University of Wollongong

Research Online

Faculty of Engineering and Information
Sciences - Papers: Part A

Faculty of Engineering and Information
Sciences

1-1-2013

Sample design using imperfect design data

Robert Graham Clark

University of Wollongong, rclark@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/eispapers>

Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Sample design using imperfect design data

Abstract

A well-designed sampling plan can greatly enhance the information that can be produced from a survey. Once a broad sample design is identified, specific design parameters such as sample sizes and selection probabilities need to be chosen. This is typically achieved using an optimal sample design, which minimizes the variance of a key statistic or statistics, expressed as a function of design parameters and population characteristics, subject to a cost constraint. In practice, only imprecise estimates of population characteristics are available, but the effects of this variability are usually ignored. A general approach to sample allocation allowing for imprecise design data is proposed and evaluated. The approach is based on the availability of two sets of design data, which can act as a check on each other. One application is stratified sampling. Optimal allocation could be used, but estimated strata variances may be highly variable. Pooling strata into groups may reduce this variability, at the possible cost of some inefficiency. Proportional allocation, ignoring differences among strata variances, could also be used. The new approach enables a data-driven compromise among all three. Simulation results based on real data show useful gains in a hypothetical farm survey, business survey, and household survey of a subpopulation.

Keywords

data, sample, design, imperfect

Disciplines

Engineering | Science and Technology Studies

Publication Details

Clark, R. Graham. (2013). Sample design using imperfect design data. *Journal of Survey Statistics and Methodology*, 1 (1), 6-23.

SAMPLE DESIGN USING IMPERFECT DESIGN DATA

ROBERT GRAHAM CLARK*

A well-designed sampling plan can greatly enhance the information that can be produced from a survey. Once a broad sample design is identified, specific design parameters such as sample sizes and selection probabilities need to be chosen. This is typically achieved using an optimal sample design, which minimizes the variance of a key statistic or statistics, expressed as a function of design parameters and population characteristics, subject to a cost constraint. In practice, only imprecise estimates of population characteristics are available, but the effects of this variability are usually ignored. A general approach to sample allocation allowing for imprecise design data is proposed and evaluated. The approach is based on the availability of two sets of design data, which can act as a check on each other. One application is stratified sampling. Optimal allocation could be used, but estimated strata variances may be highly variable. Pooling strata into groups may reduce this variability, at the possible cost of some inefficiency. Proportional allocation, ignoring differences among strata variances, could also be used. The new approach enables a data-driven compromise among all three. Simulation results based on real data show useful gains in a hypothetical farm survey, business survey, and household survey of a subpopulation.

1. INTRODUCTION

In practice, sample designs are based on limited or imprecise information. Perhaps the most common example is the optimal allocation method first proposed by [Neyman \(1934\)](#) for stratified simple random sampling

ROBERT GRAHAM CLARK is Associate Professor at the National Institute for Applied Statistics Research Australia, University of Wollongong. The author thanks Janice Scealy, David Steel, Robert Templeton, and Alan Welsh for their advice and encouragement. Three reviewers made detailed and thoughtful comments that improved the article. *Address correspondence to Robert Clark, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522; E-mail: rclark@uow.edu.au.

doi: 10.1093/jssam/smt002

© The Author 2013. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

without replacement. The variance of an estimator of the population total is equal to

$$V = \sum_{h=1}^H V_h N_h^2 n_h^{-1} \quad (1)$$

plus a term that does not depend on the strata sample sizes n_h , where V_h is the population variance and N_h is the population size of stratum h (e.g., Cochran 1977). The allocation that minimizes (1) subject to fixed total sample size satisfies $n_h \propto N_h \sqrt{V_h}$. In practice, however, the values of V_h are unknown, and estimates are substituted. This is called a *plug-in* allocation, because the optimal allocation is derived assuming knowledge of V_h , but estimates are then substituted. In some cases, differential costs apply in each stratum, and (1) may then be minimized subject to fixed cost rather than fixed sample size.

Several authors have commented that plug-in allocations are less efficient than the ideal optimal allocation. Lohr (2009, p. 90) commented that if the $\hat{V}_h$ used in allocating stratified samples are very imprecise, then the plug-in design may do worse than the proportional allocation $n_h \propto N_h$, which makes no use of $\{\hat{V}_h\}$.

Smith, Pont, and Jones (2003) highlighted the importance of allocation to strata in business surveys in the United Kingdom. They noted that “the population standard errors . . . must be estimated from previous samples. In practice these estimated samples could themselves be extremely volatile . . . Allocations based strictly on such data would be unlikely to be optimal in practice, so ‘smoothing’ would often be needed to achieve more robust results.” In the related problem of choosing the within-cluster sample size in two-stage sampling, Cochran (1977), summarizing Brooks (1955), found that a large pilot sample of around 150 units may be needed to achieve precisions within 10 percent of the ideal optimal design.

Cochran (1977, §5A.1) considered the effect of deviations from optimal allocations in stratified sampling. It was argued that the effect would often be small; for example, if the worst deviation in a stratum sample size from the optimum is 30 percent, the resulting increase of the true variance is at most 9 percent relative to the optimum, and often less than this. Cochran also cited Sukhatme (1935) and Evans (1951), who found that allocations based on a relatively small pilot study would often do better than proportional allocation. However, all three authors apparently had in mind imprecision in allocation due to sampling errors only and not due to the use of out-of-date information. Moreover, the empirical study in this article will show that the sampling errors of estimators of V_h may sometimes be much larger than supposed by these authors.

Another difficulty in applying the plug-in allocation is in estimating the variance that will be achieved, in advance of running the survey. The plug-in allocation $n = (n_1, \dots, n_H)$ depends on the estimated variances

$\hat{V} = (\hat{V}_1, \dots, \hat{V}_H)$ and is therefore itself variable. The values of $\mathbf{n}$ would normally be conditioned on once the survey is conducted, and the true variance V is then defined by (1). This quantity is normally estimated by substituting both the plug-in $\mathbf{n}$ and the estimated variances $\hat{V}$ into (1). Clearly this has the potential for bias, because the values of n_h^{-1} and $\hat{V}_h$ are likely to be negatively correlated, meaning that the pre-survey estimate of V will tend to be overly optimistic.

