Sampling with minimal strata size requirements

Stas Kolenikov* Igor Griva†

Abstract

We consider the problem of optimal stratified single stage sampling design where minimal sample size requirements are specified for all strata. We show that the problem reduces to unidimensional optimization, and present an algorithm that solves it. We discuss the substantive interpretation of the algorithm and Lagrange multipliers in terms of the sampling problem at hand. An illustrative numerical example is provided.

Keywords: stratified sampling design, survey cost, nonlinear optimization, statistical computing.

1. Introduction

In nearly all practical situations where probability sampling is used, it is used for the reasons of limited available resources for data collection. While a census data collection from a finite human or establishment population will provide an exact answer conceptually, the feasibility of a census data collection is usually ruled out for all but the specially mandated situations (such as the regular censuses required by law) or relatively small populations with readily available contact information (such as students of a university who are required to have an email in the university domain). In most other situations, a sample is taken so as not to expend the resources for the full population, and the sample size is dictated either by the statistical power calculations when the researcher or the agency collecting the data are at liberty of asking for sufficient resources, or, more often, by the available budget.

One of the founding papers of design-based inference, Neyman (1934), explicitly incorporates cost of data collection in what is now known as Neyman or Neyman-Tchuprow optimal allocation, and derives the optimal sampling design scheme that acknowledges the budget constraint. We reconsider the problem with a requirement that is often imposed by the survey stakeholders to provide minimum sample sizes in each stratum. For instance, in the U.S., Behavioral Risk Factor Surveillance Survey (BRFSS) requires an effective minimum sample size of 2,500 observations in each state (Centers for Disease Control and Prevention 2013), and some large states have sub-state data collection programs with specific sample size requirements per county—for instance, New York Expanded BRFSS requires at least 400 interviews in each county (New York State Department of Health 2014). The problem has been considered recently by Choudhry, Rao & Hidiroglou (2012). We extend on their treatment by providing the Lagrangian function and solution to the problem via an explicit algorithm that can highlight the derivation and the properties of the solution. Another interesting contribution to the problem is Wright (2012) who proposed a connection

*Abt SRBI, 8405 Colesville Road, Suite 300, Silver Spring, MD 20910. Corresponding author: kolenikovs@srbi.com.

†Department of Mathematical Sciences, 4400 University Drive, MS:3F2, George Mason University, Fairfax VA 22030.

to elections and presented another explicit algorithm that builds the sample sizes selecting, one by one, the strata to which the next available unit should be assigned.

2. Basic problem

Consider a finite population $\mathcal{U}$ divided into H strata of sizes N_h, $h = 1, \ldots, H$, $N_1 + \ldots + N_H = N$, with a variable of interest $y_{hi}; h = 1, \ldots, H; i = 1, \ldots, N_h$. Let the population variance for stratum h be S_h^2, and the cost of data collection for one completed interview be c_h. If a simple random sample with replacement (SRSWR) of size n_h is taken in each strata, and the stratified mean is given by

$$\bar{y}_{\text{str}} = \sum_{h=1}^{H} W_h \bar{y}_h, \quad W_h = N_h/N, \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \tag{1}$$

then the Neyman-Tchuprow allocation (Neyman 1934, Tchuprow 1923) is obtained as the solution to the nonlinear optimization problem

$$\begin{aligned} \mathbb{V}[\bar{y}_{\text{str}}] = &\sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h} \to \min_{\{n_h\}} \\ \text{s.t. } &\sum_h c_h n_h = C, \end{aligned} \tag{2}$$

and is given by

$$n_h \propto \frac{S_h W_h}{\sqrt{c_h}}; \quad n_h = \frac{S_h W_h/\sqrt{c_h}}{\sum_{l=1}^{H} S_l W_l/\sqrt{c_l}} \frac{C}{\sum_{l=1}^{H} S_l W_l \sqrt{c_l}} \tag{3}$$

rounding as necessary. For standalone derivations, see Thompson (1992), Section 11.7, or Hansen, Hurwitz & Madow (1953), Section 5.11. This result will also follow from our consideration of the following more general problem.

