Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

06-12

Statistical Learning In Sample Design

Robert Graham Clark

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

STATISTICAL LEARNING IN SAMPLE DESIGN

Robert Graham Clark [1]

Key Words: optimal allocation; pilot survey; robust design; stratified sampling; statistical learning; data mining

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 minimises 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 to stratified sampling, where estimated stratum variances may be highly variable. Pooling strata into groups may reduce this variability, at the possible cost of some inefficiency. Proportional allocation, ignoring differences between stratum variances, could also be used. The new approach enables a data-driven compromise between 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]Robert Clark is Associate Professor at the Centre for Statistical and Survey Methodology, University of Wollongong, NSW Australia 2522 (email rclark@uow.edu.au).

1. INTRODUCTION

In practice, sample designs are always 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 without replacement. The variance of an estimator of the population total is equal to

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

plus a term which does not depend on the stratum sample sizes n_h, where V_h is the population variance and N_h is the population size of stratum h. The total cost of the survey can often be approximated by

$$C = C_0 + \sum_{h=1}^{H} C_h n_h \tag{2}$$

where C_0 are fixed costs and C_h are cost coefficients which may be estimated based on operational information (for example interviewers' pay rates) or from cost data from previous surveys. The allocation which minimises (1) subject to the cost (2) being fixed at C_f satisfies $n_h \propto N_h\sqrt{V_h/C_h}$. In practice, however, the values of V_h are unknown, and estimates are substituted. This will be called a *plug-in* allocation, since the optimal allocation is derived assuming knowledge of V_h, but estimates are then substituted.

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 et al. (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), summarising Brooks (1955), found that a large pilot sample of around 150 units may be needed to achieve precisions within 10% of the ideal optimal design.

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 $\boldsymbol{n}$ depends on the estimated variances $\hat{\boldsymbol{V}}$ and is therefore itself variable. The values of $\boldsymbol{n}$ would normally be conditioned upon once the survey is conducted, and the achieved variance is then defined by equation (1). This quantity is normally estimated by substituting both the plug-in $\boldsymbol{n}$ and the estimated variances $\hat{\boldsymbol{V}}$ into (1). It is clear that this has the potential for bias, as 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: under what conditions is the variability of $\hat{\boldsymbol{V}}$ likely to have an appreciable effect on either the achieved variance V or the pre-survey estimation of V? and can the plug-in allocation be improved upon? These are the topic of this paper. The premise of the proposed approach is that two design datasets are available, which can act as a check upon each other. Section 2 will define a general formulation of the allocation problem and the approach of using a training and validation sample. The approach is loosely based on the statistical learning approach to model choice

(e.g. chapter 7 of Hastie et al., 2009). Theoretical results are difficult to derive, but two simple theorems will be stated. Section 3 is a simulation study. Stratum population variances for times 1, 2 and 3 are generated by multiplying auto-correlated lognormal variables for group and for stratum. The parameters of the simulation model are obtained by analysing three real datasets. Section 4 then extends the simulation study by varying these parameters, to identify when the new method provides useful gains. Section 5 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 will form the basis of the simulation study later in this paper. The aim is to design a survey to be conducted at time 3. Design data is available from a similar survey conducted at time 2. Data is also available from the survey at time 1. A stratified sample design is used, where strata naturally form into groups. The aim is to minimise

$$V_{tot(3)} = \sum_{h=1}^{H} V_{h3} n_h^{-1} \tag{3}$$

subject to fixed total sample size, where V_{ht} and $\hat{V}_{ht}$ refer to the population and estimated variances for stratum h at time t, for $t = 1, 2, 3$, and n_h depends only on $\{\hat{V}_{h2} : h = 1, \ldots, H\}$ and $\{\hat{V}_{h1} : h = 1, \ldots, H\}$. Three possible allocations are:

(i) Plug-in optimal allocation with $n_h \propto N_h \sqrt{\hat{V}_{h2}}$. This is the most common ap-

proach 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 stratum variance estimates in group k. This allocation might be used if the stratum 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.

