N-n)\bar{X}^2}{n\bar{Q}^2(1+CV_q^2)}\right]. \tag{A.42}$$

Reintroducing strata subscripts, we substitute this quantity into Equation (A.36) to yield

$$\underset{D2|D1}{\mathrm{E}}\,\mathrm{Var}\left(\bar{Y}|D2, e_1, b\right) \approx \sum_{h=1}^{H} \frac{\mathrm{E}_P(v_h)}{N^2}\left(\frac{n_h-1}{n_h-3}\right)(N_h-n_h)\left[\bar{Z}_h + \frac{\frac{n_h}{N_h}S_{xh}^2 + (N_h-n_h)\bar{X}_h^2}{n_h\bar{Q}_h^2(1+CV_{qh}^2)}\right], \tag{A.43}$$

which assumes large $\{n_h\}$.

A.3.3 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.2.2, 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 (A.16) to the current notation

yields:

$$\mathrm{E}_P(v_h) = \frac{1}{m_h - 3}\left(\sum_{i\in s1h}\frac{y_{hi}^2}{z_{hi}} - \frac{\left(\sum_{i\in s_{1h}}\frac{y_{hi}x_{hi}}{z_{hi}}\right)^2}{\sum_{i\in s_{1h}}\frac{x_{hi}^2}{z_{hi}}}\right) \tag{A.44}$$

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 $z_{hi} = x_{hi}^b$.

A.3.4 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 A.3.2. 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

$$\mathrm{E}\left(k(s_{2h}|n_h)\right) = (N_h - n_h)\bar{Z}_h + \mathrm{E}(j_h|n_h). \tag{A.45}$$

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

$$j_{rn} = \frac{(N\bar{X} - \sum_{j=1}^{n} X_{i_{rj}})^2}{\sum_{j=1}^{n} \frac{X_{i_{rj}}^2}{Z_{i_{rj}}}}, \tag{A.46}$$

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

$$\hat{\mathrm{E}}(k(s_2)|n) = (N - n)\bar{Z} + \bar{j}_{.n}, \tag{A.47}$$

where $\bar{j}_{.n} = \frac{1}{R}\sum_{r=1}^{R} j_{rn}$. The MC estimate of (A.47) 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 (A.47), 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 yields H and 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 require the use of continuous decision variables. To accommodate this disconnect, continuous decision variables can be straightforwardly handled via 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

$$\hat{\mathrm{E}}(k(s_{2h})|n_h) = (1 - frac_h)\,\hat{\mathrm{E}}(k(s_{2h})|\lfloor n_h \rfloor) + (frac_h)\,\hat{\mathrm{E}}(k(s_{2h})|\lceil n_h \rceil), \tag{A.48}$$

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.

A.4 Additional simulation detail

A.4.1 Selection of v_h to achieve target correlation

Our simulation (from §2.4 and §2.5) 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 (2.14) so as to approximately result in the desired correlations between $\{X_{hi}, Y_{hi}\}$ for the H strata.

Equation (2.14) 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 (2.14), meaning that the data-generating mechanism described earlier would approximately yield the desired correlations, on expectation, if the assumptions hold.

A.4.2 Supplementary figures and tables

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

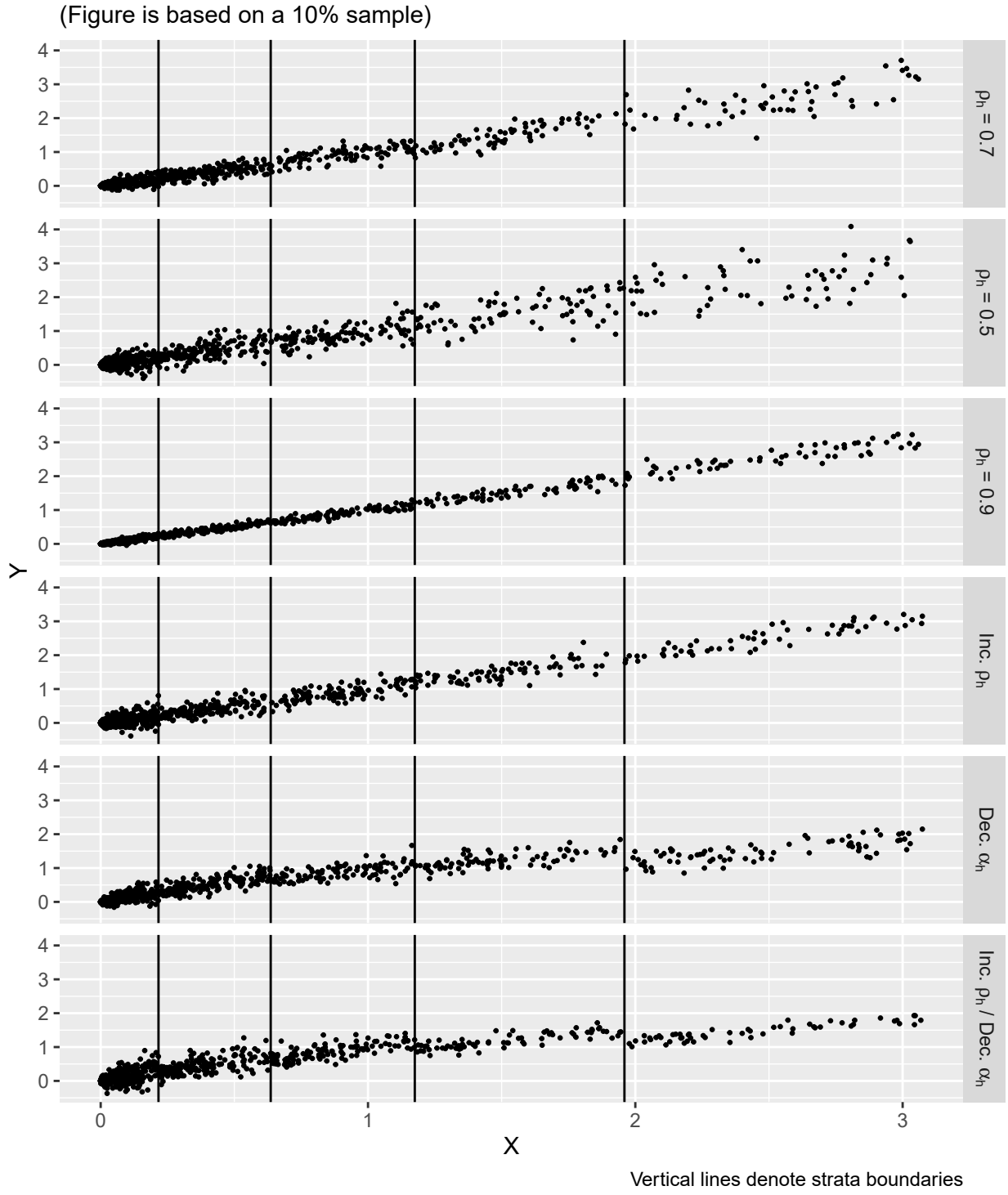


Table A.2: 1,000*RelBias by main strategy and population-generating characteristics

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

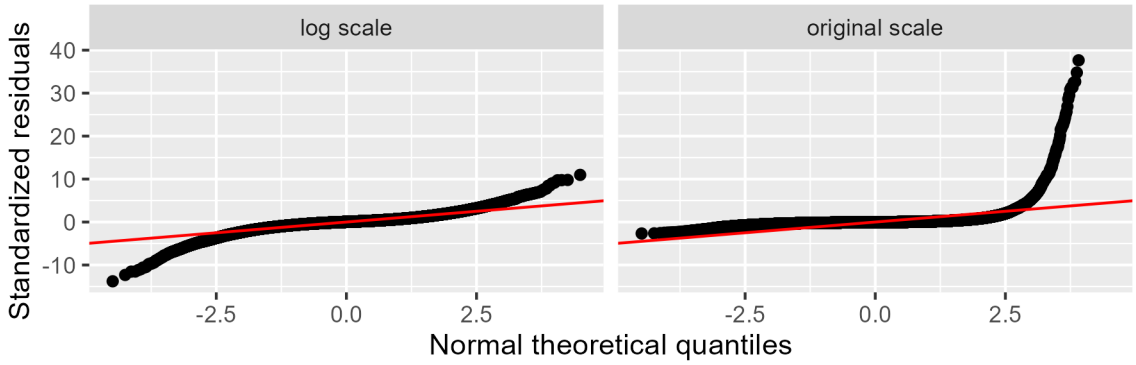
		MOS1 (most skewness)						MOS2 (mid skewness)						MOS3 (least skewness)					
		1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes	1. Baseline	2. Lower corrs	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 A.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 A.2: 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 §2.6, 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 §2.6.2, 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 A.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

Appendix B: Chapter 3 Appendices

B.1 Marginal posterior distribution for b under non-informative prior

In this appendix, we use similar notation and an equivalent set-up as in §3.3.2, except that the strata are indexed as $k = 1, ..., K$ and the variance of the error term is expressed slightly differently. More specifically, for $k = 1, ..., K$ and $i = 1, ..., n_k$, assume a model of $y_{ki} = \alpha_k x_{ki} + \varepsilon_{ki}$, where $\mathrm{E}(\varepsilon_{ki}|x_{ki}) = 0$ and $\mathrm{Var}(\varepsilon_{ki}|x_{ki}) = \frac{x_{ki}^b}{h_k}$, and where $y_{ki}|(x_{ki}, b, \alpha_k, h_k)$ is normally distributed and $\mathrm{Cov}(\varepsilon_{ki}, \varepsilon_{kj}|\{x_{ki}\}) = 0$ for $i \neq j$. As before, h_k follows the gamma distribution and the notation was chosen for consistency with Ericson (1969a). To simplify notation, the conditioning on the $\{x_{ki}\}$ will be implicit. Let $z_{ki} = x_{ki}^b$. We assume a prior of

$$\pi\left(b, \{\alpha_k, h_k\}\right) \propto \left(\prod_{k=1}^{K} h_k^{-1}\right) I_{[b_0,b_1]}(b), \tag{B.1}$$

where $\left(\prod_{k=1}^{K} h_k^{-1}\right)$ is the same as Ericson's prior (and is a Jeffreys prior) and $I_{[b_0,b_1]}(b)$ is uniform from b_0 to b_1.

To further simplify notation, let $\boldsymbol{\alpha} = \{\alpha_k\}$, $\boldsymbol{h} = \{h_k\}$, and $\boldsymbol{y_k} = \{y_{ki}\}$ be the relevant vectors for stratum k and let $\boldsymbol{y} = \{\boldsymbol{y_k}\}$ be the vector of data. Also let $\boldsymbol{\theta} = (b, \boldsymbol{\alpha}, \boldsymbol{h})$.

Given our model of $y_{ki}|b, \boldsymbol{\alpha}, \boldsymbol{h} \sim N\left(\alpha_k x_{ki}, \frac{x_{ki}^b}{h_k}\right)$, it follows that:

$$
\begin{aligned}
L(y_{ki}|\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi x_{ki}^b/h_k}} exp\left(-\frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right) \\
&\propto x_{ki}^{-b/2} h_k^{1/2} exp\left(-\frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right) \qquad \text{(B.2)} \\
L(\boldsymbol{y_k}|\boldsymbol{\theta}) &= \prod_{i=1}^{n_k} L(y_{ki}|\boldsymbol{\theta}) \\
&\propto \left(\prod_{i=1}^{n_k} x_{ki}^{-b/2}\right) h_k^{n_k/2} exp\left(-\sum_{i=1}^{n_k} \frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right) \qquad \text{(B.3)} \\
L(\boldsymbol{y}|\boldsymbol{\theta}) &= \prod_{k=1}^{K} L(\boldsymbol{y_k}|\boldsymbol{\theta}) \\
&\propto \prod_{k=1}^{K}\left[\left(\prod_{i=1}^{n_k} x_{ki}^{-b/2}\right) h_k^{n_k/2} exp\left(-\sum_{i=1}^{n_k} \frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right)\right] \\
&= \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2} \left(\prod_{k=1}^{K} h_k^{n_k}\right)^{1/2} exp\left(-\sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right) \qquad \text{(B.4)}
\end{aligned}
$$

From Equations (B.1) and (B.4), the posterior distribution is computed as:

$$
\begin{aligned}
\pi(b, \boldsymbol{\alpha}, \boldsymbol{h}|\boldsymbol{y}) &\propto \pi(b, \boldsymbol{\alpha}, \boldsymbol{h})\, L(\boldsymbol{y}|b, \boldsymbol{\alpha}, \boldsymbol{h}) \\
&\propto \left[\left(\prod_{k=1}^{K} h_k^{-1}\right) I_{[b_0,b_1]}(b)\right] \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2} \left(\prod_{k=1}^{K} h_k^{n_k}\right)^{1/2} exp\left(-\sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{(y_{ki}-\alpha_k x_{ki})^2}{2x_{ki}^b/h_k}\right) \\
&\propto I_{[b_0,b_1]}(b) \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2} \left(\prod_{k=1}^{K} h_k^{\frac{n_k}{2}-1}\right) exp\left(-\sum_{k=1}^{K} \frac{h_k}{2} \sum_{i=1}^{n_k} \frac{(y_{ki}-\alpha_k x_{ki})^2}{x_{ki}^b}\right) \qquad \text{(B.5)}
\end{aligned}
$$

Note that this is a fairly complicated expression, which would preferably be expressed as a function of $(b, \boldsymbol{\alpha}, \boldsymbol{h}, h(\boldsymbol{y}))$, for some function (e.g., sufficient statistic) of the sample $h(\boldsymbol{y})$. Hence, first, ignore the strata subscripts, and let $a_i = \frac{y_i^2}{z_i}$, $b_i = \frac{x_i y_i}{z_i}$, $c_i = \frac{x_i^2}{z_i}$, where $z_i = x_i^b$.

Also let $\bar{a} = \frac{\sum_i a_i}{n}$, $\bar{b} = \frac{\sum_i b_i}{n}$, and $\bar{c} = \frac{\sum_i c_i}{n}$. Then we have:

$$\begin{aligned}
\sum_i \frac{(y_i - \alpha x_i)^2}{z_i} &= \sum_i \left(a_i - 2\alpha b_i + \alpha^2 c_i\right) \\
&= n\bar{a} - 2\alpha n\bar{b} + \alpha^2 n\bar{c} \\
&= n\bar{c}\left(\alpha^2 - 2\alpha\frac{\bar{b}}{\bar{c}}\right) + n\bar{a} \\
&= n\bar{c}\left(\alpha - \frac{\bar{b}}{\bar{c}}\right)^2 + n\bar{a} - \frac{n\bar{b}^2}{\bar{c}} \\
&= \sum_i \frac{x_i^2}{z_i}\left(\alpha - \frac{\sum_i x_i y_i/z_i}{\sum_i x_i^2/z_i}\right)^2 + \sum_i \frac{y_i^2}{z_i} - \frac{\left(\sum_i x_i y_i/z_i\right)^2}{\sum_i x_i^2/z_i}
\end{aligned} \tag{B.6}$$

Going back to Equations (B.3) and (B.5) above, and letting $a_k = \sum_{i=1}^{n_k} \frac{y_{ki}^2}{z_{ki}}$, $b_k = \sum_{i=1}^{n_k} \frac{x_{ki}y_{ki}}{z_{ki}}$, and $c_k = \sum_{i=1}^{n_k} \frac{x_{ki}^2}{z_{ki}}$, this gives us:

$$L(\boldsymbol{y_k}|\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{n_k} x_{ki}^{-b/2}\right) h_k^{n_k/2} exp\left(-\frac{h_k}{2}\left[c_k\left(\alpha_k - \frac{b_k}{c_k}\right)^2 + a_k - \frac{b_k^2}{c_k}\right]\right) \tag{B.7}$$

$$\begin{aligned}
\pi\left(b, \boldsymbol{\alpha}, \boldsymbol{h}|\boldsymbol{y}\right) &\propto I_{[b_0,b_1]}(b)\left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2}\left(\prod_{k=1}^{K} h_k^{\frac{n_k}{2}-1}\right) exp\left(-\sum_{k=1}^{K}\frac{h_k}{2}\left[c_k\left(\alpha_k - \frac{b_k}{c_k}\right)^2 + a_k - \frac{b_k^2}{c_k}\right]\right) \\
&\propto I_{[b_0,b_1]}(b)\left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2}\prod_{k=1}^{K}\left[\sqrt{h_k}\cdot e^{-\frac{h_k c_k(\alpha_k-\mu_k)^2}{2}} h_k^{\gamma_k - 1} e^{-h_k\delta_k}\right]
\end{aligned} \tag{B.8}$$

where $\mu_k = \frac{b_k}{c_k} = \frac{\sum_i^{n_k} x_{ki}y_{ki}/z_{ki}}{\sum_i^{n_k} x_{ki}^2/z_{ki}}$, $\gamma_k = \frac{n_k - 1}{2}$, and
$\delta_k = \frac{1}{2}\left(a_k - \frac{b_k^2}{c_k}\right) = \frac{1}{2}\left\{\sum_i^{n_k} y_{ki}^2/z_{ki} - \frac{\left[\sum_i^{n_k} x_{ki}y_{ki}/z_{ki}\right]^2}{\sum_i^{n_k} x_{ki}^2/z_{ki}}\right\}$.

Noting that the $\{x_{ki}\}$ are fixed, observe that if we condition on b, then we have the joint pdf of k independent normal-gamma random variables, multiplied by a normalizing constant. More specifically, we have that $\alpha_k, h_k|\boldsymbol{y}, b \sim NG(\mu_k, c_k, \gamma_k, \delta_k)$, with independence across strata (achieved by conditioning on b, which is the only shared parameter). This can be

observed by first conditioning on b and all random variable pairs (α_j, h_j) for which $j \neq k$, which yields $\pi(\alpha_k, h_k|b, \boldsymbol{y}, \{\alpha_j, h_j : j \neq k\}) = \pi(\alpha_k, h_k|b, \boldsymbol{y}) \propto \sqrt{h_k} \cdot e^{-\frac{h_k c_k (\alpha_k - \mu_k)^2}{2}} h_k^{\gamma_k - 1} e^{-h_k \delta_k}$. Then, condition on h_k and observe that we have a pdf that is normally distributed with mean μ_k and precision $h_k c_k$, that is, $\pi(\alpha_k|\boldsymbol{y}, b, h_k) = \frac{1}{\sqrt{2\pi/h_k c_k}} exp\left(-\frac{(\alpha_k - \mu_k)^2}{2/h_k c_k}\right)$, which leads to $\pi(h_k|\boldsymbol{y}, b) = \frac{\pi(\alpha_k, h_k|\boldsymbol{y}, b)}{\pi(\alpha|\boldsymbol{y}, b, h_k)} \propto h_k^{\gamma_k - 1} exp(-h_k \delta_k)$, which is clearly a gamma distribution with shape γ_k and rate δ_k.