The following sampling problem is often encountered in practice: develop a sampling design with the total sample size n and minimal strata sizes m_h, where $m = \sum_h m_h < n$, so that additional $n - m$ units need to be freely distributed across the strata. This can be seen as a special case of the sampling problem with the varying strata costs, where $c_h = c$ in each stratum, and the overall budget constraint is replaced by the total sample size constraint. From now on, we will consider this more general problem in our derivations.

For a problem with required minimum sample sizes per stratum, let us parameterize the stratum sample size as

$$n_h = m_h + t_h \tag{4}$$

where $t_h \geq 0$. Then the sample design problem is

$$\mathbb{V}[\bar{y}_{\text{str}}] = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{m_h + t_h} \to \min_{\{t_h\}} \tag{5}$$

$$\text{s.t. } \sum_h c_h (m_h + t_h) = C, \tag{6}$$

$$t_h \geq 0 \text{ for all } h \tag{7}$$

Equations (5)–(7) describe the optimization problem in variables $\{t_h, h = 1, \ldots, H\}$. Additionally, to ensure that a non-trivial solution exists, we need to require that

$$C > \sum_h c_h m_h \tag{8}$$

Substantively, it means that there is enough budget to collect at least the minimum required samples in each stratum, and more than the minimum in some strata. For practical purposes, the sample sizes have to be integer numbers, so once real-valued t_h are obtained, they need to be rounded or truncated down to the nearest integer.

Nonlinear constrained optimization (Griva, Nash & Sofer 2008) proceeds by setting up the Lagrangian function which combines the objective function and the constraints:

$$L(\{t_h\}, \lambda, \{\nu_h\}) = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{m_h + t_h} + \lambda \Big[\sum_h c_h (m_h + t_h) - C \Big] - \sum_{h=1}^{H} \nu_h t_h \tag{9}$$

The first order (Karush-Kuhn-Tucker, KKT) conditions are

$$\frac{\partial L}{\partial t_h} = -\frac{W_h^2 S_h^2}{(m_h + t_h)^2} + \lambda c_h - \nu_h = 0 \text{ for all } h, \tag{10}$$

$$\frac{\partial L}{\partial \lambda} = \sum_h c_h (m_h + t_h) - C = 0 \tag{11}$$

$$\nu_h t_h = 0,\ \nu_h \geq 0,\ t_h \geq 0 \text{ for all } h \tag{12}$$

where $\nu_h \geq 0$ are the Lagrange multipliers corresponding to inequalities $t_h \geq 0$ for all h.

From (10), the budget constraint Lagrangian multiplier is

$$\lambda = \frac{W_h^2 S_h^2}{c_h (m_h + t_h)^2} + \frac{\nu_h}{c_h} \tag{13}$$

where the expression in the right hand side is invariant with respect to h.

This invariance is an important property, as it allows to recast the whole problem essentially as a univariate problem with respect to λ, with all other quantities derived from it. In particular, the additional sample sizes can be determined as

$$t_h(\lambda) = \max \Big[\frac{W_h S_h}{\sqrt{\lambda c_h}} - m_h, 0 \Big], \quad \lambda > 0 \tag{14}$$

The cost of the survey with these additional sample sizes is

$$C(\lambda) = \sum_h c_h \big[m_h + t_h(\lambda) \big] \tag{15}$$

which may be greater or less than the available budget C. Finally, the Lagrange multipliers ν_h for non-negativity constraints on t_h are

$$\nu_h(\lambda) = c_h \lambda - \frac{W_h^2 S_h^2}{\big[m_h + t_h(\lambda) \big]^2} \tag{16}$$

Due to (14), we have $\nu_h \geq 0$. These Lagrange multipliers are equal to zero in those strata h where the constraint $t_h \geq 0$ is not binding, i.e., $t_h > 0$. Put differently, $t_h > 0$ for some stratum h means that additional sample elements are drawn from this stratum. On the other hand, if $\nu_h > 0$, then it must be that $t_h = 0$, i.e., the stratum size in stratum h has to be limited to the required minimum m_h only. In other words, the sign of ν_h may serve as an indicator of whether additional units are to be taken from stratum h on top of m_h. A common interpretation of the Lagrange multiplier in nonlinear optimization problems is a "shadow price" of the constraint, i.e., the impact that the constraint has. A zero value indicates that the constraint is not active, and thus "costs" nothing to accommodate, in terms of adjusting other parameters of the optimization problem. Positive values of ν_h indicate that the relevant constraints are active, i.e., $t_h = 0$, and greater values additionally indicate that modifying the corresponding constraint has relatively greater impact on the value of the objective function at optimum.