The potential shortfalls of the plug-in method are clear, and suggest two questions:

- (1) Under what conditions is the variability of $\hat{V}$ likely to have an appreciable effect on either the true variance V or the pre-survey estimation of V ?
- (2) Can the plug-in allocation be improved on?

These are the topics of this article. The premise of the proposed approach is that two design data sets are available, which can act as a check on each other. The next section defines a general formulation of the allocation problem and the approach of using a training and validation sample. Theoretical results are difficult to derive, but two simple theorems will be stated. Then I present a simulation study, with parameters of the simulation model obtained by analyzing three real data sets. The article then extends the simulation study by varying these parameters, to identify when the new method provides useful gains. A final section contains conclusions.

The findings will be of interest to researchers and companies who design and carry out surveys and who must make robust design decisions using borrowed data or small pilot studies, as well as to national statistics institutes who have ready access to repeated survey data for design purposes.

2. A STATISTICAL LEARNING APPROACH TO OPTIMAL ALLOCATION

2.1 Motivating Case: Neyman Allocation with Grouping of Strata

Consider the following scenario, which forms the basis of the simulation study later in this article. A survey is run on multiple occasions. For example, many national health surveys are conducted every three or four years, and other surveys are run on an ongoing basis every month, quarter, or year. It is assumed that independent data from two previous surveys, labeled time 1 and time 2, are available to help in designing a future survey to be conducted at time 3. A stratified sample design is used in which strata naturally form into groups. For example, strata in a household survey may be defined by small geographic areas, with groups defined by larger areas, whereas strata in a business survey may be defined by detailed industry, size, and state or province, with groups defined by just broad industry and size. Let V_{ht} and $\hat{V}_{ht}$ refer

to the population and estimated variances for stratum h at time t , for $t = 1, 2, 3$. It is assumed that n_h depends only on $\{\hat{V}_{h2} : h = 1, \dots, H\}$ and $\{\hat{V}_{ht} : h = 1, \dots, H\}$. Typically $\hat{V}_{ht}$ would be a simple variance calculated from sample data in stratum h for time t . The aim is to minimize a loss function which, here, is the variance of the time 3 estimate of interest:

$$V_{\text{tot}(3)} = \sum_{h=1}^H V_{h3} N_h^2 n_h^{-1} \quad (2)$$

subject to fixed total sample size.

Three possible allocations are:

- (i) Plug-in optimal allocation with $n_h \propto N_h \sqrt{\hat{V}_{h2}}$. This is the most common approach in practice.
- (ii) Grouped optimal allocation with $n_h \propto N_h \sqrt{\hat{V}_{k2}}$ for stratum h in group k , where $\hat{V}_{kt}$ is a population-weighted average of the time t strata variance estimates in group k , for $t = 1, 2$. This allocation might be used if the strata estimates $\hat{V}_{h2}$ were thought to be too variable to be useful within groups.
- (iii) Proportional allocation with $n_h \propto N_h$. This would be used if both $\hat{V}_{h2}$ and $\hat{V}_{k2}$ were thought to be too variable to be of any use in allocation. It is convenient to express this allocation in similar form to i and ii: $n_h \propto N_h \sqrt{\hat{V}_2}$, where $\hat{V}_t$ is a population-weighted average of the time t strata variance estimates over all strata, for $t = 1, 2$.

All three allocations use data from the most recent available survey, the one conducted at time 2. Option i should be the most efficient when $\hat{V}_{h2}$ is very precise. Option ii should do well when there is considerable variability in $\hat{V}_{h2}$ but $\hat{V}_{k2}$ is at a broad enough level to be more precise, particularly if the groups capture most of the variability among the true stratum-level variances V_{h3} . Option iii would be appropriate when the design data are so unreliable that they are of virtually no use in sample design. The reality is likely to lie somewhere between these extremes. The proposed new approach is to make n_h a compromise among the three alternatives, defined by $n_h = n_h(\hat{V}_2, \boldsymbol{\lambda})$, where

$$n_h(\hat{V}_t, \boldsymbol{\lambda}) \propto N_h \sqrt{\lambda_1 \hat{V}_{ht} + \lambda_2 \hat{V}_{kt} + \lambda_3 \hat{V}_t} \quad (3)$$

($t = 1, 2$), with $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ denoting three design parameters summing to 1. The constant of proportionality in (3) is such that $\sum_{h=1}^H n_h$ equals the constraint for the total sample size n . The values of $\boldsymbol{\lambda}$ would then smoothly combine allocations i, ii, and iii, with $\boldsymbol{\lambda} = (1, 0, 0)$ corresponding to the usual plug-in allocation, $\boldsymbol{\lambda} = (0, 0, 1)$ corresponding to proportional allocation and so on.

The broad objective in the proposed approach is to choose λ so that $V_{\text{tot}(3)}$ in (2) is minimized. A naive approach would be to minimize

$$\sum_h N_h^2(n_h(\hat{V}_2, \lambda))^{-1} \hat{V}_{h2} \quad (4)$$

with respect to λ . Unfortunately, this approach will always result in $\lambda = (1, 0, 0)$, that is, the plug-in approach—this is a special case of the Neyman optimal design described earlier. The problem is that $n_h(\hat{V}_2, \lambda)$ would be expected to be positively correlated with $\hat{V}_{h2}$, unless $\lambda = (0, 0, 1)$, so that the loss function in (4) in no way penalizes the variability caused by plugging in estimates when calculating n_h . Instead, it is proposed to employ the loss function using both time 1 and time 2 data as follows:

$$\sum_h N_h^2(n_h(\hat{V}_1, \lambda))^{-1} \hat{V}_{h2}. \quad (5)$$

To borrow some terminology from the statistical learning literature (e.g., [Hastie, Tibshirani, and Friedman 2009](#)), “training” estimates $\hat{V}_1$ are used to calculate n_h given λ , and separate “validation” estimates $\hat{V}_2$ are used to assess the performance of any given value of λ . The value of λ is chosen to minimize (5). The proposed allocation at time 3 is then $n_h = n_h(\hat{V}_2, \lambda)$ with this value of λ . The minimized value of the loss function in (5) could also be used to estimate the variance that will be achieved at time 3. The values of n_h in (5) should be approximately uncorrelated with V_{h2} , thereby avoiding the bias of (4). (This is proven shortly for a more general formulation of the design problem.)

