New Stratification and Allocation Procedures for Stratified Simple Random Sampling without Replacement, with Application to Business Surveys

Naomi Baum[1], Danny Pfeffermann[2]
[1]Central Bureau of Statistics, Israel. Email: Naomib@cbs.gov.il
[2]Central Bureau of Statistics, Israel, Hebrew University of Jerusalem, Israel and University of Southampton, UK. Email: msdanny@cbs.gov.il

Abstract
The number of strata, the strata boundaries, the total sample size and the allocated number of sample units to each stratum, all affect the magnitude of the standard error of estimators of totals. The total sample size is usually given, since it is determined by the available budget. This paper considers stratified, simple random sampling without replacement (SRSWOR), when there are lower bounds to the sampling probabilities and to the number of allocated units per stratum, but the number of strata may vary. We propose to create as many strata as possible.

A new procedure for an appropriate sampling design satisfying the bounds conditions is proposed. By this procedure, the sampling frame is divided into three sub-frames of large, medium and small size units. Next, the medium size sub-frame is stratified using the equal-strata-total stratification method and the small size sub-frame is stratified using the equal-size method, in both cases into as many strata as possible. The large units are all sampled. An allocation which under a commonly used assumption coincides with Neyman's allocation is applied for the medium size sub-frame. Proportional allocation is used for the small size sub-frame, as to meet the lower bound of the sampling probabilities.

The procedure is easily programmed. A simulation study shows considerable reduction of the standard errors of estimators of totals.

Key Words: equal-strata-total stratification, geometric stratification, Neyman allocation, proportional allocation.

1. Introduction

Administrative records, such as VAT for example, provide information about the entire population, including a size variable, X, (total revenue, number of employees…) that is known for every unit. We plan the sampling design according to this known size variable X, attempting to reduce as much as possible the sampling error of its total estimator. We refer to X as the stratification size variable. The sample drawn is used to estimate the totals of many (economic) outcome variables for which the true population totals are unknown (hereafter the target variables).

In business surveys we usually expect such economic variables to be highly correlated with X (for example, high value added tax for firms with big annual revenue). Hence, we may assume that low sampling error of the estimator of the total of X will imply sufficiently accurate estimates for the totals of the target variables. An example for that is shown in section 2. Otherwise, we restrict our discussion and examples to the stratification size variable X only.

Commonly, there is a lack of robustness when having a Probability Proportional to Size (PPS) sample. This happens especially when estimating totals of variables which are not enough correlated with the variable according to which the sampling probabilities were fixed. A small growth of the very small units, being weighted by the inverse of their very small sampling probabilities, causes this lack of robustness. Customary, people tend to have stratified Simple Random Samples (SRS) in order to avoid the lack of robustness of PPS sampling. This is unnecessary. Instead, a lower positive limit of the sampling probabilities ensures the robustness and controls the variance. The very small units are no longer sampled with probabilities proportional to their size then.
We had good experience with the use of the systematic PPS method of sampling (including a lower positive limit of the sampling probabilities) in the Israel CBS up to 2003, when we transferred to the stratified Simple Random Sample WithOut Replacement (Stratified SRSWOR) method. The systematic PPS method was more complicated than the stratified SRSWOR method when we had to update the frames. Also, the standard error of different target variables is not known for systematic PPS sampling, and can be unbiasedly estimated under SRSWOR.
The variance of the estimator of the total of the stratification size variable under PPS sampling is very low (not zero when truncating the sampling probabilities under assumption 3 bellow, see section 5). We would like to have a SRS, without missing the PPS low variance.
Thus, the main idea of this paper is imitation of PPS sampling in the field of SRS: We try to achieve a similar variance for the case of stratified SRSWOR, by implementing sampling fractions close to the PPS sampling probabilities. Having many strata, as many strata as possible, helps us to get closer to the PPS probabilities. We enlarge the homogeneity within strata, thus lower the variance.
There are known unbiased estimators of totals and of the variance for stratified SRSWOR, irrespective of the number of strata. Cochran (1977) and James et al. (2005) argue that usually it is worth stratifying a frame into no more than six strata because with more strata the variances of the estimators are reduced only marginally. The authors claim also that handling sampling frames becomes too complicated when having too many strata. Our experience shows, however, that usually it pays to increase the number of strata and thus reduce the variance of the estimator of a total (Sections 2,8). We do not find it more complicated to handle many strata.

We study planning a Stratified SRSWOR, so as to minimize the CV of the estimator of the total of the stratification size variable X. The results of this paper are relevant for business surveys and for other surveys as well.

We assume:
1. Equal cost of data collection, irrespective of stratum affiliation.
2. The total sample size is given (limited by the budget available).
3. The sampling weight (=1/sampling probability) is bounded from above to control the variance and ensure robustness.
4. A minimum number of sampled units per stratum (say, 2-4) is required. Choose a minimum of say 2 units, if there is no such requirement.

Thus, we have to determine: A- the number of strata, B- the strata boundaries and C- the sample allocation to these strata.