Given that $\alpha_k, h_k|\boldsymbol{y}, b \sim NG(\mu_k, c_k, \gamma_k, \delta_k)$, then from the definition of the NG distribution, we have a joint posterior conditional pdf of:

$$\begin{aligned} \pi\left(\boldsymbol{\alpha}, \boldsymbol{h}|\boldsymbol{y}, b\right) &= \prod_{k=1}^{K} \pi(\alpha_k, h_k|\boldsymbol{y}, b) \\ &= \prod_{k=1}^{K} \left[\frac{\delta_k^{\gamma_k} \sqrt{c_k}}{\Gamma(\gamma_k)\sqrt{2\pi}} h_k^{\gamma_k - \frac{1}{2}} e^{-\delta_k h_k} e^{-\frac{h_k c_k (\alpha_k - \mu_k)^2}{2}} \right] \end{aligned} \tag{B.9}$$

(Note that after conditioning on b, the strata-specific results are aligned with those of Chapter 2, as were provided in Appendix A.2.2.2.)

We can then compute the marginal posterior of b via Equations (B.8) and (B.9) as:

$$\begin{aligned}
\pi(b|\boldsymbol{y}) &= \frac{\pi(b, \boldsymbol{\alpha}, \boldsymbol{h}|\boldsymbol{y})}{\pi(\boldsymbol{\alpha}, \boldsymbol{h}|b, \boldsymbol{y})} \\
&\propto \frac{I_{[b_0,b_1]}(b) \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2} \prod_{k=1}^{K}\left[\sqrt{h_k}\cdot e^{-\frac{h_k c_k(\alpha_k-\mu_k)^2}{2}} h_k^{\gamma_k-1} e^{-h_k\delta_k}\right]}{\prod_{k=1}^{K}\left[\frac{\delta_k^{\gamma_k}\sqrt{c_k}}{\Gamma(\gamma_k)\sqrt{2\pi}} h_k^{\gamma_k-\frac{1}{2}} e^{-\delta_k h_k} e^{-\frac{h_k c_k(\alpha_k-\mu_k)^2}{2}}\right]} \\
&= \frac{I_{[b_0,b_1]}(b) \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2}}{\prod_{k=1}^{K}\left[\frac{\delta_k^{\gamma_k}\sqrt{c_k}}{\Gamma(\gamma_k)\sqrt{2\pi}} h_k^{\gamma_k-\frac{1}{2}}\right]} \\
&\propto \frac{I_{[b_0,b_1]}(b) \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2}}{\prod_{k=1}^{K}\left[\delta_k^{\gamma_k}\sqrt{c_k}\right]} \qquad \text{(B.10)} \\
&\propto \frac{I_{[b_0,b_1]}(b) \left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} x_{ki}\right)^{-b/2}}{\prod_{k=1}^{K}\left\{\left[\sum_{i=1}^{n_k} y_{ki}^2 x_{ki}^{-b} - \frac{\left(\sum_{i=1}^{n_k} y_{ki} x_{ki}^{1-b}\right)^2}{\sum_{i=1}^{n_k} x_{ki}^{2-b}}\right]^{\frac{n_k-1}{2}} \sqrt{\sum_{i=1}^{n_k} x_{ki}^{2-b}}\right\}} \qquad \text{(B.11)}
\end{aligned}$$

Note that in Equation (B.10) above, we were able to omit the $\Gamma(\gamma_k)$ term since it is constant and could ignore the $h_k^{\gamma_k-\frac{1}{2}}$ term since it does not depend on b, although we needed to leave in functions of δ_k and c_k since these depend on b.

B.2 Computational details for proposed methods

B.2.1 Detailed overview of sampling from $f(D2|D1)$

Here, we show how to perform the outer step for the MC estimator from §3.4.1. Specifically, we show how to sample from $f(D2|D1) = \int f_1(D2|D1, b, \{\alpha_h, v_h\}) f_2(b, \{\alpha_h, v_h\}\,|D1)\,\mathrm{d}b\,\mathrm{d}\,\{\alpha_h, v_h\}$, where a given draw results in a set of sample indicators and sample data. The procedure is as follows:

1. Note that although the posterior distribution $f_2(b, \{\alpha_h, v_h\}\,|D1)$ is complicated, for stratum h we have that $\alpha_h, v_h|D1, b$ are distributed as normal-gamma; hence, we can draw from
   $f_2(b, \{\alpha_h, v_h\}\,|D1) = f_3(b|D1) f_4(\{\alpha_h, v_h\})|D1, b$ as follows:
   (a) Simulate $R1$ draws from $b|D1$, resulting in the set of draws $\{b^{(r_1)} : r_1 = 1, ..., R1\}$. Since the posterior distribution for b is complicated, this requires an efficient method for sampling from an arbitrary univariate pdf. For example, use an adaptive grid-based method for approximating the inverse-cdf function (see Appendix B.2.2 for details).
   (b) Next, draw from $\{\alpha_h, v_h\}\,|b, D1$, resulting in the set of draws $\{\alpha_h^{(r_1)}, v_h^{(r_1)} : r_1 = 1, ..., R1; h = 1, ...H\}$. This step is straightforward, since the relevant pdf is the joint distribution of h independent normal-gamma distributions. Hence, for stratum h, we draw $v_h^{(r_1)}|b^{(r_1)}, D1$ from the relevant inverse-gamma distribution and

$\alpha_h^{(r_1)}|v_h^{(r_1)}, b^{(r_1)}, D1$ from the relevant normal distribution. The normal-gamma parameters for $\alpha_h, \frac{1}{v_h}|b, D1$ can be obtained through a minor adaptation of Appendix A.2.2.2, by using pilot (rather than main) sample quantities.

2. Draw $R1$ complete sets of sample data from $D2|\{\alpha_h, v_h\}, b$ in two steps, where a given draw r_1 is drawn as follows, for $r_1 = 1, ..., R1$:

   (a) Draw a design-based sample via STSRS methods, $s_2^{(r_1)} = \left\{s_{2h}^{(r_1)} : h = 1, ..., H\right\}$, where $s_{2h}^{(r_1)} = \left\{i_{h1}^{(r_1)}, ..., i_{h\,n_h}^{(r_1)}\right\}$ refers to the set of indices of units that were sampled in stratum h for simulation r_1, for $i_{hj}^{(r_1)} \in \{1, ..., N_h\}$ with $i_{hj}^{(r_1)} \neq i_{hk}^{(r_1)}$ for $j \neq k$, and where, for definiteness, we assume $i_{h1}^{(r_1)} < i_{h2}^{(r_1)} < ... < i_{h\,n_h}^{(r_1)}$. Note that the population auxiliary data, $\{X_{hj} : h = 1, ..., H; j = 1, ..., N_h\}$, are known; therefore, knowing the set of sample indicators also results in knowing the auxiliary sample data, $\left\{X_{h\,i_{hj}^{(r_1)}} : h = 1, ..., H, j = 1, ..., n_h\right\}$. We will subsequently simplify our notation by reindexing the units so that, in stratum h, $j = 1, ..., n_h$ index the sample units while $j = n_h + 1, ..., N_h$ index the nonsampled units, with lower case to refer to sample data and upper case to refer to population values for nonsampled units. Hence, let $x_2^{(r_1)} = \{x_{hj} : h = 1, ..., H; j = 1, ..., n_h\}$ refer to the sample auxiliary data and $x_2(ns)^{(r_1)} : \{X_{hj} : h = 1, ..., H; j = n_h + 1, ..., N_h\}$ refer to the nonsample auxiliary data. Likewise, let $ns_2^{(r_1)} = \left\{ns_{2h}^{(r_1)} : h = 1, ..., H\right\}$, where $ns_{2h}^{(r_1)} = \left\{j : \left(j \not\subset s_{2h}^{(r_1)}\right) \wedge (j \in \{1, ..., N_h\})\right\}$ refers to the indices of the nonsampled units in stratum h, simulation r_1.

   (b) Given $s_2^{(r_1)}$, use the superpopulation parameters, $\left\{\alpha_h^{(r_1)}, v_h^{(r_1)}\right\}, b^{(r_1)}$, to draw the outcome data, $y_2^{(r_1)} = \left\{y_{hi}^{(r_1)} : h = 1, ..., H; i = 1, ..., n_h\right\}$, from the relevant normal

distribution, where the subscript hi is indexed by the particular sample (rather than by the population). Specifically, for sample unit hi associated with iteration r_1, we draw $y_{hi}^{(r_1)}$ from $N\left(\alpha_h^{(r_1)} x_{hi}^{(r_1)}, v_h^{(r_1)} z_{hi}^{(r_1)}\right)$, where $z_{hi}^{(r_1)} = \left(x_{hi}^{(r_1)}\right)^{b^{(r_1)}}$.

3. For draw r_1, we retain the data set, $d_2^{(r_1)} = \left(s_2^{(r_1)}, y_2^{(r_1)}\right)$.

B.2.2 Sampling from marginal posterior for b

The MC estimator of the expected posterior loss, as described in §3.4.1, entails sampling from the posterior distribution for b in two places:

- The outer step entails drawing from $b|D1$. This step entails approximating a single distribution, from which $R1$ draws are made.
- The inner step entails drawing $\{b^{(r_1,r_2)}\}$ from $b|D2$. This step entails approximating $R1$ distributions (corresponding with the $R1$ draws from $D2|D1$), from each of which $R2$ draws are made.

In each case, we need to sample from a univariate posterior distribution for b, which has a straightforward but messy posterior distribution that does not lend itself to a simple inverse-cdf. We need an efficient method for sampling from a given univariate distribution of b, given that we will be drawing from $R1 + 1$ such distributions per MC estimate of the expected posterior loss. Hence, this section outlines computational methods for sampling from a posterior distribution of b via adaptive grid-based methods. Appendix B.2.2.1 outlines how

to circumvent machine precision issues in computing the likelihood; the basic idea is to work with the log-likelihood, and then to add some constant that allows for exponentiating the resulting quantity. Using such a machine-computable likelihood function, Appendix B.2.2.2 outlines an adaptive grid-based algorithm that allows for sampling from an approximate marginal posterior distribution of b.

B.2.2.1 Computing marginal posterior for b

This section outlines how to compute $f(b|D1)$ or $f(b|D2)$ for some given pilot or main study data set, while addressing technical issues arising from machine precision limitations. Here, being able to evaluate such a posterior pdf is a necessary first step toward being able to sample from it. We use the same notation and set-up as in Appendix B.1, wherein we index the strata as $k = 1, ..., K$, let $h_k = \frac{1}{v_k}$, and assume a prior distribution of $\pi\left(b, \{\alpha_k, h_k\}\right) \propto \left(\prod_{i=1}^{K} h_k^{-1}\right) I_{[b_0,b_1]}$, where $I_{[b_0,b_1]}(b)$ is uniform from b_0 to b_1. Likewise, we let n_k refer to the sample size in stratum k and $\boldsymbol{y}$ refer to the vector of sample data. Appendix B.1 provides the derivation of the marginal posterior for $b|\boldsymbol{y}$, which, as per Equation (B.11), can be expressed as

$$\pi(b|\boldsymbol{y}) \propto \frac{I_{[b_0,b_1]}(b)\left(\prod_{k=1}^{K}\prod_{i=1}^{n_k} z_{ki}\right)^{-1/2}}{\prod_{k=1}^{K}\left\{\left[\sum_{i=1}^{n_k}\frac{y_{ki}^2}{z_{ki}} - \frac{\left(\sum_{i=1}^{n_k}\frac{y_{ki}x_{ki}}{z_{ki}}\right)^2}{\sum_{i=1}^{n_k}\frac{x_{ki}^2}{z_{ki}}}\right]^{\frac{n_k-1}{2}}\sqrt{\sum_{i=1}^{n_k}\frac{x_{ki}^2}{z_{ki}}}\right\}}, \tag{B.12}$$

where $z_{ki} = x_{ki}^b$.

In some cases, attempting to compute this directly may be problematic due to limitations of floating-point arithmetic. For example, the IEEE 754 standard (Zuras et al., 2008) on double-precision binary floating-point on a 64-bit computer (i.e., binary64) cannot be used to express numbers greater than $1.8 \cdot 10^{308} \approx e^{709.8}$ or strictly positive numbers below $5 \cdot 10^{-324}$ (with precision reduced at or below $2.2 \cdot 10^{-308} \approx e^{-708.4}$). Hence, ignoring the normalizing constant, we can compute the log likelihood as:

$$l(b|\boldsymbol{y}) = -\sum_{k=1}^{K}\left[\frac{b}{2}\sum_{i=1}^{n_k} log(x_{ki}) + \frac{n_k - 1}{2} log\left(\sum_{i=1}^{n_k}\frac{y_{ki}^2}{z_{ki}} - \frac{\left(\sum_{i=1}^{n_k}\frac{y_{ki}x_{ki}}{z_{ki}}\right)^2}{\sum_{i=1}^{n_k}\frac{x_{ki}^2}{z_{ki}}}\right) + \frac{1}{2} log\left(\sum_{i=1}^{n_k}\frac{x_{ki}^2}{z_{ki}}\right)\right] \tag{B.13}$$

if $b \in [b_0, b_1]$, and 0 otherwise. (For convenience, we use "log likelihood" to denote any function that is the logarithm of a function that is proportional to a probability density function.) We can then evaluate the right-hand term above over a simple grid to determine some sort of scaling constant that can be added to the log likelihoods as to allow for the resulting log likelihoods to be exponentiated via a machine (e.g., for binary64, we aim to compute a new log likelihood function wherein the output range is $[-708, 709]$ for any points associated with non-negligible density). For example, suppose that we evaluate $l(.)$ for each point in a set of grid points $G = \{g_1, ..., g_m\}$. (To illustrate, a simple grid could be defined via $\left\{g_i = b_0 + \frac{b_1 - b_0}{m-1}(i-1) : i = 1, ..., m\right\}$.) Then we could define

$$l_0(b|\boldsymbol{y}) = l(b|\boldsymbol{y}) - sup\left\{l(g)|g \in G\right\}. \tag{B.14}$$

As long as $sup\left\{l(g)|g \in [b_0, b_1]\right\} - sup\left\{l(g)|g \in G\right\}$ is not too big, then we should be able

to compute $f(b|\boldsymbol{y}) = \frac{exp(l_0(b|\boldsymbol{y}))}{\int_{b_0}^{b_1} exp(l_0(b|\boldsymbol{y}))db}$ without much difficulty. (Note that although the denominator in this expression could be easily computed via numerical integration methods, the adaptive grid-based sampling algorithm outlined in Appendix B.2.2.2 does not actually require this normalizing constant to be computed.)

B.2.2.2 Adaptive grid algorithm for marginal posterior for b

This section outlines an adaptive grid-based sampling method along the lines of those described by Ritter and Tanner (1991; also see Ritter & Tanner, 1992), albeit without the Gibbs sampling aspect, since the current problem only requires sampling from some marginal distribution $f(b)$. The method is grid-based, in that it uses a grid to approximate a complicated function via a piecewise linear function. Each iteration entails the use of a finer grid, and the algorithm is adaptive, in that each new iteration uses the previous approximation with the aim of expanding the grid in an efficient manner. The focus is on approximating the inverse-cdf $F_b^{-1}(q)$, which is eventually applied to $Unif(0,1)$ draws as to yield approximate draws from the target distribution. On each iteration, the grid is expanded at points that approximately correspond with equally-spaced quantiles, as to produce a grid that is finer at regions of higher density.

For purposes of this section, ignore the notation of other sections. Suppose we have some computable likelihood function, $L(x)$, within a closed interval $[b_0, b_1]$. For instance, we could define $L(b) = exp\left(l_0\left(b\right)\right)$, where $l_0(b)$ is the log-likelihood function from Equation (B.14). Let n_i denote the number of points in the grid at iteration i. Suppose we use adaptive

grid-based methods for m iterations, in each of which the grid is expanded by a factor of s, such that $n_i = s \cdot n_{i-1}$ for $i \in \{2, ..., m\}$. For the ith iteration of grid construction, let $\{x_{ij} : j = 1, ..., n_i\}$ be the points at which our likelihood is evaluated; that is, the ith iteration entails computing $L(x_{i1}), ..., L(x_{i\,n_i})$. For notational convenience, assume that in conducting any given iteration i, the points evaluated are reordered, such that $x_{i1} < x_{i2} < ... < x_{i\,n_i}$. Suppose also that the initial iteration contains the interval's endpoints, b_0, b_1, that no points are ever examined below b_0 or above b_1, and that each new iteration entails adding but not subtracting any points. It then follows that for any arbitrary iteration i, we will have that $x_{i1} = b_0$ and $x_{i\,n_i} = b_1$.

Temporarily ignore the iteration subscripts for this paragraph. For a given iteration, we compute $L(x_1), ..., L(x_n)$. Consider the $n-1$ zones $[x_1, x_2], ..., [x_{n-1}, x_n]$ and approximate the density via a uniform density within each zone wherein the density associated with a given zone is proportional to the average density of its endpoints. More formally, define $p_i = \frac{L(X_i)+L(X_{i+1})}{2}$ for $i = 1, ..., n-1$ and approximate the pdf for point $b \in (x_j, x_{j+1})$ as $\hat{f}(b) = \frac{p_j}{\sum_{i=1}^{n-1} p_i(x_{i+1}-x_i)}$. (The endpoints are not defined here since they do not need to be approximated and as they are inconsequential for purposes of estimating the cdf, due to having zero width.) Next, compute the estimated cdf at point x_j for $j \in \{2, ..., n\}$ as $\hat{F}_j = \hat{F}(x_j) = \frac{\sum_{i=1}^{j-1} p_i(x_{i+1}-x_i)}{\sum_{i=1}^{n-1} p_i(x_{i+1}-x_i)}$, and let $\hat{F}_1 = \hat{F}(x_1) = 0$ (since x_1 reflects point b_0). Then, for arbitrary point $b \in [x_j, x_{j+1}]$, we can estimate the cdf as $\hat{F}(b) = \hat{F}_j + \frac{p_j(b-x_j)}{\sum_{i=1}^{n-1} p_i(x_{i+1}-x_i)}$. Likewise, for quantile $q \in [\hat{F}_j, \hat{F}_{j+1}]$, we can estimate the inverse-cdf as $\hat{F}^{-1}(q) = x_j + \frac{q-\hat{F}_j}{\hat{F}_{j+1}-\hat{F}_j}(x_{j+1}-x_j)$.

