RESEARCH REPORT SERIES
(Statistics #2016-03)

Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition is Key

Tommy Wright

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: May 10, 2016
Report Revised: July 13, 2016

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition Is Key

Tommy Wright
U. S. Bureau of the Census

Abstract

Neyman allocation of the sample under stratified random sampling is among the top major advances and most widely used methods in probability sampling theory because it minimizes sampling variance. Neyman allocation rarely yields integer solutions.

Building on Algorithms I and II in Wright (2012) and Algorithm III in Wright (2014) which provide integer solutions and thus avoiding the need to round to integers, we present two more exact optimal sample allocation algorithms. Algorithm IV minimizes the overall sample size with a desired precision constraint, and Algorithm V seeks to minimize (or at least decrease) the sampling variance for a fixed cost constraint or budget. We actually present four variations of Algorithm V.

Remarkably, the presented simple algorithms *always* find the global optimum.

KEY WORDS: Cost constraints; Exact optimal allocation; Neyman allocation; Precision constraints; Stratification; Sampling variance decomposition.

1. INTRODUCTION

When estimating a finite population total, we consider the problem of exact optimal allocation (all positive integer values) of an overall fixed sample size n in stratified random sampling with H strata to minimize sampling variance. Wright (2012) shows for stratum h $(h = 1, 2, ..., H)$: (1) that Algorithm I gives an exact optimal allocation subject to $n_h \geq 1$ for all h; (2) that Algorithm II gives an exact optimal allocation subject to $n_h \geq 2$ for all h; and (3) that controlled rounding with Neyman allocation does not always lead to the minimum variance. Wright (2012) further shows that two problems (sample allocation and reapportionment) are special cases of the following more general problem and gives the general solution: if z_i (for $i = 1, 2, ..., I$) are positive integers such that $\sum_{i=1}^{I} z_i = z$ where z is fixed, find the values of z_i that will minimize $\sum_{i=1}^{I} \frac{r_i}{z_i}$ for any fixed positive real numbers $r_1, r_2, ..., r_I$.

Wright (2014) generalizes Algorithms I and II with a new Algorithm III that gives an exact optimal allocation subject to $0 < a_h \leq n_h \leq b_h \leq N_h$ for all h where a_h and b_h are stated positive integers and overall fixed n. Among advantages, Algorithm III avoids the possibility that $n_h > N_h$ for some stratum h, as is possible with Neyman allocation. In this short paper, we give exact optimal sample allocation algorithms (1) when n is not specified but desired precision V_o is specified (Algorithm IV) and (2) when there are overall costs or budget constraints (Algorithm V).

When $\hat{T}_Y$ is an estimator of the population total T_Y, an elementary decomposition of the sampling variance $Var(\hat{T}_Y) = Var(N\bar{y})$ under simple random sampling which

----

Tommy Wright is Chief of the Center for Statistical Research and Methodology, U. S. Bureau of the Census, Washington, DC 20233 (E-mail: tommy.wright@census.gov) and adjunct faculty in mathematics and statistics at Georgetown University. The views expressed are those of the author and not necessarily those of the U. S. Bureau of the Census.

shows how to explicitly reduce $Var(\hat{T}_Y) = N^2(\frac{N-n}{N})\frac{S^2}{n}$ step-by-step as n goes from 1 to 2 to 3 to $\cdots$ to N is given by (see Section 5)

$$Var(\hat{T}_Y) = N(N-1)S^2 - \frac{N^2S^2}{1\cdot 2} - \frac{N^2S^2}{2\cdot 3} - \cdots - \frac{N^2S^2}{(n-1)(n)}. \qquad (*)$$

This decomposition is also easily possible within each stratum under stratified random sampling (see $(**)$ below), and it is the basis for and the essence of all Algorithms I-V in this paper.

2. EXACT SAMPLE ALLOCATION

Assume a finite population of N units is partitioned into H subpopulations of $N_1, N_2, N_3, ..., N_H$ units, respectively. Thus $N = N_1 + N_2 + N_3 + \cdots + N_H$. The subpopulations are called *strata.* We assume that the values $N_1, N_2, ..., N_H$ are known.

Let Y_{hj} be the value of interest for the j^{th} unit in the h^{th} stratum ($j = 1, ..., N_h$ and $h = 1, ..., H$). Also let

$$\bar{Y}_h = \frac{\sum_{j=1}^{N_h} Y_{hj}}{N_h} \quad \text{and} \quad S_h^2 = \frac{\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)^2}{N_h - 1}.$$

In general and for the values Y_{hj}, the population total T_Y is

$$T_Y = \sum_{h=1}^{H}\sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H} T_h = \sum_{h=1}^{H} N_h\bar{Y}_h. \qquad (1)$$

To estimate T_Y under the classical design-based approach, we take (independent) simple random samples - one from each stratum - of sizes $n_1, n_2, ..., n_H$ respectively (entire process called *stratified random sampling*) and obtain the sample means $\bar{y}_1, \bar{y}_2, ..., \bar{y}_H$. Note that $n_h \geq 1$ for all h.

A natural estimator of T_Y is

$$\hat{T}_Y = \sum_{h=1}^{H} \hat{T}_h = \sum_{h=1}^{H} N_h\bar{y}_h. \qquad (2)$$

It is known that $\hat{T}_Y$ is an unbiased estimator of T_Y with sampling variance

$$Var(\hat{T}_Y) = \sum_{h=1}^{H} Var(\hat{T}_h) = \sum_{h=1}^{H} N_h^2\frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h}. \qquad (3)$$

For a given overall sample size n, there is interest in the question regarding how to allocate n among the H strata before sampling. In his landmark paper of 1934, Neyman shows that for fixed n, the allocation (known as Neyman allocation) of n that minimizes $Var(\hat{T}_Y)$ subject to the constraint $n = \sum_{h=1}^{H} n_h$ is given by

$$n_h = \frac{N_hS_h}{\sum_{i=1}^{H} N_iS_i}n \qquad h = 1, 2, 3, ..., H. \qquad (4)$$

Unfortunately, the values of n_h in Neyman allocation are rarely integers.

In this paragraph, we give some background, first for the case $H = 3$. For $H = 3$ and noting that

$$\frac{1}{n_h} = 1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - \cdots - \frac{1}{(n_h - 1)n_h}$$

for each h, it follows that $Var(\hat{T}_Y)$ in (3) can be decomposed as

$$\begin{aligned} Var(\hat{T}_Y) &= \sum_{h=1}^{3} N_h(N_h - 1)S_h^2 \\ & -\frac{N_1^2 S_1^2}{1 \cdot 2} - \frac{N_1^2 S_1^2}{2 \cdot 3} - \frac{N_1^2 S_1^2}{3 \cdot 4} - \cdots - \frac{N_1^2 S_1^2}{(n_1 - 1)(n_1)} \\ & -\frac{N_2^2 S_2^2}{1 \cdot 2} - \frac{N_2^2 S_2^2}{2 \cdot 3} - \frac{N_2^2 S_2^2}{3 \cdot 4} - \cdots - \frac{N_2^2 S_2^2}{(n_2 - 1)(n_2)} \qquad (**) \\ & -\frac{N_3^2 S_3^2}{1 \cdot 2} - \frac{N_3^2 S_3^2}{2 \cdot 3} - \frac{N_3^2 S_3^2}{3 \cdot 4} - \cdots - \frac{N_3^2 S_3^2}{(n_3 - 1)(n_3)} \end{aligned}$$

where $\sum_{h=1}^{3} N_h(N_h - 1)S_h^2$ is the sampling variance $Var(\hat{T}_Y)$ when $n_h = 1$ for all h. Each subtraction in (**) shows how much $\sum_{h=1}^{3} N_h(N_h - 1)S_h^2$ decreases each time we increase the sample size in each stratum by one additional unit until we have $n = n_1 + n_2 + n_3$. In Section 3, we observe that

$$\frac{N_h^2 S_h^2}{(m_h - 1)(m_h)}$$

is the amount by which $Var(\hat{T}_Y)$ "decreases" when the sample size for the h^{th} stratum is increased from $m_h - 1$ to m_h. The result in (**) generalizes for any H and, as follows, the result in (**) and its generalization provides a way to obtain an allocation $(n_1, n_2, ..., n_H)$ of fixed overall sample size n to integer values $n_1, n_2, ..., n_H$ which minimizes $Var(\hat{T}_Y)$.