In the present paper we propose a new stratification and allocation method, which may be applied regularly (mostly with big frames). The proposed number of strata is calculated through our method, rather than given a priory, as required for other familiar stratification methods. An example (Section 8) illustrates the effectiveness of the method in reducing the CVs of the total of the stratification variable.

In Section 2 we prove that the variance of the estimator of the total of the stratification size variable is reduced when increasing the number of strata while stratifying according to the values of this stratification size variable. In Section 3 we discuss some properties of the Neyman allocation method, while in Section 4 we review different methods of defining strata boundaries. PPS is considered in Section 5. These stratification and allocation methods helped us developing our proposed stratification and allocation method in Section 6. A special case of standard deviation proportional to mean is considered in Section 7. Finally, we illustrate in Section 8 the performance of our method and other methods proposed for variance reduction using real business data from Israel, and conclude in Section 9 with some summary remarks.

2. Increasing the Number of Strata for Variance Reduction

We start with an example. Consider a frame of enterprise data in Israel. The frame contains all enterprises belonging to a branch of activity named "architectural, engineering and other technical activities", that were active in 2010. The distribution of the enterprises' annual revenue in 2010 is listed in Table 1 (in million NIS):

Table 1: Distribution of 2010 annual revenue

Annual revenue	Number of firms	Total annual revenue	% Number of firms	% Total annual revenue
40+	37	4184	0.2	23.7
20 - 40	50	1408	0.3	8.0
10 - 20	124	1701	0.7	9.6
5 - 10	289	1973	1.6	11.2
3 - 5	388	1473	2.1	8.3
2 - 3	514	1256	2.8	7.1
1 - 2	1318	1821	7.1	10.3
0.5 - 1	2247	1573	12.1	8.9
0.15 - 0.5	6157	1774	33.3	10.1
0.00 - 0.15	7377	481	39.9	2.7
Total	18501	17646	100	100

Note that almost a quarter of the total annual revenue is due to 0.2% of the firms, whereas less than 3% of the total annual revenue is due to almost 40% of the firms. Usually, few very big firms are responsible for most of the economic activity.

The stratification size variable is the 2010 annual enterprise revenue. The frame has been stratified into 3,6,9 and 12 strata, using Lavallee & Hidiroglou (1988) optimal stratification procedure with some small changes (including a stratum sampled with probability=1, see Section 6 for more details). Stratified simple random samples without replacements (stratified SRSWOR) of 250,500,750 and 1000 enterprises were allocated

using Neyman's optimal allocation. In addition, for the case of 250 sampled enterprises, the frame was stratified into as many as 80 strata, as proposed in Section 6.

Consider Horvitz-Thompson (HT) estimation of:

- 2010 total annual revenue (the revenue is now considered as the target variable)
- Total revenue of the last 4 months of 2010 (total revenue from 09.2010 until 12.2010)
- 2010 total number of employees.

The following tables show the % coefficient of variation (%CV=100*CV) of the above mentioned estimator, based on the true population values.
Table 2 contains the %CV of the estimator of the total annual revenue.

Table 2: %CV of estimator of 2010 total annual revenue

sample size / number of strata	250	500	750	1000
3	6.30	2.76	1.85	1.43
6	2.71	1.23	0.93	0.60
9	1.64	0.71	0.47	0.36
12	1.22	0.53	0.37	0.29
80	0.35			

Table 3 contains the %CV of the estimator of the total revenue in the last 4 months of 2010. The size variable used for stratification is again the 2010 total enterprise revenue.

Table 3: %CV of estimator of 2010 third trimester total revenue

sample size / number of strata	250	500	750	1000
3	7.05	3.21	2.18	1.71
6	4.07	1.98	1.57	1.08
9	3.32	1.70	1.19	0.99
12	3.23	1.65	1.16	0.94
80	2.59			

Notice that the %CV's in Table 3 are larger than the corresponding CV's in Table 2, which could be expected since in Table 2 the stratification variable also served as the target variable.
Table 4 contains the %CV when estimating the 2010 total number of employees.

Table 4: %CV of estimator of 2010 total number of employees

sample size / number of strata	250	500	750	1000
3	11.01	5.94	4.57	3.76
6	8.42	5.02	4.04	3.23
9	7.98	4.78	3.69	3.12
12	8.04	4.76	3.61	3.04
80	7.26			

The CVs in Table 4 are even larger than in the previous tables, due to the fact that the number of employees is less correlated with the annual revenues.

What is common, however, to all the three tables is that increasing the number of strata reduces the variances of the estimators. Notice, in particular, the reduction of the CVs with 80 strata. This example illustrates the advantage of stratifying the frame into as many strata as possible. Nonetheless, this is not always the case, as the following example shows.