Hence, we can use an adaptive grid-based method as follows:

1. Iteration 1:

   (a) Let $x_{1j} = b_0 + (b_1 - b_0)\frac{j-1}{n_1-1}$ for $j = 1, ..., n_1$. This ensures that the n_1 points are equally spaced and range from the lower bound to the upper bound.

   (b) Compute $L(x_{1j}), ..., L(x_{1\,n_1})$ via the likelihood function.

   (c) Use the above likelihood in defining an approximate inverse-cdf function, denoted by $\hat{F}_1^{-1}(q)$, using methods described above, and wherein the subscript refers to the iteration.

2. Iteration i for $i = 2, ..., m$:

   (a) Assume that the likelihood was previously evaluated at n_{i-1} points. Given the scaling factor of s, iteration i entails computing the likelihood at an additional $n_i' = (s-1) \cdot n_{i-1}$ points.

   (b) Let $\xi_{ij} = \frac{j}{n_i'+1}$ for $j = 1, ..., n_i'$.

   (c) Compute $\hat{F}_{i-1}^{-1}(\xi_{ij})$ for $j = 1, ..., n_i'$. Combine the new set of points, $\left\{\hat{F}_{i-1}^{-1}(\xi_{ij}) : j = 1, ..., n_i'\right\}$, with the set of previously evaluated points, $\left\{x_{(i-1)j} : j = 1, ..., n_{i-1}\right\}$, de-duplicating any points that exist in both sets, and reorder as $\{x_{ij} : j = 1, ..., n_i\}$, such that $x_{ij} < x_{ik}$ for $j < k$, and where $n_i = n_{i-1} + n_i'$ (assuming no duplicates were removed).

   (d) Compute $L(.)$ for any newly added points.

   (e) Use the new data to define an updated function that approximates the inverse-cdf, $\hat{F}_i^{-1}(q)$.

3. Suppose that we want to simulate R approximate draws from our distribution of interest. Let $U_i \overset{iid}{\sim} Unif(0,1)$ for $i = 1, ..., R$. The ith draw is then $\hat{F}_m^{-1}(U_i)$, thereby making use of the final inverse-cdf approximation from step 2(e) above.

Note that in the algorithm above, some choices for the grid's expansion-factor s could be inefficient. For example, if $s = 2$, then roughly half of the new set of quantiles to evaluate will be very close to previously-evaluated quantiles. This issue can be mitigated to some extent by simply choosing a large enough s and incorporating a fractional component.

The resulting draws are approximate. If exact results were needed, the algorithm could be modified by adding on an accept-reject layer, wherein the piecewise linear density described above is only used as a candidate density. However, this would come at a cost to efficiency, and it may also be difficult to demonstrate that the candidate density, multiplied by a constant, is greater than the actual density at all points.

For purposes of the MC estimator in §3.4.1, we specified $(n_1 = 20, s = 3.49, m = 6)$ for outer steps (resulting in roughly 10,400 evaluations of $L(b|D1)$ to approximate a given inverse-cdf) and $(n_1 = 20, s = 3.49, m = 4)$ for inner steps (resulting in roughly 850 evaluations of $L(b|D2)$ to approximate the inverse-cdf for a given main study data set). Other uses of adaptive grid sampling in our paper generally used $(n_1 = 20, s = 3.49, m = 6)$, the only exception being that we set $m = 7$ when the inverse-cdf was needed for computing PB allocations (e.g., as an input for the $\mathrm{E}\left(\mathbf{n}^{opt}|D1\right)$ allocation described in §3.3.4).

B.2.3 Choice of R1 and R2 (detailed description)

Identifying an optimal allocation requires searching the design space, which will entail numerous evaluations of Equation (3.15). In doing so, $R1$ and $R2$ must be chosen to provide adequate accuracy within available runtime. We manage such accuracy-runtime tradeoffs by treating the choice of $R1$ and $R2$ as an intermediate optimization problem, noting parallels to two-stage sampling.

To simplify notation, let $f_{ij} = E^{(r_1,r_2)}$ and $g_{ij} = V^{(r_1,r_2)}$ (where the righthand terms were defined in §3.4.1), and temporarily use z_{ij} to denote a PSU-like quantity. Then, we can rewrite Equation (3.15) as:

$$\bar{\bar{z}} := \underset{D2|D1}{\hat{\mathrm{E}}} \mathrm{Var}(\bar{Y}|D2) = \frac{1}{R1 \cdot R2} \sum_{i=1}^{R1} \sum_{j=1}^{R2} z_{ij} \tag{B.15}$$

where $z_{ij} := g_{ij} + \frac{R2}{R2-1}\left(f_{ij} - \bar{f}_i\right)^2$ and $\bar{f}_i := \frac{1}{R2}\sum_{j=1}^{R2} f_{ij}$. Note that Equation (B.15) has the same form as the overall sample mean, $\bar{\bar{y}}$, from §10.3 of Cochran (1977), which considers two-stage sampling from a finite population. For large $R2$, $Cov(z_{ij}, z_{ij'}) \approx 0$ for $j \neq j'$, so $z_{i1}, z_{i2}, ..., z_{ik}$ are approximately independent and come from some common distribution. Given that our MC sampling process is roughly analogous to two-stage sampling from an infinite population, we can draw upon the finite population results from Cochran (1977, §10.3–§10.4) by ignoring fpc terms to obtain the variance of our MC estimator as a function of $R1$ and $R2$ (conditional on $D2|D1$, e_1, and the adaptive grid sampling specifications).

Assuming large $R2$, the variance of Equation (B.15) is approximately

$$\text{Var}\left(\bar{\bar{z}}\right) \doteq \frac{S_1^2}{R1} + \frac{S_2^2}{R1 \cdot R2}, \tag{B.16}$$

where S_1^2 denotes the variance across primary units means and S_2^2 denotes the variance among subunits within primary units, each considered with respect to the underlying superpopulation. (For example, consider some sequence of finite populations, $\{U_{IJ}\}$, where $U_{IJ} = \{Z_{ij} : i = 1, ..., I; j = 1, ..., J\}$, and where Z_{ij} is the relevant PSU-type quantity obtained via MC sampling. Then, roughly, consider S_1^2 and S_2^2 from Equation [B.16] as the limits of their finite population counterparts as $I, J \to \infty$.) We then estimate the approximate variance for given $R1$ and $R2$ by plugging high-quality estimates of S_1^2 and S_2^2 for some given allocation into Equation (B.15), yielding

$$\text{var}\left(\bar{\bar{z}}|R1, R2\right) \doteq \frac{\hat{S}_1^2}{R1} + \frac{\hat{S}_2^2}{R1 \cdot R2}. \tag{B.17}$$

For example, let $R1' = R2' = 10000$, and compute $\hat{S}_1^2 = s_1^2 - \frac{s_2^2}{R2'}$ and $\hat{S}_2^2 = s_2^2$, where $s_1^2 = \frac{1}{R1'-1} \sum_{i=1}^{R1'} \left(\bar{z}_i - \bar{\bar{z}}\right)^2$ and $s_2^2 = \frac{1}{R1'(R2'-1)} \sum_{i=1}^{R1'} \sum_{j=1}^{R2'} \left(z_{ij} - \bar{z}_i\right)^2$.

Next, consider runtime for evaluating the objective function as an approximate function of $R1$ and $R2$. We assume that the procedures in §3.4.1 approximately follow some function

$$t(R1, R2) \doteq \beta_1 \cdot R1 + \beta_2 \cdot R1 \cdot R2, \tag{B.18}$$

again implicitly conditioning on $D2|D1$, e_1, and the adaptive grid sampling specifications.

Plugging in estimates for β_1 and β_2, we can then identify the choice of $(R1, R2)$ that minimizes $\text{var}\,(\bar{\bar{z}}|R1, R2)$ subject to $\hat{t}(R1, R2) \leq C$ and for some budgeted runtime C, or we can minimize $\hat{t}(R1, R2)$ for some allowed precision of the MC estimator. When doing so, it is reasonable to apply some approximations, since the choices for $R1$ and $R2$ need not be exactly optimal for minimizing runtime. For instance, estimate S_1^2 and S_2^2 at a single allocation (e.g., the $\hat{b}_{median}$ allocation) rather than at a large number of allocations, to simplify this process.

To illustrate use of these expressions for planning stochastic optimization runs, here we briefly summarize a preliminary application. For an initial near-optimal candidate allocation, we estimated $\hat{S}_1^2$, $\hat{S}_2^2$, and $\hat{\bar{\bar{Z}}}$ via $R1' = R2' = 10000$. Runtime was estimated as being roughly $\hat{t}(R1, R2) \doteq 0.025 \cdot R1 + 0.00012 \cdot R1 \cdot R2$ seconds for a reasonably accurate grid sampling specification, based on considering 25 levels of $(R1, R2)$ (choosing each as 50, 100, 200, 400, or 800). Using a CV target of 0.005, and requiring that $R1, R2 \geq 2$, runtime was minimized at $R1 = 3895$, $R2 = 2$ (and where other evaluated design points yielded similarly small $R2$). Rather than use these quantities directly, we specified $R1 = 3000$ and $R2 = 200$, given the earlier assumption of large $R2$ (alternatively, a lower bound of $R2 \geq 200$ could be set). Compared with a naive selection of $R1 = R2 = 1000$, the resulting choice was anticipated to yield a two-thirds reduction in variance for similar estimated runtime, while allowing for an estimated 588 objective function evaluations per day with an estimated CV of 0.00565.

Given that this analysis involved some simplifying assumptions, the choice of $(R1, R2)$ could be considered iteratively, if needed. For example, after making some initial choice, refine $(R1, R2)$ as needed to achieve faster runtime and/or increased precision. If runtime is still

too large and precision too small, consider this choice in conjunction with other parameter choices affecting runtime and accuracy of the MC estimator.

B.3 Alternative MC estimation approach: sampling from predictive distribution

Appendix B.3.1 outlines a general MC sampling approach for averaging the posterior loss over the predictive distribution for a given allocation, with details provided in Appendix B.3.2. The main difference (in comparison with §3.4.1) is that the alternative approach described below changes the inner MC step to obtain direct predictions from $f(\bar{Y}|D2)$. Given that this alternative approach does not make use of distributional information specific to the finite population mean, it is more easily adapted toward other inferential objectives, but at a potentially considerable computational cost for large populations.

B.3.1 Overview of alternative approach

Suppose that we aimed to estimate $\hat{\mathrm{E}}_{D2|D1} \mathrm{Var}(\bar{Y}|D2)$, but without making use of the analytic results from Equations (3.4) and (3.5). Our interest is in simulating

$$\mathrm{E}_{D2|D1} \mathrm{Var}(\bar{Y}|D2) = \int \mathrm{Var}\left(\bar{Y}|D2\right) f(D2|D1)\, \mathrm{d}D2 \tag{B.19}$$

for a given design vector $e_1 = \{n_h\}$. Let $\phi = (b, \boldsymbol{\alpha}, \boldsymbol{v})$ refer to the list of superpopulation parameters, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_H)$ and $\boldsymbol{v} = (v_1, ..., v_H)$. We can simulate Equation (B.19) via two nested steps, wherein the outer step entails drawing $R1$ main study data sets from the distribution $f(D2|D1)$ and the inner step entails drawing $R2$ predictions of $\bar{Y}$ from $f(\bar{Y}|D2)$. As in §3.3.2, we assume a diffuse prior of $\pi\left(b, \left\{\alpha_h, \frac{1}{v_h}\right\}\right) \propto \prod_{h=1}^{H} v_h I_{[b_0,b_1]}(b)$, where $I_{[b_0,b_1]}(b)$ is an indicator of whether $b_0 \leq b \leq b_1$. Given that the pilot data are used only for study design and not for ultimate inferences, this prior is applied at each phase (i.e., whether considering $f(\phi|D1)$ or $f(\phi|D2)$). A general MC sampling procedure is summarized below (see Appendix B.3.2 for detail):

1. Draw R1 main study data sets via $f(D2|D1) = \int f_1(D2|\phi)\, f_2(\phi|D1)\, \mathrm{d}\phi$, using the same procedures summarized as the outer step in §3.4.1.1 and detailed in Appendix B.2.1.

2. For each main study data set, draw $R2$ predictions of the finite population mean, $\bar{Y}$, via $f\left(\bar{Y}|D2\right) = \int f_1\left(\bar{Y}|\phi\right) f_2(\phi|D2)\, \mathrm{d}\phi$. In doing so, a given prediction, $\bar{Y}^{(r_1,r_2)}$, is obtained by sampling from the associated posterior distribution for the superpopulation parameters, $(b, \{v_h, \alpha_h\}) \,|\, \left(D2 = d_2^{(r_1)}\right)$, using these sampled parameters to compute predictions for the set of $N_h - n_h$ nonsampled units, and then combining these predictions with information about the sampled units to predict the finite population mean.

After applying the above steps, our estimator for the quantity $\zeta = \underset{D2|D1}{\mathrm{E}} \mathrm{Var}\left(\bar{Y}|D2\right)$ is

$$\hat{\zeta} = \frac{1}{R1} \sum_{r_1=1}^{R1} V^{(r_1)}, \tag{B.20}$$

where

$$V^{(r_1)} = \frac{1}{R2-1} \sum_{r_2=1}^{R2} \left(\bar{Y}^{(r_1,r_2)} - \bar{\bar{Y}}^{(r_1)}\right)^2, \tag{B.21}$$

$\bar{\bar{Y}}^{(r_1)} := \frac{1}{R2} \sum_{r_2=1}^{R2} \bar{Y}^{(r_1,r_2)}$, and where, as before, $\bar{Y}^{(r_1,r_2)}$ denotes the (r_2)nd prediction for the finite population mean associated with the (r_1)st draw of D2, for $r_1 = 1, ..., R1$ and $r_2 = 1, ..., R2$.

B.3.2 Detailed overview of alternative approach

This section provides additional detail on the procedure described in §B.3.1 for sampling from $\underset{D2|D1}{\mathrm{E}} \mathrm{Var}(\bar{Y}|D2) = \int \mathrm{Var}\left(\bar{Y}|D2\right) f(D2|D1)\, \mathrm{d}D2$. The two steps are as follows:

1. Simulate $R1$ draws from $D2|D1$, using the procedures outlined in Appendix B.2.1, and where a given draw results in a set of sample indicators and sample data.

2. For a given draw r_1 from step (1) above, draw $R2$ complete sets of predictions for the nonsampled units from $f(Y_{ns}|D2) = \int f_1(Y_{ns}|b, \{\alpha_h, v_h\}) f_2(b, \{\alpha_h, v_h\}\,|D2)\, \mathrm{d}b\, \mathrm{d}\,\{\alpha_h, v_h\}$. For each complete set of predictions for the nonsampled data, compute the predicted finite population mean. This results in a set of $R2$ predictions $\left\{\bar{Y}^{(r_1,r_2)} : r_2 = 1, ..., R2\right\}$,

to be associated with the data set, $d_2^{(r_1)}$, from step (1) above. Specific steps are as follows:

(a) Draw a set of superpopulation parameters $\left(b^{(r_1,r_2)}, \left\{\alpha_h^{(r_1,r_2)}, v_h^{(r_1,r_2)}\right\}\right)$ from the posterior distribution $b, \{\alpha_h, v_h\} \,|D2$, with the sample fixed at $\left(s_2^{(r_1)}, d_2^{(r_1)}\right)$. This is analogous to step 1 of Appendix B.2.1, except that the relevant posterior distributions are conditioned on the (simulated) main study sample data rather than the (actual) pilot sample data. In particular, we obtain $R2$ draws from $b|d_2^{(r_1)}$ and then use these to sample from $\{\alpha_h, v_h\}|b^{(r_1,r_2)}, d_2^{(r_1)}$, the latter step making use of the distributional properties and formulas from Appendix A.2.2.2 (i.e., given that $(\alpha_h, \frac{1}{v_h})$ are conditionally normal-gamma).

(b) Use these superpopulation parameters and sample unit indices to fill out the Y's for the nonsampled units, $\left(ns_2^{(r_1)}\right)$, which are associated with the auxiliary data, $x_2(ns)^{(r_1)}$. Specifically, for non-sampled unit hi within simulation r_1, for $h = 1, ..., H$ and $i = n_h + 1, ..., N_h$, draw $Y_{hi}^{(r_1,r_2)}$ from $N\left(\alpha_h^{(r_1,r_2)} X_{hi}^{(r_1)}, v_h^{(r_1,r_2)} Z_{hi}^{(r_1,r_2)}\right)$ where $Z_{hi}^{(r_1,r_2)} = \left(X_{hi}^{(r_1)}\right)^{b^{(r_1,r_2)}}$.

(c) Compute the simulated population mean as $\bar{Y}^{(r_1,r_2)} = \frac{n}{N}\bar{Y}_s^{(r_1)} + \frac{N-n}{N}\bar{Y}_{ns}^{(r_1,r_2)}$, where $n = \sum_{h=1}^{H} n_h$ and $N = \sum_{h=1}^{H} N_h$. Specifically:

- Compute the sample mean via $\bar{Y}_s^{(r_1)} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi}^{(r_1)}$. (Note that this does not need to be repeated $R2$ times, since it is constant within each outer step.)
- Compute the mean for non-sampled units as $\bar{Y}_{ns}^{(r_1,r_2)} = \frac{1}{N-n}\sum_{h=1}^{H}\sum_{i=n_h+1}^{N_h} Y_{hi}^{(r_1,r_2)}$.
- Note that this last step (2c) is specific to our particular inferential objec-

tive. Other inferential quantities of interest could be straightforwardly accommodated through modifications to this last step, especially if they reflect readily-computed functions of the sample data.