We will suppose that n_h is to be a compromise between these three alternatives:

$$n_h \propto N_h\sqrt{\lambda_1\hat{V}_{h2} + \lambda_2\hat{V}_{k2} + \lambda_3\hat{V}_2} \tag{4}$$

where $\hat{V}_2$ is a population-weighted mean of $\hat{V}_{h2}$ over all h, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. The values of λ would then smoothly interpolate between 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 value of $\boldsymbol{\lambda}$ should be chosen so as to give a low value of $V_{tot(3)}$ in (3). A naive approach would be to minimise

$$\sum_h N_h^2 n_h\left(\hat{\boldsymbol{V}}_{\boldsymbol{2}}, \boldsymbol{\lambda}\right)^{-1}\hat{V}_{h2} \tag{5}$$

with respect to $\boldsymbol{\lambda}$, writing $n_h = n_h\left(\hat{\boldsymbol{V}}_{\boldsymbol{2}}, \boldsymbol{\lambda}\right)$ to emphasise that n_h is determined by $\hat{\boldsymbol{V}}_{\boldsymbol{2}}$ and $\boldsymbol{\lambda}$. Unfortunately, this approach will always result in $\boldsymbol{\lambda} = (1, 0, 0)$, i.e. the plug-in approach - this is a special case of the Neyman optimal design described in the first paragraph of the paper. The problem is that $n_h\left(\hat{\boldsymbol{V}}_{\boldsymbol{2}}, \boldsymbol{\lambda}\right)$ would be expected to be positively correlated with $\hat{V}_{h2}$, unless $\boldsymbol{\lambda} = (0, 0, 1)$, so that the loss function in no way penalizes the variability caused by plugging in estimates when calculating

n_h. The loss function

$$\sum_h N_h^2 n_h \left(\hat{\boldsymbol{V}}_{\mathbf{1}}, \boldsymbol{\lambda}\right)^{-1} \hat{V}_{h2} \tag{6}$$

is proposed, to remove (or at least reduce) this difficulty by using separate training estimates $\hat{\boldsymbol{V}}_{\mathbf{1}}$ and validation estimates $\hat{\boldsymbol{V}}_{\mathbf{2}}$. The value of $\boldsymbol{\lambda}$ is chosen to minimize (6). The allocation $n_h = n_h\left(\hat{\boldsymbol{V}}_{\mathbf{2}}, \boldsymbol{\lambda}\right)$ would then be implemented for the time 3 survey with this $\boldsymbol{\lambda}$. An estimator similar to (6) could also be used to estimate the variance that will be achieved in time 3. The values of n_h in (6) should be approximately uncorrelated with $\boldsymbol{V_{h2}}$, thereby avoiding the bias of (5). (This will be proven in subsection 2.3 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 Neyman 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, p330). The design parameters are the first phase sample size and the second phase stratum 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, p277) 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 & Steel, 2000), as can two stage sampling with unequal probabilities of selection for each cluster and different within-cluster sample sizes (Clark, 2009).

Equation (4) defined one choice of the smoothing function $\boldsymbol{n}\left(\boldsymbol{\lambda}, \boldsymbol{v}\right)$. There are a

variety of other ways that this function could be defined. For example, (4) 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 (4) could be extended to include $\hat{\boldsymbol{V}}_{\boldsymbol{l2}}$ as well as $\hat{\boldsymbol{V}}_{\boldsymbol{k2}}$. Another example might be a stratified repeated business survey, where 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 $\boldsymbol{\lambda}$ consisting of these weights.

To allow for all of these possibilities, a more general formulation of the allocation problem is needed. Subsection 2.2 will now define this, and propose 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 set of design parameters $\boldsymbol{n} = (n_1, \ldots, n_H)$ which specify the sampling method. The aim will be to choose $\boldsymbol{n}$ to achieve a low value of $L(\boldsymbol{n}; \boldsymbol{V})$ subject to a cost constraint $C_f = C(\boldsymbol{n})$ for a known function $C(.)$. Typically $\boldsymbol{n}$ would be subject to some range constraints also, for example $n_h > 0$. The function $L(.,.)$ is a loss function which would typically be the variance of a statistic of interest, or a linear combination of several such variances. The vector $\boldsymbol{V}$ would be a set of population parameters. For example, $\boldsymbol{n}$ might be the sample sizes for each of H strata in a stratified sample design, the cost constraint might be a fixed total sample size $n = \sum_{h=1}^{N} n_h$ and $\boldsymbol{V}$ might be the population variances in each of H strata, or $\boldsymbol{n}$ might contain probabilities of selection in an unequal probability design, or sample sizes by stage or phase in

multi-stage or multiphase designs. The ideal design is obviously:

$$\boldsymbol{n}_{ideal} = \underset{\boldsymbol{n}:C(\boldsymbol{n})=C_f}{\arg\min}\ L\left(\boldsymbol{n};\boldsymbol{V}\right) \tag{7}$$

but in practice $\boldsymbol{V}$ would be unknown. Instead, the current practice is to use the plug-in allocation:

$$\boldsymbol{n}_{plugin} = \underset{\boldsymbol{n}:C(\boldsymbol{n})=C_f}{\arg\min}\ L\left(\boldsymbol{n};\hat{\boldsymbol{V}}\right). \tag{8}$$