The frame now consists of the smallest 10,000 enterprises of the previous frame. Again, it is sorted by total enterprise revenue in 2010. We estimate the total annual revenue of these 10,000 enterprises, using a proportionate stratified sample of size 100. We get %CV=3.50% when the frame is divided into two strata consisting of the 4000 largest enterprises, and the other 6000 smaller enterprises (with sample sizes 40 and 60 respectively). If the frame is divided into three strata with the largest 9400 units in one stratum, and 300 units in each of the other two strata (with sample size 94, 3 and 3 respectively), we get %CV=6.54%. Enlarging the number of strata does not lower the CV in this case. The reason is, of course, inefficient stratification in the second case.

Cochran (1977, p.98) shows that increasing the number of strata reduces the variance, for ignorable sampling rates and proportional allocation, irrespective of the method of stratification. The sampling rates are usually not ignorable in business surveys. In Theorem 1 below we prove that the variance of the estimator of the population total of the stratification variable is reduced, when dividing a stratum into two strata according to the stratification variable, and using proportional allocation. It thus follows that splitting any given stratum into two or more strata further reduces the variance of the resulting estimator.

Theorem 1

Let $F = F_1 \cup F_2$ define a division of the frame into two exclusive and exhaustive strata; $F_1 \neq \emptyset, F_2 \neq \emptyset, F_1 \cap F_2 = \emptyset$, and X denote the stratification variable, such that $X_i < b < X_j$ for every $i \epsilon F_1, j \epsilon F_2$, for some $b \in \mathbb{R}$. Denote by N, N_1, N_2 the number of units in F, F_1, F_2 respectively.

Consider a SRSWOR from F of size n, and another proportionate stratified SRSWOR from F_1, F_2 of sizes n_1, n_2 respectively, such that $n = n_1 + n_2$, $\frac{n}{N} = \frac{n_1}{N_1} = \frac{n_2}{N_2} = f$.

Let $\bar{x}, \bar{x}_1, \bar{x}_2$ define the sample means in F, F_1, F_2 and $\hat{X} = N\bar{x}, \hat{X}_1 = N_1\bar{x}_1, \hat{X}_2 = N_2\bar{x}_2$ define the estimators of the totals X, X_1, X_2 over F, F_1, F_2 respectively.

Then, $VAR(\hat{X}) \geq VAR(\hat{X}_1) + VAR(\hat{X}_2)$

Proof

Let $Z_i = X_i - b$. Then $Z_i < 0 < Z_j$ for any $i \in F_1, j \in F_2$.

Let $\bar{Z}, \bar{Z}_1, \bar{Z}_2, S^2, S_1^2, S_2^2$ represent the frame means and variances of Z over F, F_1, F_2 respectively.

Clearly, $\bar{Z}_1 < 0 < \bar{Z}_2$, $\bar{Z} = (N_1\bar{Z}_1 + N_2\bar{Z}_2)/N$.

Let $\hat{Z} = \hat{X} - Nb, \hat{Z}_1 = \hat{X}_1 - N_1 b, \hat{Z}_2 = \hat{X}_2 - N_2 b$.

Then, since the sampling fraction f is the same over F, F_1, F_2 (proportional allocation),

$VAR(\hat{X}) = VAR(\hat{Z}) = (1/f - 1)NS^2$,

$VAR(\hat{X}_1) = VAR(\hat{Z}_1) = (1/f - 1)N_1S_1^2$, $VAR(\hat{X}_2) = VAR(\hat{Z}_2) = (1/f - 1)N_2S_2^2$.

Hence, it remains to prove that: $NS^2 \geq N_1S_1^2 + N_2S_2^2$

$N_1S_1^2 + N_2S_2^2 = \left(\sum_{i \in F_1} Z_i^2 - N_1\bar{Z}_1^2\right)N_1/(N_1 - 1) + \left(\sum_{j \in F_2} Z_j^2 - N_2\bar{Z}_2^2\right)N_2/(N_2 - 1)$