With the above definitions (14)–(16) as functions of a single parameter λ, the optimization problem is that of finding such λ^* that $C(\lambda^*) = C$. Then the additional sample sizes can be evaluated from (14).

Note that from (13), $\lambda > 0$, as the first term is strictly positive, and the second one is non-negative. Two characteristic values of the main Lagrange multiplier are

$$\overline{\lambda} = \max_h \frac{W_h^2 S_h^2}{c_h m_h^2}, \quad \underline{\lambda} = \min_h \frac{W_h^2 S_h^2}{c_h m_h^2} \tag{17}$$

corresponding to the range of the minimum sample size requirements plugged into (13). They serve as natural bounds for the Lagrange multiplier as established later in Lemma 2.

2.1 Optimization algorithm

The following procedure can be implemented to find the optimal design parameters in practice using bisection method.

Algorithm 1.

1. Set the convergence criteria ϵ (e.g., $\epsilon = C \cdot 10^{-6}$).
2. Find the upper bound $\overline{\lambda}$ using (17).
3. Find the lower bound $\underline{\lambda}$ using (17).
4. If $C(\underline{\lambda}) \leq C$, none of the constraints in (7) are binding, and the optimal allocation is the Neyman-Tchuprow allocation, as demonstrated by the Lemma 1 below.
5. If $C(\underline{\lambda}) > C$, set $\lambda_u^{(k)} \leftarrow \overline{\lambda}$, $\lambda_l^{(k)} \leftarrow \underline{\lambda}$, $k \leftarrow 1$.
6. Set $\lambda^{(k)} \leftarrow (\lambda_l^{(k)} + \lambda_u^{(k)})/2$.
7. Compute $t_h(\lambda^{(k)}), h = 1, \ldots, H$.

8. Evaluate the budget constraint $C(\lambda^{(k)})$.

9. If $|C - C^{(k)}| < \epsilon$, go to step 13.

10. If the sample size is too large, and the tentative design based on the k-th iteration is over budget ($C(\lambda^{(k)}) > C$), increase λ: set $\lambda_l^{(k)} \leftarrow \lambda^{(k)}, k \leftarrow k + 1$.

11. If the sample size is too small, and the tentative design based on the k-th iteration is under budget ($C(\lambda^{(k)}) < C$), decrease λ: set $\lambda_u^{(k)} \leftarrow \lambda^{(k)}, k \leftarrow k + 1$.

12. Re-iterate to step 6.

13. Set $t_h = t_h(\lambda^{(k)})$, rounding down to the integer part as needed. Exit.

2.2 Properties of the proposed algorithm

The following lemmas establish the properties of the algorithm.

Lemma 1. *If $C(\underline{\lambda}) \leq C$, then $\lambda^* \leq \underline{\lambda}$, none of the constraints in (7) are binding, and the optimal allocation is the Neyman-Tchuprow allocation (3).*

In other words, before embarking on the actual optimization via Algorithm 1, this simple check can be conducted to see if the budget is sufficient to support the minimum sample sizes for all strata.

Proof of Lemma 1.

From the definitions of $t_h(\lambda)$ and $C(\lambda)$, it follows that they are monotonic in λ. In particular, $t_h(\lambda)$ are non-increasing in λ, and are strictly decreasing if $\lambda < (W_h^2 S_h^2)/(c_h m_h^2)$. Hence, $C(\lambda)$ is strictly decreasing as long as some $t_h(\lambda) > 0$ for a given λ. Therefore, since $C(\underline{\lambda}) \leq C = C(\lambda^*)$, we have $\lambda^* \leq \underline{\lambda}$.

