RESEARCH REPORT SERIES
(Statistics #2014-07)

A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns

Tommy Wright

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

Report Issued: August 21, 2014

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.

A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns

Tommy Wright
U. S. Bureau of the Census

Abstract

While making a surprising observation linking Neyman (1934) sample allocation in probability sampling and the current method used to allocate the seats in the U. S. House of Representatives called equal proportions, Wright (2012) provides an exact optimal allocation $[n_1, ..., n_h, ..., n_H]$ of the fixed overall sample size n among H strata under stratified random sampling that minimizes the sampling variance $Var(\hat{T}_Y)$ of an estimator of a total $\hat{T}_Y$ subject to the constraint $n = \sum_{h=1}^{H} n_h$. The exact optimal allocation avoids the need to round to integer values, as is the case with Neyman allocation. Neyman allocation with rounded integers does not always lead to the optimal allocation.

In this paper, we demonstrate a very easy extension and generalization of the result in Wright (2012) to the problem of finding an exact optimal allocation $[n_1, ..., n_h, ..., n_H]$ to minimize the sampling variance subject to $n = \sum_{h=1}^{H} n_h$ and additional mixed constraint patterns $0 < a_h \leq n_h \leq b_h \leq N_h$, where n, N_h, a_h, and b_h are fixed integers and N_h is the size of the h^{th} stratum. Avoiding the costly tendency to round up to ensure minimum sampling variance, the exact optimal allocation is especially useful in applications where H is very large and there are minimum and maximum size constraints on the allocated sample sizes n_h, as is the case with the Census Bureau's Service Annual Survey which has $H = 391$ sampling strata.

The presented methods are what some might call "greedy" algorithms, and the methods solve a nonlinear optimization problem with linear constraints over a space of integer values. While it's common that greedy algorithms are easy to compute, they are not guaranteed to find the global optimum. Remarkably, the presented simple algorithms *always* find the global optimum.

KEY WORDS: Exact optimal allocation; Mixed constraint patterns; Neyman allocation; Stratification.

1. INTRODUCTION

In perhaps the most significant advance in probability sampling theory, Neyman (1934) discusses the desire to have a "representative method" when sampling from a finite population; specifically, he considers the "...two different aspects of the representative method. One of them is called the method of random sampling and the other the method of purposive selection." Neyman argues in favor of random sampling, or more specifically, stratified random sampling. In such cases, a random sample is selected independently from each subpopulation called a stratum.

There are several reasons why one might want to stratify before sample selection (Cochran, 1977; Fuller, 2009; Lohr, 2010): (1) estimates of stated precision are desired for each stratum as well as for the overall population; (2) sampling and data collection objectives, operations, costs, and challenges may differ greatly in different parts of the population (e.g., city vs farm areas) or for different data collection modes (e.g., mail, telephone, Internet, face-to-face, administrative records); and (3) stratification may produce a gain in precision in the estimates of characteristics of the entire population, especially when it is possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous.

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.

Under stratified random sampling with a fixed overall sample size n, one might desire to allocate n proportionally among the strata according to stratum sizes, i.e., larger strata should have more sample units than smaller strata. Proportional allocation ignores the variability within each stratum. While considering stratum sizes as well as the variability within each stratum, Neyman's approach allocates n with a goal of minimizing the sampling error of the estimate of a total of the overall population. However, his result faces the issue of noninteger solutions, and we have the issue of what to do with the fractional parts. Integer programming could be used as Fuller (2009) notes. Concern over fractional parts has received limited attention, and it is a focus of this paper.

In the allocation of the overall sample to the various strata, one may frequently need to round to integer values. The issue is often handled by controlled rounding. This is done by sorting fractional parts (non-integer remainders) from the largest to smallest and assigning the desired number of additional units to the strata with the largest fractional parts. We will illustrate controlled rounding in Section 2.3.

In this paper, we consider the problem of exact optimal allocation of an overall sample size n in stratified random sampling. Specifically, we demonstrate: (1) that Algorithm I gives an exact optimal allocation where $n_h \geq 1$ for all h; (2) that Algorithm II gives an exact optimal allocation where $n_h \geq 2$ for all h; (3) that controlled rounding with Neyman allocation does not always lead to the optimum allocation; and (4) that Algorithm III gives an exact optimal allocation where $0 < a_h \leq n_h \leq b_h \leq N_h$ for all h where a_h and b_h are stated positive integers.