$NS^2 = \left(\sum_{i\in F_1} Z_i^2 + \sum_{j\in F_2} Z_j^2 - (N_1\bar{Z}_1 + N_2\bar{Z}_2)^2/N\right)N/(N-1) =$
$\left(\sum_{i\in F_1} Z_i^2 + \sum_{j\in F_2} Z_j^2\right)N/(N-1) - \left({N_1}^2\bar{Z}_1^2 + 2N_1N_2\bar{Z}_1\bar{Z}_2 + {N_2}^2\bar{Z}_2^2\right)/(N-1)$
$= N_1S_1^2 + \sum_{i\in F_1} Z_i^2\left(\frac{N}{N-1} - \frac{N_1}{N_1-1}\right) + {N_1}^2\bar{Z}_1^2\left(-\frac{1}{N-1} + \frac{1}{N_1-1}\right) + N_2S_2^2 + \sum_{j\in F_2} Z_j^2\left(\frac{N}{N-1} - \frac{N_2}{N_2-1}\right) + {N_2}^2\bar{Z}_2^2\left(-\frac{1}{N-1} + \frac{1}{N_2-1}\right) - 2N_1N_2\bar{Z}_1\bar{Z}_2/(N-1)$.
Then, since $N = N_1 + N_2$, $NS^2 \geq N_1S_1^2 + N_2S_2^2$ iff
$\left({N_1}^2\bar{Z}_1^2 - \sum_{i\in F_1} Z_i^2\right)N_2/\left((N-1)(N_1-1)\right) +$
$\left({N_2}^2\bar{Z}_2^2 - \sum_{j\in F_2} Z_j^2\right)N_1/\left((N-1)(N_2-1)\right) \geq 2N_1N_2\bar{Z}_1\bar{Z}_2/(N-1)$.
But ${N_1}^2\bar{Z}_1^2 - \sum_{i\in F_1} Z_i^2 = 2\sum_{i,j\in F_1, i<j} Z_iZ_j > 0$ since $Z_i, Z_j < 0$ for all $i,j \in F_1$ and
${N_2}^2\bar{Z}_2^2 - \sum_{j\in F_2} Z_j^2 = 2\sum_{i,j\in F_2, i<j} Z_iZ_j > 0$ since $Z_i, Z_j > 0$ for all $i,j \in F_2$
Also $\bar{Z}_1\bar{Z}_2 < 0$, which completes the proof.
Q.E.D.

Remark 1. The use of Neyman's optimal allocation instead of proportional allocation would further reduce the variance.

3. Neyman Sample Allocation with modifications

Let the frame $F = \cup F_h$ be stratified, and let $N_h, n_h, X_h, \bar{X}_h, \bar{x}_h, S_h$ define respectively the frame size, the sample size, the total, the mean, the sample mean and the standard deviation of the stratification size variable X in stratum h. Define $\hat{X}_h = N_h\bar{x}_h$ the unbiased estimator of the total X_h under SRSWOR in each stratum, and $\hat{X} = \sum_h \hat{X}_h$. Let $n = \sum n_h$ be the total desired sample size. The optimal Neyman allocation for which $VAR(\hat{X})$ is minimized is: $n_h = nN_hS_h/\sum_k N_kS_k$.

It may happen that the Neyman allocation in a given stratum is bigger than the stratum size. Usually, in business surveys, these are the strata containing the large units, and the units in such strata are sampled with probability=1. Let N_L be the total number of units in these strata.

Suppose that the sampling weights (N_h/n_h) are restricted by a fixed value w_{max} (assumption 3 mentioned in the introduction).

Sometimes, Neyman's allocation rule yields sampling weights $N_h/n_h = \sum_k N_kS_k/nS_h > w_{max}$. In business surveys, these are usually the strata containing the very small units. Staying as close as possible to Neyman's optimal allocation, these strata should be given additional allocated sampling units to meet the restriction of the sampling probability i.e. $n_h \approx N_h/w_{max}$. Hence, all of these strata will have about the same sampling probability $(\approx 1/w_{max})$. In proportional allocation too, all strata have the same sampling probability. Thus, the restriction of the sampling probability implies that the strata with the smaller size units should be proportionally allocated. Let N_S be the total number of units in these strata. Then the number of units sampled in these strata, after correction because of the restriction of the sampling probability, is $n_S \approx N_S/w_{max}$.

The other strata, for which $1 \leq N_h/n_h = \sum_k N_kS_k/nS_h \leq w_{max}$, are usually the strata of the medium size units. These strata should be sampled using the optimal Neyman allocation. The strata that are sampled with probability 1 are allocated less than the original Neyman allocation to them. The strata that meet the sampling probability restriction are allocated more than the original Neyman allocation to them. Hence, we should recalculate the Neyman allocation for all the strata of the medium size units, with a different total sample size $n_M = n - N_L - n_S$.

The above considerations suggest grouping the strata into three different sub-groups:

F_L - The strata that are sampled with probability 1, usually containing the large size units,

F_M - The strata allocated by the Neyman allocation rule, usually containing the medium size units,

F_S - The strata that are proportionally allocated with sampling probability $\cong 1/w_{max}$, usually containing the small size units.

Notice that we wanted to have optimal Neyman allocation over the entire frame, but end up with the use of Neyman allocation only in the sub-frame of the medium size units. The strata of the large units or of the very small units cannot be Neyman allocated.

4. Stratification Methods

Having outlined in Section 3 an appropriate sample allocation for given strata, the next and more difficult question to consider is how to define the strata. In this section we assume a given number *H* of strata into which we have to stratify the frame. The aim is to minimize the CV of the estimator of the population total of the stratification size variable, *X*, given the total sample size or conversely, to minimize the total sample size given the CV. Since the values of *X* are known for every sampling unit, we first sort the frame by these values. The sorting is necessary in order to maximize homogeneity within the strata and thus lower the variance of the estimator. For a given number of strata, stratifying the frame becomes then a question of determining the strata boundaries. Several methods have been proposed in the literature:

- $cum\sqrt{f}$ (Dalenius and Hodges, 1959). The strata boundaries are determined so that they define equal intervals of the cumulative square roots of the frequency f of the stratification variable. The method assumes SRS with replacement.