Use the simulated data above to compute the quantities from Equations (B.20–B.21).

B.4 SPSA modification for implicitly defined bound constraints

This section provides additional detail to supplement §3.4.5.

B.4.1 Approximate unbiasedness of modified gradient estimator

Assume use of the same notation and assumptions from §3.4.5. We aim to show that Equation (3.24) is an approximately unbiased estimate for $L'_d(\theta_k)$.

As per §3.4.5, a first-order Taylor series approximation yields $\mathrm{E}\left(\hat{g}_{kd}(\theta_k)|\theta_k\right) \doteq L'_d(\theta_k) + \sum_{i \neq d} L'_i(\theta_k)\mathrm{E}\left(\frac{\Delta_{ki}}{\Delta_{kd}}\right)$. This can be rewritten as

$$\mathrm{E}\left(\hat{g}_{kd}(\theta_k)|\theta_k\right) \doteq L'_d(\theta_k)(1-\rho) + D\rho\bar{L}'_k, \tag{B.22}$$

where $\bar{L}'_k := \frac{1}{D}\sum_{d=1}^{D} L'_d(\theta_k)$ is the average term of the (true, unknown) gradient when eval-

uated at θ_k. Therefore,

$$\mathrm{E}\left(\frac{\hat{g}_{kd}(\theta_k) - D\rho\bar{L}'_k}{1-\rho}\middle|\theta_k\right) \doteq L'_d(\theta_k). \tag{B.23}$$

In the lefthand side above, ρ is straightforward to compute, and all other terms are known except for $\bar{L}'_k$, which must be estimated. A naïve but badly biased estimate of $\bar{L}'_k$ can be computed by averaging the D components from Equation (3.22) via $\frac{1}{D}\sum_d^D \hat{g}_{kd}(\theta_k)$. Through D applications of (B.22), we have that

$$\begin{aligned}\mathrm{E}\left(\sum_{d=1}^{D}\hat{g}_{kd}(\theta_k)\middle|\theta_k\right) &\doteq D\bar{L}'_k(1-\rho) + D^2\rho\bar{L}'_k \\ &= D\bar{L}'_k(1-\rho+D\rho).\end{aligned} \tag{B.24}$$

It follows that $\hat{\bar{g}}_k(\theta_k) = \frac{\sum_{d=1}^{D}\hat{g}_{kd}(\theta_k)}{D(1-\rho+D\rho)}$ is an approximately unbiased estimator of $\bar{L}'_k$. From there, it follows that $\hat{\tilde{g}}_{kd}(\theta_k) = \frac{\hat{g}_{kd}(\theta_k) - D\rho\hat{\bar{g}}_k(\theta_k)}{1-\rho}$ is an approximately unbiased estimator of $L'_d(\theta_k)$.

B.4.2 Correlation between perturbation components under modified procedure

Assume use of the same notation and assumptions from §3.4.5. Additionally, assume that Δ_k is drawn via the two-step procedure suggested in §3.4.5 (i.e., draw scalar $B_k \sim$ symmetric Bernoulli(± 1); then, randomly assign $\lfloor\frac{D-1}{2}\rfloor$ elements of Δ_k to each receive the value B_k, with elements of the complementary set each receiving the value of $-B_k$). For practical

purposes, we also assume that $D > 2$, which avoids some illogical results (e.g., perfectly correlated decision variables for $D = 2$).

Note that for fixed D, the correlation between different components of Δ_k is $\rho_{\Delta_{ki},\Delta_{kj}} = \mathrm{E}(\Delta_{ki}\Delta_{kj})$, for $i \neq j$, since the marginal distributions have mean 0 and standard deviation 1. For brevity, we write the correlation as ρ, noting the exchangeability of components. From marginal distributional properties, this becomes $\rho = \Pr(\Delta_{ki} = \Delta_{kj}) - \Pr(\Delta_{ki} \neq \Delta_{kj}) = 2 \cdot \Pr(\Delta_{ki} = \Delta_{kj}) - 1$, again assuming $i \neq j$.

If D is odd, then through basic probability theory, $\Pr(\Delta_{ki} = \Delta_{kj} | i \neq j) = \frac{(D-1)/2}{D} \cdot \frac{(D-3)/2}{D-1} + \frac{(D+1)/2}{D} \cdot \frac{(D-1)/2}{D-1} = \frac{D-1}{2D}$, and so $\rho = \frac{-1}{D}$. Likewise, if D is even, then $\Pr(\Delta_{ki} = \Delta_{kj} | i \neq j) = \frac{\frac{D}{2}-1}{D} \cdot \frac{\frac{D}{2}-2}{D-1} + \frac{\frac{D}{2}+1}{D} \cdot \frac{\frac{D}{2}}{D-1} = \frac{D^2-2D+4}{2D(D-1)}$, and so $\rho = \frac{-D+4}{D(D-1)} = \frac{-1}{D} + \frac{3}{D(D-1)}$.

B.5 Modified Nelder–Mead overview

The Nelder–Mead algorithm (Nelder & Mead, 1965) is a commonly used simplex-based optimization method in which the geometry of the simplex adapts to the function's response surfaces through a series of geometric operations. Each iteration of the algorithm entails a sequence of decision rules that typically involve replacing the worst point of the simplex with a new point on the line containing the previous worst point and the centroid of the remaining points; occasionally, the simplex is contracted at the previous best point. Details on the procedure are well documented elsewhere (e.g., Conn et al., 2009; Kelley, 1999b). Although there are some functions for which Nelder–Mead can fail to converge (Dennis & Woods,

1987) or can converge to a nonstationary point (McKinnon, 1998), its enduring popularity has been attributed to factors such as the simplicity with which it can be implemented and the fact that it can perform well in many practical situations, often producing rapid improvements in parameter evaluations without requiring an inordinate number of objective function evaluations (e.g., Conn et al., 2009; Lagarias et al., 1998). In situations where the simplex geometry deteriorates or progress stagnates, the procedure can be restarted with a new simplex (e.g., Conn et al., 2009; Kelley, 1999a). Although Nelder–Mead is an unconstrained optimization method, it can be adapted to constrained settings with minor modifications; other modifications can be made to address certain limitations or improve its performance in particular settings. For instance, Huan and Marzouk (2013) used a modified version of Nelder–Mead in the context of optimal Bayesian experimental design; they suggested that by using only a relative ordering, it may have a natural advantage in noisy optimization settings over methods that involve estimating the gradient.

For this study, the starting point for our optimization methods was Conn et al.'s modern interpretation of Nelder–Mead (2009, pp. 141–145), which addresses some ambiguities in the original algorithm with respect to inequalities and tie-breaking. We handle constraints in a similar manner as is used in Box's (1965) constrained simplex method: a point violating an explicit constraint (i.e., constant upper or lower bound on a decision variable) is moved to the nearest feasible point inside the simplex, whereas a point violating an implicit constraint (i.e., bound on a function of the decision variables) is repeatedly moved toward the centroid of the already selected points until a feasible point is obtained. We use Barton and Ivey's (1996) S9 + RS modifications to the shrink step, which aim to prevent the simplex from retaining

points that reflect unusually good measurements and from contracting too fast toward such points, and which they found to improve performance for noisy functions at a moderate cost to efficiency. We further incorporate two types of simplex restarts for scenarios where the simplex stagnates or the geometry collapses. First, we use Kelley's (1999a) method for detecting and remedying stagnation via oriented restarts, although we allowed the algorithm to continue (rather than terminating) after reaching the maximum oriented restarts, since this led to noticeably improved performance for one of Kelley's test cases. (In addition, we also modified the oriented restart to incorporate the earlier S9 + RS modifications, since oriented restarts are a shrink-like operation.) When the simplex geometry collapses (e.g., due sometimes to Box's method for handling constraints, as noted by Le Floc'h, 2012), we also allow a second type of restart, where we retain the previous best vertex and randomly select additional vertices from feasible points in its vicinity.

There are many methods for choosing a starting simplex (e.g., Baudin, 2010, pp. 16–20), and which may affect the speed of the algorithm (e.g., Parkinson & Hutchinson, 1972). We employed the axis-by-axis method (e.g., Baudin, 2010, pp. 18–19), with constant step size. This axis-by-axis method has some similarities to the default initial simplex method used in Matlab R2019a (i.e., Pfeffer's method), except that the latter incorporates scaling in determining new points, which could cause some problematic initial simplices in some sample allocation settings. For purposes of stopping rules, we terminate the modified Nelder–Mead algorithm once the simplex has approximately converged to a point (based on the simplex diameter) or once the maximum number of iterations has been reached.

B.6 Additional simulation figures and tables

For each of the pseudo-Bayesian and Bayesian allocations, Table B.1 summarizes simulation results for each of the three inferential methods described in §3.5.2.

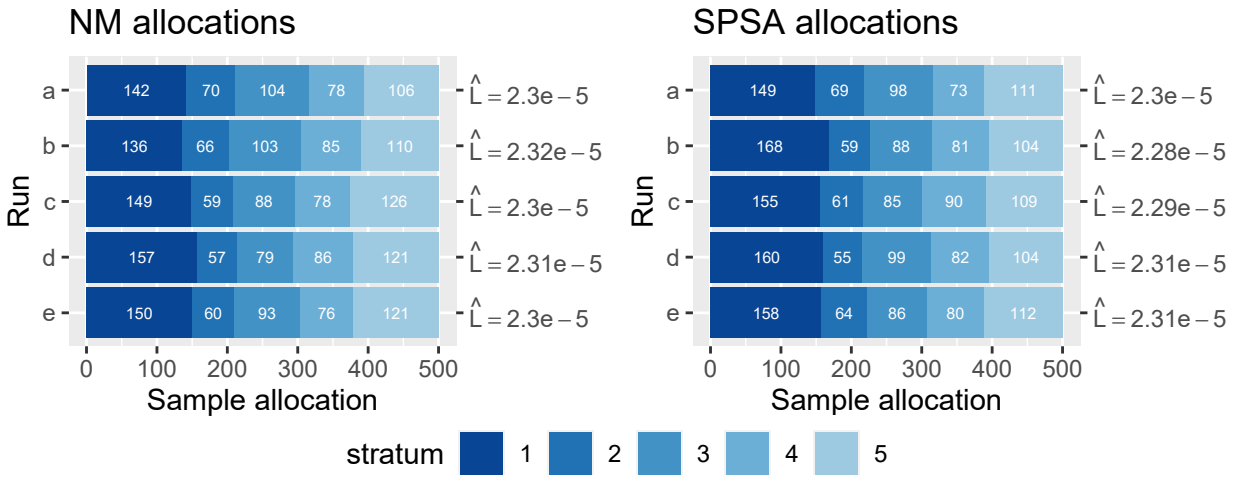
Table B.1: Simulation results by detailed strategy and metric (for PB and B allocations)

Allocation	Inferential method	Relative RMSE	1000 * RelBias	CI Coverage (%)	1000 * CI RelWidth
PB-MAP	1. Conditional posterior moments	1.006	0.54	95.1	41.0
	2. Unconditional posterior moments	1.006	0.54	95.1	41.1
	3. Unconditional posterior sampling	1.006	0.54	95.1	41.1
PB-mean	1. Conditional posterior moments	1.003	0.60	95.0	40.9
	2. Unconditional posterior moments	1.003	0.60	95.1	41.0
	3. Unconditional posterior sampling	1.003	0.60	95.1	41.1
PB-median	1. Conditional posterior moments	1.000	0.61	95.2	41.0
	2. Unconditional posterior moments	1.000	0.61	95.3	41.1
	3. Unconditional posterior sampling	1.000	0.61	95.2	41.1
PB-nopt	1. Conditional posterior moments	1.004	0.80	95.1	40.9
	2. Unconditional posterior moments	1.004	0.80	95.1	41.0
	3. Unconditional posterior sampling	1.004	0.80	95.1	41.0
B-NM	1. Conditional posterior moments	0.999	0.47	95.1	41.0
	2. Unconditional posterior moments	0.999	0.47	95.2	41.1
	3. Unconditional posterior sampling	0.999	0.47	95.2	41.1
B-SPSA	1. Conditional posterior moments	1.000	0.65	95.1	40.9
	2. Unconditional posterior moments	1.000	0.65	95.2	41.0
	3. Unconditional posterior sampling	1.000	0.65	95.2	41.0

Note: RMSE is displayed relative to that of the B-SPSA / unconditional posterior sampling strategy.

Figure B.1 provides the sample allocations (and associated estimated losses) from repeated applications of NM or SPSA to the first pilot from the preliminary simulation (as described in §3.5.3.3). In this example, although there were modest differences in sample allocations across repeated applications of a given method (e.g., NM allocations to stratum 3 range from 79 to 104), there were no meaningful differences in estimated loss.

Figure B.1: Preliminary simulation: allocations by method and run for Pilot 1



Note. In each panel, the rows are annotated on the righthand side with the corresponding MC-estimated losses (which had estimated CVs averaging 0.5%).

B.7 Simulation results for b = 1 population

Table B.2: b = 1 population: Simulation results by detailed strategy and metric (for PB and B allocations)

Allocation	Inferential method	Relative RMSE	1000 * RelBias	CI Coverage (%)	1000 * CI RelWidth
PB-MAP	1. Conditional posterior moments	1.004	0.03	95.0	41.4
	2. Unconditional posterior moments	1.004	0.03	95.0	41.5
	3. Unconditional posterior sampling	1.005	0.03	95.0	41.5
PB-mean	1. Conditional posterior moments	0.999	-0.10	95.2	41.4
	2. Unconditional posterior moments	0.999	-0.09	95.2	41.5
	3. Unconditional posterior sampling	0.999	-0.09	95.2	41.5
PB-median	1. Conditional posterior moments	1.002	-0.04	95.0	41.4
	2. Unconditional posterior moments	1.002	-0.04	95.0	41.5
	3. Unconditional posterior sampling	1.002	-0.04	95.0	41.5
PB-nopt	1. Conditional posterior moments	1.005	-0.06	94.8	41.3
	2. Unconditional posterior moments	1.005	-0.06	94.9	41.4
	3. Unconditional posterior sampling	1.005	-0.06	94.9	41.4
B-NM	1. Conditional posterior moments	0.996	-0.07	95.2	41.3
	2. Unconditional posterior moments	0.996	-0.07	95.2	41.4
	3. Unconditional posterior sampling	0.996	-0.07	95.3	41.4
B-SPSA	1. Conditional posterior moments	1.000	0.03	95.1	41.3
	2. Unconditional posterior moments	1.000	0.03	95.1	41.4
	3. Unconditional posterior sampling	1.000	0.03	95.1	41.4

Note: RMSE is displayed relative to that of the B-SPSA / unconditional posterior sampling strategy.

Figure B.2: b = 1 population: Estimated loss box plots by pilot-based allocation method

Figure B.3: b = 1 population: Estimated loss box plots by rule of thumb-based allocation method (relative to B-SPSA)

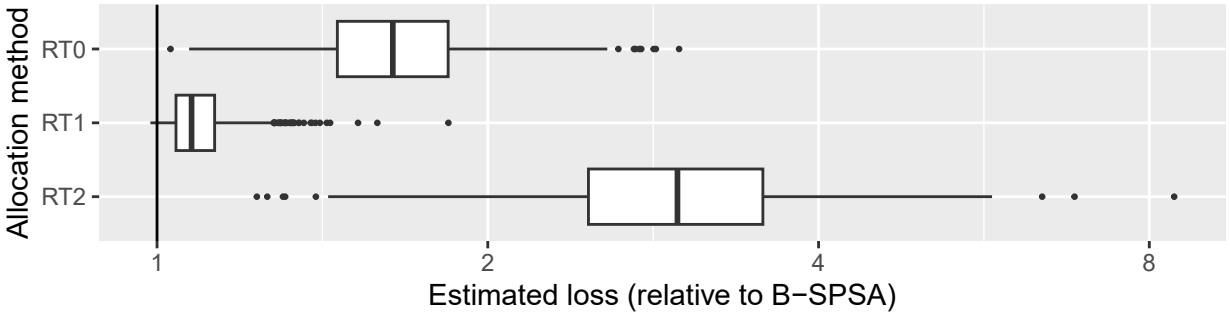


Table B.3: b = 1 population: Simulation results

Strategy	Relative RMSE	1000 * RelBias	CI Coverage (%)	1000 * CI RelWidth
N-HT	1.430	0.14	94.8	58.9
C-SR	1.015	0.01	94.8	41.4
PB-MAP	1.005	0.03	95.0	41.5
PB-mean	0.999	-0.09	95.2	41.5
PB-median	1.002	-0.04	95.0	41.5
PB-nopt	1.005	-0.06	94.9	41.4
B-NM	0.996	-0.07	95.3	41.4
B-SPSA	1.000	0.03	95.1	41.4
RT0-SR	1.234	-0.06	94.7	50.5
RT1-SR	1.007	-0.11	94.5	40.9
RT2-SR	1.574	-0.10	92.8	61.8

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

Figure B.4: b = 1 population: Average allocations to strata by method, across simulations