2. SAMPLE ALLOCATION

2.1 Neyman 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. These subpopulations are disjoint and together their union gives the entire population. 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.

The general setup where Y_{hj} is the value of interest for the j^{th} unit in the h^{th} stratum ($j = 1, ..., N_h$ and $h = 1, ..., H$) is

Stratum 1	Stratum 2	$\cdots$	Stratum h	$\cdots$	Stratum H
N_1	N_2	$\cdots$	N_h	$\cdots$	N_H
$\bar{Y}_1$	$\bar{Y}_2$	$\cdots$	$\bar{Y}_h$	$\cdots$	$\bar{Y}_H$
S_1^2	S_2^2	$\cdots$	S_h^2	$\cdots$	S_H^2

where $\bar{Y}_h$ and S_h^2 are the mean and variance for the population values in the h^{th} stratum, respectively. Specifically

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

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

$$T_Y = \sum_{h=1}^{H} \sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H} N_h \bar{Y}_h. \tag{1}$$

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.

Each $\bar{y}_h$ can be considered a random variable. Indeed $\bar{y}_1, \bar{y}_2, ..., \bar{y}_H$ are independent random variables. Hence a natural estimator for T_Y is

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

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

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

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

$$n_h = \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} n \qquad h = 1, 2, 3, ..., H. \tag{4}$$

Neyman obtains this result (4) by noting that $Var(\hat{T}_Y)$ in (3) can be written as in (5).

$$Var(\hat{T}_Y) = \frac{N-n}{n} \sum_{h=1}^{H} N_h S_h^2 + \sum_{h=1}^{H} n_h (\frac{N_h S_h}{n_h} - \frac{\sum_{i=1}^{H} N_i S_i}{n})^2 - \frac{N}{n} \sum_{h=1}^{H} N_h (S_h - \frac{\sum_{i=1}^{H} N_i S_i}{N})^2 \tag{5}$$

From the middle sum of (5) and by noting that the first and third sums of (5) are fixed relative to n_h, the result (4) follows. It turns out that Tschuprow (1923) had obtained the result over a decade earlier, showing that the result follows as a special case of a more general problem.

Because the n_h determined as just noted in (4) are almost never positive integers, we might round up in practice with the resulting overall sample size being possibly near $n + H$ instead of the fixed n. If H is large or if resource limitations dictate that the stated n not be exceeded, this is a concern.

2.2 Exact Optimal Sample Allocation

Is it possible to obtain an exact allocation of fixed n that minimizes $Var(\hat{T}_Y)$ in which all n_h are positive integers and $n = \sum_{h=1}^{H} n_h$? In this section, we show that the answer is yes.

Noting that $Var(\hat{T}_Y)$ can be written as

$$Var(\hat{T}_Y) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2, \tag{6}$$

that the overall sample size n is fixed, and that $\sum_{h=1}^{H} N_h S_h^2$ is a constant relative to n_h , it is easy to see that finding n_h to minimize $Var(\hat{T}_Y)$ for fixed n is equivalent to finding n_h to minimize

$$\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} \tag{7}$$

for fixed $n = \sum_{h=1}^{H} n_h$.

From

$$\begin{aligned} 1 - \frac{1}{n_h} &= (1 - \frac{1}{2}) + (\frac{1}{2} - \frac{1}{3}) + \cdots + (\frac{1}{n_h - 1} - \frac{1}{n_h}) \\ &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(n_h - 1) \cdot (n_h)} \ , \end{aligned}$$

we see that

$$\frac{1}{n_h} = 1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - \frac{1}{3 \cdot 4} - \cdots - \frac{1}{(n_h - 1) \cdot (n_h)} \ . \tag{8}$$

By substituting (8) into (7), we proceed to determine $n_1, n_2, ..., n_H$ to minimize

$$\sum_{h=1}^{H} N_h^2 S_h^2 (\frac{1}{n_h}) \ = \ \sum_{h=1}^{H} N_h^2 S_h^2 - \sum_{h=1}^{H} (\frac{N_h^2 S_h^2}{1 \cdot 2} + \frac{N_h^2 S_h^2}{2 \cdot 3} + \cdots + \frac{N_h^2 S_h^2}{(n_h - 1)(n_h)}). \tag{9}$$