The difficulty with the plug-in methodology is that it leads to more volatile designs, where the variability in $\hat{\boldsymbol{V}}$ results in a loss greater than $L\left(\boldsymbol{n};\boldsymbol{V}\right)$. To manage the variability in $\hat{\boldsymbol{V}}$, the following approach is proposed. The design variables $\boldsymbol{n}$ based on an estimate $\boldsymbol{v}$ of $\boldsymbol{V}$ will be defined to equal $\boldsymbol{n}\left(\boldsymbol{\lambda},\boldsymbol{v}\right)$, where $\boldsymbol{n}(.,.)$ is a function chosen by the sample designer such that $\boldsymbol{\lambda}$ is a p-vector controlling the "complexity" of the design. The details of how this complexity is defined would 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 $\boldsymbol{\lambda}$, say $\boldsymbol{\lambda_0}$, would result in $\boldsymbol{n}\left(\boldsymbol{\lambda_0},\boldsymbol{v}\right)$ being the plug-in allocation, while on the other extreme, $\boldsymbol{\lambda_1}$ would be such that $\boldsymbol{n}\left(\boldsymbol{\lambda_1},\boldsymbol{v}\right)$ does not depend on $\boldsymbol{v}$ at all. For example, in the special case described in subsection 2.1, $\boldsymbol{\lambda_0} = (1,0,0)$ and $\boldsymbol{\lambda_1} = (0,0,1)$.

The aim is to choose a suitable value of $\boldsymbol{\lambda}$, which results in something close to the plug-in design when $\hat{\boldsymbol{V}}$ is highly precise, and in a simpler design when $\hat{\boldsymbol{V}}$ is a poor estimate. Ideally, $\boldsymbol{\lambda}$ should minimise $L\left(\boldsymbol{n};\boldsymbol{V}\right)$, or to be precise

$$L\left(\boldsymbol{n}\left(\boldsymbol{\lambda},\hat{\boldsymbol{V}}\right);\boldsymbol{V}\right). \tag{9}$$

The trouble is, we can't reliably estimate this loss using a single estimate of $\boldsymbol{V}$, since

$$L\left(\boldsymbol{n}\left(\boldsymbol{\lambda},\hat{\boldsymbol{V}}\right);\hat{\boldsymbol{V}}\right) \tag{10}$$

is by definition minimised by $\boldsymbol{\lambda} = \mathbf{0}$, giving the plug-in allocation. The problem is that (10) is negatively biased for (9), because the same estimates $\hat{\boldsymbol{V}}$ used to calculate $\boldsymbol{n}\left(\boldsymbol{\lambda}, \hat{\boldsymbol{V}}\right)$ are then used to evaluate this allocation.

To improve on the plug-in design, it will be assumed that independent training and validation estimates of $\boldsymbol{V}$, $\hat{\boldsymbol{V}}_{\boldsymbol{train}}$ and $\hat{\boldsymbol{V}}_{\boldsymbol{valid}}$ are available. For example, these could consist of estimates of $\boldsymbol{V}$ calculated using two previous instances of a repeated survey. Then we calculate $\boldsymbol{\lambda}$ to minimise

$$L\left(\boldsymbol{n}\left(\boldsymbol{\lambda}, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right); \hat{\boldsymbol{V}}_{\boldsymbol{valid}}\right) \tag{11}$$

and implement an allocation based on this $\boldsymbol{\lambda}$. This approach will be shown by simulation to give sensible allocations with lower values of (9).

The preceding development is similar to the discussion of model choice in chapter 7 of Hastie et al's 2009 book on statistical learning. Models should be chosen with an appropriate size, and this is achieved by fitting the model using a training dataset, but evaluating its predictive performance using an independent validation dataset. The loss estimates based on the validation dataset are used to choose tuning parameters which control the model's size or complexity. While the concepts and motivation are similar, the approach discussed here differs from the statistical learning literature in several respects. In particular, in modelling: the loss function is usually a measure of the model's predictive performance, the complexity parameters are based on measures of model size or complexity, and training and validation datasets are often constructed as repeated partitions of a single dataset. In contrast, the training and validation datasets here will typically consist of estimates of $\boldsymbol{V}$ obtained from surveys run at different times, because changes in the population over time may be a major source of error in $\hat{\boldsymbol{V}}$.