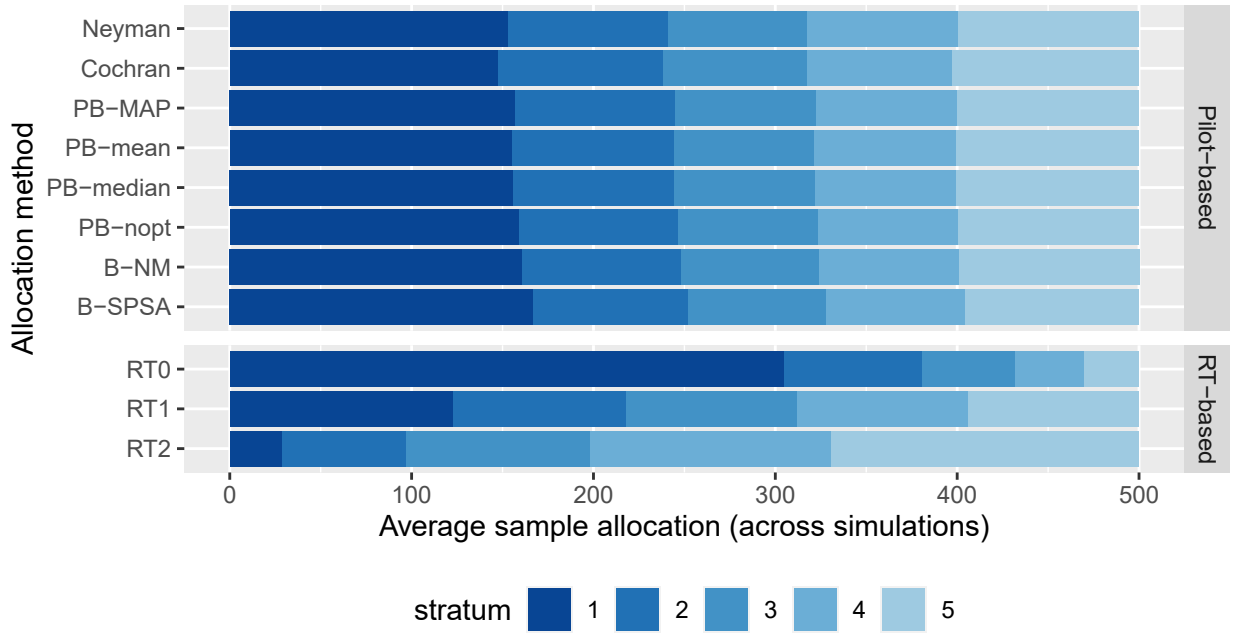
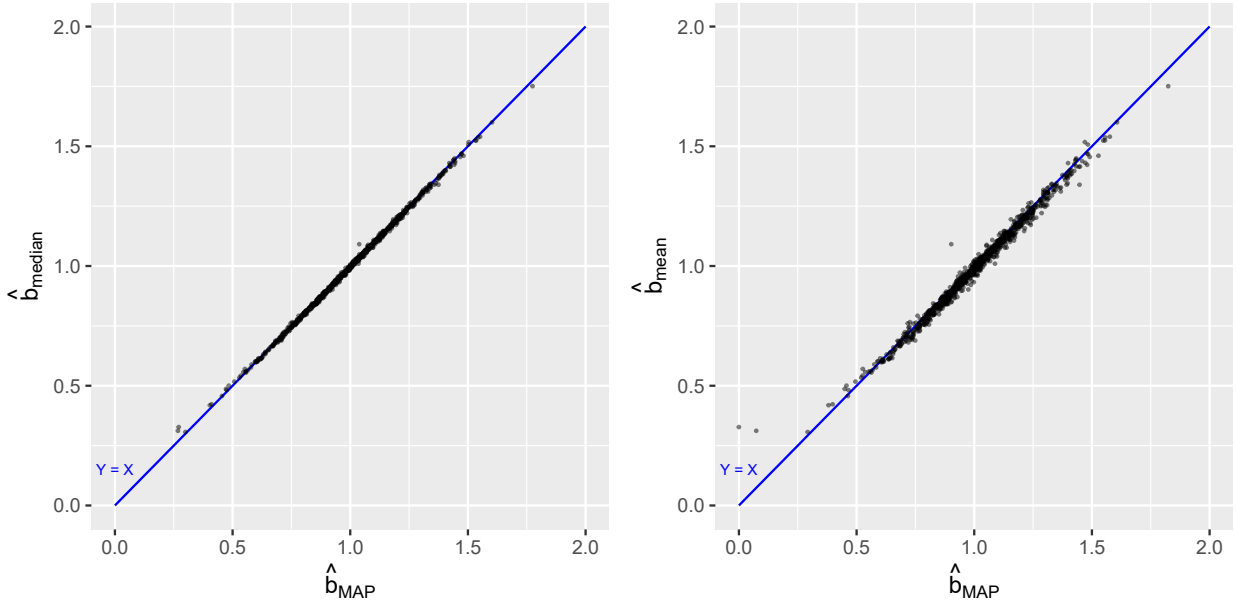
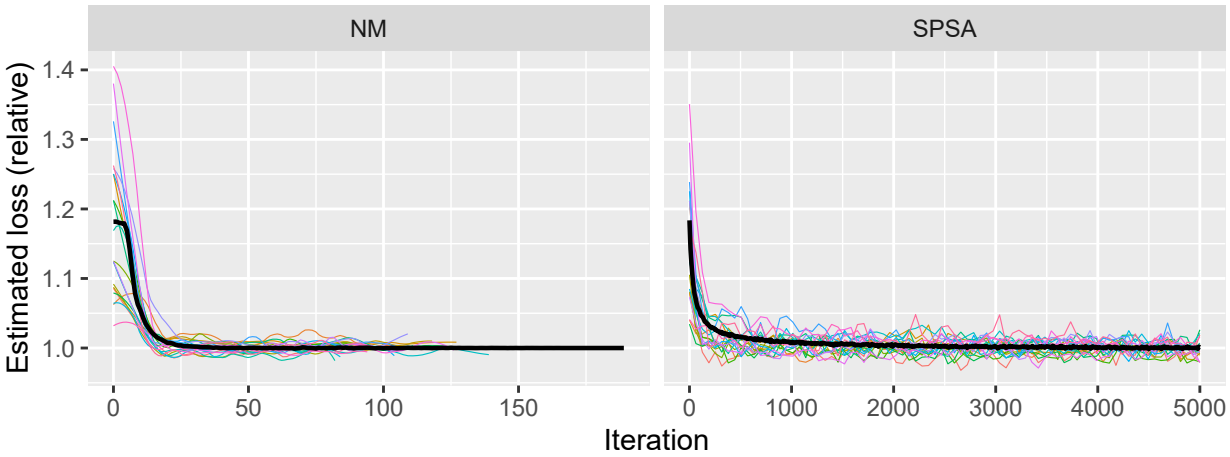


Figure B.5: b = 1 population: Estimated b by method: MAP versus posterior median and mean.



Note. Estimates are based on pilot data; each point denotes results for one pilot.

Figure B.6: b = 1 population: Estimated loss at candidate allocation by iteration and method, overall and for randomly selected pilots



Note. Estimated losses were computed following §3.5.4 (second-to-last paragraph), with estimated CVs of 3.4% for SPSA and 1.4% for NM, and were considered relative to the high quality estimated loss at the final allocation (via §3.5.3.1). Overall results (black line) denote averages across all 1000 pilots; pilot-specific results (thin colored lines) are LOESS-smoothed curves, shown for 20 randomly selected pilots. Given that the number of iterations varied across NM runs, for purposes of averaging results across pilots, the estimated loss for iterations after convergence was assumed to be equal to the high quality estimated loss at the final allocation.

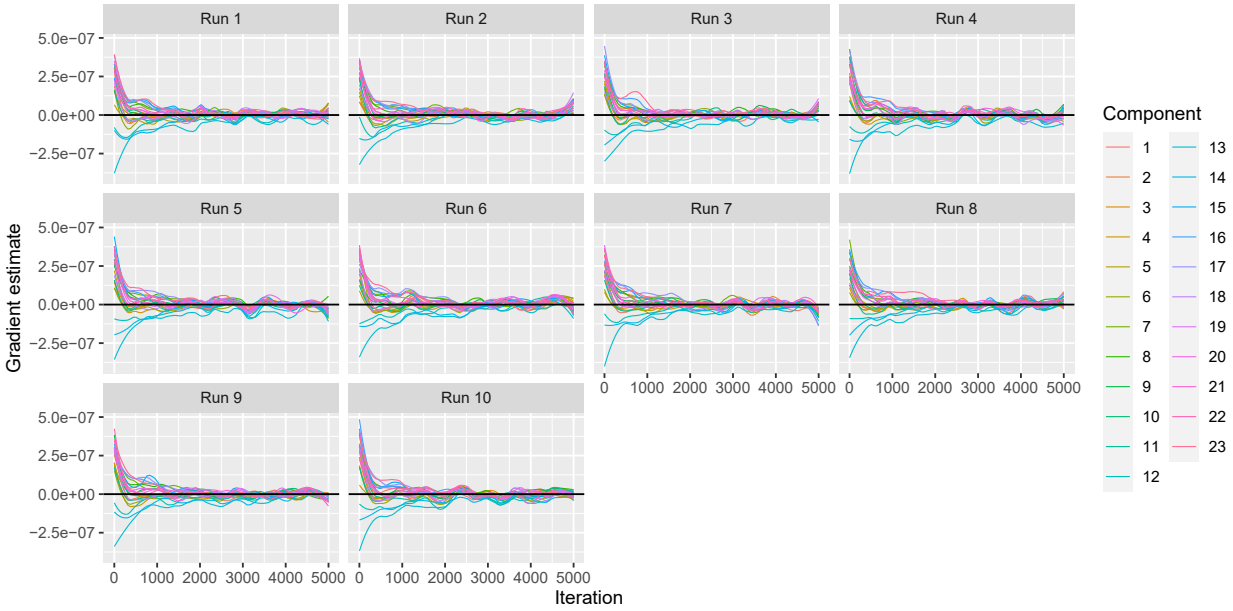
B.8 Additional NCCS application figures and tables

Table B.4: NCCS sample allocations by stratum and method

h	Nonprofit Sector	Revenue Class	Neyman	Cochran	PB-MAP	PB-mean	PB-median	PB-nopt	B-NM*	B-SPSA*	RT0	RT1	RT2
1	Arts, culture, and humanities	Under $300k	55	73	73	73	73	73	66	74	84	79	74
2	Arts, culture, and humanities	$300k-$1.2m	42	36	37	37	37	37	44	38	62	62	61
3	Arts, culture, and humanities	$1.2m+	61	65	65	66	66	66	76	67	37	39	41
4	Education	Under $300k	51	84	83	83	83	83	101	82	72	68	64
5	Education	$300k-$1.2m	44	61	62	61	62	61	64	62	75	75	74
6	Education	$1.2m+	130	81	80	81	81	81	84	82	101	108	115
7	Environment	Under $300k	16	23	24	24	24	24	28	25	34	32	30
8	Environment	$300k-$1.2m	20	26	28	28	28	28	32	28	28	28	28
9	Environment	$1.2m+	32	41	42	41	41	41	49	42	17	18	19
10	Health	Under $300k	30	53	54	54	54	54	54	55	58	55	52
11	Health	$300k-$1.2m	51	74	74	74	74	74	86	76	72	72	71
12	Health	$1.2m+	373	130	128	125	126	125	120	122	150	163	178
13	Human services	Under $300k	290	440	430	429	429	429	337	436	255	240	226
14	Human services	$300k-$1.2m	114	190	187	187	187	187	184	173	254	251	248
15	Human services	$1.2m+	260	142	140	143	142	143	131	137	227	241	255
16	International	Under $300k	13	21	22	22	22	22	29	23	15	14	13
17	International	$300k-$1.2m	12	17	18	18	18	18	21	19	12	12	12
18	International	$1.2m+	16	14	16	16	16	16	23	16	12	12	13
19	Public and societal benefit	Under $300k	42	63	64	64	64	64	66	65	53	50	47
20	Public and societal benefit	$300k-$1.2m	30	43	44	44	44	44	54	45	49	49	48
21	Public and societal benefit	$1.2m+	46	31	33	33	33	33	37	33	37	39	41
22	Religion	Under $300k	35	47	48	48	48	48	53	49	49	46	43
23	Religion	$300k-$1.2m	19	33	34	35	34	35	36	36	32	31	31
24	Religion	$1.2m+	18	12	14	14	14	14	26	15	15	16	16

* B-NM and B-SPSA allocations are averaged across the ten runs.

Figure B.7: Smoothed SPSA-estimated gradient components by run and iteration



Curves denote LOESS-smoothed SPSA-estimated gradient components (span of .2). Each panel represents a run, with separate colors for the different decision variables, so that we can examine whether all components approach zero for individual runs. We observe that for each run, by the last iteration, the gradient is approximately zero (in contrast to the early iterations), suggesting we have reached a relatively flat area of the response surface, as would occur at an optimum.

Figure B.8: B-NM centroid by stratum, run, and iteration

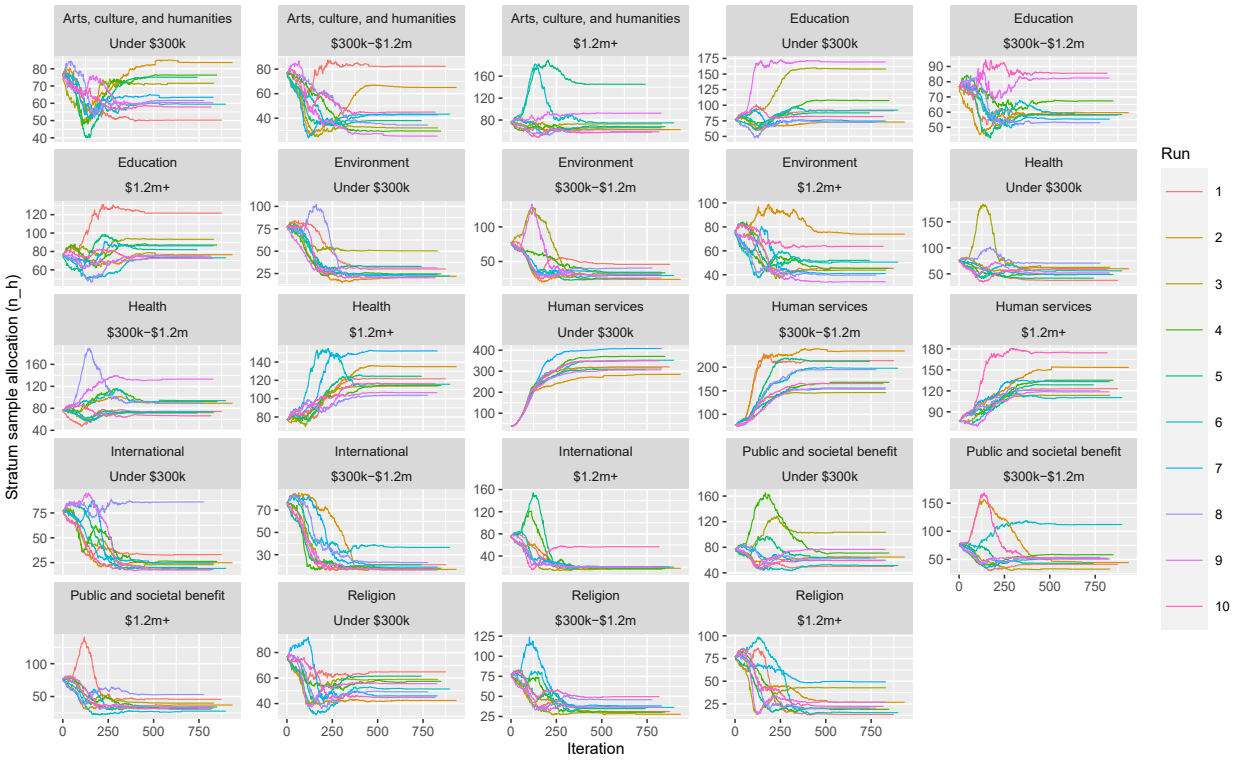
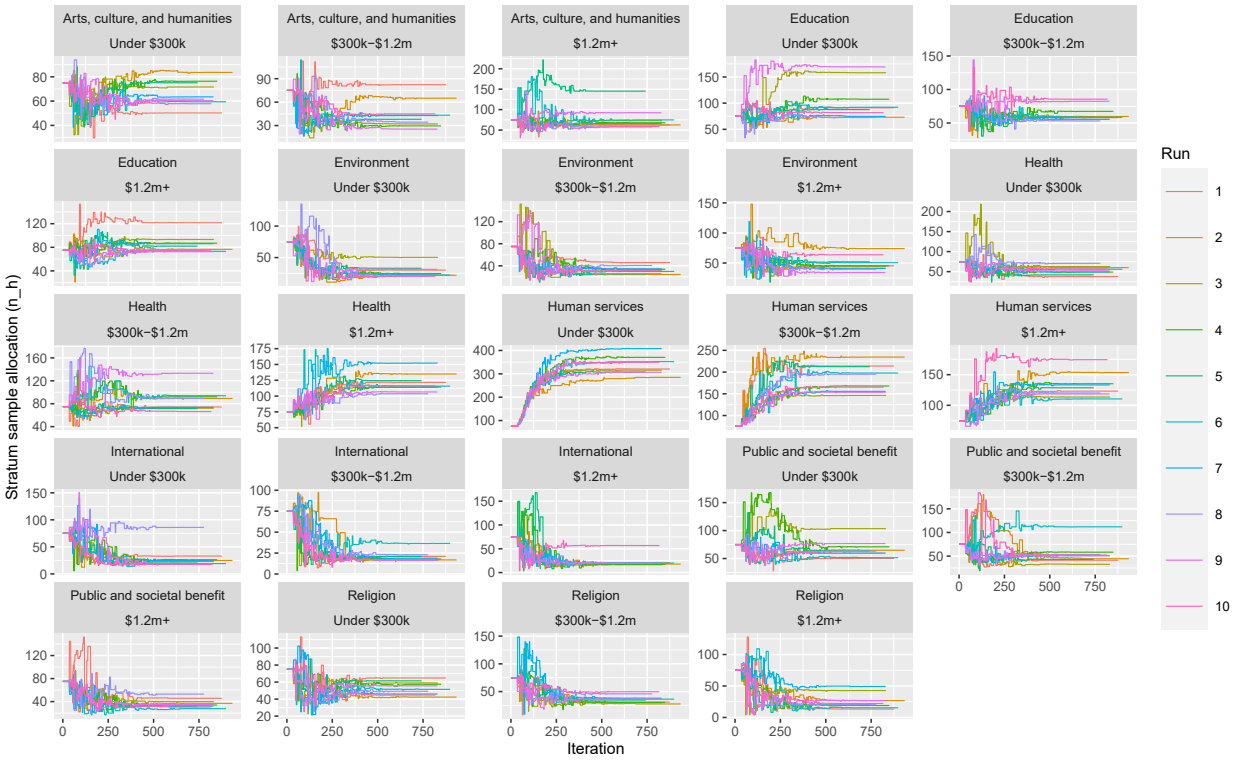


Figure B.9: B-NM best point by stratum, run, and iteration



Appendix C: Chapter 4 Appendices

C.1 Truncated binomial distribution

In modeling the number of respondents in a given stratum, and assuming constant response propensities within stratum, it would be natural to assume use of a binomial distribution, if not for the fact that this assigns some mass to a scenario with zero respondents, inhibiting analysis. Therefore, we assume use of a (zero-) truncated binomial distribution (e.g., Rider, 1955; Stephan, 1945), which assigns zero mass to this scenario but that otherwise is approximately binomial under reasonable conditions.