Because $\sum_{h=1}^{H} N_h^2 S_h^2$ in (9) is a constant for given values $N_1, ..., N_H, S_1^2, ..., S_H^2$, we proceed to determine $n_1, n_2, ..., n_H$ to maximize the sum of H sums

$$\sum_{h=1}^{H} (\frac{N_h^2 S_h^2}{1 \cdot 2} + \frac{N_h^2 S_h^2}{2 \cdot 3} + \cdots + \frac{N_h^2 S_h^2}{(n_h - 1)(n_h)}) =$$

$$(\frac{N_1^2 S_1^2}{1 \cdot 2} + \frac{N_1^2 S_1^2}{2 \cdot 3} + \cdots + \frac{N_1^2 S_1^2}{(n_1 - 1)(n_1)}) + \cdots + (\frac{N_H^2 S_H^2}{1 \cdot 2} + \frac{N_H^2 S_H^2}{2 \cdot 3} + \cdots + \frac{N_H^2 S_H^2}{(n_H - 1)(n_H)}) \tag{10}$$

subject to the constraint $\sum_{h=1}^{H} n_h = n$. By careful inspection, we see that (10) will be maximized if we pick the $n - H$ largest terms in the sum of H sums, subject to the constraint.

Perhaps, a clearer way to see how to determine $n_1, ..., n_H$ to maximize (10) is to imagine the $(H) \times (n - H)$ array where the h^{th} row consists of the terms of the h^{th} sum in (10):

$$\frac{N_h^2 S_h^2}{1 \cdot 2}, \quad \frac{N_h^2 S_h^2}{2 \cdot 3}, \quad \cdots, \quad \frac{N_h^2 S_h^2}{(n - H) \cdot (n - H + 1)}.$$

Note that the h^{th} sum has $n_h - 1$ strictly decreasing terms. So to maximize the sum of H sums in (10) subject to the constraint $\sum_{h=1}^{H} n_h = n$, we pick the $(n_1 - 1) + (n_2 - 1) + \cdots (n_h - 1) = n - H$ largest terms in the $(H) \times (n - H)$ array.

Because n is fixed and each stratum must have at least one sample unit, we see that (10) will be maximized, our constraint will be satisfied, and each stratum will have at least one sample unit if

(i) each stratum is assigned one unit in the sample, and

(ii) each stratum receives an additional sample unit each time it has a term in (10) that appears among the $n - H$ largest terms.

Because the quantities $N_h S_h$ and n_h are all positive, selecting the $n - H$ largest terms among

$$\frac{N_h^2 S_h^2}{1 \cdot 2}, \ \frac{N_h^2 S_h^2}{2 \cdot 3}, ..., \frac{N_h^2 S_h^2}{(n_h - 1)(n_h)} \tag{11}$$

for $h = 1, 2, ..., H$ is the same as selecting the $n - H$ largest terms among

$$\frac{N_h S_h}{\sqrt{1 \cdot 2}}, \ \frac{N_h S_h}{\sqrt{2 \cdot 3}}, ..., \frac{N_h S_h}{\sqrt{(n_h - 1)(n_h)}} \tag{12}$$

for $h = 1, 2, ..., H$. Because the values of the terms in (11) can be quite large, in practice, we prefer to use the equivalent values in (12).

Hence it should be clear that our solution for the sample allocation problem is as follows.

Exact Optimal Allocation Algorithm I (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, we assume that the $N_h S_h$ values are ordered so that $N_1 S_1 \geq N_2 S_2 \geq \cdots \geq N_H S_H$):

$$\begin{array}{cccc}
\frac{N_1 S_1}{\sqrt{1 \cdot 2}} & \frac{N_1 S_1}{\sqrt{2 \cdot 3}} & \frac{N_1 S_1}{\sqrt{3 \cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_h S_h}{\sqrt{1 \cdot 2}} & \frac{N_h S_h}{\sqrt{2 \cdot 3}} & \frac{N_h S_h}{\sqrt{3 \cdot 4}} & \cdots \\
 & \vdots & & \\
\frac{N_H S_H}{\sqrt{1 \cdot 2}} & \frac{N_H S_H}{\sqrt{2 \cdot 3}} & \frac{N_H S_H}{\sqrt{3 \cdot 4}} & \cdots
\end{array}$$

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.