2.3 Two Theorems

This subsection states two theorems with proofs in the appendix. The first result is that (6) is unbiased subject to regularity conditions. The second is that (5) is negatively biased. Strong assumptions are made in each theorem, to enable clear and interpretable results. Real surveys are likely to be messier, and this is reflected in the simulations in sections 4 and 5, which do not satisfy the assumptions of these theorems, but which still show useful gains from the new approach.

Theorem 1. If

(A1) $E\left[\hat{\boldsymbol{V}}_{\boldsymbol{train}} - \boldsymbol{V}\right] = E\left[\hat{\boldsymbol{V}}_{\boldsymbol{valid}} - \boldsymbol{V}\right] = \mathbf{0}$;

(A2) $E\left[\hat{\boldsymbol{V}}_{\boldsymbol{valid}} - \boldsymbol{V} \,\middle|\, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right] = \mathbf{0}$;

(A3) $E\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}\right) - L\left(\boldsymbol{n}, \boldsymbol{V}\right)\right] = 0$ for any constant $\boldsymbol{n}$ and any $\hat{\boldsymbol{V}}$ satisfying $E\left[\boldsymbol{v} - \boldsymbol{V}\right] = \mathbf{0}$; and

(A4) $\boldsymbol{n} = \boldsymbol{n}\left(\boldsymbol{\lambda}, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right)$

then $E\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{valid}}\right) - L\left(\boldsymbol{n}, \boldsymbol{V}\right)\right] = 0$ for any $\boldsymbol{\lambda}$.

Condition (A1) of Theorem 1 states that the training and validation estimates of $\boldsymbol{V}$ are both unbiased, which seems reasonable, and would be satisfied if these datasets were obtained by stratified simple random sampling and the simple stratum variances were used. Condition (A3) states that the expected value of loss function using unbiased estimates of $\boldsymbol{V}$ equals the loss function with $\boldsymbol{V}$, when $\boldsymbol{n}$ are non-random. This is satisfied for theNeyman function (1). Condition (A2) is that the training and validation datasets are independent in a particular way, and condition (A4) is that the allocation be based on the training sample only.

The Theorem means that (6) is unbiased for any value of $\boldsymbol{\lambda}$ and hence can be used in order to optimize $\boldsymbol{\lambda}$. In contrast, Theorem 2 states that the plug-in estimator (5) is negatively biased for all $\boldsymbol{\lambda}$ except the value or values of $\boldsymbol{\lambda}$ corresponding to not using any of the design data in the allocation.

Theorem 2. Conditions (A1) - (A4) from Theorem 1 are assumed as well as

(A5) $\boldsymbol{n}(\boldsymbol{\lambda}, \boldsymbol{v}) = \arg\min_{\boldsymbol{n}} L_{\boldsymbol{\lambda}}(\boldsymbol{n}, \boldsymbol{v})$ for some function $L_{\boldsymbol{\lambda}}(.,.)$.

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

Assumption (A5) is that $\boldsymbol{n}$ can be expressed as the minimiser of a loss function depending on $\boldsymbol{\lambda}$. This is the case with the loss function used in 2.1, 3 and 4.

Theorem 2 means that the usual plug-in approach is biased in favour of values of $\boldsymbol{\lambda}$ such that $\boldsymbol{n}$ is more dependent on the design data. This results in a sub-optimal choice of $\boldsymbol{\lambda}$ and in the variance that will be achieved being under-estimated 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 Stratum Variances

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

group stratum factor B_{ht} as follows:

$$\left.\begin{array}{l} log\left(V_{ht}\right) = \mu + A_{kt} + B_{ht} \text{ where stratum h belongs to group k} \\ \boldsymbol{A} \sim N\left(\boldsymbol{0}, \sigma^2_{group}\boldsymbol{R_{group}}\right) \\ \boldsymbol{B} \sim N\left(\boldsymbol{0}, \sigma^2_{stratum}\boldsymbol{R_{stratum}}\right) \\ \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ independent} \end{array}\right\} \quad (12)$$

where $\boldsymbol{A}$ is the vector of all A_{kt} over k and t, and $\boldsymbol{B}$ is the vector of all B_{ht} over h and t. The correlation matrix $\boldsymbol{R_{group}}$ 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 $corr\left[A_{kt_1}, A_{kt_2}\right] = \rho^{|t_2-t_1|}_{group}$. Similarly, the correlation matrix $\boldsymbol{R_{stratum}}$ 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 $corr\left[A_{ht_1}, A_{ht_2}\right] = \rho^{|t_2-t_1|}_{stratum}$.