This motivating case may seem to be very specific, but in practice variances for many statistics of interest from many possible designs are special cases of the variance function (1). For example, the variance under two-phase sampling for stratification is of this form, with a linear cost model often being assumed ([Cochran 1977](#), p. 330). The design parameters are the first phase sample size and the second phase strata sample sizes. Two-stage sampling is another special case. A sample of n_1 clusters (e.g., areas) is selected, followed by a sample of n_2 units (e.g., households or people) within selected clusters. The variance is then of the form (1) and a linear cost model is often assumed ([Cochran 1977](#), p. 277) although more complex cost models are sometimes used. Stratified two-stage sampling can also be written in the form of (1), where the design parameters are the sample sizes of clusters and of units in each stratum ([Clark and Steel 2000](#)), as can two-stage sampling with unequal probabilities of selection for each cluster and different within-cluster sample sizes ([Clark 2009](#)).

Equation (3) defined one choice of the smoothing function $n(\lambda, \nu)$. There are a variety of other ways that this function could be defined. For example, (3) could be generalized to allow a hierarchy of groupings of strata. If strata were industry by size by state (h), then groups could be defined by industry by size (k), and broader groups by just size (l). The smoothing function (3) could be extended to

include $\hat{V}_l$ as well as $\hat{V}_k$. Another example might be a stratified repeated business survey, in which allocations may be calculated using many previous instances of the survey. An efficient design might be obtained by estimating V_h using a weighted average over many past surveys, with λ consisting of these weights.

To allow for all of these possibilities, a more general formulation of the allocation problem is needed. The next section defines this and proposes a general statistical learning approach.

2.2 General Formulation of Allocation Problem and Proposed Approach

The allocation problem is defined as the choice of a vector of design parameters $n = (n_1, \dots, n_H)$ which specify the sampling method. The aim will be to choose n to minimize $L(n; V)$ subject to a cost constraint $C_f = C(n)$ for a known function $C(\cdot)$ and to range constraints such as $n_h > 0$. $L(\cdot, \cdot)$ is a loss function that would typically be the variance of a statistic of interest, or a linear combination of several such variances. The vector V would be a set of population parameters. The previous section gave one example of this framework. Let $n(v) = \arg \min_{n: C(n)=C_f} L(n; v)$. The ideal design is obviously $n(V)$, but in practice V would be unknown. Instead, the current practice is to use the plug-in allocation $n(\hat{V})$, which would be expected to be more variable and result in a higher loss. To manage the variability in $\hat{V}$, the following approach is proposed. The design parameters n based on an estimate $\hat{V}$ of V will be defined to equal $n(\lambda, \hat{V})$, where $n(\cdot, \cdot)$ is a function chosen by the sample designer such that λ is a p-vector controlling the “complexity” of the design. The details of how this complexity is defined depend on the particular survey, but it would generally be related to the variation in the probabilities of selection across the population. The function would be such that one value of λ , say, λ_0 , would result in $n(\lambda_0, \hat{V})$ being the plug-in allocation, and on the other extreme, λ_1 would be such that $n(\lambda_1, \hat{V})$ does not depend on $\hat{V}$ at all. For example, in the special case described earlier, $\lambda_0 = (1, 0, 0)$ and $\lambda_1 = (0, 0, 1)$.

The aim is to choose a suitable value of λ , which results in something close to the plug-in design when $\hat{V}$ is highly precise, and in a simpler design (in the sense of less variable probabilities of selection) when $\hat{V}$ is a poor estimate. Ideally, λ should minimize $L(n; V)$, or to be precise,

$$L(n(\lambda, \hat{V}); V). \quad (6)$$

We can't reliably estimate this loss using a single estimate of V , because

$$L(n(\lambda, \hat{V}); \hat{V}) \quad (7)$$

is by definition minimized by $\lambda = \lambda_0$, giving the plug-in allocation. The problem is that (7) is negatively biased for (6), as will be demonstrated in the next subsection.

To improve on the plug-in design, it is assumed that independent training and validation estimates of V , $\hat{V}_{\text{train}}$ and $\hat{V}_{\text{valid}}$, are available. Then we calculate λ to minimize $L(\mathbf{n}(\lambda, \hat{V}_{\text{train}}); \hat{V}_{\text{valid}})$ and implement an allocation based on this λ .

The preceding development was inspired by the discussion of model choice in chapter 7 of Hastie, Tibshirani, and Friedman's book on statistical learning (Hastie, Tibshirani, and Friedman 2009), where loss estimates based on a validation data set are used to choose tuning parameters that control the model's size or complexity. The problem here is only weakly analogous to model choice, however, with very different interpretations of both the loss function and the complexity of the design/model.

2.3 Two Theorems

This subsection states two theorems with proofs in the Appendix. The first result is that (5) is unbiased subject to regularity conditions. The second is that (4) is negatively biased. To enable clear and interpretable results, strong assumptions are made. Real surveys are likely to be messier, and this is reflected in the simulations that appear later, which do not satisfy the assumptions of these theorems but still show useful gains from the new approach.