Note that for (8) – (12) to be valid, we must have $n_h \geq 2$. Because each stratum gets at least one sample unit, we will only need to consider strata to see if they get additional sample units, and in these cases $n_h \geq 2$. For convenience, we could alternately define n'_h to be the additional sample units to be assigned to stratum h beyond the first one which it receives by *Step 1.* In this case, $n_h = n'_h + 1$ and the notation in (12) becomes

$$\frac{N_h S_h}{\sqrt{1 \cdot 2}}, \ \frac{N_h S_h}{\sqrt{2 \cdot 3}}, ..., \ \frac{N_h S_h}{\sqrt{n'_h \cdot (n'_h + 1)}} \tag{13}$$

In the notation of (13), the priority value $\frac{N_h S_h}{\sqrt{1 \cdot 2}}$ can be viewed as stratum h has one sample unit; and if this priority value is among the largest $n - H$ values, it would mean that stratum h gets a second sample unit. Similarly, the priority value $\frac{N_h S_h}{\sqrt{2 \cdot 3}}$ can be viewed as stratum h has two sample units; and if this priority value is among the largest $n - H$ values, it would mean that stratum h gets a third sample unit. All priority values can be viewed in a similar manner.

Note that the size of the array in Algorithm I is not greater than $(H) \times (n - H)$ because it is possible that all $n - H$ sample units in Step 3 of Algorithm I could come from the first stratum.

Unbiased estimation of $Var(\hat{T}_Y)$ requires the selection of at least two units from each stratum in addition to the requirement that $n = \sum_{h=1}^{H} n_h$. In such cases and considering the comments of the previous paragraph, the following modification minimizes $Var(\hat{T}_Y)$ subject to both requirements.

Exact Optimal Allocation Algorithm II (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 $N_1 S_1 \geq N_2 S_2 \geq \cdots \geq N_H S_H$):

$$\begin{array}{cccc}
\frac{N_1 S_1}{\sqrt{2 \cdot 3}} & \frac{N_1 S_1}{\sqrt{3 \cdot 4}} & \frac{N_1 S_1}{\sqrt{4 \cdot 5}} & \cdots \\
 & \vdots & & \\
\frac{N_h S_h}{\sqrt{2 \cdot 3}} & \frac{N_h S_h}{\sqrt{3 \cdot 4}} & \frac{N_h S_h}{\sqrt{4 \cdot 5}} & \cdots \\
 & \vdots & & \\
\frac{N_H S_H}{\sqrt{2 \cdot 3}} & \frac{N_H S_H}{\sqrt{3 \cdot 4}} & \frac{N_H S_H}{\sqrt{4 \cdot 5}} & \cdots
\end{array}$$

Note that the array of priority values in *Step* $2'$ is the same as the previous array in *Step 2* of Rule 1 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 at least two units from each stratum.

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.

If three units are required from each stratum, by careful consideration of the definition of n_h and n'_h in equations (8) – (13), it is clear that the optimal allocation of n would follow in a similar way by first assigning three units to be selected from each stratum and then considering an array that only has priority values with the following values in the denominator $\sqrt{3 \cdot 4}$, $\sqrt{4 \cdot 5}$, $\sqrt{5 \cdot 6}$,...

Generalizations beyond 3 sample units per stratum follow in a similar manner.

Note that the size of the array in Algorithm II is not greater than $(H) \times (n - 2H)$ because it is possible that all $n - 2H$ sample units in Step 3 of Algorithm II could come from the first stratum.

2.3 Example: Exact Optimal Sample Allocation

Assume a stratified population of $N = 149$ units distributed among $H = 3$ strata as noted below with the desire to select a stratified random sample to estimate the unknown value of the total T_Y.

Stratum 1	Stratum 2	Stratum 3
$N_1 = 47$	$N_2 = 61$	$N_3 = 41$
$S_1^2 = 100$	$S_2^2 = 36$	$S_3^2 = 16$

For $n = 10$, the optimum allocation can be found by applying Algorithm I as follows.