Next, we present Algorithms I and II. Specifically, for a given overall sample size n, Wright (2012) shows that it is possible to obtain an exact allocation $(n_1, n_2, n_3, ..., n_H)$ of fixed n that minimizes $Var(\hat{T}_Y)$ subject to $n = \sum_{h=1}^{H} n_h$ where n_h is a positive integer for all h, and it is given by Algorithm I.

Exact Optimal Allocation Algorithm I [$n_h \geq 1$] (Wright, 2012)

Step 1: First, assign one unit to be selected for the sample from each stratum.

Step 2: Compute the array of *priority values* where each row corresponds to one of the strata (For simplicity, assume $N_1S_1 \geq N_2S_2 \geq \cdots \geq N_HS_H$):

$$\begin{array}{cccc}
\frac{N_1S_1}{\sqrt{1\cdot 2}} & \frac{N_1S_1}{\sqrt{2\cdot 3}} & \frac{N_1S_1}{\sqrt{3\cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_hS_h}{\sqrt{1\cdot 2}} & \frac{N_hS_h}{\sqrt{2\cdot 3}} & \frac{N_hS_h}{\sqrt{3\cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_HS_H}{\sqrt{1\cdot 2}} & \frac{N_HS_H}{\sqrt{2\cdot 3}} & \frac{N_HS_H}{\sqrt{3\cdot 4}} & \cdots
\end{array}$$

Step 3: Pick the $n - H$ largest priority values from the above array in *Step 2* along with the associated strata. Each stratum is allocated an additional sample unit each time one of its priority values is among the $n - H$ largest values.

Unbiased estimation of $Var(\hat{T}_Y)$ requires the selection of at least two units from each stratum in addition to the requirement that $n = \sum_{h=1}^{H} n_h$, and Wright (2012) gives the following modification which minimizes $Var(\hat{T}_Y)$ in Algorithm II.

Exact Optimal Allocation Algorithm II [$n_h \geq 2$] (Wright, 2012)

Step 1: First, assign two units to be selected from each stratum.

Step 2: Compute the array of *priority values* where each row corresponds to one of the strata (Assume $N_1S_1 \geq N_2S_2 \geq \cdots \geq N_HS_H$):

$$\begin{array}{cccc}
\frac{N_1S_1}{\sqrt{2\cdot 3}} & \frac{N_1S_1}{\sqrt{3\cdot 4}} & \frac{N_1S_1}{\sqrt{4\cdot 5}} & \cdots \\
 & \vdots & & \\
\frac{N_hS_h}{\sqrt{2\cdot 3}} & \frac{N_hS_h}{\sqrt{3\cdot 4}} & \frac{N_hS_h}{\sqrt{4\cdot 5}} & \cdots \\
 & \vdots & & \\
\frac{N_HS_H}{\sqrt{2\cdot 3}} & \frac{N_HS_H}{\sqrt{3\cdot 4}} & \frac{N_HS_H}{\sqrt{4\cdot 5}} & \cdots
\end{array}$$

The array in *Step 2* is the same as the previous array in *Step 2* of Algorithm I except the first column of priority values has been removed. Only priority values with the following values in the denominator $\sqrt{2\cdot 3}$, $\sqrt{3\cdot 4}$, $\sqrt{4\cdot 5}$,... are in the array when we require $n_h \geq 2$.

Step 3: Pick the $n - 2H$ largest priority values from the above array in *Step* 2 along with the associated strata. Each stratum is allocated an additional sample unit each time one of its priority values is among the $n - 2H$ largest values.

3. INTERPRETATION OF THE PRIORITY VALUES $\frac{N_h S_h}{\sqrt{(m_h-1)(m_h)}}$

We give an interpretation of the priority values in Algorithms I and II.

Assume a simple random sample of size m_h from the h^{th} stratum and let $\bar{y}_{m_h}$ be the sample mean based on the m_h sample units. Similarly, let $\bar{y}_{m_h-1}$ be the sample mean from the h^{th} stratum based on $m_h - 1$ sample units. When the sample size for the h^{th} stratum is *"increased"* from $m_h - 1$ to m_h, the associated sampling variance for the h^{th} stratum *"decreases"* by

$$Var(N_h\bar{y}_{m_h-1}) - Var(N_h\bar{y}_{m_h}) = \frac{N_h^2 S_h^2}{(m_h-1)(m_h)} = \left(\frac{N_h S_h}{\sqrt{(m_h-1)(m_h)}}\right)^2. \tag{5}$$

The result in (5) is also the amount by which the overall sampling variance $Var(\hat{T}_Y)$ *"decreases"* when the sample size for the h^{th} stratum is increased from $m_h - 1$ to m_h. When the $n - H$ largest terms are selected sequentially as stated in Algorithm I, each selection decreases the $Var(\hat{T}_Y)$ by an associated squared priority value from a stratum which is the largest amount possible at that point. Also by picking the value $\frac{N_h^2 S_h^2}{(m_h-1)(m_h)}$ (or equivalently $\frac{N_h S_h}{\sqrt{(m_h-1)(m_h)}}$), it is clear that up to and including that point, we have a sample size of m_h from the h^{th} stratum.

4. EXACT OPTIMAL SAMPLE ALLOCATION FOR MIXED CONSTRAINTS, DESIRED PRECISION, or COST CONSTRAINTS

4.1 Varying Strata Minimum and Maximum Sample Size Constraints

Given the intrepretation in Section 3, Wright (2014) notes it is possible to set varying minimum sample sizes for the strata as well as varying maximum sample sizes for the strata. That is, for example, we may want to have mixed constraints $2 \leq n_1 \leq 7$; $3 \leq n_2 \leq 6$; $1 \leq n_3 \leq 10$; etc. More generally, let a_h and b_h be integers with the following constraints:

$$n = \sum_{h=1}^{H} n_h; \tag{6}$$

$$0 < a_h \leq n_h \leq b_h \leq N_h. \tag{7}$$

We refer to the combination of constraints in (6) - (7) as a *mixed constraint pattern.*

From the discussion in Section 3, it follows immediately that the following Algorithm III yields an exact optimal sample allocation $(n_1, n_2, ..., n_H)$ of fixed n under the mixed constraint pattern in (6) - (7) to minimize $Var(\hat{T}_Y)$.

Algorithm III is a generalization of Algorithms I and II.

Exact Optimal Allocation Algorithm III [Mixed Constraints] (Wright, 2014)

Step 1: Determine an array as given assuming $N_1S_1 \geq N_2S_2 \geq \cdots \geq N_HS_H$.

$$\begin{array}{cccc}
\frac{N_1S_1}{\sqrt{1\cdot 2}} & \frac{N_1S_1}{\sqrt{2\cdot 3}} & \frac{N_1S_1}{\sqrt{3\cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_hS_h}{\sqrt{1\cdot 2}} & \frac{N_hS_h}{\sqrt{2\cdot 3}} & \frac{N_hS_h}{\sqrt{3\cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_HS_H}{\sqrt{1\cdot 2}} & \frac{N_HS_H}{\sqrt{2\cdot 3}} & \frac{N_HS_H}{\sqrt{3\cdot 4}} & \cdots
\end{array}$$

Step 2: Assume the mixed constraints in (6) - (7). On the h^{th} row (stratum) of the array, remove the values in all of the columns less than or equal to the $(a_h - 1)^{th}$ column and the values in all of the columns greater than the $(b_h - 1)^{th}$ column for $h = 1, ..., H$.