Theorem 1 Let $\mathbf{n} = \mathbf{n}(\lambda, \hat{V}_{\text{train}})$. If

A1. $E[\hat{V}_{\text{valid}} - V | \hat{V}_{\text{train}}] = \mathbf{0}$; and

A2. $L(\mathbf{n}, V)$ is a linear function of V ,

then $E[L(\mathbf{n}, \hat{V}_{\text{valid}}) - L(\mathbf{n}, V)] = 0$ for any λ .

Condition A1 of Theorem 1 states that the validation estimate of V is unbiased, conditional on the training data. This can be interpreted as meaning that $\hat{V}_{\text{valid}}$ is unbiased and is independent of $\hat{V}_{\text{train}}$ in a particular way. If the training and validation data come from different time periods, it would also require constancy of the stratum variances over time. Note that A1 implies that $E[\hat{V}_{\text{valid}} - V] = \mathbf{0}$. Condition A2 is satisfied for the Neyman function (1).

Because (5) is unbiased for any value of λ , it is reasonable to use it to optimize λ . In contrast, the plug-in estimator (4) is shown in Theorem 2 to be negatively biased for all λ except the value(s) of λ corresponding to not using any of the design data in the allocation.

Theorem 2 Conditions A1 and A2 from Theorem 1 are assumed as well as

A3. $\mathbf{n}(\lambda, \mathbf{v}) = \arg \min_{\mathbf{n} \in C(\lambda)} L(\mathbf{n}, \mathbf{v})$ for any $\mathbf{v}$, where $\mathbf{n} \in C(\lambda)$ represents a constraint on $\mathbf{n}$ depending on λ .

Then $E[L(\mathbf{n}, \hat{V}_{\text{train}}) - L(\mathbf{n}, V)] \leq 0$ for any λ , with strict inequality except when $L(\mathbf{n}(\lambda, \mathbf{v}), \hat{V}_{\text{train}})$, does not depend on $\mathbf{v}$, which would normally only occur if λ was such that no design data were used in calculating $\mathbf{n}$.

Assumption A3 is that $\mathbf{n}$ can be expressed as the minimizer of a loss function subject to a constraint depending on λ . For example, the plug-in design would correspond to an unconstrained $\mathbf{n}$, and proportional allocation is equivalent to constraining n_h fully by requiring them to be proportional to N_h .

Theorem 2 means that the usual plug-in approach is biased in favor of values of λ such that $\mathbf{n}$ is more dependent on the design data. This results in a suboptimal choice of λ and in the variance that will be achieved being underestimated at the design stage. Theorem 1 suggests that the proposed statistical learning approach can solve both problems.

3. SIMULATION STUDY WITH PARAMETERS BASED ON SURVEY DATA

3.1 Model for Population and Sample Strata Variances

Population and sample variances for each stratum within groups of strata for three time periods are generated by using an assumed model. Various allocation methods are calculated and evaluated for each generated set of strata variances. The population variances V_{ht} will be the product of a group factor A_{kt} and a within-group stratum factor B_{ht} as follows:

$$\left. \begin{aligned} \log(V_{ht}) &= \mu + A_{kt} + B_{ht} \text{ where stratum } h \text{ belongs to group } k \\ \mathbf{A} &\sim N(\mathbf{0}, \sigma_{\text{group}}^2 \mathbf{R}_{\text{group}}) \\ \mathbf{B} &\sim N(\mathbf{0}, \sigma_{\text{stratum}}^2 \mathbf{R}_{\text{stratum}}) \\ \mathbf{A} \text{ and } \mathbf{B} &\text{ independent} \end{aligned} \right\}, \quad (8)$$

where $\mathbf{A}$ is the vector of all A_{kt} over k and t , and $\mathbf{B}$ is the vector of all B_{ht} over h and t . The correlation matrix $\mathbf{R}_{\text{group}}$ has 1s on the diagonal and is assumed to be such that A_{k_1t} and A_{k_2t} are independent when $k_1 \neq k_2$, and $\text{corr}[A_{k_1t}, A_{k_2t}] = \rho_{\text{group}}^{|t_2 - t_1|}$. Similarly, the correlation matrix $\mathbf{R}_{\text{stratum}}$ has 1s on the diagonal and is assumed to be such that B_{h_1t} and B_{h_2t} are independent when $h_1 \neq h_2$, and $\text{corr}[A_{h_1t}, A_{h_2t}] = \rho_{\text{stratum}}^{|t_2 - t_1|}$.

Given the strata population variances, V_{ht} , the strata sample variances are generated as independent scaled chi-squared distributions, with

$$\hat{V}_{ht} \sim V_{ht} \chi_d^2 / d, \quad (9)$$

where d are the degrees of freedom.

There are six parameters that can be varied in the simulation: μ , σ_{group}^2 , $\sigma_{\text{stratum}}^2$, ρ_{group} , ρ_{stratum} , and d , as well as the number of groups, K , and the number of strata per group, H_1 . The first of these, μ , is just assumed to be 0, because it affects variances and estimated variances as a simple multiplicative factor, and so does not affect the relative performance of the different allocation methods.

Strata population sizes N_{kht} were set to be inversely proportional to the true variances, $\bar{V}_{kh} = \sum_{t=1}^3 V_{kht} / 3$, because this is approximately the case in some

equal aggregate stratification methods. (Equal strata population sizes were also simulated with similar conclusions. For details see the appendix in the online Supplementary Materials.)

It is worth noting that the assumed model does not satisfy the assumptions of the theorems presented earlier. In particular, the training and validation estimates $\hat{V}_{h2}$ and $\hat{V}_{h3}$ follow a stochastic model, rather than being assumed to be unbiased and independent estimates of the true stratum variances V_{h3} . The theorem required strong assumptions to obtain a simple theoretical result, whereas the simulation study provides a test of the new approach in more realistic conditions.

3.2 Empirical Fitting of Model for Strata Population Variances for Three Data Sets

The five parameters that can be varied in the simulation give rise to an astronomical number of plausible scenarios that could be simulated. This subsection describes the fitting of model (8) and (9) to three data sets. The following subsection describes simulations based on these estimated model parameters.