Step 1: First, assign one unit to be selected from each stratum.
Step 2: Compute the 3×7 array of *priority values* as noted below.

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
470	**332.34**	**191.88**	**135.68**	105.10	85.81	72.52	62.81	**4**
366	**258.80**	**149.42**	**105.66**	81.84	66.82	56.48	48.91	**4**
164	**115.97**	66.95	47.34	36.67	29.94	25.31	21.92	**2**

Step 3: The $n - 3 = 7$ largest priority values from the above array in *Step 2* are noted in bold. The optimum allocation that gives minimum variance is given in the last column of the above table.

Thus the exact optimal sample allocation is $n_1 = 4$, $n_2 = 4$, and $n_3 = 2$, and it can be shown that $Var_{ExOpt}(\hat{T}_Y) = 94,610$.

Applying (4), it is easy to show that the Neyman allocation yields $n_1 = 4.70$, $n_2 = 3.66$, and $n_3 = 1.64$. Applying controlled rounding as mentioned at the end of Section 1 of this paper, we see that strata 1, 2, and 3 get 4, 3, and 1 units allocated respectively by considering the integer parts of n_1, n_2, and n_3. But $4 + 3 + 1 = 8$ units, and 2 more units are needed to bring the overall sample size to $n = 10$. The largest fractional part is .70 which is associated with stratum 1. So this brings the sample size for stratum 1 to 5 ($= 4 + 1$), and the overall sample count to 9. The next largest fractional part is .66 which is associated with stratum 2. This brings the sample size for stratum 2 to 4. Because the overall sample size is $n = 10$, this leads to the allocation $n_1 = 5$, $n_2 = 4$, and $n_3 = 1$ with $Var_{Ney}(\hat{T}_Y) = 97,013$.

Because $Var_{ExOpt}(\hat{T}_Y) < Var_{Ney}(\hat{T}_Y)$ in this example, we see that controlled rounding with Neyman allocation does not always lead to optimum allocation. Also, by applying the Exact

Optimal Allocation Algorithm II, it is easy to show that the optimum allocation is $n_1 = 4$, $n_2 = 4$, and $n_3 = 2$, if the desire is to have $n_h \geq 2$ for $h = 1$, 2, and 3.

3. INTREPRETATION OF THE VALUES $\frac{N_h^2 S_h^2}{(m_h - 1)(m_h)}$ FOR $m_h = 2, ..., N_h$

In this section, we give an intrepretation of the values in (11), or equivalently the values in (12).

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. Then the contribution to the estimator $\hat{T}_Y$ from the h^{th} stratum is $N_h \bar{y}_{m_h}$ and

$$\begin{aligned} Var(N_h \bar{y}_{m_h}) &= N_h^2 \left(\frac{N_h - m_h}{N_h} \right) \frac{S_h^2}{m_h} \\ &= \frac{N_h^2 S_h^2}{m_h} - N_h S_h^2 \\ &= N_h^2 S_h^2 \left(1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - \cdots - \frac{1}{(m_h - 1)(m_h)} \right) - N_h S_h^2 \\ &= (N_h^2 S_h^2 - N_h S_h^2) - \left(\frac{N_h^2 S_h^2}{1 \cdot 2} + \frac{N_h^2 S_h^2}{2 \cdot 3} + \cdots + \frac{N_h^2 S_h^2}{(m_h - 2)(m_h - 1)} + \frac{N_h^2 S_h^2}{(m_h - 1)(m_h)} \right) \end{aligned} \tag{14}$$

is the associated sampling error from the h^{th} stratum with m_h sampling units. Similarly the associated sampling error from the h^{th} stratum based on $m_h - 1$ sampling units is

$$\begin{aligned} Var(N_h \bar{y}_{m_h - 1}) &= N_h^2 \left(\frac{N_h - (m_h - 1)}{N_h} \right) \frac{S_h^2}{m_h - 1}. \\ &= (N_h^2 S_h^2 - N_h S_h^2) - \left(\frac{N_h^2 S_h^2}{1 \cdot 2} + \frac{N_h^2 S_h^2}{2 \cdot 3} + \cdots + \frac{N_h^2 S_h^2}{(m_h - 2)(m_h - 1)} \right). \end{aligned} \tag{15}$$