- *Equal-strata-total*: The strata boundaries are determined such that $X_h = X/H$ for every stratum *h*, where X_h is the total of the size variable in stratum h and $X = \sum_h X_h$. This criterion is well known, already mentioned by Dalenius and Hodges (1959) and we discuss it further latter on.

- *Geometric method* (Gunning and Horgan, 2004, Horgan, 2006). Denote the extreme values of *X* by X_{min}, X_{max}. Gunning and Horgan propose the use of geometric strata boundaries defined as $X_{max}(X_{min}/X_{max})^{h/H}, h = 0, \dots, H$ (with $h = H$ defining the lowest boundary and $h = 0$ the highest). The authors compare this method with the $cum\sqrt{f}$ method, showing the better performance of the geometric method. Kozak and Verma (2006) show that the Gunning and Horgan geometric method is not satisfactory, however they apply it over the entire frame. Clearly, the method is very sensitive to the values of the largest and smallest units only, and it may thus happen that most of the units are gathered in very few strata. Gunning & Horgan's geometric method is not suitable when there are too extreme values of the size variable, which is often the case in business surveys. It can be applied in the sub-frame, F_M, of the medium size units.

- *LH method* (Lavallee and Hidiroglou, 1988). The authors propose an iterative method to determine optimal strata boundaries, assuming the size variable is continuous, with no restriction on the sampling probability. Applying the LH procedure over the whole frame does not yield an optimal solution in the case where the sampling probability is restricted. Also, LH assume the same allocation rule over the entire frame, but as argued in Section 3, it is better to have two different allocation rules: Neyman's allocation for F_M, and proportional for F_S. It could be a good idea to run the LH method over the sub-frame of the medium size units F_M with the Neyman optimal rule as the target, but our experience shows that the method often fails to converge with many strata.

- *Equi-distance*: This method too is mentioned in Dalenius & Hodges (1959). The strata boundaries are $X_{max} - h(X_{max} - X_{min})/H\,, h = 0, \dots, H$ (with $h = H$ defining the lowest boundary and $h = 0$ the highest).
- *Equal-strata-size*: All the strata contain an equal number of frame units.

The different stratification methods listed above are compared in Section 8.

All the above mentioned stratification methods assume a given number of strata and do not address the question of what is the desired number of strata that minimizes the variance. As illustrated in Section 2, increasing the number of strata reduces the CV. Thus, fixing the number of strata in advance and applying one of the above stratification methods will not generally minimize the CV (or total sample size).
In Section 6 we propose simultaneously determining the number of strata, the strata boundaries and the sample allocation.

5. Probability Proportional to Size

In this section we discuss briefly some features of the Probability Proportional to Size (PPS) method of sampling (without replacement). We borrow some ideas from these features for our proposed stratified SRSWOR method of sampling (Section 6), with the aim of reducing the variance of the estimators.
Assume a frame F with size variable values $X_i, i \epsilon F$ and a total $X = \sum_{i \epsilon F} X_i$. PPS sampling means that each unit i is sampled with probability nX_i/X, where n is the total sample size. Notice that the expected number of sampled units from a sub-frame F_h with a total size X_h is, $n_h = \sum_{i \epsilon F_h} nX_i/X = nX_h/X$.

The unbiased Horvitz-Thompson (1952) estimator of the total of the size variable is $\hat{X} = \sum_{i \epsilon S} \frac{X_i}{nX_i/X} \equiv X$, where S is the selected sample. Clearly, the variance of $\hat{X}$ over all possible sample selections is null. The null variance is achieved only for the estimator of the total of the size variable X, according to which the sampling probabilities are defined. As stated in Section 1, we expect that the null variance of $\hat{X}$ would imply sufficiently accurate estimators of the totals of other target variables, depending on their correlation with X.

As already discussed in Section 3, in business surveys there are typically large enterprises which should always be in the sample (with probability 1). In PPS sampling, for these businesses $nX_i/X > 1$. On the other hand, the sampling probability is usually bounded from below (assumption 3 mentioned in the introduction) in order to ensure robustness and control the variance of total estimators of variables not well correlated with X. Thus for the many very small enterprises we find that $nX_i/X < 1/w_{max}$, where w_{max} is the upper bound of the sampling weights ($1/w_{max}$ is the lower bound of the sampling probability). These characteristics suggest dividing the frame into the three different sub-frames F_L, F_M, F_S with sampling schemes,
- F_L , Large units, sampled with probability 1.
- F_M, the Medium size units, sampled with probability proportional to size.
- F_S , the very Small units, sampled with probability=$1/w_{max}$.

Notice that we end up using PPS sampling probability only in the sub-frame of the medium size units F_M. The very big units F_L or the very small units F_S cannot be sampled with probability proportional to their size.