The first data set was extracted from the Australian Agriculture and Grazing Industries Survey 1991–95, conducted by the Australian Bureau of Agricultural and Resource Economics. Sample data from the two largest industries (cropping specialists and mixed cropping and sheep) were used. Strata were defined by industry, size, and state (five largest states in Australia). Three size categories were defined for each industry by year cell, such that the sums of the square root of total land cleared over farms were equal for each size. (This is an example of an equal aggregate approach to size stratification—see, for example, Valliant, Dorfman, and Royall [2000], section 6.5.2.) This meant that the size categorization changed from year to year, although only slightly. The aim is assumed to be to estimate the population total of annual total cash income from crops. Groups were defined as industry by size, because these are thought to be more important than state, although state also matters due to differences in climate, accessibility and remoteness, and so on. Strata with a sample size of less than six were excluded, leaving a total of 3,189 farms in 39 strata in 9 groups over five years.

The second data set consisted of data on enterprises with up to 100 employees from the Business Longitudinal Survey, conducted by the Australian Bureau of Statistics for the financial years 1994–95, 1995–96, 1996–97, and 1997–98. Only the industries metal product manufacturing, machinery and equipment manufacturing, machinery and motor vehicle wholesaling, and personal and household goods wholesaling were used, as these had large enough sample sizes to avoid very small strata. Strata were defined by industry, size, and type of legal organization (incorporated/unincorporated). This last variable is often used as a stratifying variable, but is thought to be less distinguishing than industry and size. Hence, groups were defined as industry by size. The

size variable was defined based on total employment from 1993, with equal aggregates of the square root of employment in each industry. The variable of interest was assumed to be annual total business income. This gave a sample of 7,328 businesses in 32 strata in 16 groups over four years.

Finally, 2001 and 2006 Counts of the Pacific Population by meshblock were obtained from the 2006 New Zealand (NZ) Census. The Pacific population consists of people born in the Pacific islands or who identify themselves as part of this population. The 2006 NZ meshblock file records a total of 4.03 million usual residents in approximately 41,000 meshblocks (small areas containing on average about 100 people), of whom approximately 6.6 percent are recorded to be in the Pacific population. It is assumed that the aim is to estimate means and proportions for the Pacific population, using a sample from the general population, stratified by meshblock. In this situation, subject to some assumptions, the relevant population stratum variance is proportional to the proportion of the meshblock population who belong to the Pacific population (Kalton and Anderson 1986). Strata are grouped into area units (larger areas containing on average about 2,000 people). To facilitate computation, a sample of 250 area units and 10 meshblocks from each were selected for analysis. This is a somewhat unrealistic scenario, as there are too many meshblocks for these to be an appropriate stratifier. In reality, two-stage sampling could be used, with meshblocks as primary sampling units. However, the resulting variance expression is approximately equivalent to the one assumed here (Clark 2009), so applying model (8) to this data set still gives a useful insight into which parameter values are likely to crop up in real surveys.

The first step in the empirical analysis was to fit model (8) by maximum likelihood, assuming that $\hat{V}_{hi}$ is equal to V_{hi} . This was done because of the difficulty of specifying a model for $\hat{V}_{hi}$ that would fit these data sets, due to the variation in strata sample sizes and the skewed, heavy-tailed distributions of the farm and business data sets. To correct for this, an empirical bootstrap bias correction was applied with 30 resamples (e.g., Chernick 2008, pp. 26–27; Davison and Hinkley 1997, pp. 15, 48). The correlation parameters ρ_{stratum} and ρ_{group} were first transformed using the hyperbolic arctangent transformation (also known as the Fisher z transformation), to ensure that the corrected estimates lay between -1 and 1 . Bias correction was not required for the census data set. Table 1 shows the parameter estimates with bootstrap standard errors in brackets. For more detail on the bootstrap approach, see the appendix in the online Supplementary Materials.

It is notable from the table that population variances vary greatly across strata, as shown by the values of σ_{group}^2 and $\sigma_{\text{stratum}}^2$. In the farms survey and the NZ Pacific population, the variation among strata within groups is much greater than the variation among groups; the reverse is true for the business survey. The coefficient of variation (CV) of the group and stratum factors ($e^{A_{ki}}$ and $e^{B_{khi}}$) are equal to $cv_{\text{group}} = \sqrt{e^{\sigma_{\text{group}}^2} - 1}$ and $cv_{\text{stratum}} = \sqrt{e^{\sigma_{\text{stratum}}^2} - 1}$

Table 1. Estimates of Parameters of Model (8) for Three Data Sets

Data Set	σ^2_{group}	$\sigma^2_{\text{stratum}}$	ρ_{group}	ρ_{stratum}	CV (%) ($e^{A_{kt}}$)	CV (%) ($e^{B_{kht}}$)
Farms survey	1.27 (0.26)	2.21 (0.38)	1.00 (0.00)	0.89 (0.03)	160	285
Business survey	3.76 (0.27)	2.28 (0.25)	1.00 (0.00)	0.99 (0.00)	647	296
NZ Pacific population	0.88	1.57	0.98	0.44	119	195

Table 2. Estimates of Parameter d in (9) and Associated Information

Survey	$\hat{P}[Y = 0]$	SD(log (Y)) (σ)	Estimated d.f. $\hat{d}$			Estimated CV% of $\hat{V}$		
			$n = 6$	$n = 10$	$n = 20$	$n = 6$	$n = 10$	$n = 20$
Farms survey	0.00	0.65	3.2	5.7	11.7	79	59	41
Business survey	0.32	1.14	1.7	2.9	5.8	109	83	59

(e.g., Johnson, Kotz, and Balakrishnan 1994, p. 212). These values are well above 100 percent. This is perhaps not surprising, particularly in the surveys of financial variables for farms and businesses, where data are known to be right-skewed and heavy tailed. The stratum and group factors are generally stable over time, with autocorrelations of around 0.9 and 1, respectively.