Step 3: From the h^{th} row (stratum), at least a_h units and no more than b_h units are to be included in the sample. So from the remaining values in the array from Step 2, select the largest $n - \sum_{h=1}^{H} a_h$ values to complete the overall allocation of n among the H strata. Each stratum is allocated an additional sample unit each time one of its priority values is among the $n - \sum_{h=1}^{H} a_h$ largest values from the array in Step 2.

4.2 When n Is Unspecified and Precison V_o Is Specified

When n is not given or specified, assume the investigator expresses the desire to determine minimum n and allocation $(n_1, ..., n_H)$ to meet a specified sampling variance $V_o = Var(\hat{T}_Y)$. We assume that

$$V_o < \sum_{h=1}^{H} N_h(N_h - 1)S_h^2 \tag{8}$$

which is the sampling variance when $n_h = 1$ for all h.

We present the following solution, and refer to it as Algorithm IV.

Exact Optimal Allocation Algorithm IV [Unspecified n, Specified Precision V_o]

Step 1: First, assign 1 unit to be selected for the sample from each stratum.

Step 2: Compute the array of *priority values* as given in Algorithm I.

Step 3: Compute $Var(\hat{T}_Y) = Var(\hat{T}_Y|n_{11}, ..., n_{H1})$ where $n_{11} = 1, ..., n_{H1} = 1$. Because we assumed $Var(\hat{T}_Y|n_{11}, ..., n_{H1}) > V_o$, go to Step 4.
[Note $Var(\hat{T}_Y|n_{11}, ..., n_{H1}) = \sum_{h=1}^{H} N_h(N_h - 1)S_h^2$.]

Step 4: Pick the largest priority value, refer to it as the "1st selected priority value", associate it with its stratum, and increase that stratum sample size by 1. Our new sample sizes are $n_1 = n_{12}, n_2 = n_{22}, ..., n_H = n_{H2}$ where exactly one of $n_{12}, n_{22}, ..., n_{H2}$ is 2 and the rest are each 1. Next, compute $Var(\hat{T}_Y|n_{12}, ..., n_{H2})$. If $Var(\hat{T}_Y|n_{12}, ..., n_{H2}) \leq V_o$, stop with $n_1 = n_{12}, ..., n_H = n_{H2}$, and $n = n_{12} + \cdots + n_{H2}$. If $Var(\hat{T}_Y|n_{12}, ..., n_{H2}) > V_o$, go to Step 5.
[Note $Var(\hat{T}_Y|n_{12}, ..., n_{H2}) = Var(\hat{T}_Y|n_{11}, ..., n_{H1})$ - (1st selected priority value)^2.]

Step 5: Pick the next largest priority value, refer to it as the "2nd selected priority value", associate it with its stratum, and increase that stratum sample size by 1. Our new sample sizes are $n_1 = n_{13}, n_2 = n_{23}, ..., n_H = n_{H3}$. Next, compute $Var(\hat{T}_Y|n_{13}, ..., n_{H3})$. If $Var(\hat{T}_Y|n_{13}, ..., n_{H3}) \leq V_o$, stop with $n_1 = n_{13}, ..., n_H = n_{H3}$, and $n = n_{13} + \cdots + n_{H3}$. If $Var(\hat{T}_Y|n_{13}, ..., n_{H3}) > V_o$, go to the next Step.
[Note $Var(\hat{T}_Y|n_{13}, ..., n_{H3}) = Var(\hat{T}_Y|n_{12}, ..., n_{H2})$ - (2nd selected priority value)^2.]

NEXT Steps: Altogether, continue through J steps sequentially until $Var(\hat{T}_Y|n_{1J}, ..., n_{HJ}) \leq V_o$ and the final sample size and allocation $(n_{1J}, n_{2J}, ..., n_{HJ})$ are such that

$$n = n_{1J} + n_{2J} + \cdots + n_{HJ}. \tag{9}$$

Remark 1: Note that n is the minimum overall sample size that achieves V_o.

4.3 When n Is Unspecified and Overall Cost (or Budget) Is Specified

All of the allocation algorithms considered so far in this paper, including Neyman allocation, implicitly assume that the cost of one sample unit observation from stratum h is the same as the cost of one sample unit observation from stratum h' for $h \neq h'$ and $h, h' = 1, ..., H$. Assume that c_h is the cost of one sample unit observation from stratum h for $h = 1, ..., H$ and that C ($> \sum_{h=1}^{H} c_h$) is the overall given available cost or budget. Thus for allocation $(n_1, n_2, ..., n_H)$, we have the cost constraint

$$\sum_{h=1}^{H} c_h n_h \leq C. \tag{10}$$

Assume that $N_1S_1 \geq N_2S_2 \geq \cdots \geq N_HS_H$. We actually present four variations of algorithms when there is a cost constraint or fixed budget. We first consider Algorithm *V-a*.

Exact Optimal Allocation Algorithm V-a [Unspecified n, Specified Cost C]

```
Step 1: First, assign 1 unit to be selected for the sample from each stratum.
Step 2: Compute the array of priority values as given in Algorithm I.
Step 3: Pick the largest priority value in the array that has not already been picked; associate it with its stratum; and increase that stratum sample size by 1 if the new cost including this 1 new sample unit is $\sum_{h=1}^{H} c_h n_h \leq C$. Otherwise (i.e., $\sum_{h=1}^{H} c_h n_h > C$), stop without increasing the sample size by 1 for the associated stratum; cost of sample when we stop is $C_a = \sum_{h=1}^{H} c_h n_h$.
Step 4: Go to Step 3.
```

Remark 2: When we stop with Algorithm *V-a*, the resulting allocation $(n_1, n_2, ..., n_H)$ gives minimum variance (3) for the overall realized sample size $n = \sum_{h=1}^{H} n_h$; and because $C_a = \sum_{h=1}^{H} c_h n_h \leq C$, we stay within budget.

Remark 3: When we stop with Algorithm *V-a*, it is possible that $C - C_a$ is sufficiently large that we could add a few more sample units from some strata without exceeding the budget C. This motivates Algorithm *V-b*.

Exact Optimal Allocation Algorithm V-b [Unspecified n, Specified Cost C]

```
Step 1: First, assign 1 unit to be selected for the sample from each stratum.
Step 2: Compute the array of priority values as given in Algorithm I.
Step 3: Compute $C^* = \sum_{h=1}^{H} c_h n_h$, the cost for the current stratum sample sizes.
Step 4: For each row of the array, $h$, compute $C^* + c_h$. If $C^* + c_h > C$, then remove all the unpicked values from the $h^{th}$ row of the array.
Step 5: When all values have been removed from the array, stop.
Step 6: Pick the largest remaining priority value in the array; associate it with its stratum; and increase that stratum sample size by 1. Remove the priority value which was chosen from the array.
Step 7: Go to Step 3.
```

Remark 4: When we stop with Algorithm *V-b*, the resulting $n_1, n_2, ..., n_H$ further decrease the variance that would be obtained by Algorithm *V-a*; and the amount of the budget spent for the sample $C_b = \sum_{h=1}^{H} c_h n_h$ is closer than C_a to C without exceeding it.