Given the stratum population variances, V_{ht}, the stratum sample variances will be generated as independent scaled chi-square distributions, with

$$\hat{V}_{ht} \sim V_{ht}\chi^2_d/d \quad (13)$$

where d are the degrees of freedom.

There are 6 parameters which can be varied in the simulation: μ, σ^2_{group}, $\sigma^2_{stratum}$, ρ_{group}, $\rho_{stratum}$ and d, as well as the number of groups, K, and the number of strata per group, H_1 The first of these, μ, will just be assumed to be zero, since it affects variances and estimated variances as a simple multiplicative factor, and so does not affect the relative performance of the different allocation methods.

Stratum 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 stratum population sizes were also simulated with similar conclusions, but are not shown in this paper.)

3.2 Empirical Fitting of Model for Stratum Population Variances for Three Datasets

The 5 parameters which can be varied in the simulation give rise to an astronomical number of plausible scenarios which could be simulated. This subsection describes the fitting of model (12) and (13) to three datasets. Subsection 3.3 will then describe simulations based on these estimated model parameters.

The first dataset was extracted from the Australian Agriculture and Grazing Industries Survey 1991-1995, 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 (5 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 were equal for each size. (This is an example of an equal aggregate approach to size stratification - see for example Valliant et al., 2000, section 6.5.2). 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, since these are thought to be more important explanators than state, although state also matters due to differences in climate, accessibility and remoteness etc. Strata with less than a sample size of 6 were excluded, leaving a total of 3189 farms in 39 strata in 9 groups over five years.

The second dataset 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, 95-96, 96-97 and 97-98. Only industries Metal Product Manufacturing, Machinery and Equipment Manufacturing, Machinery and Motor Vehicle Wholesaling and Personal and Household Good 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 organisation (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 7328 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 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% 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 & Anderson, 1986). Strata are grouped into Area Units (larger areas containing on average about 2000 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 (12) to this dataset 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 (12) by maximum likelihood, assuming that $\hat{V}_{ht}$ is equal to V_{ht}. This was done because of the difficulty of specifying a model for $\hat{V}_{ht}$ which would fit these datasets, due to the variation in stratum sample sizes and the skewed, heavy-tailed distributions of the farm and business datasets. To correct for this, an empirical bootstrap bias-correction was applied with 30 resamples (e.g. Chernick, 2008, pp.26-27). The correlation parameters $\rho_{stratum}$ and ρ_{group}, were first transformed using the hyperbolic arctangent transformation, to ensure that the corrected estimates lay between -1 and 1. Bias correction was not required for the Census dataset. Table 1 shows the parameter estimates with bootstrap standard errors in brackets.

Table 1: Estimates of Parameters of Model (12) for Three Datasets

Dataset	σ^2_{group}	$\sigma^2_{stratum}$	ρ_{group}	$\rho_{stratum}$	$CV(\%)$ $(e^{A_{kt}})$	$CV(\%)$ $(e^{B_{kht}})$
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 Pop'n	0.88	1.57	0.98	0.44	119	195

It is notable from the table that population variances vary greatly across strata, as shown by the values of σ^2_{group} and $\sigma^2_{stratum}$. The coefficient of variation (CV) of the group and stratum factors ($e^{A_{kt}}$ and $e^{B_{kht}}$) are equal to $cv_{group} = \sqrt{e^{\sigma^2_{group}} - 1}$ and $cv_{stratum} = \sqrt{e^{\sigma^2_{stratum}} - 1}$ (e.g. Johnson et al., 1994, p.212). These values are well above 100%. 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 (13) was estimated by a small parametric simulation based on the farms and business datasets discussed in Section 3.2. In both cases, the unit values of the variable of interest (which will be denoted Y_{khti} for unit i in stratum h, group k and time t) can reasonably be modelled as a mixture of zero values and lognormally 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 (13) 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 $cv_{est} = \sqrt{2/d}$ and the estimated CVs are also shown in Table 2.

Table 2: Estimates of Parameter d in (13) 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

3.3 Simulation Results

For each simulation, the statistical learning allocation, defined by (4), was calculated, with $\boldsymbol{\lambda}$ chosen to minimise (6). The plug-in Neyman, plug-in grouped Neyman and proportional allocations were also calculated. All simulations and em-

pirical 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 in subsection 3.3 and section 4 can be fully reproduced using the programs which are contained in the supplementary material, but the datasets used in subsection 3.2 are not available because of confidentiality restrictions.