The parameter d in (9) was estimated by a small parametric simulation based on the farms and business data sets discussed earlier. In both cases, the unit values of the variable of interest (which are denoted Y_{khti} for unit i in stratum h , group k , and time t) can reasonably be modeled as a mixture of zero values and log-normally distributed values (this was confirmed visually using q-q plots). It will be assumed that $P[Y_{khti} = 0] = p$, and $\log(Y_{khti}) \sim N(\alpha_{kht}, \gamma^2)$ conditional on $Y_{khti} > 0$. To obtain an estimate of d , 10,000 samples each containing n values of Y were generated from the fitted model, where n was 6, 10, or 20, these being reasonably typical stratum sample sizes. Observations of Y were truncated at the 97.5th percentile, reflecting that business surveys normally use some form of outlier correction, for example, winsorization. 10,000 observations of $\hat{V}$ were then calculated, and fit against (9) by matching of the first two moments, with the true variance and d being unknown parameters. Table 2 shows the resulting estimates of d . The estimates of p and γ are also shown. The model implies that the CV of $\hat{V}_{ht}$ given V_{ht} is $cv_{est} = \sqrt{2/d}$; the estimated CVs are also shown in table 2.

3.3 Simulation Results

One thousand sets of strata sample variances were simulated for three time periods using models (8) and (9). The parameters were obtained from the analyses described earlier and tabulated in tables 1 and 2. It is assumed that survey data from time 1 and 2 are to be used in the design of a survey to be conducted at time 3. For the simulations based on the farms and business survey, the number of groups (K) was set to 20 and the number of strata per group (H_1) was set to 5. For the NZ Pacific population example, $K = 5$ and $H_1 = 20$ were used.

For each simulation, the statistical learning allocation, defined by (3), was calculated, with λ chosen to minimize (5). The plug-in Neyman, plug-in grouped Neyman, and proportional allocations were also calculated, using only the time 2 data. An ideal statistical learning allocation was also evaluated. In this allocation, $\mathbf{n}$ was defined by (3), but λ was calculated to minimize the variance over all simulations, effectively assuming omniscience in the choice of λ . All simulations and empirical analyses were conducted in the R statistical environment (R Development Core Team 2012). The mvtnorm package was also used (Genz et al. 2012). The tables and figures here and in the next section can be fully reproduced using the programs contained in the Supplementary Materials, but the data sets used in the “Empirical Fitting” subsection are not available because of confidentiality restrictions. The new allocation method is implemented in a publicly available R package (Clark 2013).

Table 3 shows the true variances of the different allocation methods in (2), relative to proportional allocation. The statistical learning allocation is the best option in all cases. Particularly strong gains are apparent in the business survey example with 6 or 10 units per stratum (35 percent and 13 percent reduction in the variance compared to the plug-in Neyman allocation), in the farms example

Table 3. Variance Achieved by Different Allocation Methods Relative to Proportional Allocation for Simulations with Parameters Based on Three Data Sets and Several Values of n_h

Data set and assumed n_h in design data sets	Allocation method			
	Plug-in Neyman	SL	Plug-in grouped Neyman	SL with ideal λ
Farms with $n_h = 6$	0.730	0.654	0.774	0.651
Farms with $n_h = 10$	0.648	0.623	0.759	0.621
Farms with $n_h = 20$	0.628	0.611	0.750	0.610
Businesses with $n_h = 6$	0.821	0.535	0.557	0.531
Businesses with $n_h = 10$	0.589	0.514	0.536	0.512
Businesses with $n_h = 20$	0.501	0.485	0.514	0.483
New Zealand Pacific population	0.934	0.822	0.850	0.820

when the design data sets have 6 units per stratum (10 percent reduction in variance), and in the NZ Pacific example (12 percent reduction in variance).

It is notable that the statistical learning allocation with λ optimized using the training and validation data sets is very nearly as efficient as when the ideal λ is used. This suggests that a close-to-optimal value for λ can be obtained even using imprecise design data. This may be partly due to the fact that there are 100 strata in the study, all of which contribute training and validation estimates of V_h , so there is apparently ample information to guide the choice of λ .

An appendix included in the online Supplementary Materials shows that the simulation standard deviations associated with table 3 are small relative to the differences highlighted.

It is of interest to estimate the variance that will be achieved in the time 3 survey at the design stage. The usual variance estimator is (4), which Theorem 2 states is biased. An alternative would be to use (5), which is unbiased provided the assumptions of Theorem 1 are satisfied. However, one of the assumptions of Theorem 1 is that the V_t are constant over time, whereas in practice there may be systematic movement over time, for example, due to inflation in financial variables. The following estimator allows for this to some extent by use of a ratio adjustment:

$$\left\{ \sum_h N_h^2 (n_h(\hat{V}_2, \lambda))^{-1} \hat{V}_{h1} \right\} \left\{ \sum_h N_h \hat{V}_{h2} \right\} / \left\{ \sum_h N_h \hat{V}_{h1} \right\}. \tag{10}$$

Table 4 shows the ratio of the mean of the estimated variance (over 1,000 simulations) to the true variance given by (2), for the naive variance estimator

Table 4. Expected Value of Variance Estimate Divided by True Value for Naive and Proposed Variance Estimators, Defined by (4) and (10), for Simulations with Parameters Based on Real Data Sets and Several Values of n_h , for Three Different Allocation Methods

Data set and assumed n_h	Naive variance estimator			Proposed variance estimator		
	Plug-in Neyman	SL	Plug-in grouped Neyman	Plug-in Neyman	SL	Plug-in grouped Neyman
Farms with $n_h = 6$	0.574	0.692	0.849	1.008	1.004	1.008
Farms with $n_h = 10$	0.695	0.752	0.877	1.021	1.018	1.019
Farms with $n_h = 20$	0.733	0.776	0.885	1.016	1.016	1.014
Businesses with $n_h = 6$	0.425	0.782	0.899	0.993	0.990	0.994
Businesses with $n_h = 10$	0.659	0.826	0.933	0.998	0.998	0.998
Businesses with $n_h = 20$	0.833	0.889	0.969	1.007	1.004	1.002
New Zealand Pacific population	0.655	0.836	0.983	0.998	1.002	1.005

and the improved version (10). The parameter settings are the same as in table 3. The naive variance estimator has substantial negative bias, of up to 58 percent, particularly when Neyman allocation is used and n_h is small in the design datasets. The proposed variance estimator is very nearly unbiased in all cases.