For the last two variations of algorithms involving cost constraints, we consider a different function other than $Var(\hat{T}_Y)$. Now assume that n_h is at least 1 for $h = 1, 2, ..., H$ and that $\frac{N_1 S_1}{\sqrt{c_1}} \geq \cdots \geq \frac{N_H S_H}{\sqrt{c_H}}$ for convenience. For mathematical convenience, the desire is to determine an allocation $n_1, n_2, ..., n_H$ that minimizes a new objective function, the *weighted sum of stratum sampling variances*, below without exceeding the budget

$$\sum_{h=1}^{H} Var(N_h\bar{y}_h)\frac{1}{c_h} = \sum_{h=1}^{H} N_h^2(\frac{N_h - n_h}{N_h})\frac{S_h^2}{n_h}\frac{1}{c_h}. \tag{11}$$

Determining an allocation $(n_1, n_2, ..., n_H)$ to minimize (11) without exceeding the budget is equivalent to finding the allocation to minimize

$$\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{c_h}\frac{1}{n_h}. \tag{12}$$

Following the derivation in Wright (2012) with $\frac{N_h^2 S_h^2}{c_h}$ in place of $N_h^2 S_h^2$ leads to the following solution which we call Algorithm *V-c*.

Exact Optimal Allocation Algorithm V-c [Unspecified n, Specified Cost C]

Step 1: First, assign 1 unit to be selected for the sample from each stratum.

Step 2: Compute the array of *priority values* where each row corresponds to one of the strata (assume $\frac{N_1 S_1}{\sqrt{c_1}} \geq \frac{N_2 S_2}{\sqrt{c_2}} \geq \cdots \geq \frac{N_H S_H}{\sqrt{c_H}}$):

$$\begin{array}{cccc} \frac{(N_1 S_1/\sqrt{c_1})}{\sqrt{1\cdot 2}} & \frac{(N_1 S_1/\sqrt{c_1})}{\sqrt{2\cdot 3}} & \frac{(N_1 S_1/\sqrt{c_1})}{\sqrt{3\cdot 4}} & \cdots \\ & \vdots & & \\ \frac{(N_h S_h/\sqrt{c_h})}{\sqrt{1\cdot 2}} & \frac{(N_h S_h/\sqrt{c_h})}{\sqrt{2\cdot 3}} & \frac{(N_h S_h/\sqrt{c_h})}{\sqrt{3\cdot 4}} & \cdots \\ & \vdots & & \\ \frac{(N_H S_H/\sqrt{c_H})}{\sqrt{1\cdot 2}} & \frac{(N_H S_H/\sqrt{c_H})}{\sqrt{2\cdot 3}} & \frac{(N_H S_H/\sqrt{c_H})}{\sqrt{3\cdot 4}} & \cdots \end{array} \tag{13}$$

Step 3: Pick the largest priority value in the array that has not already been picked; associate it with its stratum; and increase that stratum sample size by 1 if the new cost including this 1 new sample unit is $\sum_{h=1}^{H} c_h n_h \leq C$. Otherwise (i.e., $\sum_{h=1}^{H} c_h n_h > C$), stop without increasing the sample size by 1 for the associated stratum; cost of sample when we stop is $C_c = \sum_{h=1}^{H} c_h n_h$.

Step 4: Go to Step 3.

Remark 5: When we stop with Algorithm *V-c*, the resulting allocation $(n_1, n_2, ..., n_H)$ gives minimum *weighted sum of stratum sampling variances* (11) for the overall realized sample size $n = \sum_{h=1}^{H} n_h$; and because $C_c = \sum_{h=1}^{H} c_h n_h \leq C$, we stay within budget.

Remark 6: As with Algorithm *V-a*, when we stop with Algorithm *V-c*, it is possible that $C - C_c$ is sufficiently large that we could add a few more sample units from some strata without exceeding the budget C. This motivates Algorithm *V-d*.

Exact Optimal Allocation Algorithm V-d [Unspecified n, Specified Cost C]

Step 1: First, assign 1 unit to be selected for the sample from each stratum.
Step 2: Compute the array of *priority values* as in Algorithm *V-c*.
Step 3: Compute $C^* = \sum_{h=1}^{H} c_h n_h$, the cost for the current stratum sample sizes.
Step 4: For each row of the array, h, compute $C^* + c_h$. If $C^* + c_h > C$, then remove all the unpicked values from the h^{th} row of the array.
Step 5: When all values have been removed from the array, stop.
Step 6: Pick the largest remaining priority value in the array; associate it with its stratum; and increase that stratum sample size by 1. Remove the priority value which was chosen from the array.
Step 7: Go to Step 3.

Remark 7: With the minimization of (11) in mind, when we stop with Algorithm *V-d*, the final realized value of n in Algorithm *V-d* is $n = \sum_{h=1}^{H} n_h$; the objective function in (11) is decreased further than it would be with Algorithm *V-c*; and the amount of the budget spent for the sample $C_d = \sum_{h=1}^{H} c_h n_h$ is closer than C_c to C without exceeding C.

We give two examples to illustrate Algorithms *V-a*, *V-b*, *V-c*, and *V-d.*

Example 1: With fixed budget $C = \$55$, assume a finite population of $N = 149$ units stratified into three strata with the following parameters:

h	N_h	S_h	c_h	$N_h S_h$
1	47	10	\$9	470
2	61	6	\$4	366
3	41	4	\$1	164

We start with the following Array 0.

c_h	$N_h S_h$	$\frac{1}{\sqrt{1 \cdot 2}}$	$\frac{1}{\sqrt{2 \cdot 3}}$	$\frac{1}{\sqrt{3 \cdot 4}}$	$\frac{1}{\sqrt{4 \cdot 5}}$	$\frac{1}{\sqrt{5 \cdot 6}}$	$\frac{1}{\sqrt{6 \cdot 7}}$	$\frac{1}{\sqrt{7 \cdot 8}}$	$\cdots$
\$9	470	332.34	191.98	135.68	105.10	85.81	72.52	62.81	$\cdots$
\$4	366	258.80	149.42	105.66	81.84	66.82	56.48	48.91	$\cdots$
\$1	164	115.97	66.95	47.34	36.67	29.94	25.31	21.92	$\cdots$

Array 0: Array of Priority Values for Algorithms *V-a* and *V-b*.

Proceeding through Array 0 according to Algorithm V-*a*, the amount of the budget being spent accumulates as shown in parentheses in Array 1 as we sequentially pick the largest priority values (first 332.34; next 258.80; then 191.98;...; finally 105.66), resulting in final allocation $(n_1, n_2, n_3) = (4, 4, 2)$. Note that we use only $C_a = \$54$ of the fixed budget $C = \$55$. Along the way applying Algorithm *V-a*, we note the $Var(\hat{T}_Y)$ for the allocation in brackets [ ] for the allocation at that point. The double brackets [[ ]] gives the weighted sum of stratum sampling variances (11) for the allocation at that point. Note further that the final allocation gives a sampling variance of $Var(\hat{T}_Y) = 94,610$. Specifically, we describe the first set of entries in Array 1 as follows:

- 332.34 = priority value obtained by $470 \times (\frac{1}{\sqrt{1 \cdot 2}})$.
- $(2, 1, 1) = (n_1, n_2, n_3)$ gives the allocation for $n = 4$ after the first largest priority value 332.34 (stratum 1) is picked.
- \$23 = the cost of the sample allocation (2,1,1).
- $Var(\hat{T}_Y) = \sum_{h=1}^{3} Var(N_h \bar{y}_h) = 263,750$ for the sample allocation (2,1,1).
- $\sum_{h=1}^{3} Var(N_h \bar{y}_h) \frac{1}{c_h} = 70,930$ for the sample allocation (2,1,1).