Assume, without loss of generality, that the strata are numbered in the increasing order of the quantity $(W_h^2 S_h^2)/(c_h m_h^2)$. Then the minimum of $(W_h^2 S_h^2)/(c_h m_h^2)$ is achieved in the first stratum, meaning that $\underline{\lambda} = (W_1^2 S_1^2)/(c_1 m_1^2) \leq (W_h^2 S_h^2)/(c_h m_h^2)$ for all h. Then

$$
\begin{aligned}
t_1(\underline{\lambda}) &= \max\Big[\frac{W_1 S_1}{\sqrt{c_1 \underline{\lambda}}} - m_1, 0\Big] = \max(m_1 - m_1, 0) = 0, \\
t_h(\underline{\lambda}) &= \max\Big[\frac{W_h S_h}{\sqrt{c_h \underline{\lambda}}} - m_h, 0\Big] = \max\Big[\frac{W_h S_h \sqrt{c_1}}{W_1 S_1 \sqrt{c_h}} m_1 - m_h, 0\Big] \\
&= \frac{W_h S_h \sqrt{c_1}}{W_1 S_1 \sqrt{c_h}} m_1 - m_h \geq t_1(\underline{\lambda}) = 0, \quad h > 1
\end{aligned}
$$

As $t_h(\lambda)$ are strictly decreasing in λ for $0 < \lambda < \underline{\lambda}$, all of $t_h(\lambda)$ are strictly positive in this interval. In particular, $t_h(\lambda^*) > 0$ for all h, including $h = 1$, since $\lambda^* < \underline{\lambda}$.

□

The proof of Lemma 1 demonstrates that the value $\underline{\lambda}$ is the lowest value of λ at which some strata are constrained by the minimum sample size requirements. However, if the cost $C(\underline{\lambda})$ is too high, then more strata sizes may need to be constrained. This more general case is treated in the following lemma.

Lemma 2. *If $C(\overline{\lambda}) < C < C(\underline{\lambda})$, the optimal Lagrange multiplier for the problem (5)–(7) is contained between the upper and lower bounds of Algorithm 1.*

Proof of Lemma 2.

As was done in the proof of Lemma 1, assume, without loss of generality, that the strata are numbered in the increasing order of $(W_h^2 S_h^2)/(c_h m_h^2)$, so that the maximum of $W_h^2 S_h^2/c_h m_h^2$ is achieved in the last stratum, $\overline{\lambda} = (W_H^2 S_H^2)/(c_H m_H^2) \geq (W_h^2 S_h^2)/(c_h m_h^2)$. Then

$$t_h(\overline{\lambda}) = m_h \max\Big[\frac{W_h S_h}{m_h \sqrt{\overline{\lambda} c_h}} - 1, 0\Big] = m_h \max\Big[\frac{W_h S_h}{m_h \sqrt{c_h}} \frac{m_H \sqrt{c_H}}{W_H S_H} - 1, 0\Big] = 0 \quad \text{for all } h$$

since the first argument of the maximum is non-positive for all h (and is identically zero for $h = H$). Hence

$$C(\overline{\lambda}) = \sum_h c_h m_h < C$$

according to the assumption (8). On the other hand, $C(\underline{\lambda}) > C$ as stated in the assumptions of the Lemma, and as assured on step 4 of Algorithm 1. Since $C(\lambda)$ is a continuous function of λ, the optimal point such that $C(\lambda^*) = C$ is contained in $[\underline{\lambda}, \overline{\lambda}]$. □

Lemma 3. *If $C < C(\overline{\lambda})$, then no solution can be found.*

Proof of Lemma 3.