4. FURTHER SIMULATIONS

As before, 1,000 design data sets for three time periods were generated using models (8) and (9). The previous section described the results when the parameters of the simulation models were based on three survey data sets. This section investigates the range of possible outcomes by varying each of the parameters $cv_{\text{group}} = \sqrt{e^{\sigma_{\text{group}}^2} - 1}$, $cv_{\text{stratum}} = \sqrt{e^{\sigma_{\text{stratum}}^2} - 1}$, ρ_{group} , ρ_{stratum} , $cv_{\text{est}} = \sqrt{2/d}$, K , and H_1 . To reduce computation and simplify presentation of results, a baseline scenario was based on the empirical results from before: $cv_{\text{group}} = 1.5$, $cv_{\text{stratum}} = 1$, $\rho_{\text{group}} = 1$, $\rho_{\text{stratum}} = 0.9$, $cv_{\text{est}} = 0.5$, $K = 10$, and $H_1 = 5$. One parameter at a time was varied relative to the baseline, in the ranges: $cv_{\text{group}} \in 0, 0.2, 0.4, \dots, 3$, $cv_{\text{stratum}} \in 0, 0.2, 0.4, \dots, 2$, $\rho_{\text{group}} = 1$, $\rho_{\text{stratum}} = 0.9$, $cv_{\text{est}} \in 0, 0.1, 0.2, \dots, 1.2$, $K \in 2, 3, \dots, 30$, and $H_1 = 5$.

Figure 1 shows the average true variance, defined by (2), for three allocation methods, divided by the value for proportional allocation. Each plot in the figure plots the three relative efficiencies against the value of one simulation parameter. Generally Neyman does better than grouped Neyman, except when cv_{stratum} is small (< 0.75), cv_{est} is high (> 0.5), or ρ_{stratum} is less than 0.7. Otherwise grouped Neyman does better, and Neyman can do spectacularly poorly.

It is striking that the new statistical learning (SL) method is superior to Neyman in all cases and to grouped Neyman in almost all cases. Thus, the method is able to choose an appropriate interpolation among the Neyman, grouped Neyman, and proportional allocations and do better than any of them in almost all cases. The improvement of SL over Neyman (which would be the usual approach in practice) is most dramatic when either: strata in the same group are homogeneous (low cv_{stratum}) and hence the signal to noise ratio in $\hat{V}_{kht}$ is poor, the CV of $\hat{V}_{kht}$ given V_{kht} (cv_{est}) is high, or the lag 1 correlation (ρ_{stratum}) is small.

Figure 2 contains ternary composition plots produced by the compositions package in R (van den Boogaart, Tolosana, and Bren 2011). The values of $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ as estimated by 200 simulations are shown, with two sets of values of K and H_1 . These coefficients are non-negative and sum to 1, with $\lambda_1 = 1$ indicating that the SL allocation is the same as Neyman, $\lambda_2 = 1$ indicating SL = grouped Neyman, and $\lambda_3 = 1$ indicating proportional. The vertices of the triangles represent these three extremes, with the closeness to each vertex reflecting the corresponding element of λ . The value of λ_3 is generally small and becomes smaller as K increases. The SL allocation is generally closer to Neyman than to grouped Neyman. There is some variability in the values of

Figure 1. Efficiency (Estimated from 1,000 Simulations) Relative to Proportional Allocation of Three Allocation Methods: Plug-in Neyman Allocation, the New Statistical Learning Method (SL), and Plug-in Grouped Neyman. Each Plot Varies One Parameter at a Time from Baseline Level. Baseline Levels of Parameters Indicated by Vertical Dashed Lines.

Figure 2. Ternary Composition Plots Showing λ_1 , λ_2 , and λ_3 in the Statistical Learning Allocation as Estimated from 200 Simulations. Best Possible Values of λ , Based on all 200 Simulations, Plotted as *. Number of Groups (K) and Strata per Group (H_1) Varied. All Other Parameters Held at Baseline Level.

λ obtained by the SL method using the time 1 and time 2 data, roughly distributed about the best possible λ . The variability is smaller in the right plot, where K is larger.

5. DISCUSSION

When two sets of design data are available, a statistical learning approach to optimal allocation can be adopted. In simulations based on real data sets, gains of up to 35 percent in the variance were achieved, compared to the usual plug-in Neyman allocation. The gains are greatest when the autocorrelations of the true strata variances are weak or the stratum degrees of freedom are small. Moreover, the new allocation method is more robust and would be closer to most survey designers' judgment, because it reduces the variability of sampling rates caused by volatile design data.

Standard pre-survey estimation of variances that will be achieved are negatively biased by 15–55 percent. This bias is virtually removed by the use of a second design data set and an ad hoc adjustment.

If only one design data set is available, the new approach could be applied by repeatedly randomly splitting the data set in two. However, this would only allow for the sampling variability of the design data and not for population change over time. The usefulness of such an approach requires further study.

One feature of the new approach is the somewhat arbitrary definition of a compromise allocation, such as (3). A referee suggested that a Bayesian approach might lead to a more natural derivation from first principles, which would presumably be more efficient. No Bayesian approach to this problem has been suggested, and this could be a topic for future research. This would require more detailed modelling work, including the formulation of priors and a unit-level model allowing for change in population quantities over time.