When the sample size for the h^{th} stratum is increased from $m_h - 1$ to m_h, the associated sampling error for the h^{th} stratum *"decreases"* by

$$Var(N_h \bar{y}_{m_h - 1}) - Var(N_h \bar{y}_{m_h}) = \frac{N_h^2 S_h^2}{(m_h - 1)(m_h)} \tag{16}$$

using (14) and (15). The result in (16) is also the amount by which the overall sampling error $Var(\hat{T}_Y)$ *"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 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 for $h = 1, 2, ..., H$.

4. EXACT OPTIMAL SAMPLE ALLOCATION FOR ANY MIXED CONSTRAINT PATTERNS

4.1 Varying Strata Minimum and Maximum Sample Size Constraints

Algorithm I gives the exact optimal allocation of fixed n assuming the constraint $n_h \geq 1$ for $h = 1, ..., H$ and under the additional constraint $n = \sum_{h=1}^{H} n_h$; clearly $n \geq H$. Algorithm II gives the exact optimal allocation of fixed n assuming the constraint $n_h \geq 2$ for $h = 1, ..., H$ and under the additional constraint $n = \sum_{h=1}^{H} n_h$; clearly $n \geq 2H$. Each algorithm sets the same minimum for the number of sample units to be selected from each stratum [Algorithm I calls for at least 1 sample unit from each stratum; Algorithm II calls for at least 2 sample units from each stratum].

Given the intrepretation in Section 3, 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 such that we have the following constraints for $h = 1, 2, ..., H$:

$$n = \sum_{h=1}^{H} n_h; \tag{17}$$

$$0 < a_h \leq n_h \leq b_h \leq N_h. \tag{18}$$

We refer to the combination of constraints in (17) - (18) as a *mixed constraint pattern* where for $h = 1, ..., H$, (1) the overall sample size is n; and (2) n_h can be any integer between a_h and b_h inclusive.

From the discussion in Section 3, it follows immediately that the following Algorithm III yields an exact optimal sample allocation of fixed n under the mixed constraint pattern in (17) - (18) to minimize $Var(\hat{T}_Y)$.

Exact Optimal Allocation Algorithm III

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 constraint pattern in (17) - (18). 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 resulting array from Step 2.

Note that the size of the array in Algorithm III need not be greater than $(H) \times (max\{b_h\} + 1)$.

4.2 Example to Illustrate Algorithm III

Recall the stratified population of $N = 149$ units in Section 2.3. Again, assume the constraint $n = \sum_{h=1}^{3} n_h = 10$. Assume the additional mixed constraint pattern: $1 \leq n_1 \leq 5$; $2 \leq n_2 \leq 6$; and $3 \leq n_3 \leq 4$. Note that $max\{b_h\} + 1 = 7$. Then the exact optimal sample allocation can be found by applying the Exact Optimal Alocation Algorithm III as follows.

Step 1: Compute the 3×7 array of *priority values* as noted below.

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
470	332.34	191.88	135.68	105.10	85.81	72.52	62.81	
366	258.80	149.42	105.66	81.84	66.82	56.48	48.91	
164	115.97	66.95	47.34	36.67	29.94	25.31	21.92	

Step 2: Applying Step 2 of Algorithm III that reflects the mixed constraint pattern, the 3×7 array reduces to:

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
470	332.34	191.88	135.68	105.10	–	–	–	
366	–	149.42	105.66	81.84	66.82	–	–	
164	–	–	47.34	–	–	–	–	

Step 3: The $n - \sum_{h=1}^{3} a_h = 10 - (1 + 2 + 3) = 4$ largest priority values from the above array in *Step 2* are noted in bold below. The optimal allocation which gives minimum variance subject to all constraints is given in the last column of the array below, and it is $n_1 = 4, n_2 = 3$, and $n_3 = 3$.