As was done in proofs of Lemmas 1 and 2, assume without loss of generality that the strata are numbered in the increasing order of $(W_h^2 S_h^2)/(c_h m_h^2)$, so that the maximum of $W_h^2 S_h^2/c_h m_h^2$ is achieved in the last stratum, $\overline{\lambda} = (W_H^2 S_H^2)/(c_H m_H^2) \geq (W_h^2 S_h^2)/(c_h m_h^2)$. Then

$$t_H(\overline{\lambda}) = \max\Big[\frac{W_H S_H}{\sqrt{c_H \overline{\lambda}}} - m_H, 0\Big] = \max(m_H - m_H, 0) = 0,$$

$$t_h(\overline{\lambda}) = \max\Big[\frac{W_h S_h}{\sqrt{c_h \overline{\lambda}}} - m_h, 0\Big] = \max\Big[\frac{W_h S_h \sqrt{c_H}}{W_H S_H \sqrt{c_h}} m_H - m_h, 0\Big] = 0, \quad h < H$$

since $\frac{W_h S_h \sqrt{c_H}}{W_H S_H \sqrt{c_h}} m_H - m_h \leq 0$.

For all $\lambda \geq \overline{\lambda}$, all of the constraints are binding, $t_h(\lambda) = 0$, $n_h = m_h$, $C(\lambda) = \sum_h c_h m_h = C(\overline{\lambda})$. On the other hand, since $t_H(\lambda) > 0$ when $\lambda < \overline{\lambda}$, the overall budget $C(\lambda)$ is a strictly decreasing function of λ in that interval. Hence $C(\lambda) > C(\overline{\lambda}) > C$ for $\lambda < \overline{\lambda}$. Therefore there is no λ^* such that $C(\lambda^*) = C$.

□

In terms of the sampling design problem, the condition $C < C(\overline{\lambda})$ of Lemma (3) means that the budget constraint (8) is violated. Thus the lemma establishes the existence of a feasible sampling design given the available budget.

Combining these lemmas together, the general properties of the optimal design can be established depending on the relation between the available budget C and the characteristics values $\underline{\lambda}$, $\overline{\lambda}$:

1. If $C < C(\overline{\lambda})$, no solutions exist, as the budget is insufficient even for the required minimum sample sizes.

2. If $C(\overline{\lambda}) \leq C \leq C(\underline{\lambda})$, then the solution exists, and the constraints on the minimum sample sizes are active in at least one stratum.

3. If $C(\underline{\lambda}) < C$, none of the constraints on the minimum sample sizes are active, and the optimal allocation is Neyman-Tchuprow.

We also need to establish some technical conditions required for the proof that the algorithm converges to the optimal point.

Lemma 4. *The optimal values t_h^* that solve the optimization problem (5)–(7) exists and is unique.*

Proof of Lemma 4.

The existence of the solution follows from the fact that the objective function (5) being minimized is continuous in its arguments t_h, and the feasible set (i.e., the set of values t_h such that the conditions (6)–(7) are satisfied) is nonempty and compact (in fact, it is a bounded closed polytop).

The objective function (5) being minimized is a strictly convex function of its parameters t_h. The constraints are linear. Uniqueness of the solution $\{t_h^*, h = 1, \ldots, H\}$ follows from the standard convex optimization theory results (Griva et al. 2008).

□

Lemma 5. *The Lagrange multipliers λ^* and ν_h^*, $h = 1, ..., H$ that satisfy the first order optimality KKT conditions (6)–(7) exist and are unique.*

Proof of Lemma 5. The existence of the Lagrange multipliers follows from the fact the problem (5)–(7) satisfies the constraint qualification (e.g. the Slaters condition; see Griva et al. (2008)). To prove the uniqueness we note that due to (14) and (15), $C(\lambda)$ is a monotonically decreasing function of λ. Moreover, by (8) there exists a stratum number $h_0 \in \{1, \ldots, H\}$ such that $t_{h_0}^* > 0$. Since $t_{h_0}^*$ must satisfy (14) for any optimal λ^*, $C(\lambda)$ is strictly monotonically decreasing function of λ in the neighborhood of any optimal λ^*. As $C(\lambda)$ is nonincreasing for any λ and is strictly monotonic in the neighborhood of an optimal λ^*, we conclude that there cannot be $\lambda_1^* \neq \lambda_2^*$ such that $C(\lambda_1^*) = C(\lambda_2^*)$. Therefore there is only one λ^* that satisfies the KKT conditions.

Due to Lemma 4, the solution to the problem t_h^*, $h = 1, ..., H$ is uniquely defined. Therefore the uniqueness of ν_h^*, $h = 1, ..., H$ follows from the uniqueness of t_h^*, $h = 1, ..., H$, λ^*, and (16).