The methodology developed here can be applied to any loss function of interest and is therefore very versatile. The simulation study and examples were based on the Neyman loss function with a linear cost constraint. This case is worth special consideration, because stratified simple random sampling remains one of the most versatile and widely used sample designs in practice, and imprecision in estimated variances can be significant. Moreover, the great majority of other sample designs used in practice, including multistage and multiphase sampling, have variances of the same algebraic form as (1). Future research will focus on applying the new approach to more complex design problems.

Supplementary Materials

Supplementary materials are available online at http://www.oxfordjournals.org/our_journals/jssam/.

Appendix

Proof of Theorem 1

$E[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{valid}}) - L(\mathbf{n}, \mathbf{V})] = EE[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{valid}}) - L(\mathbf{n}, \mathbf{V}) | \hat{\mathbf{V}}_{\text{train}}]$. Now, $\mathbf{n}$ is a constant conditional on $\hat{\mathbf{V}}_{\text{train}}$ and A1 states that $E[\hat{\mathbf{V}}_{\text{valid}} - \mathbf{V} | \hat{\mathbf{V}}_{\text{train}}] = \mathbf{0}$. Hence, A2 implies that $E[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{valid}}) - L(\mathbf{n}, \mathbf{V}) | \hat{\mathbf{V}}_{\text{train}}] = 0$. The result follows.

Proof of Theorem 2

Note that $\boldsymbol{\lambda} = \mathbf{n}(\boldsymbol{\lambda}, \hat{\mathbf{V}}_{\text{train}})$. Assumption A3 means that

$$L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}}) = L(\mathbf{n}(\boldsymbol{\lambda}, \hat{\mathbf{V}}_{\text{train}}), \hat{\mathbf{V}}_{\text{train}}) \leq L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \hat{\mathbf{V}}_{\text{train}})$$

with equality only if $L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}}) = L(\mathbf{n}, \mathbf{V})$. Taking expectations gives

$$E[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}})] \leq E[L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \hat{\mathbf{V}}_{\text{train}})] \quad (11)$$

with strict inequality except when $L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}})$ does not depend on the value of $\hat{\mathbf{V}}_{\text{train}}$. Assumptions A1 and A2 then imply that the right side of (11) is equal to $E[L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \mathbf{V})]$, so (11) becomes

$$E[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}})] \leq E[L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \mathbf{V})]. \quad (12)$$

Assumption A3 implies that the right side of (12) satisfies

$$L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \mathbf{V}) \leq L(\mathbf{n}(\boldsymbol{\lambda}, \hat{\mathbf{V}}_{\text{train}}), \mathbf{V})$$

for any $\hat{\mathbf{V}}_{\text{train}}$ and therefore $E[L(\mathbf{n}(\boldsymbol{\lambda}, \mathbf{V}), \mathbf{V})] \leq E[L(\mathbf{n}(\boldsymbol{\lambda}, \hat{\mathbf{V}}_{\text{train}}), \mathbf{V})]$. Hence (12) leads to

$$E[L(\mathbf{n}, \hat{\mathbf{V}}_{\text{train}})] \leq E[L(\mathbf{n}(\boldsymbol{\lambda}, \hat{\mathbf{V}}_{\text{train}}), \mathbf{V})] = E[L(\mathbf{n}, \mathbf{V})], \quad (13)$$

which gives the result.

References

- Brooks, S. H. (1955), "The Estimation of an Optimum Subsampling Number," *Journal of the American Statistical Association*, 50, 398–415.
- Chernick, M. (2008), *Bootstrap Methods: A Guide for Practitioners and Researchers*, Wiley Series in Probability and Statistics, Wiley-Interscience.
- Clark, R. G. (2009), "Sampling of Subpopulations in Two-stage Surveys," *Statistics in Medicine*, 28, 3697–3717.

- . (2013), “robustallocation: Robust Allocation of Stratum Sample Sizes Allowing for Imperfect Design Data,” R package version 1.1, available at www.uow.edu.au/~rclark/rpackages/robustallocation.
- Clark, R. G., and D. G. Steel (2000), “Optimum Allocation of Sample to Strata and Stages with Simple Additional Constraints,” *Journal of the Royal Statistical Society, Series D: The Statistician*, 49, 197–207.
- Cochran, W. (1977), *Sampling Techniques*, Wiley series in probability and mathematical statistics: Applied probability and statistics (3rd ed.), Wiley.
- Davison, A., and D. Hinkley (1997), *Bootstrap Methods and Their Application*, Cambridge University Press.
- Evans, W. (1951), “On Stratification and Optimum Allocations,” *Journal of the American Statistical Association*, 46, 95–104.
- Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2012), *mvtnorm: Multivariate Normal and t Distributions*, R package version 0.9-9992. Available at <http://CRAN.R-project.org/package=mvtnorm>.
- Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.
- Johnson, N., S. Kotz, and N. Balakrishnan (1994), *Continuous Univariate Distributions*, Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley.
- Kalton, G., and D. W. Anderson (1986), “Sampling Rare Populations,” *Journal of the Royal Statistical Society Series A*, 149, 65–82.
- Lohr, S. (2009), *Sampling: Design and Analysis*, Advanced Series, Brooks/Cole.
- 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, 558–625.
- R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.
- Smith, P., M. Pont, and T. Jones (2003), “Developments in Business Survey Methodology in the Office for National Statistics, 1994–2000,” *Journal of the Royal Statistical Society, Series D (The Statistician)*, 52, 257–295.
- Sukhatme, P. (1935), “Contribution to the Theory of the Representative Method,” *Supplement to the Journal of the Royal Statistical Society*, 2, 253–268.
- Valliant, R., A. Dorfman, and R. Royall (2000), *Finite Population Sampling and Inference: A Prediction Approach*, Wiley Series in Probability and Statistics: Survey Methodology Section, John Wiley.
- van den Boogaart, K. G., R. Tolosana, and M. Bren (2011), *compositions: Compositional Data Analysis*, R package version 1.10-2.