Note: All subsequent arrays are defined similarly noting that priority values for Algorithms *V-a* and *V-b* differ from priority values for Algorithms *V-c* and *V-d*.

c_h	$N_h S_h$	$\frac{1}{\sqrt{1 \cdot 2}}$	$\frac{1}{\sqrt{2 \cdot 3}}$	$\frac{1}{\sqrt{3 \cdot 4}}$	$\frac{1}{\sqrt{4 \cdot 5}}$	$\frac{1}{\sqrt{5 \cdot 6}}$	$\frac{1}{\sqrt{6 \cdot 7}}$	$\frac{1}{\sqrt{7 \cdot 8}}$	n_h
\$9	470	332.34 $(2,1,1)$ (\$23) $[263,750]$ $[[70,930]]$	191.98 $(3,2,1)$ (\$36) $[159,955]$ $[[50,095]]$	135.68 $(4,3,1)$ (\$49) $[119,221]$ $[[42,468]]$	105.10	85.81	72.52	62.81	4
\$4	366	258.80 $(2,2,1)$ (\$27) $[196,772]$ $[[54,186]]$	149.42 $(3,3,1)$ (\$40) $[137,629]$ $[[44,513]]$	105.66 $(4,4,2)$ (\$54) $[94,610]$ $[[26,229]]$	81.84	66.82	56.48	48.91	4
\$1	164	115.97 $(4,3,2)$ (\$50) $[105,773]$ $[[29,020]]$	66.95	47.34	36.67	29.94	25.31	21.92	2
									$n = 10$

Array 1: Application of Algorithm V-*a*; Final $(n_1, n_2, n_3) = (4, 4, 2)$ costs $C_a = \$54$.

Noting that we have an unspent $C - C_a = \$1$ in Array 1 with the final allocation $(n_1, n_2, n_3) = (4, 4, 2)$, below in Array 2, we apply Algorithm V-*b* which leads to the allocation $(n_1, n_2, n_3) = (4, 4, 3)$; this allocation makes use of the full fixed budget $C_b = C = \$55$; and the sampling variance is further decreased now to $Var(\hat{T_Y}) = 90,127$.

c_h	N_hS_h	$\frac{1}{\sqrt{1\cdot2}}$	$\frac{1}{\sqrt{2\cdot3}}$	$\frac{1}{\sqrt{3\cdot4}}$	$\frac{1}{\sqrt{4\cdot5}}$	$\frac{1}{\sqrt{5\cdot6}}$	$\frac{1}{\sqrt{6\cdot7}}$	$\frac{1}{\sqrt{7\cdot8}}$	n_h
\$9	470	332.34 $(2, 1, 1)$ (\$23) $[263, 750]$ $[[70, 930]]$	191.98 $(3, 2, 1)$ (\$36) $[159, 955]$ $[[50, 095]]$	135.68 $(4, 3, 1)$ (\$49) $[119, 221]$ $[[42, 468]]$	–	–	–	–	4
\$4	366	258.80 $(2, 2, 1)$ (\$27) $[196, 772]$ $[[54, 186]]$	149.42 $(3, 3, 1)$ (\$40) $[137, 629]$ $[[44, 513]]$	105.66 $(4, 4, 2)$ (\$54) $[94, 610]$ $[[26, 229]]$	–	–	–	–	4
\$1	164	115.97 $(4, 3, 2)$ (\$50) $[105, 773]$ $[[29, 020]]$	66.95 $(4, 4, 3)$ (\$55) $[90, 127]$ $[[21, 746]]$	–	–	–	–	–	3
									$n = 11$

Array 2: Application of Algorithm V-b; Final $(n_1, n_2, n_3) = (4, 4, 3)$ costs $C_b = \$55$.

Next, we consider Algorithm V-c with the following parameters (we renumber the strata so that $\frac{N_1S_1}{\sqrt{c_1}} \geq \frac{N_2S_2}{\sqrt{c_2}} \geq \frac{N_3S_3}{\sqrt{c_3}}$ for convenience):

h	N_h	S_h	c_h	$\frac{N_hS_h}{\sqrt{c_h}}$
1	61	6	\$4	183
2	41	4	\$1	164
3	47	10	\$9	$\frac{470}{3}$

We begin with Array 00:

c_h	$\frac{N_hS_h}{\sqrt{c_h}}$	$\frac{1}{\sqrt{1\cdot2}}$	$\frac{1}{\sqrt{2\cdot3}}$	$\frac{1}{\sqrt{3\cdot4}}$	$\frac{1}{\sqrt{4\cdot5}}$	$\frac{1}{\sqrt{5\cdot6}}$	$\frac{1}{\sqrt{6\cdot7}}$	$\frac{1}{\sqrt{7\cdot8}}$	$\cdots$
\$4	183	129.40	74.71	52.83	40.92	33.41	28.24	24.45	$\cdots$
\$1	164	115.97	66.95	47.34	36.67	29.94	25.31	21.92	$\cdots$
\$9	$\frac{470}{3}$	110.78	63.96	45.23	35.03	28.60	24.17	20.94	$\cdots$

Array 00: Array of Priority Values for Algorithms *V-c* and *V-d*.

From application of Algorithm V-c in Array 3, we have the final allocation $(n_1, n_4, n_3) = (4, 4, 3)$. Here, we use only $C_c = \$47$ of the fixed budget $C = \$55$; and this results in a sampling variance $Var(\hat{T}_Y) = 106,294$.

c_h	$\frac{N_h S_h}{\sqrt{c_h}}$	$\frac{1}{\sqrt{1\cdot 2}}$	$\frac{1}{\sqrt{2\cdot 3}}$	$\frac{1}{\sqrt{3\cdot 4}}$	$\frac{1}{\sqrt{4\cdot 5}}$	$\frac{1}{\sqrt{5\cdot 6}}$	$\frac{1}{\sqrt{6\cdot 7}}$	$\frac{1}{\sqrt{7\cdot 8}}$	n_h
$\$4$	183	129.40 $(2,1,1)$ $(\$18)$ $[307,222]$ $[[66,458]]$	74.71 $(3,2,2)$ $(\$32)$ $[160,998]$ $[[35,156]]$	52.83 $(4,3,3)$ $(\$46)$ $[108,535]$ $[[23,792]]$	40.92	33.41	28.24	24.45	4
$\$1$	164	115.97 $(2,2,1)$ $(\$19)$ $[293,774]$ $[[53,009]]$	66.95 $(3,3,2)$ $(\$33)$ $[156,515]$ $[[30,673]]$	47.34 $(4,4,3)$ $(\$47)$ $[106,294]$ $[[21,551]]$	36.67	29.94	25.31	21.92	4
$\$9$	$\frac{470}{3}$	110.78 $(2,2,2)$ $(\$28)$ $[183,324]$ $[[40,738]]$	63.96 $(3,3,3)$ $(\$42)$ $[119,698]$ $[[26,583]]$	45.23	35.03	28.60	24.17	20.94	3
									$n = 11$

Array 3: Application of Algorithm V-c; Final $(n_1, n_2, n_3) = (4, 4, 3)$ costs $C_c = \$47$.