□

The main result is thus the following.

Theorem 1. *Algorithm 1 generates sequence $\{t_h(\lambda^{(k)}), k = 1, 2, \ldots\}$ that converges to the optimal solution as the algorithm is iterated without stopping,* $\lim_{k\to\infty} t_h(\lambda^{(k)}) = t_h^*$.

Proof of Theorem 1. The result follows from the existence and uniqueness of the solution $\{t_h^*, \lambda^*, \nu_h^*, h = 1, \dots H\}$ that must satisfy the optimality conditions (10)–(12) (Lemmas 4 and 5), and the fact that the algorithm generates a sequence $\{t_h(\lambda^{(k)}), \lambda^{(k)}, \nu_h(\lambda^{(k)})\}$ that in the limit satisfies the first order conditions (10)–(12).

□

Note that conditions (10) and (12) are satisfied exactly by definitions of $t_h(\lambda)$ (14) and $\nu_h(\lambda)$ (16), while the condition (11) is being satisfied in the limit due to the general properties of the bisection algorithm.

In terms of the practical implementation of the algorithm, since at each step the length of the interval $(\lambda_l^{(k)}, \lambda_u^{(k)})$ is being cut in half, the convergence condition of step (9) will be satisfied after at most $K = \lceil \log_2(\overline{\lambda}/\epsilon) \rceil$, i.e., rounded up to the nearest integer.

2.3 Interpretation

In substantive terms, the Neyman-Tchuprow allocation is the optimal allocation with a generous enough budget, $C(\underline{\lambda}) \leq C$. In fact Algorithm 1 can find this allocation when the initial lower bound is set to zero or any value $\tilde{\lambda}$ such that $C(\tilde{\lambda}) > C$, rather than to $\underline{\lambda}$. In terms of the proof of existence (Lemma 2), by taking an arbitrarily small $\tilde{\lambda}$, the budget $C(\tilde{\lambda})$ can be made arbitrarily large to ensure $C(\tilde{\lambda}) > C$. As setting $\lambda = 0$ makes the cost function $C(0)$ go to infinity, in practice the initial lower bound $\lambda_l^{(1)}$ can be set to an arbitrary small value such as a small multiple of the machine precision. When the algorithm is implemented that way, the step 4 of Algorithm 1 is superfluous, as Neyman-Tchuprow allocation will be found as a special case of the more general problem that it solves.

3. Example: ethnicity in the North-Eastern region of the U.S.

In this example, we shall demonstrate the technique in an application to the proportion of persons of Hispanic ethnicity in the Northeastern region of the United States. The population parameters obtained from the 2014 American Community Survey data are given in Table 1, and states are considered to be the sampling strata. The population variances are those of the binary indicator of being Hispanic, $S_h^2 = p_h(1 - p_h)$. Larger states with bigger cities have higher proportions of Hispanics. This population shows both large differences in strata sizes and strata variances.

3.1 Allocations that account for unequal variances

Suppose that a sample of size $n = 1,000$ is to be taken from this population, and the data collection costs are the same across strata: $C = 1,000$; $c_h = 1$ for all h. The starting point is the Neyman-Tchuprow allocation, which can also be thought of as the allocation with the minimal sample size requirement of $m_h = 1$. The smallest sample size is $n_9 = 5$ for the smallest state of Vermont. Note that in this case, the solution can be obtained via the proposed algorithm, albeit $\lambda^* < \underline{\lambda}$.

Table 1: Hispanic ethnicity, Northeastern region of the U.S.