A procedure of dividing the frame and deriving the sampling probabilities is as follows:

Procedure 1:
Begin:
- Let $0 \le \delta \ll 1 - 1/w_{max} < 1$
- Let $\lambda = n/X$, where n is the total desired sample size and X is the total of the size variable

Main loop:
- Calculate $p_i = \lambda X_i$ for $i \in F$
- Denote:

F_L	the sub-frame of units for which $p_i \ge 1 - \delta$
N_L	number of units in F_L
F_S	the sub-frame of units for which $p_i \le 1/w_{max}$
N_S	number of units in F_S
$n_S = N_S/w_{max}$	minimum possible sample size from sub-frame F_S
F_M	the sub-frame of units for which $1/w_{max} < p_i < 1 - \delta$
$X_M = \sum_{i \in F_M} X_i$	total size of units in F_M
$n_M = n - N_L - n_S$	maximum possible sample size from sub-frame F_M

- Redefine $\lambda = n_M/X_M$.
- Repeat the main loop a few times, each time with the last calculated λ.

Finish:
- Define $p_i = 1$ for $i \in F_L$
- Define $p_i = 1/w_{max}$ for $i \in F_S$
- Each unit is now affiliated to either F_L, F_M or F_S

Remark 2. In practice, it helps to set $\delta \sim 0.1$ so that the process converges more rapidly.

Remark 3. It may happen that $n_M < 0$, for example, when $n < N/w_{max}$. In this case the total sample size or the restriction of the sampling probability should be changed before sampling.

Remark 4. In General, one should iterate over the main loop until convergence, at which stage F_L, F_M, F_S are stable. Our experience in business surveys shows that convergence is achieved after 3-5 iterations, provided that the frame is sufficiently big. We never encountered an application where procedure 1 did not converge (with big enough frames). However, it is not claimed that procedure 1 always converges. Convergence, although obviously desirable, is not crucial, and one can stop at any iteration and define F_L, F_M, F_S and sampling probabilities as obtained at that iteration.

Procedure 1 ends with all units sampled in F_L, with PPS sampling probabilities in F_M, and with equal probabilities defined by the sampling probability limit $(1/w_{max})$ for units in F_S. The (expected) total sample size n is as required.

Neyman's Allocation too ends with splitting the frame into F_L, F_M, F_S, with all units sampled in F_L, and with the units in F_S sampled with probability $(1/w_{max})$ (Section 3).

Thus, we propose to first pretend as if we are sampling using the PPS method, and based on that, continue with stratified SRSWOR. This is discussed in section 6.

6. Proposed Stratification and Allocation Procedure for Stratified SRSWOR

We would like to have a SRS, without losing the PPS low variance. As mentioned in Section 5, the variance of the HT estimator of the total of the size variable under PPS sampling is null (or close to null if the sampling probability is bounded). We now attempt to reduce the variance of the estimator of the total of the size variable as much as possible

under stratified SRSWOR. We do so by "converting" the PPS sampling over $F = F_L \cup F_M \cup F_S$ into stratified SRSWOR. This is done by stratifying the frame into as many strata as possible. Each of the strata has a sampling fraction close to the PPS sampling probability of its units. Having many strata helps us to get closer to the PPS sampling scheme.
Hereafter, we add the assumption (4 in the introduction) of a given minimum n_{min} of sampled units per stratum.

We propose the following extended procedure for stratification and sample allocation:

Procedure 2: Stratification and Allocation for Stratified SRSWOR

- Apply Procedure 1 (Section 5) to obtain a division of the frame into the three sub-frames, F_L, F_M, F_S. Denote, as before, by N_L, N_S the number of units in F_L and F_S, by X_M the total size of units in F_M and let $n_M = n - N_L - N_S/w_{max}$.
- Stratify F_S into $N_S/(n_{min}w_{max})$ strata, based on the ascending values of X such that each stratum h contains $N_h = n_{min}w_{max}$ frame units.
- Stratify F_M into n_M/n_{min} strata, based on the ascending values of X.
- The strata boundaries should be defined such that each stratum has equal total size $X_h = X_M\, n_{min}/n_M$, thus satisfying the equal-strata-total criterion.
- Sample all the units in F_L.
- Allocate $n_h = n_{min}$ units to the sample for each stratum in F_S, F_M and draw a stratified SRSWOR.

Notice that we split the frame *units* into F_L, F_M, F_S *before* stratifying the frame. Notice also that we used different stratification methods for F_S (equal-strata-size) and for F_M (equal-strata-total).

Slight changes should be made in Procedure 2 in situations where N_S turns out to be too small and/or n_M/n_{min} is not an integer number, etc. Notice that the frame is stratified into the largest possible number of strata, as discussed in Section 2 (Theorem 1.) The number of strata in F_M and F_S is calculated by Procedure 2, rather than given, as under the stratification methods reviewed in Section 4. An example in Section 8 illustrates the potential big reduction in variance by use of Procedure 2.