1000 sets of stratum sample variances were simulated using models (12) and (13). The parameters were obtained from the analyses described in 3.2. For the simulations based on the farms and business survey, K and H_1 were set to 20 and 5 respectively. For the NZ Pacific population example, $K = 5$ and $H_1 = 20$ were used. Table 3 shows the achieved variances of the different allocation methods, (3), 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 (13% and 35% reduction in the variance compared to the plug-in Neyman allocation), in the farms example when the design datasets have 6 units per stratum (10% reduction in variance), and in the NZ Pacific example (12% reduction in variance).

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 (5), which Theorem 2 states is biased. An alternative would be to use (6), which is unbiased provided the assumptions of Theorem 1 are satisfied. However, one of the assumptions of Theorem 1 is that $\boldsymbol{V_t}$ are stationary over time, whereas in practice there may be systematic movement over time, for example due to inflation in financial variables.

Table 3: Variance achieved by Different Allocation Methods relative to Proportional Allocation for Simulations with Parameters based on Three Datasets and Several Values of n_h

Dataset and Assumed n_h in Design Datasets	Allocation Method			
	Plug-in Neyman	SL	Plug-in Grouped Neyman	SL with Ideal $\boldsymbol{\lambda}$
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 Popn	0.934	0.822	0.850	0.820

The following estimator allows for this to some extent by use of a ratio adjustment.

$$\left\{\sum_h N_h^2 n_h^{-1} \hat{V}_{h1}\right\} \left\{\sum_h N_h \hat{V}_{h2}\right\} / \left\{\sum_h N_h \hat{V}_{h1}\right\} \tag{14}$$

Table 4 shows the ratio of the mean of the estimated variance (over 1000 simulations) to the true variance given by (3), for the naive variance estimator and the improved version (14). The parameter settings are the same as in Table 3. The naive variance estimator has substantial negative bias, of up to 58%, 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 in subsection 3.3, 1000 design datasets for three time periods were generated using models (12) and (13). The previous section described the results when the parameters of the simulation models were based on three survey datasets. This section investigates the range of possible outcomes by varying each of the parameters

Table 4: Expected Value of Variance Estimate divided by True Value for Naive and Proposed Variance Estimators, defined by (5) and (14), for Simulations with Parameters based on Real Datasets and Several Values of n_h, for Three Different Allocation Methods

Dataset 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 Popn	0.655	0.836	0.983	0.998	1.002	1.005

$cv_{group} = \sqrt{e^{\sigma^2_{group}} - 1}$, $cv_{stratum} = \sqrt{e^{\sigma^2_{stratum}} - 1}$, ρ_{group}, $\rho_{stratum}$, $cv_{est} = \sqrt{2/d}$, K and H_1. To reduce computation and simplify presentation of results, a baseline scenario was based on the empirical results from subsection 3.2: $cv_{group} = 1.5$, $cv_{stratum} = 1$, $\rho_{group} = 1$, $\rho_{stratum} = 0.9$, $cv_{est} = 0.5$, $K = 10$ and $H_1 = 5$. One parameter at a time was varied relative to the baseline, in the ranges: $cv_{group} \in \{0, 0.2, 0.4, \ldots, 3\}$, $cv_{stratum} \in \{0, 0.2, 0.4, \ldots, 2\}$, $\rho_{group} = 1$, $\rho_{stratum} = 0.9$, $cv_{est} \in \{0, 0.1, 0.2, \ldots, 1.2\}$, $K \in \{2, 3, \ldots, 30\}$ and $H_1 = 5$.

Figure 1 shows the average achieved variance, defined by (3), 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 $\rho_{stratum}$ is less than 0.7. Otherwise Grouped Neyman does better, and Neyman can do spectacularly poorly.