	Total pop	Hispanic pop	% Hispanic	S_h^2
Connecticut (CT)	3,592,053	512,795	14.28%	0.12238
Maine (ME)	1,328,535	18,592	1.40%	0.01380
Massachusetts (MA)	6,657,291	681,824	10.24%	0.09193
New Hampshire (NH)	1,321,069	40,301	3.05%	0.02958
New Jersey (NJ)	8,874,374	1,649,784	18.59%	0.15134
New York (NY)	19,594,330	3,559,644	18.17%	0.14866
Pennsylvania (PA)	12,758,729	784,562	6.15%	0.05771
Rhode Island (RI)	1,053,252	139,832	13.28%	0.11514
Vermont (VT)	626,358	10,226	1.63%	0.01606

The solutions for $m_h = 20, 50$ and 100 are also demonstrated in the table. As the sample size requirements increase, the constraints become binding for the smallest states, with their respective $t_h(\lambda)$ values becoming zeroes. Finally, for the most demanding allocation problem with $m_h = 100$, only $n - \sum_h m_h = 100$ cases can be freely allocated, and they are all allocated to the largest state of New York. Note that, as a fraction of $\overline{\lambda}$, the optimal value of the Lagrange multiplier λ^* moves from being a tiny fraction of $\overline{\lambda}$ for the Neyman-Chuprow $m_h = 1$ allocation to about 1/4 of the value of $\underline{\lambda}$ for $m_h = 100$. In terms of the interpretation of the Lagrange multiplier as the "shadow price" of a constraint, growing values of λ^* reflect that the constraint features more and more prominently in the optimization problem as the increasing minimal strata sample size requirements become more restrictive. This is also highlighted by the "% free" row that shows the sample sizes in the states that are freely allocated with nonzero values of t_h. Finally, the design effect row provides the comparison against the variance attained by the Neyman-Tchuprow allocation. Higher values of the design effects indicate the sacrifices that the sampling design makes in order to satisfy the minimal sample size constraints.

3.2 Allocations that do not account for unequal variances

If the tentative survey is an omnibus, the strata variances can be set to be equal (say to 1), and the resulting allocation without constraints is the proportional allocation. For this particular population, it provides larger sample sizes to the states that have the proportion of Hispanic population that is lower than the overall population one.

If the same strata-specific sample sizes are imposed as in the previous section, the resulting optimal designs are provided in Table 3. Design effect DFFF1 measures efficiency losses relative to the proportional allocation, and design effect DEFF2, relative to Neyman-Tchuprow allocation.

Table 2: Sampling designs with minimal sample size requirements.

	Neyman-Tchuprow ($m_h = 1$)		$m_h = 20$		$m_h = 50$		$m_h = 100$	
	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h
CT	67.87	69	46.49	67	7.81	58	0	100
ME	7.55	9	0	20	0	50	0	100
MA	109.62	111	86.80	107	42.86	93	0	100
NH	11.45	13	0	20	0	50	0	100
NJ	188.21	190	162.67	183	108.84	159	0	100
NY	413.06	415	379.73	400	297.58	348	99.61	200
PA	166.98	168	142.17	163	91.01	142	0	100
RI	18.59	20	0	20	0	50	0	100
VT	3.35	5	0	20	0	50	0	100
Total		1000		1000		1000		1000
% free		100%		92%		80%		10%
DEFF	1		1.023		1.152		1.688	
$\underline{\lambda}$	$2.02 \cdot 10^{-6}$		$5.06 \cdot 10^{-9}$		$8.09 \cdot 10^{-10}$		$2.02 \cdot 10^{-10}$	
λ^*	$1.069 \cdot 10^{-7}$		$1.15 \cdot 10^{-7}$		$1.52 \cdot 10^{-7}$		$4.60 \cdot 10^{-7}$	
$\overline{\lambda}$	0.0183		$4.58 \cdot 10^{-5}$		$7.33 \cdot 10^{-6}$		$1.83 \cdot 10^{-6}$	

Table 3: Sampling designs with minimal sample size requirements.