We want the PPS and the SRSWOR sampling probabilities to be as close as possible, so that the PPS low variance is (almost) achieved by SRSWOR too. The sampling probabilities of units in F_L, F_S remain unchanged in both methods (PPS & SRS, when running procedures 1 and 2 respectively). Consider the units of stratum h in F_M. Their average PPS sampling probability is, $\frac{n_M}{X_M}\frac{\sum_{i\in h} X_i}{N_h} = \frac{n_M X_h}{X_M N_h}$. The corresponding SRSWOR sampling probability is n_{min}/N_h. Hence, we impose $\frac{n_M X_h}{X_M N_h} = \frac{n_{min}}{N_h}$, or equivalently, $X_h = X_M\, n_{min}/n_M$, which is the same for every stratum h. This provides a justification for stratifying F_M according to the equal-strata-total criterion. The homogeneity within the strata (achieved by stratifying based on the ascending values of the size variable X), guarantees that as the number of strata increases, $X_i \approx X_h/N_h = \bar{X}_h$ for all units i in stratum h. The SRS sampling probability is then $n_h/N_h = n_{min}/N_h = n_M X_h/(X_M N_h) = n_M \bar{X}_h/X_M \approx n_M X_i/X_M$, which is the same as the sampling probability if we draw a PPS sample. Thus, a good way to check the stratification and allocation is to check whether the PPS probability $n_M X_i/X_M$ is sufficiently close to n_h/N_h for all the units in stratum h. Also, we may look at $S_h n_M/X_M$ and $N_h S_h n_M/X_M$ as the average and total distance

respectively between the PPS probabilities $(n_M X_i/X_M)$ and the SRS probabilities $(n_M \bar{X}_h/X_M)$ in stratum h.

Remark 5. It is not claimed that stratifying F_M according to the equal-strata-total criterion will always result with the minimum variance of the total size estimator. Instead, one could try stratifying F_M according to the geometric method. Then, since all values of the size variable are known, one could calculate the true variance under both methods (equal-strata-total and geometric), and choose the method which yields the lower variance (notice the negligible difference between the methods when applied over F_M only, in the example of Section 8). As mentioned before, our experience shows that the LH method does not converge when stratifying into many strata.

Remark 6. Neyman's allocation of n_M units to the n_M/n_{min} strata (given by Procedure 2), necessarily allocates less than the required n_{min} units to some of the strata, unless all strata are allocated exactly n_{min} units as in Procedure 2. If Neyman's allocation turns out to be different than that of Procedure 2 it is recommended to (slightly) change the strata boundaries.

Remark 7. An optimal stratification method in F_M, although difficult to apply, should achieve equal $N_h S_h$ in all strata. This is discussed in Section 7.

7. The Case Where the Standard Deviation is Proportional to the Mean

In this section we restrict our attention to the medium size strata in F_M. Let $F_M = \cup F_h$ be stratified, and let $N_h, n_h, X_h, \bar{X}_h, S_h$ define respectively the frame size, sample size, total, mean and the standard deviation of the stratification size variable X in stratum h. Also let $X_M = \sum X_h$, $n_M = \sum n_h$ the desired sample size from F_M and a given minimum n_{min} of sampled units per stratum (introduction, assumption 4).

Appling Procedure 2 stratifies F_M into $(n_M/n_{min} =)$ the maximum possible number of strata with $n_h = n_{min}$ sampled units allocated to each stratum h. On the other hand, application of Neyman allocation requires that $n_h = n_M N_h S_h / \sum_k N_k S_k$. Thus, Procedure 2 is optimal iff $n_{min} = n_M N_h S_h / \sum_k N_k S_k$ for all h, i.e. equal $N_h S_h$ in all strata.

Procedure 2 proposes to apply the equal-strata-total stratification method so that X_h is equal in all strata. This means that Procedure 2 is optimal iff $N_h S_h = c X_h$ or instead $S_h/\bar{X}_h = c$ for all h, in which case $n_{min} = n_M N_h S_h / \sum_k N_k S_k = n_M X_h / X_M$ or $X_h = X_M \, n_{min}/n_M$ same for all strata h. Thus, the equal-strata-total criterion is optimal when $S_h/\bar{X}_h = c$.

Dalenius & Hodges (1959) already mention the observation that for many populations the relative variance does not vary much from stratum to stratum. Indeed, experience shows that this property is approximately true for business surveys, especially when excluding the very big businesses. Actually, they claim that $S_h/\bar{X}_h \approx c$ for all h.

This provides another justification for the use of the equal-strata-total criterion as part of the proposed Procedure 2 in Section 6. Application of this criterion will be close to optimal when $S_h/\bar{X}_h \approx c$ and we stratify into the maximum possible number of strata.

If $N_h S_h$ are not sufficiently similar, it may result from not satisfying the property $S_h/\bar{X}_h \approx c$, and it may indicate that the equal-strata-total criterion is not optimal then. The sampling procedure can usually be improved by slightly changing the strata boundaries in this case, and transfer sampling units to "neighbor" strata if possible, in order to even $N_h S_h$.