Noting an unspent $C - C_c = \$8$ in Array 3 with final allocation $(n_1, n_2, n_3) = (4, 4, 3)$, apply Algorithm V-$d$ as given in Array 4 which leads to the allocation $(n_1, n_2, n_3) = (5, 8, 3)$; this allocation makes use of the full fixed budget $C_d = C = \$55$; from Array 3 to Array 4, the sampling variance is further decreased now to $Var(\hat{T}_Y) = 96,235$. (Note the renumbering of strata.)

c_h	$\frac{N_hS_h}{\sqrt{c_h}}$	$\frac{1}{\sqrt{1\cdot 2}}$	$\frac{1}{\sqrt{2\cdot 3}}$	$\frac{1}{\sqrt{3\cdot 4}}$	$\frac{1}{\sqrt{4\cdot 5}}$	$\frac{1}{\sqrt{5\cdot 6}}$	$\frac{1}{\sqrt{6\cdot 7}}$	$\frac{1}{\sqrt{7\cdot 8}}$	$\frac{1}{\sqrt{8\cdot 9}}$	n_h
\$4	183	129.40 $(2,1,1)$ (\$18) $[307,222]$ $[[66,458]]$	74.71 $(3,2,2)$ (\$32) $[160,998]$ $[[35,156]]$	52.83 $(4,3,3)$ (\$46) $[108,535]$ $[[23,792]]$	40.92 $(5,4,3)$ (\$51) $[99,596]$ $[[19,876]]$	–	–	–	–	5
\$1	164	115.97 $(2,2,1)$ (\$19) $[293,774]$ $[[53,009]]$	66.95 $(3,3,2)$ (\$33) $[156,515]$ $[[30,673]]$	47.34 $(4,4,3)$ (\$47) $[106,294]$ $[[21,551]]$	36.67 $(5,5,3)$ (\$52) $[98,251]$ $[[18,531]]$	29.94 $(5,6,3)$ (\$53) $[97,355]$ $[[17,635]]$	25.31 $(5,7,3)$ (\$54) $[96,715]$ $[[16,994]]$	21.92 $(5,8,3)$ (\$55) $[96,235]$ $[[16,514]]$	–	8
\$9	$\frac{470}{3}$	110.78 $(2,2,2)$ (\$28) $[183,324]$ $[[40,738]]$	63.96 $(3,3,3)$ (\$42) $[119,698]$ $[[26,583]]$	–	–	–	–	–	–	3
										$n = 16$

Array 4: Application of Algorithm V-d; Final $(n_1, n_2, n_3) = (5, 8, 3)$ costs $C_d = \$55$.

Example 2: With fixed budget $C = \$61$, assume a finite population of $N = 149$ units stratified into three strata with the following parameters:

h	N_h	S_h	c_h	N_hS_h
1	47	10	\$4	470
2	61	6	\$2	366
3	41	4	\$9	164

Applying Algorithms V-a and V-b as in Example 1, we obtain the corresponding Array 1^* and Array 2^*.

According to Algorithm V-a, the amount of the budget being spent accumulates as shown in parentheses in Array 1^* as we sequentially pick the largest priority values, resulting in final allocation $(n_1, n_2, n_3) = (7, 5, 2)$. Note that we use only $C_a = \$56$ of the fixed budget $C = \$61$. Along the way, we note the $Var(\hat{T}_Y)$ for the allocation in brackets [ ] for the allocation at that point; the weighted sum of stratum sampling variances (11) is given by double brackets [[ ]]. Note further that the final allocation gives a sampling variance of $Var(\hat{T}_Y) = 64,244$.

c_h	N_hS_h	$\frac{1}{\sqrt{1\cdot 2}}$	$\frac{1}{\sqrt{2\cdot 3}}$	$\frac{1}{\sqrt{3\cdot 4}}$	$\frac{1}{\sqrt{4\cdot 5}}$	$\frac{1}{\sqrt{5\cdot 6}}$	$\frac{1}{\sqrt{6\cdot 7}}$	$\frac{1}{\sqrt{7\cdot 8}}$	n_h
\$4	470	332.34 $(2,1,1)$ (\$19) $[263,750]$ $[[95,233]]$	191.98 $(3,2,1)$ (\$25) $[159,955]$ $[[52,540]]$	135.68 $(4,3,1)$ (\$31) $[119,221]$ $[[36,775]]$	105.10 $(5,4,2)$ (\$46) $[83,565]$ $[[26,938]]$	85.81 $(6,4,2)$ (\$50) $[76,202]$ $[[25,097]]$	72.52 $(7,5,2)$ (\$56) $[64,244]$ $[[20,433]]$	62.81	7
\$2	366	258.80 $(2,2,1)$ (\$21) $[196,772]$ $[[61,744]]$	149.42 $(3,3,1)$ (\$27) $[137,629]$ $[[41,377]]$	105.66 $(4,4,2)$ (\$42) $[94,610]$ $[[29,699]]$	81.84 $(6,5,2)$ (\$52) $[69,504]$ $[[21,748]]$	66.82	56.48	48.91	5
\$9	164	115.97 $(4,3,2)$ (\$40) $[105,773]$ $[[32,281]]$	66.95	47.34	36.67	29.94	25.31	21.92	2
									$n = 14$

Array 1*: Application of Algorithm V-a; Final $(n_1, n_2, n_3) = (7, 5, 2)$ costs $C_b = \$56$.

Noting that we have an unspent $C - C_a = \$5$ in Array 1* with final allocation $(n_1, n_2, n_3) = (7, 5, 2)$, below in Array 2*, we apply Algorithm V-b which leads to the allocation $(n_1, n_2, n_3) = (7, 7, 2)$; this allocation makes use of $C_b = \$60$ of the full fixed budget $C = \$61$; the sampling variance is further decreased now to $Var(\hat{T_Y}) = 56,590$.

c_h	N_hS_h	$\frac{1}{\sqrt{1\cdot 2}}$	$\frac{1}{\sqrt{2\cdot 3}}$	$\frac{1}{\sqrt{3\cdot 4}}$	$\frac{1}{\sqrt{4\cdot 5}}$	$\frac{1}{\sqrt{5\cdot 6}}$	$\frac{1}{\sqrt{6\cdot 7}}$	$\frac{1}{\sqrt{7\cdot 8}}$	n_h
\$4	470	332.34 $(2,1,1)$ (\$19) $[263,750]$ $[[95,233]]$	191.98 $(3,2,1)$ (\$25) $[159,955]$ $[[52,540]]$	135.68 $(4,3,1)$ (\$31) $[119,221]$ $[[36,775]]$	105.10 $(5,4,2)$ (\$46) $[83,565]$ $[[26,938]]$	85.81 $(6,4,2)$ (\$50) $[76,202]$ $[[25,097]]$	72.52 $(7,5,2)$ (\$56) $[64,244]$ $[[20,433]]$	–	7
\$2	366	258.80 $(2,2,1)$ (\$21) $[196,772]$ $[[61,744]]$	149.42 $(3,3,1)$ (\$27) $[137,629]$ $[[41,377]]$	105.66 $(4,4,2)$ (\$42) $[94,610]$ $[[29,699]]$	81.84 $(6,5,2)$ (\$52) $[69,504]$ $[[21,748]]$	66.82 $(7,6,2)$ (\$58) $[59,779]$ $[[18,201]]$	56.48 $(7,7,2)$ (\$60) $[56,590]$ $[[16,606]]$	–	7
\$9	164	115.97 $(4,3,2)$ (\$40) $[105,773]$ $[[32,281]]$	–	–	–	–	–	–	2
									$n = 16$

Array 2*: Application of Algorithm V-b; Final $(n_1, n_2, n_3) = (7, 7, 2)$ costs $C_b = \$60$.