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
470	**332.34**	**191.88**	**135.68**	105.10	–	–	–	**4**
366	–	**149.42**	105.66	81.84	66.82	–	–	**3**
164	–	–	47.34	–	–	–	–	**3**

For each feasible allocation (n_1, n_2, n_3) of $n = 10$ among the three strata subject to the mixed constraint pattern $(1 \leq n_1 \leq 5; 2 \leq n_2 \leq 6; 3 \leq n_3 \leq 4)$, Table 1 gives the associated values of $Var(\hat{T}_Y)$. We see that the minimum variance does indeed occur with $n_1 = 4, n_2 = 3$, and $n_3 = 3$ as obtained above.

Table 1: Feasible Allocations of $n = 10$ Under Mixed Constraint Pattern and Associated $Var(\hat{T}_Y)$

n_1	n_2	n_3	$Var(N_1\bar{y}_1)$	$Var(N_2\bar{y}_2)$	$Var(N_3\bar{y}_3)$	$Var(\hat{T}_Y)$
1	6	3	216,200.00	20,130.00	8,309.33	244,639.33
2	5	3	105,750.00	24,595.20	8,309.33	138,654.53
3	4	3	68,933.33	31,293.00	8,309.33	108,535.66
4	**3**	**3**	**50,525.00**	**42,456.00**	**8,309.33**	**101,290.33**
5	2	3	39,480.00	64,782.00	8,309.33	112,571.33
1	5	4	216,200.00	24,595.20	6,068.00	246,863.20
2	4	4	105,750.00	31,293.00	6,068.00	143,111.00
3	3	4	68,933.33	42,456.00	6,068.00	117,457.33
4	2	4	50,525.00	64,782.00	6,068.00	121,375.00

Consider rows 1 and 2 of Table 1. Note that as n_1 increases from 1 to 2, $Var(N_1\bar{y}_1)$ decreases by 216,200 - 105,750 = 110,450 $= \dfrac{N_1^2 S_1^2}{1 \cdot 2}$. (Note that $\sqrt{\dfrac{N_1^2 S_1^2}{1 \cdot 2}} \approx 332.34$ is the entry in the first row and first column of the 3×7 array in Step 1.) Also as n_2 decreases from 6 to 5, $Var(N_2\bar{y}_2)$ increases by 24,595.2 - 20,130 = 4,465.2 $= \dfrac{N_2^2 S_2^2}{5 \cdot 6}$. The sample size n_3 is unchanged between rows 1 and 2. Thus the overall $Var(\hat{T}_Y)$ changes from 244,639.33 to 138,654.53 (= 244,639.33 - 110,450 + 4,465.2).

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 strata optimal sample sizes, as desired, while yielding minimum sampling variance.

In this paper, we also generalize the exact result to varying sample size constraints on the different strata. The results of this paper are simple, exact, and optimal.

This presentation seeks to reach a wide audience. In a more efficient implementation of the algorithms, one need not precompute every priority value in a rectangular array as described. Instead, in each step for each stratum, one computes the currently applicable priority value. So initially you compute H priority values, inserting each into a sorted list as you go. This approach cuts down on the computations considerably.

The most interesting aspect of this paper is that we have given algorithms that are simple in concept and execution to solve nonlinear optimization problems with a linear constraint over a space of integer values. We understand that the algorithms could be described as "greedy" algorithms (Cormen, Leiserson, Rivest, and Stein, 2009) because they work by allocating sample units to strata one by one, and in each step, we choose the allocation which minimizes the objective function $Var(\hat{T}_Y)$. We understand further that it's common that greedy algorithms are easy to compute but are not guaranteed to find the global optimum. It is worth noting that the algorithms presented in this paper are greedy algorithms that *do* find the global optimum, *always*.

ACKNOWLEDGMENTS

The author is grateful to his colleagues Pat Cantwell, Deanna Weidenhamer, and Josh Tokle whose reviews greatly improved clarity. It should be noted that the wording of the comments in the Conclusion on greedy algorithms and a more efficient implementation of the algorithms are due totally to Josh Tokle.

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques (Third Edition)*, New York, NY: John Wiley & Sons.

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2009). "Greedy Algorithms" in *Introduction to Algorithms*, (Revised/Expanded ed.), Cambridge, MA: MIT Press.

Fuller, W. A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis (Second Edition)*, Boston, MA: 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, Vol. 97*, 558-606.

Tschuprow, A. A. (1923). "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations", *Metron, Vol 2*, 461-493, 646-683.

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.