Suppose that X is a random variable with pmf of

$$p(k,n,p) = \Pr(X = k; n, p) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{1-(1-p)^n} \propto \binom{n}{k} p^k (1-p)^{n-k}, \tag{C.1}$$

for $k = 1, 2, ..., n$, and with zero mass, otherwise. Call this a *truncated binomial distribution* with parameters (n, p) and write the above as $X \sim TBinom(n, p)$. The pmf is proportional to that of the binomial distribution, except for $k = 0$, where it has zero mass; we can see that it is a valid distribution since the denominator is the probability that a $Binom(n, p)$

random variable has value greater than 0.

The moments of X can be obtained by relating them to moments of the $Binom(n, p)$ distribution, which has a similar form. Specifically, letting $Y \sim Binom(n, p)$, we have:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{k=1}^{n} k \frac{\binom{n}{k} p^k (1-p)^{n-k}}{1-(1-p)^n} \\
&= \frac{1}{1-(1-p)^n} \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{1}{1-(1-p)^n} \mathrm{E}(Y) \\
&= \frac{np}{1-(1-p)^n}.
\end{aligned} \tag{C.2}$$

Through analogous steps, we obtain

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{k=1}^{n} k^2 \frac{\binom{n}{k} p^k (1-p)^{n-k}}{1-(1-p)^n} \\
&= \frac{\mathrm{E}(Y^2)}{1-(1-p)^n} \\
&= \frac{np(1-p+np)}{1-(1-p)^n},
\end{aligned} \tag{C.3}$$

leading to

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{np(1-p+np)}{1-(1-p)^n} - \left(\frac{np}{1-(1-p)^n} \right)^2 \\
&= \frac{np}{1-q^n} \left(1-p+np - \frac{np}{1-q^n} \right),
\end{aligned} \tag{C.4}$$

where $q = 1-p$. Assuming that $q^n \doteq 0$, then $\mathrm{E}(X) \doteq np$ and $\mathrm{Var}(X) \doteq npq$, meaning that

the moments approximate those of the corresponding binomial distribution.

In a sampling context, we will subsequently analyze assume a model of $r_h \sim TBinom(n_h, \bar{\phi}_h)$, to ensure that the poststratified estimator is well defined, even though a preferable population model would be $r_h \sim Binom(n_h, \bar{\phi}_h)$. Loosely speaking, a sort of (relative) misspecification bias could be computed as the relative change in the expected first moment from assuming that r_h follows the truncated rather than standard binomial distribution. We define this quantity as

$$RB(n_h, \bar{\phi}_h) = \frac{\mathrm{E}\left(TBinom(n_h, \bar{\phi}_h)\right)}{\mathrm{E}\left(Binom(n_h, \bar{\phi}_h)\right)} - 1 = \frac{1}{1 - (1 - \bar{\phi}_h)^n} - 1. \tag{C.5}$$

This can be used to inform a rough rule of thumb for how small the the target number of respondents in each stratum can be to use the resulting approximation. For example, assuming $0.01 \leq \bar{\phi}_h \leq 1$, we find that $n_h \bar{\phi}_h \geq 7$ ensures that $0 \leq RB(n_h, \bar{\phi}_h) < .001$.

If these assumptions can't be made, or if it is necessary that $\mathrm{E}(r_h) = n_h \bar{\phi}_h$ be exactly met, a different approach would be to select $p'|(n, p)$ such that $\frac{p'}{1-(1-p')^n} = p$, and which results in $\mathrm{E}(TBinom(n, p')) = np$. Therefore, define a *reparameterized truncated binomial distribution* with parameters (n, p), written as $RTBinom(n, p)$, as having the same pmf of a $TBinom(n, p')$ random variable. The value of p' for given (n, p) can be identified via Newton–Raphson as the root to $f(x) = p\,(1 - (1 - x)^n) - x$, with $x_0 = p$.

C.2 Comparison with alternative allocations

Here, we will examine the efficiency of our proposed allocation from §4.2, relative to alternative allocations. From (4.12), the approximately optimal allocation can be written as

$$n_h^{opt} \propto \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}, \tag{C.6}$$

where we define $c_h := \frac{\mathrm{E}(C_h)}{n_h} = c_{NR_h}\left(\bar{\phi}_h(\tau_h - 1) + 1\right)$ as the expected per-invitee cost in stratum h, and where we have assumed that the strata are adequately sized (in terms of expected respondents) such that $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$. Assuming a fixed total sample size of n^{opt}, the stratum h allocation can be written as

$$n_h^{opt} = \frac{\frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}}{\sum_h \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}} n^{opt}. \tag{C.7}$$

Let $AV\left(\hat{\bar{Y}} | \{n_h^0\}\right) := \frac{1}{N^2}\sum_h \frac{N_h^2 S_h^2}{n_h^0 \bar{\phi}_h}$ denote the approximate variance from (4.8), when evaluated at some arbitrary allocation $\{n_h\} = \{n_h^0\}$, under the assumptions of a negligible fpc (i.e., $\frac{r_h}{N_h}$ close to zero) and again assuming $\zeta_h(n_h, \bar{\phi}_h) \doteq 1$. Substituting (C.7) into this expression yields an approximate variance of

$$AV\left(\hat{\bar{Y}} | \{n_h^{opt}\}\right) = \frac{1}{n^{opt} N^2}\left(\sum_h \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}\right)\left(\sum_h \frac{N_h S_h \sqrt{c_h}}{\sqrt{\bar{\phi}_h}}\right). \tag{C.8}$$

Next, consider some arbitrary alternative allocation $\{n_h^{alt}\}$, with total sample size of $n^{alt} =$

$\sum_h n_h^{alt}$, and where we define $\alpha_h = \frac{n_h^{alt}}{n_h^{opt}}$ for $h = 1, 2, ..., H$ and where $\alpha_h > 0$. The alternative allocation can be written as

$$n_h^{alt} = \frac{\frac{N_h S_h \alpha_h}{\sqrt{\bar{\phi}_h c_h}}}{\sum_h \frac{N_h S_h \alpha_h}{\sqrt{\bar{\phi}_h c_h}}} n^{alt} \tag{C.9}$$

and has an approximate variance of

$$AV\left(\hat{\bar{Y}} | \{n_h^{alt}\}\right) = \frac{1}{n^{alt} N^2} \left(\sum_h \frac{N_h S_h \alpha_h}{\sqrt{\bar{\phi}_h c_h}}\right) \left(\sum_h \frac{N_h S_h \sqrt{c_h}}{\alpha_h \sqrt{\bar{\phi}_h}}\right). \tag{C.10}$$

The ratio of the approximate variance under the alternative allocation compared with that of the proposed allocation is then

$$\frac{AV\left(\hat{\bar{Y}} | \{n_h^{alt}\}\right)}{AV\left(\hat{\bar{Y}} | \{n_h^{opt}\}\right)} = \frac{n^{opt}}{n^{alt}} \frac{\left(\sum_h \frac{N_h S_h \alpha_h}{\sqrt{\bar{\phi}_h c_h}}\right) \left(\sum_h \frac{N_h S_h \sqrt{c_h}}{\alpha_h \sqrt{\bar{\phi}_h}}\right)}{\left(\sum_h \frac{N_h S_h}{\sqrt{\bar{\phi}_h c_h}}\right) \left(\sum_h \frac{N_h S_h \sqrt{c_h}}{\sqrt{\bar{\phi}_h}}\right)}. \tag{C.11}$$

Suppose further that $\{n_h^{alt}\}$ and $\{n_h^{opt}\}$ both have total expected costs of C'. In that case, the total sample sizes are

$$n^{opt} = C' \frac{\sum_h N_h S_h / \sqrt{\bar{\phi}_h c_h}}{\sum_h N_h S_h \sqrt{c_h / \bar{\phi}_h}}, \text{and} \tag{C.12}$$

$$n^{alt} = C' \frac{\sum_h N_h S_h \alpha_h / \sqrt{\bar{\phi}_h c_h}}{\sum_h N_h S_h \alpha_h \sqrt{c_h / \bar{\phi}_h}}, \tag{C.13}$$

in which case (C.11) simplifies to

$$\frac{AV\left(\hat{\bar{Y}}|\{n_h^{alt}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{\left(\sum_h T_h\alpha_h\right)\left(\sum_h \frac{T_h}{\alpha_h}\right)}{\left(\sum_h T_h\right)^2}, \tag{C.14}$$

where $T_h = \frac{N_h S_h \sqrt{c_h}}{\sqrt{\bar{\phi}_h}}$.

As a special case, consider the scenario of $\tau_h = 1$ and $c_{NR_h} = K$ for $h = 1, ..., H$ and constant K, in which case, cost per invitee is constant across strata and can be ignored. Under this scenario, our proposed allocation becomes $n_h^{opt} \propto \frac{N_h S_h}{\sqrt{\bar{\phi}_h}}$. As alternatives, consider allocations of $n_h^{Ninv} \propto N_h S_h$ and $n_h^{Nresp} \propto \frac{N_h S_h}{\bar{\phi}_h}$, where the superscript conveys that these refer to Neyman allocations of invitees and respondents, respectively, the latter being common practice. In applying (C.14), observe that the $\{n_h^{Ninv}\}$ allocation is equivalent to setting $\alpha_h = \sqrt{\bar{\phi}_h}$, whereas the $\{n_h^{Nresp}\}$ allocation is equivalent to setting $\alpha_h = \frac{1}{\sqrt{\bar{\phi}_h}}$, leading to equivalent variance ratios, which are

$$\frac{AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{opt}\}\right)} = \frac{\left(\sum_h N_h S_h\right)\left(\sum_h \frac{N_h S_h}{\bar{\phi}_h}\right)}{\left(\sum_h \frac{N_h S_h}{\sqrt{\bar{\phi}_h}}\right)^2}. \tag{C.15}$$

An interesting feature of this analysis is that it illustrates that, when allocating a fixed number of invitees, adjusting by the inverse of the expected response rates does not actually reduce the resulting (approximate) variance.

C.3 Comparison between Neyman allocations of invitees and respondents

The approximate variance ratio of the Neyman allocation of invitees to that of respondents is

$$VR_{Nresp}^{Ninv} := \frac{AV\left(\hat{\bar{Y}}|\{n_h^{Ninv}\}\right)}{AV\left(\hat{\bar{Y}}|\{n_h^{Nresp}\}\right)} = \frac{VR_{Ninv}}{VR_{Nresp}} = \frac{\left(\sum_h N_h S_h c_h\right)\left(\sum_h \frac{N_h S_h}{\bar{\phi}_h}\right)}{\left(\sum_h N_h S_h \frac{c_h}{\bar{\phi}_h}\right)\left(\sum_h N_h S_h\right)}, \tag{C.16}$$

where, assuming constant $c_{NR_h} = c_{NR}$ and $\tau_h = \tau$ across strata, the expected per-invitee costs are $c_h = c_{NR}\left(\bar{\phi}_h\left(\tau - 1\right) + 1\right)$. We aim to show that $VR_{Nresp}^{Ninv} \geq 1$ when $\tau \geq 1$, which entails considering the partial derivative of (C.16) with respect to τ via use of the quotient rule. For purposes of this section, define $W_h := N_h S_h$ and rewrite the numerator and denominator of (C.16) as $f(\tau) = \left(\sum_h W_h c_h\right) K_1$ and $g(\tau) = \left(\sum_h W_h \frac{c_h}{\bar{\phi}_h}\right) K_2$, respectively, for quantities $K_1 := \sum_h \frac{W_h}{\bar{\phi}_h}$ and $K_2 := \sum_h W_h$, the latter of which do not depend on τ. Further, write (C.16) as $h(\tau) = \frac{f(\tau)}{g(\tau)}$.

Since $\frac{\partial c_h}{\partial \tau} = c_{NR}\bar{\phi}_h$, we find that

$$f'(\tau) := \frac{\partial f(\tau)}{\partial \tau} = K_1 c_{NR} \sum_h W_h \bar{\phi}_h \tag{C.17}$$

$$g'(\tau) := \frac{\partial g(\tau)}{\partial \tau} = K_2 c_{NR} \sum_h W_h. \tag{C.18}$$

Letting $h'(\tau) := \frac{\partial h(\tau)}{\partial \tau}$, the quotient rule indicates that $h'(\tau) = \frac{f'(\tau)g(\tau) - g'(\tau)f(\tau)}{g(\tau)^2}$. From (C.17–

C.18), the numerator of $h'(\tau)$ is

$$\begin{aligned}
f'(\tau)g(\tau) - g'(\tau)f(\tau) &= K_1 K_2 c_{NR} \left[\left(\sum_h W_h \bar{\phi}_h \right) \left(\sum_h W_h \frac{c_h}{\bar{\phi}_h} \right) - \left(\sum_h W_h \right) \left(\sum_h W_h c_h \right) \right] \\
&= K_1 K_2 c_{NR} \left[\sum_{h_1=1}^{H} \sum_{h_2=1}^{H} W_{h_1} W_{h_2} c_{h_2} \left(\frac{\bar{\phi}_{h_1}}{\bar{\phi}_{h_2}} - 1 \right) \right] \\
&= K_1 K_2 c_{NR} \left[\sum \sum_{h_1<h_2} W_{h_1} W_{h_2} \left[c_{h_2} \left(\frac{\bar{\phi}_{h_1} - \bar{\phi}_{h_2}}{\bar{\phi}_{h_2}} \right) + c_{h_1} \left(\frac{\bar{\phi}_{h_2} - \bar{\phi}_{h_1}}{\bar{\phi}_{h_1}} \right) \right] \right] \\
&= K_1 K_2 c_{NR} \left[\sum \sum_{h_1<h_2} W_{h_1} W_{h_2} \left(\frac{\left(\bar{\phi}_{h_1} - \bar{\phi}_{h_2} \right) \left(c_{h_2} \bar{\phi}_{h_1} - c_{h_1} \bar{\phi}_{h_2} \right)}{\bar{\phi}_{h_1} \bar{\phi}_{h_2}} \right) \right] \\
&= K_1 K_2 c_{NR}^2 \left[\sum \sum_{h_1<h_2} W_{h_1} W_{h_2} \left(\frac{\left(\bar{\phi}_{h_1} - \bar{\phi}_{h_2} \right)^2}{\bar{\phi}_{h_1} \bar{\phi}_{h_2}} \right) \right], \qquad \text{(C.19)}
\end{aligned}$$

and where the last line took advantage of the fact that

$$\begin{aligned}
c_{h_2} \bar{\phi}_{h_1} - c_{h_1} \bar{\phi}_{h_2} &= c_{NR} \left(\bar{\phi}_{h_2} (\tau - 1) + 1 \right) \bar{\phi}_{h_1} - c_{NR} \left(\bar{\phi}_{h_1} (\tau - 1) + 1 \right) \bar{\phi}_{h_2} \\
&= c_{NR} \left(\bar{\phi}_{h_1} - \bar{\phi}_{h_2} \right), \qquad \text{(C.20)}
\end{aligned}$$

since the other terms cancel out. The individual terms in (C.19) are all nonnegative, and so $f'(\tau)g(\tau) - g'(\tau)f(\tau) \geq 0$ (the latter inequality being strict if the response rates vary at all across strata). Additionally,

$$\begin{aligned}
g(\tau) &= K_2 \sum_h W_h \frac{c_{NR} \left(\bar{\phi}_h (\tau - 1) + 1 \right)}{\bar{\phi}_h} \qquad \text{(C.21)} \\
&= K_2 C_{NR} \sum_h W_h \frac{\bar{\phi}_h (\tau - 1) + 1}{\bar{\phi}_h} \qquad \text{(C.22)}
\end{aligned}$$

is strictly positive for $\tau \geq 1$, as is $g(\tau)^2$, the latter being the denominator of $h'(\tau)$. As a

result, $\tau \geq 1$ implies that $h'(\tau) \geq 0$. In conjunction with the fact that $h(1) = 1$, we see that $\tau \geq 1$ leads to $VR_{Nresp}^{Ninv} \geq 1$, the latter inequality being strict as long as $\tau > 1$ and the response rates vary across strata.