Next, as in Example 1, we consider Algorithm V-c with the following parameters (we renumber the strata so that $\frac{N_1 S_1}{\sqrt{c_1}} \geq \frac{N_2 S_2}{\sqrt{c_2}} \geq \frac{N_3 S_3}{\sqrt{c_3}}$ for convenience):

h	N_h	S_h	c_h	$\frac{N_h S_h}{\sqrt{c_h}}$
1	61	6	$\$2$	$\frac{366}{\sqrt{2}}$
2	47	10	$\$4$	235
3	41	4	$\$9$	$\frac{164}{3}$

We have the following Array 3* and Array 4* applying Algorithms V-c and V-d:

	$\frac{N_h S_h}{\sqrt{c_h}}$	$\frac{1}{\sqrt{1 \cdot 2}}$	$\frac{1}{\sqrt{2 \cdot 3}}$	$\frac{1}{\sqrt{3 \cdot 4}}$	$\frac{1}{\sqrt{4 \cdot 5}}$	$\frac{1}{\sqrt{5 \cdot 6}}$	$\frac{1}{\sqrt{6 \cdot 7}}$	$\frac{1}{\sqrt{7 \cdot 8}}$	n_h
$\$2$	$\frac{366}{\sqrt{2}}$	183.00 $(2,1,1)$ ($\$17$) $[307,222]$ $[[89,357]]$	105.66 $(3,2,1)$ ($\$23$) $[174,446]$ $[[50,581]]$	74.71 $(4,3,1)$ ($\$29$) $[126,466]$ $[[35,795]]$	57.87 $(5,4,1)$ ($\$35$) $[101,360]$ $[[27,844]]$	47.25 $(6,5,1)$ ($\$41$) $[85,850]$ $[[22,851]]$	39.93 $(7,6,1)$ ($\$47$) $[75,297]$ $[[19,415]]$	34.58	7
$\$4$	235	166.17 $(2,2,1)$ ($\$21$) $[196,772]$ $[[61,744]]$	95.94 $(3,3,1)$ ($\$27$) $[137,629]$ $[[41,377]]$	67.84 $(4,4,1)$ ($\$33$) $[108,058]$ $[[31,193]]$	52.55 $(5,5,1)$ ($\$39$) $[90,315]$ $[[25,083]]$	42.90 $(6,6,1)$ ($\$45$) $[78,486]$ $[[21,010]]$	36.26 $(7,7,2)$ ($\$60$) $[56,590]$ $[[16,606]]$	31.40	7
$\$9$	$\frac{164}{3}$	38.66 $(7,6,2)$ ($\$56$) $[61,849]$ $[[17,921]]$	22.32	15.78	12.22	9.98	8.44	7.31	2
									$n = 16$

Array 3*: Application of Algorithm V-c; Final $(n_1, n_2, n_3) = (7,7,2)$ costs $C_c = \$60$.

c_h	$\frac{N_hS_h}{\sqrt{c_h}}$	$\frac{1}{\sqrt{1\cdot 2}}$	$\frac{1}{\sqrt{2\cdot 3}}$	$\frac{1}{\sqrt{3\cdot 4}}$	$\frac{1}{\sqrt{4\cdot 5}}$	$\frac{1}{\sqrt{5\cdot 6}}$	$\frac{1}{\sqrt{6\cdot 7}}$	$\frac{1}{\sqrt{7\cdot 8}}$	n_h
\$2	$\frac{366}{\sqrt{2}}$	183.00 $(2,1,1)$ (\$17) $[307,222]$ $[[89,357]]$	105.66 $(3,2,1)$ (\$23) $[174,446]$ $[[50,581]]$	74.71 $(4,3,1)$ (\$29) $[126,466]$ $[[35,795]]$	57.87 $(5,4,1)$ (\$35) $[101,360]$ $[[27,844]]$	47.25 $(6,5,1)$ (\$41) $[85,850]$ $[[22,851]]$	39.93 $(7,6,1)$ (\$47) $[75,297]$ $[[19,415]]$	–	7
\$4	235	166.17 $(2,2,1)$ (\$21) $[196,772]$ $[[61,744]]$	95.94 $(3,3,1)$ (\$27) $[137,629]$ $[[41,377]]$	67.84 $(4,4,1)$ (\$33) $[108,058]$ $[[31,193]]$	52.55 $(5,5,1)$ (\$39) $[90,315]$ $[[25,083]]$	42.90 $(6,6,1)$ (\$45) $[78,486]$ $[[21,010]]$	36.26 $(7,7,2)$ (\$60) $[56,590]$ $[[16,606]]$	–	7
\$9	$\frac{164}{3}$	38.66 $(7,6,2)$ (\$56) $[61,849]$ $[[17,921]]$	–	–	–	–	–	–	2
									$n = 16$

Array 4*: Application of Algorithm V-*d*; Final $(n_1, n_2, n_3) = (7, 7, 2)$ costs $C_d = \$60$.

Note from Array 3* that $Var(\hat{T}_Y) = 56,590$ when we stop with final allocation $(n_1, n_2, n_3) = (7, 7, 2)$; also from Array 4*, that $Var(\hat{T}_Y) = 56,590$ when we stop with the same final allocation.

Remark 8: As in (5), when the sample size for the h^{th} stratum is *"increased"* from $m_h - 1$ to m_h in Algorithms *V-c* and *V-d*, the associated $Var(N_h\bar{y}_h)\frac{1}{c_h}$ for the h^{th} stratum *"decreases"* by

$$\frac{Var(N_h\bar{y}_{n_h-1})}{c_h} - \frac{Var(N_h\bar{y}_{n_h})}{c_h} = \frac{(N_h^2S_h^2/c_h)}{(m_h-1)(m_h)} = \left(\frac{(N_hS_h/\sqrt{c_h})}{\sqrt{(m_h-1)(m_h)}}\right)^2. \qquad (14)$$

The result in (14) is also the amount by which the overall expression in (11) *"decreases"* when the sample size for the h^{th} stratum is increased from $m_h - 1$ to m_h.

Remark 9: The quantity in (14) is the reduction in $Var(\hat{T}_Y)$ per unit cost by picking a unit for the sample from stratum h bringing that stratum's sample size to m_h. At each step in Algorithm *V-c*, we pick a unit from the stratum which reduces $Var(\hat{T}_Y)$ by the largest amount per unit cost at that point in the sequence, which is very reasonable and desirable.

Remarks 8 and 9 explain why Algorithm *V-c* results in an allocation of the realized n that minimizes (11) the *weighted sum of stratum sampling variances* subject to (10).

Algorithm *V-d* reduces (11) further than Algorithm *V-c* does while still subject to (10).

Remark 10: By design, the allocation $(n_1, ..., n_H)$ of $n = \sum_{h=1}^{H} n_h$ that is observed and associated with $C_c = \sum_{h=1}^{H} c_h n_h$ minimizes (11) the *weighted sum of stratum sampling variances* subject to $C_c \leq C$. The following theorem which is inspired by Kadane (2005) shows that the allocation $(n_1, ..., n_H)$ of $n = \sum_{h=1}^{H} n_h$ that is observed and associated with $C_c = \sum_{h=1}^{H} c_h n_c$ minimizes $Var(\hat{T}_Y)$ the *unweighted sum of stratum sampling variances* over all other allocations that cost less than or equal to C_c.

Before stating and proving the Theorem, note that by picking the largest priority values $\frac{N_h S_h/\sqrt{c_h}}{\sqrt{(m_h-1)m_h}}$ in Algorithm *V-c* we are equivalently picking the largest squared priority values $\frac{N_h^2 S_h^2/c_h}{(m_h-1)m_h}$. Because these largest priority values (also largest squared priority values) are all positive, there is a positive real number k such that

$$\frac{N_h^2 S_h^2/c_h}{(m_h-1)m_h} \geq k$$

for all priority values that are picked to obtained $C_c = \sum_{h=1}^{H} c_h n_h$.