	Proportional ($m_h = 1$)		$m_h = 20$		$m_h = 50$		$m_h = 100$	
	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h	$t_h(\lambda)$	n_h
CT	63.14	64	43.52	64	5.60	56	0	100
ME	22.72	23	3.49	24	0	50	0	100
MA	117.88	118	97.71	118	53.05	104	0	100
NH	22.59	23	3.36	24	0	50	0	100
NJ	157.47	158	136.92	157	87.37	138	0	100
NY	348.89	349	326.47	347	253.32	304	181.07	182
PA	226.83	227	205.60	226	147.50	198	117.91	118
RI	17.81	18	0	20	0	50	0	100
VT	10.18	11	0	20	0	50	0	100
Total		1000		1000		1000		1000
% free		100%		96%		80%		20%
DEFF1	1		1.004		1.096		1.573	
DEFF2	1.055		1.063		1.191		1.730	
$\underline{\lambda}$	$1.26 \cdot 10^{-4}$		$3.15 \cdot 10^{-7}$		$5.04 \cdot 10^{-8}$		$1.26 \cdot 10^{-8}$	
λ^*	$1.01 \cdot 10^{-6}$		$1.03 \cdot 10^{-6}$		$1.34 \cdot 10^{-6}$		$3.76 \cdot 10^{-6}$	
$\overline{\lambda}$	0.1233		$3.08 \cdot 10^{-4}$		$4.93 \cdot 10^{-5}$		$1.23 \cdot 10^{-5}$	

4. Discussion

The presented work builds a foundation for a number of extensions.

First, sample size requirements may be given by the stakeholders for some but not all strata. To accommodate such a set of sampling design requirements, dummy constraints $m_h = 1$ can be introduced for such strata, as demonstrated by the first column of Tables 2–3.

Second, most practical sampling designs are those in which sampling is taken without replacement (SRSWOR). The differences between SRSWR and SRSWOR are immaterial when sampling fractions are small. In the optimization problem considered here, incorporating sampling with replacement would require modifying the objective function (5) to

$$\sum_{h=1}^{H} W_h^2 S_h^2 \Big(\frac{1}{n_h} - \frac{1}{N_h} \Big),$$

and boils down to an additive constant that does not affect the solution. In other words, the case of SRSWOR with strata, or even a mix of SRSWR in some strata and SRSWOR in others, is fully covered by the results and methods of this paper. We chose to only present SRSWR simply because the formulae are more compact.

Finally, more complex sampling designs within strata can be incorporated if they produce contributions to the variance that scale exactly as $\mathcal{V}_h/n_h$ for some fixed effective variance $\mathcal{V}_h$. That is to say, if the sampling designs within each stratum are cluster sampling designs, multiple frame designs, unequal probability of selection designs, two-phase sampling designs such as those screening for a rare population, etc., that produce a sample with a fixed design effect that does not depend on the overall nominal sample size n_h (at least for a given variable of interest), then the proposed approach and Algorithm 1 can be utilized to arrive at the optimal allocation. This covers the important case of BRFSS mentioned in the introduction, where the two frames are the landline and cell phone random digit dialing. (ignoring the cell phone cases that reside in states other than those that they were dialed in). Even if the complex designs are not scalable with a single design effect per stratum, the presentation of the optimization problem and the algorithm as given in this paper lay out the groundwork for the steps necessary to solve this problem.

References

Centers for Disease Control and Prevention (2013), Comparability of data BRFSS 2013, Technical report, Atlanta, GA. Available from http://www.cdc.gov/brfss/annual_data/2013/pdf/compare_2013.pdf.

Choudhry, G. H., Rao, J. N. K. & Hidiroglou, M. A. (2012), 'On sample allocation for efficient domain estimation', *Survey Methodology* **38**(1), 23–29.

Griva, I., Nash, S. G. & Sofer, A. (2008), *Linear and Nonlinear Optimization, Second Edition*, 2 edn, Society for Industrial Mathematics, Philadelphia.

Hansen, M., Hurwitz, W. N. & Madow, W. G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons, New York.

New York State Department of Health (2014), *Expanded Behavioral Risk Factor Surveillance System*, Albany, NY. Available at https://www.health.ny.gov/statistics/brfss/expanded/.

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* **109**, 558–606.

Tchuprow, A. A. (1923), 'On the mathematical expectation of the moments of frequency distributions in the case of correlated observations', *Metron* **2**, 646–680.

Thompson, S. K. (1992), *Sampling*, John Wiley and Sons, New York.

Wright, T. (2012), 'The equivalence of Neyman optimum allocation for sampling and equal proportions for apportioning the U.S. House of Representatives', *The American Statistician* **66**(4), 217–224.