C.4 2016 PEVS-ADM example: detailed allocations

Table C.1: Detailed PEVS-ADM allocations under constant cost per invitee scenario

h	Service	Pay grade	Age	Region	Sex	$\hat{N}_h$	$\hat{\bar{\phi}}_h$ (%)	n_h^{Ninv}	n_h^{Nresp}	n_h^{p-A}	n_h^{p-E}	ζ_h^{p-E}
1	Army	E1-E5	18-24	US	M	109,214	1.89	4,269	12,975	8,309	8,267	1.01
2	Army	E1-E5	18-24	US	F	18,921	3.64	740	1,164	1,036	1,041	1.03
3	Army	E1-E5	18-24	Overseas	M	18,465	2.14	722	1,935	1,320	1,333	1.04
4	Army	E1-E5	18-24	Overseas	F	3,312	4.31	129	172	167	179	1.17
5	Army	E1-E5	25-29	US	M	48,348	4.87	1,890	2,225	2,289	2,280	1.01
6	Army	E1-E5	25-29	US	F	8,680	6.19	339	314	364	370	1.05
7	Army	E1-E5	25-29	Overseas	M/F	10,762	5.66	421	426	473	477	1.04
8	Army	E1-E5	30-34	US	M/F	21,352	8.91	835	537	747	746	1.01
9	Army	E1-E5	30-34	Overseas	M/F	3,739	9.04	146	93	130	135	1.09
10	Army	E1-E5	35+	US	M/F	9,055	18.81	354	108	218	219	1.02
11	Army	E1-E5	35+	Overseas	M/F	1,840	20.54	72	20	42	44	1.11
12	Army	E6-E9	18-29	US/Over.	M/F	13,845	7.61	541	408	524	526	1.02
13	Army	E6-E9	30-34	US	M/F	27,166	11.49	1,062	529	837	834	1.01
14	Army	E6-E9	30-34	Overseas	M/F	3,943	10.70	154	83	126	130	1.08
15	Army	E6-E9	35+	US	M/F	49,462	21.75	1,934	509	1,108	1,101	1.00
16	Army	E6-E9	35+	Overseas	M/F	8,060	19.91	315	91	189	189	1.02
17	Army	W1-W5	18-34	US/Over.	M/F	4,821	13.51	188	80	137	139	1.05
18	Army	W1-W5	35+	US/Over.	M/F	9,764	20.81	382	105	224	224	1.02
19	Army	O1-O3	18-24	US	M/F	7,719	17.40	302	99	193	194	1.03
20	Army	O1-O3	18-24	Overseas	M/F	873	15.94	34	12	23	26	1.29
21	Army	O1-O3	25-29	US	M/F	15,515	16.41	607	212	400	399	1.01
22	Army	O1-O3	25-29	Overseas	M/F	2,288	16.85	89	30	58	61	1.10
23	Army	O1-O3	30-34	US	M/F	11,360	24.20	444	105	241	241	1.01
24	Army	O1-O3	30-34	Overseas	M/F	1,643	18.23	64	20	40	43	1.14
25	Army	O1-O3	35+	US	M/F	7,554	30.50	295	55	143	143	1.02
26	Army	O1-O3	35+	Overseas	M/F	1,300	24.59	51	12	27	29	1.15
27	Army	O4-O6	30-34	US/Over.	M/F	2,389	20.71	93	26	55	57	1.08
28	Army	O4-O6	35+	US	M/F	21,373	36.59	836	131	369	367	1.00
29	Army	O4-O6	35+	Overseas	M/F	4,622	32.00	181	32	85	86	1.03
30	Navy	E1-E5	18-24	US	M	74,440	2.32	2,910	7,173	5,100	5,080	1.01
31	Navy	E1-E5	18-24	US	F	22,212	3.84	868	1,297	1,185	1,188	1.02
32	Navy	E1-E5	18-24	Overseas	M/F	10,451	2.43	409	964	701	717	1.06
33	Navy	E1-E5	25-29	US	M	40,053	5.52	1,566	1,624	1,780	1,774	1.01
34	Navy	E1-E5	25-29	US	F	9,671	6.73	378	322	389	394	1.04
35	Navy	E1-E5	25-29	Overseas	M/F	5,970	5.76	233	232	260	267	1.07
36	Navy	E1-E5	30+	US/Over.	M/F	21,373	8.22	836	583	779	778	1.01
37	Navy	E6-E9	18-29	US/Over.	M/F	13,225	6.89	517	430	526	529	1.03
38	Navy	E6-E9	30-34	US	M/F	22,642	11.33	885	448	703	701	1.01
39	Navy	E6-E9	30-34	Overseas	M/F	2,761	13.48	108	46	79	82	1.10
40	Navy	E6-E9	35+	US/Over.	M/F	38,288	16.60	1,497	517	982	976	1.01
41	Navy	O1-O3	18-24	US/Over.	M/F	5,332	13.92	208	86	149	151	1.05
42	Navy	O1-O3	25-29	US/Over.	M	9,210	20.20	360	102	214	214	1.02
43	Navy	O1-O3	25-29	US/Over.	F	2,595	32.97	101	18	47	48	1.05
44	Navy	W1-O3	30-34	US	M/F	8,086	21.73	316	83	181	182	1.02
45	Navy	W1-O3	30-34	Overseas	M/F	1,046	18.51	41	13	25	28	1.24
46	Navy	W1-O3	35+	US/Over.	M/F	8,207	27.45	321	67	164	164	1.02
47	Navy	O4-O6	25+	US/Over.	M/F	19,145	40.08	748	107	316	314	1.00
48	MC	E1-E5	18-24	US	M/F	86,520	2.81	3,382	6,903	5,394	5,368	1.01
49	MC	E1-E5	18-24	Overseas	M/F	16,790	3.85	656	978	894	900	1.03
50	MC	E1-E5	25+	US	M/F	19,701	9.43	770	468	670	670	1.01
51	MC	E1-E5	25+	Overseas	M/F	3,650	7.29	143	112	141	147	1.11
52	MC	E6-E9	18-29	US/Over.	M/F	4,871	11.41	190	96	151	154	1.06

Table C.1: Detailed PEVS-ADM allocations under constant cost per invitee scenario *(continued)*

h	Service	Pay grade	Age	Region	Sex	$\hat{N}_h$	$\hat{\bar{\phi}}_h$ (%)	n_h^{Ninv}	n_h^{Nresp}	n_h^{p-A}	n_h^{p-E}	ζ_h^{p-E}
53	MC	E6-E9	30-34	US	M/F	9,090	20.64	355	99	209	209	1.02
54	MC	E6-E9	30-34	Overseas	M/F	1,656	13.15	65	28	48	52	1.19
55	MC	E6-E9	35+	US	M/F	9,683	24.12	379	90	206	206	1.02
56	MC	E6-E9	35+	Overseas	M/F	1,825	21.61	71	19	41	43	1.11
57	MC	W1-W5	25+	US/Over.	M/F	2,023	23.53	79	19	44	45	1.09
58	MC	O1-O3	18-29	US/Over.	M/F	7,277	14.89	284	109	197	198	1.03
59	MC	O1-O3	30+	US/Over.	M/F	4,542	19.71	178	52	107	108	1.04
60	MC	O4-O6	30+	US/Over.	M/F	6,201	30.52	242	46	117	117	1.02
61	AF	E1-E5	18-24	US	M	57,817	10.44	2,260	1,240	1,869	1,858	1.00
62	AF	E1-E5	18-24	US	F	14,443	16.93	565	191	367	366	1.01
63	AF	E1-E5	18-24	Overseas	M	15,408	11.19	602	308	481	481	1.02
64	AF	E1-E5	18-24	Overseas	F	3,151	12.55	123	56	93	96	1.09
65	AF	E1-E5	25-29	US	M	36,898	14.82	1,442	558	1,001	996	1.01
66	AF	E1-E5	25-29	US	F	8,939	12.59	349	159	263	265	1.03
67	AF	E1-E5	25-29	Overseas	M	12,376	14.80	484	187	336	336	1.02
68	AF	E1-E5	25-29	Overseas	F	2,271	15.27	89	33	61	63	1.11
69	AF	E1-E5	30+	US	M/F	17,848	19.29	698	207	425	423	1.01
70	AF	E1-E5	30+	Overseas	M/F	5,782	21.78	226	59	129	130	1.03
71	AF	E6-E9	18-34	US/Over.	M/F	27,881	22.35	1,090	279	616	613	1.01
72	AF	E6-E9	35+	US	M/F	32,796	30.21	1,282	243	623	619	1.00
73	AF	E6-E9	35+	Overseas	M	7,920	30.64	310	58	149	149	1.02
74	AF	E6-E9	35+	Overseas	F	1,536	31.34	60	11	29	30	1.09
75	AF	O1-O3	18-24	US/Over.	M/F	5,385	37.52	211	32	92	92	1.02
76	AF	O1-O3	25-29	US	M/F	12,914	26.94	505	107	260	259	1.01
77	AF	O1-O3	25-29	Overseas	M/F	2,274	25.00	89	20	48	49	1.07
78	AF	O1-O3	30-34	US	M/F	8,862	36.57	346	54	153	153	1.01
79	AF	O1-O3	30-34	Overseas	M/F	1,724	30.25	67	13	33	34	1.08
80	AF	O1-O3	35+	US/Over.	M/F	3,450	38.33	135	20	58	59	1.03
81	AF	O4-O6	25-34	US/Over.	M/F	3,830	30.46	150	28	72	73	1.03
82	AF	O4-O6	35+	US	M	14,619	42.35	571	77	235	233	1.01
83	AF	O4-O6	35+	US	F	3,003	52.17	117	13	43	44	1.02
84	AF	O4-O6	35+	Overseas	M/F	3,690	41.25	144	20	60	60	1.03
85	CG	E1-E5	18-24	US/Over.	M/F	7,121	12.36	278	129	212	214	1.04
86	CG	E1-E5	25-29	US/Over.	M/F	7,221	18.75	282	86	174	175	1.03
87	CG	E1-E5	30+	US/Over.	M/F	5,161	23.87	202	48	110	111	1.03
88	CG	E6-E9	18+	US/Over.	M/F	10,641	32.94	416	72	194	193	1.01
89	CG	W1-W5	25+	US/Over.	M/F	1,731	61.07	68	6	23	23	1.03
90	CG	O1-O3	18+	US/Over.	M/F	3,845	28.65	150	30	75	76	1.04
91	CG	O4-O6	30+	US/Over.	M/F	2,563	41.38	100	14	42	42	1.04
Total						1,279,021	13.59	50,000	50,000	50,000	50,000	

Note: n_h^{p-A} and n_h^{p-E} denote the approximate and exact proposed allocations; ζ_h^{p-E} denotes $\zeta_h(n_h^{p-E}, \bar{\phi}_h)$. Allocations assume that S_h is constant across strata and that $n = 50,000$.

References

AAPOR. (2016). *Standard definitions: final dispositions of case codes and outcome rates for surveys, 9th edition.* (https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf [Accessed 8 May 2023].)

Amzal, B., Bois, F. Y., Parent, E., & Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, *101*(474), 773–785. https://doi.org/10.1198/016214505000001159

Andridge, R., & Thompson, K. J. (2015). Assessing nonresponse bias in a business survey: proxy pattern-mixture analysis for skewed data. *The Annals of Applied Statistics*, *9*(4), 2237–2265. https://doi.org/10.1214/15-aoas878

Barton, R. R., & Ivey, J. S., Jr. (1996). Nelder-Mead simplex modifications for simulation optimization. *Management Science*, *42*(7), 954–973. https://doi.org/10.1287/mnsc.42.7.954

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 202–242). Toronto: Holt, Rinehart and Winston.

Battaglia, M. P., Dillman, D. A., Frankel, M. R., Harter, R., Buskirk, T. D., McPhee, C. B., ... Yancey, T. (2016). Sampling, data collection, and weighting procedures for address-based sample surveys. *Journal of Survey Statistics and Methodology*, *4*(4), 476–500. https://doi.org/10.1093/jssam/smw025

Baudin, M. (2010). Nelder-Mead user's manual. *Consortium Scilab*. (https://www.scilab.org/sites/default/files/neldermead.pdf)

Bielza, C., Müller, P., & Insua, D. R. (1999). Decision analysis by augmented probability simulation. *Management Science*, *45*(7), 995–1007. https://doi.org/10.1287/mnsc.45.7.995

Box, M. (1965). A new method of constrained optimization and a comparison with other methods. *The Computer Journal*, *8*(1), 42–52. https://doi.org/10.1093/comjnl/8.1.42

Brereton, M., & Bowers, D. (2021). *Insights and Analytics Market and Top 50 Report.* (https://www.insightsassociation.org/sites/default/files/misc_files/ia_-_top_50_report_us_2021_v5print.pdf [Accessed 9 November 2021])

Brewer, K. R. (2002). *Combined survey sampling inference: weighing Basu's elephants.* Arnold.

Brewer, K. R. (2013). Three controversies in the history of survey sampling. *Survey Method-*

ology, *39*(2), 249–262.

Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly*, *75*(5), 872–888. https://doi.org/10.1093/poq/nfr045

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329–353. https://doi.org/10.2478/jos-2013-0026

Brick, J. M., Andrews, W. R., & Mathiowetz, N. A. (2016). Single-phase mail survey design for rare population subgroups. *Field Methods*, *28*(4), 381–395. https://doi.org/10.1177/1525822X15616926

Brick, J. M., Contos, G., Masken, K., & Nord, R. (2010). Response mode and bias analysis in the IRS Individual Taxpayer Burden Survey. *Survey Practice*, *3*(5). https://doi.org/10.29115/SP-2010-0025

Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59. https://doi.org/10.1177/0002716212456834

Buskirk, T. D., & Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods*, 1–17. https://doi.org/10.13094/SMIF-2015-00003

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304. https://doi.org/10.1214/ss/1177009939

Chen, S., & Kalton, G. (2015). Geographic oversampling for race/ethnicity using data from the 2010 U.S. population census. *Journal of Survey Statistics and Methodology*, *3*(4), 543–565. https://doi.org/10.1093/jssam/smv023

Clark, R. G. (2013). Sample design using imperfect design data. *Journal of Survey Statistics and Methodology*, *1*(1), 6–23. https://doi.org/10.1093/jssam/smt002

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, Inc.

Coffey, S., & Elliott, M. R. (2023). Optimizing data collection interventions to balance cost and quality in a sequential multimode survey. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smad007

Coffey, S., West, B. T., Wagner, J., & Elliott, M. R. (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. https://doi.org/10.12758/mda.2020.05

Conn, A. R., Scheinberg, K., & Vicente, L. N. (2009). *Introduction to derivative-free optimization*. SIAM. https://doi.org/10.1137/1.9780898718768

Dalenius, T., & Hodges, J. L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, *54*(285), 88–101. https://doi.org/10.1080/01621459.1959.10501501

Dennis, J., & Woods, D. J. (1987). Optimization on microcomputers: the Nelder-Mead simplex algorithm. *New computing environments: microcomputers in large-scale computing*, 116–122.

DesRoches, C. M., Barrett, K. A., Harvey, B. E., Kogan, R., Reschovsky, J. D., Landon, B. E., … Rich, E. C. (2015). The results are only as good as the sample: assessing three national physician sampling frames. *Journal of General Internal Medicine*, *30*(3),

595–601. https://doi.org/10.1007/s11606-015-3380-9

Dillman, D. A. (2017). The promise and challenge of pushing respondents to the web in mixed-mode surveys. *Survey Methodology*, *43*(1), 3–31.

Dorfman, A. H., & Valliant, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, *16*(2), 139–154.

Draper, N. R., & Guttman, I. (1968a). Bayesian stratified two-phase sampling results: k characteristics. *Biometrika*, *55*(3), 587–589. https://doi.org/10.1093/biomet/55.3.587

Draper, N. R., & Guttman, I. (1968b). Some Bayesian stratified two-phase sampling results. *Biometrika*, *55*(1), 131–139. https://doi.org/10.1093/biomet/55.1.131

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*(383), 535–543. https://doi.org/10.2307/2288115

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Ericson, W. A. (1965). Optimum stratified sampling using prior information. *Journal of the American Statistical Association*, *60*(311), 750–771. https://doi.org/10.1080/01621459.1965.10480825

Ericson, W. A. (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. https://doi.org/10.1111/j.2517-6161.1967.tb00712.x

Ericson, W. A. (1967b). Optimal sample design with nonresponse. *Journal of the American Statistical Association*, *62*(317), 63–78. https://doi.org/10.1080/01621459.1967.10482888

Ericson, W. A. (1969a). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 195–233. https://doi.org/10.1111/j.2517-6161.1969.tb00782.x

Ericson, W. A. (1969b). Subjective Bayesian models in sampling finite populations: stratification. In H. Smith & N. L. Johnson (Eds.), *New developments in survey sampling* (pp. 326–357). Wiley.

Ericson, W. A. (1972). *Subjective Bayesian models in sampling finite populations: stratification* (Tech. Rep. No. 15). Ann Arbor, Michigan: University of Michigan, Department of Statistics.

Ericson, W. A. (1973). *A Bayesian approach to two-stage sampling* (Tech. Rep. No. 26). Ann Arbor, Michigan: University of Michigan, Department of Statistics.

Ericson, W. A. (1975). *A Bayesian approach to two-stage sampling* (Tech. Rep. No. AFFDL-TR-75-145). Wright-Patterson AFB, Ohio: Air Force Dynamics Laboratory.

Ericson, W. A. (1983a). *Response bias and error in sampling finite populations* (Tech. Rep. No. 122). Ann Arbor, Michigan: University of Michigan, Department of Statistics.

Ericson, W. A. (1983b). *Survey sampling: modeling, randomization, and robustness: a subjective Bayesian approach* (Tech. Rep. No. 121). Ann Arbor, Michigan: University of Michigan, Department of Statistics.