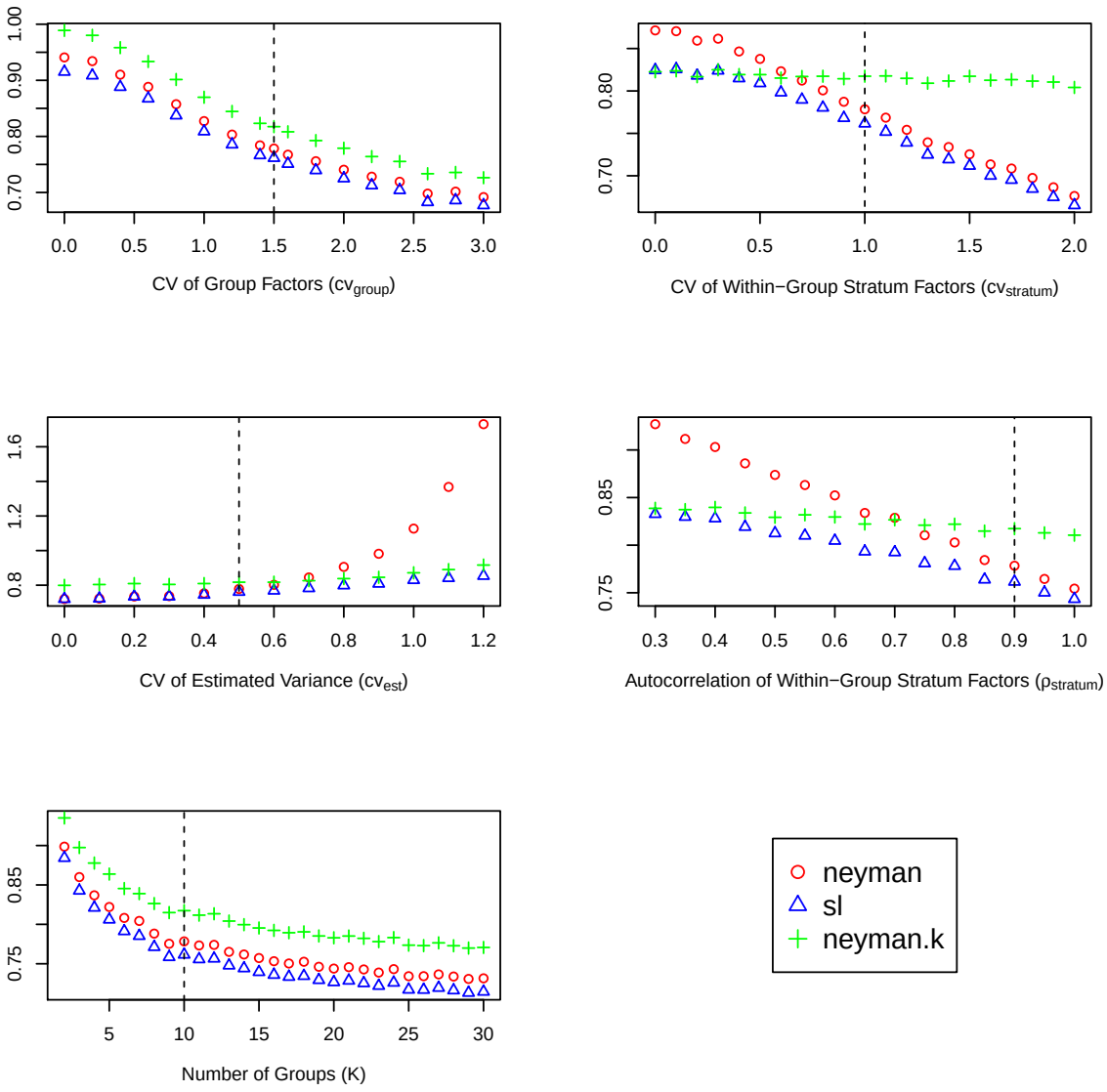

Figure 1: Efficiency (estimated from 1000 simulations) relative to proportional allocation of three allocation methods: plug-in Neyman allocation, the new statistical learning method (SL), and plug-in grouped Neyman. Each plot varies one parameter at a time from baseline level. Baseline levels of parameters indicated by vertical dashed lines.

It is striking that the new statistical learning method (SL) 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 between the Neyman, grouped Neyman and proportional allocation, and to 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 homogenous (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 ($\rho_{stratum}$) is small.