8. Example

We continue the example of Section 2. The frame contains all enterprises in Israel in the architectural, engineering and other technical activities, which were active in 2010 (about 18000 enterprises). The size variable is the enterprises' annual revenue.

We assume a total sample size of n=250 and maximum sampling weight $w_{max} = 100$.

We proposed (Procedure 2) to apply the equal-strata-total stratifying method over F_M. In the following example we compare the stratifying method mentioned in Section 4 over six different sampling schemes:

a. Procedure 2 (but stratifying F_M) is applied with $n_{min} = 3$ in which case we obtain 82 strata
b. Procedure 2 (but stratifying F_M) is applied with $n_{min} = 8$ in which case we obtain 31 strata
c. Neyman allocation over the entire frame with 31 strata (and $n_{min} = 3$)
d. Neyman allocation over the entire frame with 12 strata (and $n_{min} = 3$)
e. Neyman allocation over the entire frame with 82 strata and $n_{min} = 3$
f. Neyman allocation over the entire frame with 31 strata and $n_{min} = 8$

We use the different stratification methods over the entire frame in cases c. and d., and over F_M only in cases a. and b. We ask for $n_{min} = 3$ in the rare strata where Neyman allocation concludes with less then n_{min} sampled units in cases c. and d. In cases e. and f., where in fact we tried using the maximum number of strata without applying Procedure 2, we couldn't sample 250 units only (as desired) with any of the methods, and the LH method did not converge.

The %CV of the estimator of the total revenue under the different stratification methods is shown in Table 5. The shaded cells are the outcomes when running our proposed Procedure 2.

Table 5:

%CV of total revenue estimator under different stratification and allocation methods

case	a	b	c	d	e	f
n_{min}	3	8	3	3	3	8
number of strata	82	31	31	12	82	31
Split into F_L, F_M, F_S	yes	yes	no	no	no	no
Equal-strata-total	0.36	0.96	0.53	1.35	---	---
Geometric method	0.37	1.03	1.80	3.97	---	---
LH	---	---	1.02	1.35	---	---
Equidistant	2.07	4.84	5.01	10.76	---	---
Equal-strata-size	6.87	8.12	14.82	24.91	---	---

The main conclusions from Table 5 are as follows:

- Procedure 2, (split into F_L, F_M, F_S, maximum possible number of strata, Equal-strata-total stratification over F_M), is most effective and achieves the lowest CVs.
- There is not much difference between the equal-strata-total method and the geometric method when splitting the frame and stratifying F_M only, accordingly (see Remark 5). However, the equal-strata-total method performs much better than the geometric method when not splitting the frame. In Section 6 we mentioned that

both methods may be applied when stratifying F_M, in order to choose the stratification which yields the lower CVs.

- Splitting the frame into subgroups of large, medium and small size units and enlarging the number of strata reduces the CV very significantly.
- The LH method does not perform any better than the equal-strata-total method, and in fact performs much worse when applied with many strata. However, when not splitting the frame, the LH method performs much better than the geometric method.
- The use of the equidistant and the equal-strata-size stratification methods yields much higher CVs than the other methods.

9. Summary

In this paper we considered stratified SRSWOR samples drawn from a given finite population. For a given total sample size (as determined by budget availability), we propose a new procedure for determining the number of strata, the strata boundaries and the number of sample units to be drawn from each stratum. Our proposed procedure imitates PPS method in the field of SRSWOR, so as to gain the PPS low variance on the one hand, and the many advantages of SRS (update frames, calculate or unbiasedly estimate variance…) on the other hand. We split the frame into subgroups of large, medium and small size units. The large size units are all sampled. An allocation which coincides with Neyman's allocation, whenever the standard deviation is proportional to the mean, is applied for the medium size sub-frame. Proportional allocation is used for the small size sub-frame. The main result established and illustrated is that by stratifying into as many strata as possible, one can reduce the CV very drastically. We also justify the use of the equal-strata-total method for determining the strata boundaries of the medium size subgroup and illustrate its superiority over other stratification methods proposed in the literature.

Acknowledgements

This paper is part of a PhD dissertation titled "Statistical Inference from Dependent Samples on Periodic Change, with Application to Business Surveys", written by Naomi Baum at the Hebrew University of Jerusalem. I am grateful to my PhD supervisor, Prof. Danny Pfeffermann for his help and encouragement during the years of my study. Special thanks are due to my former director, Mr. Abraham Burstein from the Israel Central Bureau of Statistics (ICBS), who was my teacher of sampling in business surveys.

References

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

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, **54**, 88-101.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, **30**, 159-166.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, **74**, 67-76.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

James G., Pont M., and Sova, M. (2005). Aspects of sample allocation in business surveys. Office for National Statistics (ONS), Newport, UK.

Kozak M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: a comparison efficiency. *Survey Methodology*, **32**, 157-163.

Lavallee P., and Hidiroglou, M.A. (1988). On the Stratification of Skewed Population. *Survey Methodology*, **14**, 33-43.