Theorem: For fixed budget C, let E_1 be the subset of largest priority values in the array of Algorithm *V-c* such that for some $k \in R^+$

(i) $\quad \frac{N_h^2 S_h^2/c_h}{(m_h-1)m_h} \geq k$ $\quad$ for each priority value in E_1;

(ii) $\quad \frac{N_h^2 S_h^2/c_h}{(m_h-1)m_h} < k$ $\quad$ for each priority value in E_1^c, the complement of E_1; and

(iii) $\quad \sum_{E_1} c_h + \sum_{h=1}^{H} c_h = C_c$ $\quad$ where C_c is as defined in Algorithm *V-c*.

Let E_2 be another subset of priority values in the array of Algorithm *V-c* such that $C_{E_2} \leq C_c$, where C_{E_2} is the cost of a sample corresponding to E_2. Keep in mind that C_{E_2} includes the total cost of one unit from each stratum as well as for the additional units included as a result of picking the largest priority values from the array. Then

$$\sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h} \geq \sum_{E_2} \frac{N_h^2 S_h^2}{(m_h-1)m_h}. \tag{15}$$

(*Comment:* From the decomposition of $Var(\hat{T}_h)$ in Section 5 or $(**)$,

$$Var(\hat{T}_Y) = \sum_{h=1}^{H} Var(\hat{T}_h) = \sum_{h=1}^{H} N_h(N_h-1)S_h^2 - \sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h}$$

where E_1 is the set of priority values (and associated allocation $(n_1, ..., n_H)$) leading to $C_c = \sum_{h=1}^{H} c_h n_h$. Thus by showing that

$$\sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h}$$

is maximized in (15), we have that Algorithm *V-c* which stops at C_c with allocation $(n_1, ..., n_H)$ minimizes $Var(\hat{T}_Y)$, the unweighted sum of stratum sampling variances, for any allocation whose associated cost does not exceed C_c.)

Proof: It is enough to show

$$\sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2} \frac{N_h^2 S_h^2}{(m_h-1)m_h} \geq 0.$$

Note that $E_1 = (E_1 \cap E_2) \cup (E_1 \cap E_2^c)$ and $E_2 = (E_2 \cap E_1) \cup (E_2 \cap E_1^c)$. Now

$$\sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2} \frac{N_h^2 S_h^2}{(m_h-1)m_h}$$

$$= \sum_{E_1 \cap E_2} \frac{N_h^2 S_h^2}{(m_h-1)m_h} + \sum_{E_1 \cap E_2^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2 \cap E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2 \cap E_1^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h}$$

$$= \sum_{E_1 \cap E_2^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2 \cap E_1^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} \tag{16}$$

By assumption (i),

$$\frac{N_h^2 S_h^2 / c_h}{(m_h-1)m_h} \geq k \tag{17}$$

for all priority values in E_1 and hence for all priority values in $E_1 \cap E_2^c$. Thus

$$\frac{N_h^2 S_h^2}{(m_h-1)m_h} \geq k c_h \tag{18}$$

for all priority values in $E_1 \cap E_2^c$, and we have

$$\sum_{E_1 \cap E_2^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} \geq k \sum_{E_1 \cap E_2^c} c_h. \tag{19}$$

By assumption (ii), for all priority values in E_1^c and hence for all priority values in $E_2 \cap E_1^c$

$$\frac{N_h^2 S_h^2/c_h}{(m_h-1)m_h} < k$$

Thus

$$\frac{N_h^2 S_h^2}{(m_h-1)m_h} < kc_h$$

for all priority values in $E_2 \cap E_1^c$, and we have

$$\sum_{E_2\cap E_1^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} < k \sum_{E_2\cap E_1^c} c_h \tag{20}$$

So from (16), (19), and (20), we have

$$\begin{aligned}
\sum_{E_1} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2} \frac{N_h^2 S_h^2}{(m_h-1)m_h} &= \sum_{E_1\cap E_2^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} - \sum_{E_2\cap E_1^c} \frac{N_h^2 S_h^2}{(m_h-1)m_h} \\
&\geq k(\sum_{E_1\cap E_2^c} c_h - \sum_{E_2\cap E_1^c} c_h) \\
&= k(\sum_{E_1\cap E_2^c} c_h + \sum_{E_1\cap E_2} c_h - \sum_{E_2\cap E_1} c_h - \sum_{E_2\cap E_1^c} c_h) \\
&= k(\sum_{E_1} c_h - \sum_{E_2} c_h) \\
&= k([\sum_{E_1} c_h + \sum_{h=1}^{H} c_h] - [\sum_{E_2} c_h + \sum_{h=1}^{H} c_h]) \\
&= k(C_c - C_{E_2}) \geq 0
\end{aligned}$$

Thus the theorem has been shown.

5. CONCLUSION

Perhaps the most important advance in probability sampling theory is Neyman's 1934 paper in which he provides arguably the most widely used and known concept of stratification and optimal allocation of the sample. The exact result in Wright (2012) improves upon the method by Neyman and guarantees integers for all stratum optimal sample sizes, as desired, while yielding minimum sampling variance.

In this paper, we consider a solution with a precision constraint (Algorithm IV), and we consider various options with a cost constraint (Algorithms *V-a*, *V-b*, *V-c*, *V-d*). The results of this paper are simple, clear, exact, optimal, and practical.

The key to all five algorithms is revealed by considering a very simple decomposition of $Var(\hat{T}_h)$ for stratum h. For a simple random sample of size n_h within stratum h which has N_h units, note the following decomposition of $Var(\hat{T}_h)$

$$Var(\hat{T}_h) = N_h^2(\frac{N_h - n_h}{N_h})\frac{S_h^2}{n_h} = N_h(N_h-1)S_h^2 - \frac{N_h^2 S_h^2}{1\cdot 2} - \frac{N_h^2 S_h^2}{2\cdot 3} - \cdots - \frac{N_h^2 S_h^2}{(n_h-1)(n_h)}.$$

Specifically, for all possible values of n_h, we see the following cumulative decreases:

for $n_h = 1$, $Var(\hat{T}_h) = N_h(N_h-1)S_h^2$;

for $n_h = 2$, $Var(\hat{T}_h) = N_h(N_h-1)S_h^2 - \frac{N_h^2 S_h^2}{1\cdot 2}$;

for $n_h = 3$, $Var(\hat{T}_h) = N_h(N_h-1)S_h^2 - \frac{N_h^2 S_h^2}{1\cdot 2} - \frac{N_h^2 S_h^2}{2\cdot 3}$;

for $n_h = 4$, $Var(\hat{T}_h) = N_h(N_h-1)S_h^2 - \frac{N_h^2 S_h^2}{1\cdot 2} - \frac{N_h^2 S_h^2}{2\cdot 3} - \frac{N_h^2 S_h^2}{3\cdot 4}$;

$\vdots$

for $n_h = N_h$, $Var(\hat{T}_h) = 0$.

ACKNOWLEDGMENTS

The author is grateful to a reviewer who suggested the seeds for Algorithm IV, to Pat Hunley and Andrew Perry (two interns) who helped in many discussions leading to the variations of Algorithm V, to his colleagues Deanna Weidenhamer, Gauri Datta, Josh Tokle, and Robert Ashmead whose reviews greatly improved clarity and to his colleague Michael Leibert who helped with the computations and read many drafts.

REFERENCES

Kadane, J. (2005). "Optimal Dynamic Sample Allocation Among Strata", *Journal of Official Statistics, Vol 21, No. 4*, 531-541.

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, Vol. 97*, 558-606.

Wright, T. (2012). "The Equivalence of Neyman Optimum Allocation for Sampling & Equal Proportions for Apportioning the U.S. House of Representatives", *The American Statistician, Vol. 66, No.4*, 217-224.

Wright, T. (2014). "A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Pattern", *Research Report Series (Statistics #2014-07)*, Center for Statistical Research and Methodology, U. S. Bureau of the Census, Washington, D.C.