Ericson, W. A. (1985). *Bayesian inference in finite populations* (Tech. Rep. No. 129). Ann

Arbor, Michigan: University of Michigan, Department of Statistics.
Ericson, W. A. (1988). Bayesian inference in finite populations. In *Handbook of statistics 6: Sampling* (pp. 213–246). Elsevier. https://doi.org/http://dx.doi.org/10.1016/S0169-7161(88)06011-0
Evans, W. D. (1951). On stratification and optimum allocations. *Journal of the American Statistical Association*, *46*(253), 95–104. https://doi.org/10.1080/01621459.1951.10500772
Fay, R. E., & Riddles, M. K. (2017). One-versus two-step approaches to survey nonresponse adjustments. In *ASA Proceedings of the Survey Research Methods Section.* (http://www.asasrms.org/Proceedings/y2017/files/593858.pdf [Accessed 23 April 2023])
Federal Voting Assistance Program. (2017a). *2015–2016 Active Duty Military dataset.* Federal Voting Assistance Program, U.S. Department of Defense. (https://www.fvap.gov/uploads/FVAP/Surveys/PEV51601_Pub.dta [Accessed 17 January 2023])
Federal Voting Assistance Program. (2017b). *Post-Election Voting Surveys: Active Duty Military: Technical report 2016* (Tech. Rep.). Federal Voting Assistance Program, U.S. Department of Defense. (https://www.fvap.gov/uploads/FVAP/Reports/PEVS_ADM_TechReport_Final.pdf [Accessed 17 January 2023])
Fernández-i-Marín, X. (2016). ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, *70*(9), 1–20. https://doi.org/10.18637/jss.v070.i09
Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, *1*(1), 1–40. https://doi.org/10.1214/06-BA101
Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review/Revue Internationale de Statistique*, 237–253. https://doi.org/10.2307/1403784
Foreman, E. (1991). *Survey sampling principles.* New York: Marcel Dekker.
Frankel, M. R., & Frankel, L. R. (1987). Fifty years of survey sampling in the United States. *Public Opinion Quarterly*, *51*, S127–S138. https://doi.org/10.1093/poq/51.4_PART_2.S127
Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.2307/2289776
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC Boca Raton, FL, USA. https://doi.org/10.1201/b16018
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC Boca Raton, FL, USA. https://doi.org/10.1201/9780429258480
Ghosh, M. (2009). Bayesian developments in survey sampling. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics 29b: Sample surveys: Inference and analysis* (pp. 155–189). Elsevier. https://doi.org/10.1016/s0169-7161(09)00229-6
Ghosh, M., & Meeden, G. (1997). *Bayesian methods for finite population sampling.* CRC

Press. https://doi.org/10.1201/9781315138169
Godambe, V., & Joshi, V. (1965). Admissibility and Bayes estimation in sampling finite populations - i. *The Annals of Mathematical Statistics*, *36*(6), 1707–1722. https://doi.org/10.1214/aoms/1177699799
Green, D. P., & Gerber, A. S. (2006). Can registration-based sampling improve the accuracy of midterm election forecasts? *Public Opinion Quarterly*, *70*(2), 197–223. https://doi.org/10.1093/poq/nfj022
Grosh, D. L. (1972). A Bayes sampling allocation scheme for stratified finite populations with hyperbinomial prior distributions. *Technometrics*, *14*(3), 599–612. https://doi.org/10.1080/00401706.1972.10488949
Groves, R. M. (1989). *Survey errors and survey costs*. John Wiley & Sons. https://doi.org/10.1002/0471725277
Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457. https://doi.org/10.1111/j.1467-985x.2006.00423.x
Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, *95*(2), 347–361. https://doi.org/10.1162/rest_a_00288
Han, L., & Neumann, M. (2006). Effect of dimensionality on the Nelder–Mead simplex method. *Optimization Methods and Software*, *21*(1), 1–16.
Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, *2*(2), 180–190. https://doi.org/10.1214/ss/1177013352
Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory (vols. 1 and 2)*. John Wiley.
Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion and rejoinder). *Journal of the American Statistical Association*, *78*(384), 776–807. https://doi.org/10.1080/01621459.1983.10477018
Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., ... Zukerberg, A. (2016). *AAPOR report: Address-based sampling*. (https://aapor.org/wp-content/uploads/2022/11/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-2.pdf [Accessed 8 May 2023].)
Hartley, H. O., & Rao, J. (1968). A new estimation theory for sample surveys. *Biometrika*, *55*(3), 547–557. https://doi.org/10.1093/biomet/55.3.547
Henry, K. A., & Valliant, R. (2009). Comparing sampling and estimation strategies in establishment populations. *Survey Research Methods*, *3*(1), 27–44. https://doi.org/10.18148/srm/2009.v3i1.72
Hitchcock, D. B. (2003). A history of the Metropolis–Hastings algorithm. *The American Statistician*, *57*(4), 254–257. https://doi.org/10.1198/0003130032413
Hoadley, B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*, *64*(325), 216–229. https://doi.org/10.1080/01621459.1969.10500965
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–

685. https://doi.org/10.1080/01621459.1952.10483446
Huan, X., & Marzouk, Y. (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. *International Journal for Uncertainty Quantification*, *4*(6). https://doi.org/10.1615/int.j.uncertaintyquantification.2014006730
Huan, X., & Marzouk, Y. M. (2013). Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, *232*(1), 288–317. https://doi.org/10.1016/j.jcp.2012.08.013
Iannacchione, V. G. (2011). The changing role of address-based sampling in survey research. *Public Opinion Quarterly*, *75*(3), 556–575. https://doi.org/10.1093/poq/nfr017
Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*(377), 89–96. https://doi.org/10.1080/01621459.1982.10477770
Jackson, M. T., McPhee, C. B., & Lavrakas, P. J. (2020). Using response propensity modeling to allocate noncontingent incentives in an address-based sample: Evidence from a national experiment. *Journal of Survey Statistics and Methodology*, *8*(2), 385–411. https://doi.org/10.1093/jssam/smz007
Jackson, M. T., Medway, R. L., & Megra, M. W. (2023). Can appended auxiliary data be used to tailor the offered response mode in cross-sectional studies? Evidence from an address-based sample. *Journal of Survey Statistics and Methodology*, *11*(1), 47–74. https://doi.org/10.1093/jssam/smab023
Jinn, J., Sedransk, J., & Smith, P. (1987). Optimal two-phase stratified sampling for estimation of the age composition of a fish population. *Biometrics*, *43*(2), 343–353. https://doi.org/10.2307/2531817
Johnson, S. G. (2014). *The NLopt nonlinear-optimization package.*
Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 175–188. https://doi.org/10.2307/1402747
Kalton, G. (2020). *Introduction to survey sampling* (2nd ed.) (No. 35). Sage Publications.
Kelley, C. T. (1999a). Detection and remediation of stagnation in the Nelder–Mead algorithm using a sufficient decrease condition. *SIAM Journal on Optimization*, *10*(1), 43–55. https://doi.org/10.1137/s1052623497315203
Kelley, C. T. (1999b). *Iterative methods for optimization.* SIAM. https://doi.org/10.1137/1.9781611970920
Khan, M. (1976a). Optimum allocation in Bayesian stratified two-phase sampling (a general case). *Journal of the Indian Statistical Association*, *14*, 65–74.
Khan, M. (1976b). Optimum allocation in Bayesian stratified two-phase sampling when there are m-attributes. *Metrika*, *23*(1), 211–219. https://doi.org/10.1007/bf01902867
Kiefer, J., & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, *23*(3), 462–466. https://doi.org/10.1214/aoms/1177729392
Kish, L. (1965). *Survey sampling.* Wiley.
Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, *8*(2), 183–200.
Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error

in simulation-based statistical analyses. *The American Statistician*, *63*(2), 155–162. https://doi.org/10.1198/tast.2009.0030

Kushner, H. (2010). Stochastic approximation: a survey. *WIREs Computational Statistics*, *2*(1), 87–96. https://doi.org/10.1002/wics.57

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, *9*(1), 112–147. https://doi.org/10.1137/s1052623496303470

Lavrakas, P., Benson, G., Blumberg, S., Buskirk, T., Cervantes, I. F., Christian, L., ... Shuttles, C. (2017). *AAPOR report: The future of U.S. general population telephone survey research.* (https://aapor.org/wp-content/uploads/2022/11/Future-of-Telephone-Survey-Research-Report.pdf [Accessed 8 May 2023].)

Le Floc'h, F. (2012). Issues of Nelder-Mead simplex optimisation with constraints. *Available at SSRN*. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2097904.) https://doi.org/10.2139/ssrn.2097904

Lindley, D. V. (1972). *Bayesian statistics: a review.* Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970654

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys. *Public Opinion Quarterly*, *72*(1), 6–27. https://doi.org/10.1093/poq/nfn003

Link, M. W., & Burks, A. T. (2013). Leveraging auxiliary data, differential incentives, and survey mode to target hard-to-reach groups in an address-based sample design. *Public Opinion Quarterly*, *77*(3), 696–713. https://doi.org/10.1093/poq/nft018

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546–556. https://doi.org/10.1198/016214504000000467

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

Little, R. J. (2022). A note about the definition of response propensity for survey nonresponse. *Journal of Survey Statistics and Methodology*, *10*(4), 1098–1106.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

Lohr, S. L. (2009). *Sampling: design and analysis* (2nd ed.).

Lohr, S. L. (2021). *Sampling: design and analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429298899

Lohr, S. L., & Brick, J. M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology*, *2*(4), 388–409. https://doi.org/10.1093/jssam/smu016

Long, Q., Scavino, M., Tempone, R., & Wang, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, *259*, 24–39. https://doi.org/10.1016/j.cma.2013.02.017

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: results of an international study across countries and surveys.

Journal of Official Statistics, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025

Mahalanobis, P. (1952). Some aspects of the design of sample surveys. *Sankhyā: The Indian Journal of Statistics*, 1–7.

Mason, R., Wheeless, S., George, B., Dever, J., Riemer, R., & Elig, T. (1995). Sample allocation for the Status of the Armed Forces Surveys. In *ASA Proceedings of the Survey Research Methods Section* (Vol. 2, pp. 769–774). (http://www.asasrms.org/Proceedings/papers/1995_133.pdf [Accessed 1 February 2023])

McKinnon, K. I. (1998). Convergence of the Nelder–Mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, *9*(1), 148–158. https://doi.org/10.1137/s1052623496303482

Mendenhall, W., & Lehman, E. (1960). An approximation to the negative moments of the positive binomial useful in life testing. *Technometrics*, *2*(2), 227–242. https://doi.org/10.1080/00401706.1960.10489896

Mersmann, O. (2021). microbenchmark: Accurate timing functions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=microbenchmark (R package version 1.4.9)

Müller, P., & Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, *90*(432), 1322–1330. https://doi.org/10.1080/01621459.1995.10476636

Müller, P., Sansó, B., & De Iorio, M. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, *99*(467), 788–798. https://doi.org/10.1198/016214504000001123

National Academies of Sciences, Engineering, and Medicine. (2018). *Reengineering the Census Bureau's annual economic surveys*. National Academies Press. https://doi.org/10.17226/25098

National Research Council. (2013). *Nonresponse in social science surveys: a research agenda*. National Academies Press. https://doi.org/10.17226/18293

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

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. https://doi.org/10.2307/2342192

Office of Management and Budget. (2017). *Statistical programs of the United States Government, Fiscal Year 2017*. (https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/information_and_regulatory_affairs/statistical-programs-2017.pdf [Accessed 23 April 2017])

Office of People Analytics. (2022). *2021 Workplace and Gender Relations Survey of Military Members - Reserve Component: Statistical methodology report* (OPA Report No. 2022-137). Office of People Analytics, U.S. Department of Defense. (https://apps.dtic.mil/sti/pdfs/AD1178879.pdf [Accessed 17 January 2023].)

Oh, H. L., & Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys, vol. 2: Theory and bibliographies* (pp. 143–184). New York: Academic Press.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., ... Wagner, J. (2021). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, *9*(3), 381–411. https://doi.org/10.1093/jssam/smz062

Olson, K., Wagner, J., & Anderson, R. (2020). Survey costs: Where are we and what is the way forward? *Journal of Survey Statistics and Methodology*, *9*(5), 921–942. https://doi.org/10.1093/jssam/smaa014

O'Malley, A. J., & Zaslavsky, A. M. (2016). Optimal small-area estimation and design when nonrespondents are subsampled for follow-up. *Journal of Survey Statistics and Methodology*, *4*(1), 2–21. https://doi.org/10.1093/jssam/smv036

Overstall, A. M., & McGree, J. (2020). Bayesian design of experiments for intractable likelihood models using coupled auxiliary models and multivariate emulation. *Bayesian Analysis*, *15*(1), 103–131. https://doi.org/10.1214/19-ba1144

Overstall, A. M., & Woods, D. C. (2017). Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, *59*(4), 458–470. https://doi.org/10.1080/00401706.2016.1251495

Overstall, A. M., Woods, D. C., & Parker, B. M. (2020). Bayesian optimal design for ordinary differential equation models with application in biological science. *Journal of the American Statistical Association*, *115*(530), 583–598. https://doi.org/10.1080/01621459.2019.1617154

Parkinson, J., & Hutchinson, D. (1972). An investigation into the efficiency of variants on the simplex method. *Numerical Methods for Nonlinear Optimization*, 115–135.

Peytchev, A. (2013). Consequences of survey nonresponse. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 88–111. https://doi.org/10.1177/0002716212461748

Plummer, M. (2017). Jags version 4.3.0 user manual [computer software manual]. *Retrieved from sourceforge. net/projects/mcmc-jags/files/Manuals/4.x*.

Powell, M. J. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez & J. Hennart (Eds.), *Advances in optimization and numerical analysis* (pp. 51–67). Springer. https://doi.org/10.1007/978-94-015-8330-5_4

Presser, S., & McCulloch, S. (2011). The growth of survey research in the United States: government-sponsored surveys, 1984–2004. *Social Science Research*, *40*(4), 1019–1024. https://doi.org/10.1016/j.ssresearch.2011.04.004

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raiffa, H., & Schlaifer, R. (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. https://doi.org/10.1214/10-sts346

Rao, J. N. K., & Ghangurde, P. D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association*, *67*(338), 439–443. https://doi.org/10.1080/01621459.1972.10482406

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons. https://doi.org/10.1002/9781118735855

Rider, P. R. (1955). Truncated binomial and negative binomial distributions. *Journal of the American Statistical Association*, *50*(271), 877–883. https://doi.org/10.1080/01621459.1955.10501973

Ritter, C., & Tanner, M. (1991). *The griddy Gibbs sampler* (Tech. Rep. No. 878). Madison, WI: University of Wisconsin, Madison.

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, *87*(419), 861–868. https://doi.org/10.1080/01621459.1992.10475289

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, *28*(2), 191–198.

Rizzo, L., Kalton, G., & Brick, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, *22*(1), 43–53.

Roshwalb, A. (1987). The estimation of a heteroscedastic linear model using inventory data. In *ASA Proceedings of the Business and Economic Statistics Section* (pp. 321–326).

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387. https://doi.org/10.1093/biomet/57.2.377

Ruppert, D. (1983). Kiefer–Wolfowitz procedure. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (Vol. 3, pp. 379–381). Wiley, New York.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*(1), 128–154. https://doi.org/10.1111/insr.12107

Ryan, E. G., Drovandi, C. C., Thompson, M. H., & Pettitt, A. N. (2014). Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics & Data Analysis*, *70*, 45–60. https://doi.org/10.1016/j.csda.2013.08.017

Sadegh, P. (1997, 5). Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. *Automatica*, *33*(5), 889-892. https://doi.org/10.1016/s0005-1098(96)00230-0

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., & Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*, *6*(4), 431–464. https://doi.org/10.1093/jssam/smy012

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, *24*(4), 495–506.

Sekkappan, R. (1981a). Generalization of Ericson's subjective Bayesian models in sampling finite populations for k characteristics. *J. Ind. Statist. Assoc.*, *19*, 159–164.

Sekkappan, R. (1981b). Subjective Bayesian multivariate stratified sampling from finite populations. *Metrika*, *28*(1), 123–132. https://doi.org/10.1007/bf01902884

Singh, B., & Sedransk, J. (1978). Sample size selection in regression analysis when there is nonresponse. *Journal of the American Statistical Association*, *73*(362), 362–365.

https://doi.org/10.1080/01621459.1978.10481583
Singh, B., & Sedransk, J. (1984). Bayesian inference and sample design for regression analysis when there is nonresponse. *Biometrika*, *71*(1), 161–170. https://doi.org/10.1093/biomet/71.1.161
Smith, P., Pont, M., & Jones, T. (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. https://doi.org/10.1111/1467-9884.03571
Smith, P. J., & Sedransk, J. (1982). Bayesian optimization of the estimation of the age composition of a fish population. *Journal of the American Statistical Association*, *77*(380), 707–713. https://doi.org/10.1080/01621459.1982.10477875
Smith, T. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society: Series A (General)*, *139*(2), 183–204. https://doi.org/10.2307/2345174
Soland, R. M. (1967). Optimal multivariate stratified sampling with prior information. *Scandinavian Actuarial Journal*, *1967*(1-2), 31–39. https://doi.org/10.1080/03461238.1967.10406205
Solomon, H., & Zacks, S. (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. https://doi.org/10.1080/01621459.1970.10481115
Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proceedings of the 1987 American Control Conference* (pp. 1161–1167).
Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, *37*(3), 332–341. https://doi.org/10.1109/9.119632
Spall, J. C. (1998). Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, *34*(3), 817–823. https://doi.org/10.1109/7.705889
Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. *The Annals of Mathematical Statistics*, *16*(1), 50–61. https://doi.org/10.1214/aoms/1177731170
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. https://doi.org/10.1093/jssam/smv038
Szeitl, B., & Rudas, T. (2022). Reducing variance with sample allocation based on expected response rates in stratified sample designs. *Journal of Survey Statistics and Methodology*, *10*(4), 1107–1120. https://doi.org/10.1093/jssam/smab021
U.S. Census Bureau. (2022). *2021 National Survey of Children's Health: Methodology report* (Tech. Rep.). U.S. Census Bureau. (https://www2.census.gov/programs-surveys/nsch/technical-documentation/methodology/2021-NSCH-Methodology-Report.pdf [Accessed 17 January 2023])
Valliant, R. (2023). Hansen Lecture 2022: The evolution of the use of models in survey sampling. *Journal of Survey Statistics and Methodology*, smad021. https://doi

.org/10.1093/jssam/smad021
Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. Springer. https://doi.org/10.1007/978-1-4614-6449-5
Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical tools for designing and weighting survey samples* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-93632-1
Valliant, R., Dever, J. A., & Kreuter, F. (2020). PracTools: tools for designing and weighting survey samples [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=PracTools (R package version 1.2.2)
Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*. John Wiley.
Valliant, R., & Gentle, J. E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, *25*(3), 337–360. https://doi.org/10.1016/s0167-9473(97)00007-8
Valliant, R., Hubbard, F., Lee, S., & Chang, C. (2014). Efficient use of commercial lists in us household sampling. *Journal of Survey Statistics and Methodology*, *2*(2), 182–209. https://doi.org/10.1093/jssam/smu006
Wagner, J., West, B. T., Elliott, M. R., & Coffey, S. (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. https://doi.org/10.2478/jos-2020-0043
Wald, A. (1950). *Statistical decision functions*. Wiley.
Walker, J. J. (1984). Improved approximations for the inverse moments of the positive binomial distribution. *Communications in Statistics-Simulation and Computation*, *13*(4), 515–519. https://doi.org/10.1080/03610918408812393
Wang, I.-J., & Spall, J. C. (2008). Stochastic optimisation with inequality constraints using simultaneous perturbations and penalty functions. *International Journal of Control*, *81*(8), 1232–1238. https://doi.org/10.1080/00207170701611123
West, B. T., Wagner, J., Coffey, S., & Elliott, M. R. (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*. https://doi.org/10.1093/jssam/smab036
Williams, D., & Brick, J. M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2). https://doi.org/10.1093/jssam/smx019
Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, *78*(384), 879–884. 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. https://doi.org/10.1080/00401706.1970.10488639
Zangeneh, S. Z., & Little, R. J. (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. https://doi.org/10.1093/jssam/smv002
Zuras, D., Cowlishaw, M., Aiken, A., Applegate, M., Bailey, D., Bass, S., ... others (2008).

IEEE standard for floating-point arithmetic. *IEEE Std. 754-2008*. https://doi.org/10.1109/ieeestd.2008.4610935