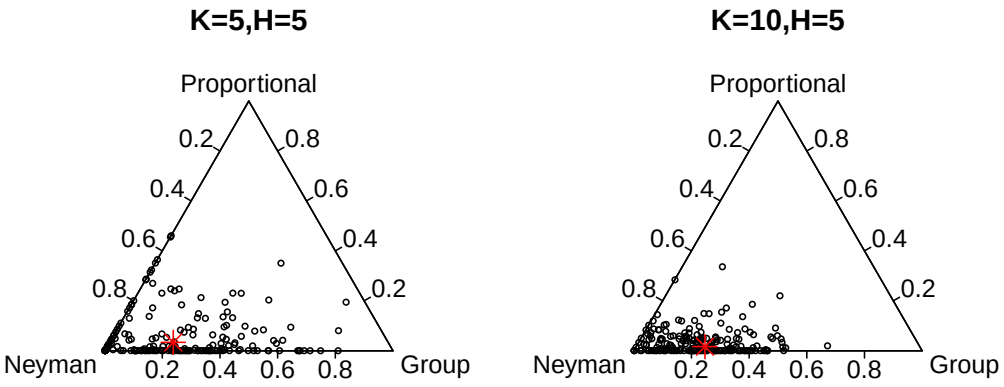

Figure 2: Ternary composition plots showing λ_1, λ_2 and λ_3 in the statistical learning allocation as estimated from 200 simulations. Best possible values of $\boldsymbol{\lambda}$, based on all 200 simulations, plotted as *. Number of groups (K) and strata per group (H_1) varied. All other parameters held at baseline level

Figure 2 contains ternary composition plots produced by the compositions package in R (Boogaart et al., 2011). The values of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ as estimated by 200 simulations are shown, with 2 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 $\boldsymbol{\lambda}$. 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 $\boldsymbol{\lambda}$ obtained by the SL method using the time 1 and time 2 data, roughly

distributed about the best possible $\boldsymbol{\lambda}$. The variability is smaller in the right hand 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 datasets, gains of up to 35% in the variance were achieved. The gains are greatest when the autocorrelations of the true stratum variances are weak or the stratum degrees of freedom are small. Moreover, the new allocation method is more subjectively robust, and would be closer to most survey designers' judgement, as 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%. This bias was virtually removed by the use of a second design dataset.

If only one design dataset is available, the new approach could be applied by repeatedly randomly splitting the dataset 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.

The statistical learning approach 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 multi-stage and multi-phase sampling, have variances

of the same algebraic form as (1). Future research will focus on applying the new approach to more complex design problems.

APPENDIX: PROOF OF THEOREMS

Proof of Theorem 1:

$$E\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{valid}}\right) - L\left(\boldsymbol{n}, \boldsymbol{V}\right)\right] = EE\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{valid}}\right) - L\left(\boldsymbol{n}, \boldsymbol{V}\right) \middle| \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right]$$

(A4) implies that $\boldsymbol{n}$ is a constant conditional on $\hat{\boldsymbol{V}}_{\boldsymbol{train}}$ and (A2) states that $E\left[\hat{\boldsymbol{V}}_{\boldsymbol{valid}} - \boldsymbol{V} \middle| \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right] = \mathbf{0}$. Hence (A3) implies that $E\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{valid}}\right) - L\left(\boldsymbol{n}, \boldsymbol{V}\right) \middle| \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right] = 0$. The result follows.

Proof of Theorem 2: Assumption (A5) means that

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

with equality obtaining only if $L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right) = L\left(\boldsymbol{n}, \boldsymbol{V}\right)$. Taking expectations, we get

$$E\left[L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right)\right] \leq E\left[L\left(\boldsymbol{n}\left(\boldsymbol{\lambda}, \boldsymbol{V}\right), \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right)\right] \tag{15}$$

with strict inequality except when $L\left(\boldsymbol{n}, \hat{\boldsymbol{V}}_{\boldsymbol{train}}\right)$ does not depend on the value of $\hat{\boldsymbol{V}}_{\boldsymbol{train}}$. Assumptions (A1) and (A2) then imply that the right hand side of (15) is zero, giving the result.

REFERENCES

Boogaart, K. G. van den, Tolosana, R., & Bren, M. (2011). compositions: Compositional data analysis [Computer software manual]. Available from `http://CRAN.R-project.org/package=compositions` (R package version 1.10-2)

Brooks, S. H. (1955). The estimation of an optimum subsampling number. *Journal of the American Statistical Association*, *50*(270), pp. 398-415.

Chernick, M. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley-Interscience.

Clark, R. G. (2009). Sampling of subpopulations in two-stage surveys. *Statistics in Medicine*, *28*(29), 3697–3717.

Clark, R. G., & Steel, D. G. (2000). Optimum allocation of sample to strata and stages with simple additional constraints. *Journal of the Royal Statistical Society, Series D: The Statistician*, *49*(2), 197–207.

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). Wiley.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., et al. (2012). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Available from `http://CRAN.R-project.org/package=mvtnorm` (R package version 0.9-9992)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer. Available from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Wiley & Sons.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society Series A*, *149*(1), 65–82.

Lohr, S. (2009). *Sampling: Design and Analysis*. 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 [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org/` (ISBN 3-900051-07-0)

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), pp. 257-295.

Valliant, R., Dorfman, A., & Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley.