SAMPLE EXPANSION FOR PROBABILITY PROPORTIONAL TO SIZE WITHOUT REPLACEMENT SAMPLING

Lawrence R. Ernst
Ernst.Lawrence@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212-0001

KEY WORDS: PPS sampling, Sample expansion, Tillé's method, National Compensation Survey

1. Introduction

There exist numerous methods for selecting a fixed size sample with probability proportional to size (PPS), without replacement. Brewer and Hanif (1983) describe 50 such procedures and detail the properties of each of these procedures.

The properties described are with respect to the selection of a single sample directly from the universe. However, sometimes it may be necessary to either select a subsample of the original sample or to expand the original sample to a larger sample from the same universe, where in either case the new sample is also to be a fixed size sample selected PPS, without replacement from the original universe with the original measures of size. In the case of subsampling, if the procedure used to select the original sample is strictly PPS, without replacement, then subsampling with the desired properties can always be done.

However, in the case of sample expansion it appears that for most, if not all, of the procedures in Brewer and Hanif (1983), an expansion of the original sample satisfying the required conditions is not possible.

Tillé (1996) presents a new PPS, without replacement sampling procedure. The general idea of this procedure is to begin with the universe and eliminate one unit at a time until the number of remaining units is equal to the desired sample size. The set of remaining units then becomes the original sample. Although Tillé does not discuss applying his procedure to the sample expansion problem, the method of doing so is essentially immediate provided a record had been kept of the order of elimination of the units not in the original sample. A key result in this paper is that it is also always possible to perform a sample expansion with the required properties when the original sample had been selected using Tillé's method and no record had been kept of the order of elimination of the units.

In Section 2 of this paper, we introduce some notation and terminology. We also explain precisely what we mean by a PPS, without replacement sample; in particular, how certainty units are handled in such a sample. In Section 3, we illustrate by means of examples using the Brewer-Durbin procedure (Cochran 1977, Sec. 9A.8) why it is generally not possible to expand a sample that had been selected PPS, without replacement, if it is required that the expanded sample also be selected PPS, without replacement. In Section 4, we review Tillé's procedure. In Section 5, we establish that it is always possible to expand a sample if the original sample had been selected using Tillé's method, even if no record had been keep of the order of elimination of the units not in the original sample.

One of the more commonly used PPS sampling procedures is systematic PPS sampling, with units in the universe ordered by the measure of size (Brewer and Hanif 1983, p. 21-22). In addition to being a simple method to implement operationally, it yields an "implicit stratification." Tillé's method, like most other PPS sampling methods, largely lacks this implicit stratification property. In Section 6, we discuss a possible modification of Tillé's method that allows it to attain the advantages of implicit stratification to a greater extent, both for the original and the expanded sample.

This work was motivated in part by a sample expansion issue that arose for the National Compensation Survey (NCS) conducted by the Bureau of Labor Statistics (BLS). When that survey was first introduced several years ago, a smaller than desired sample was selected because of resource limitations. Later it became operationally feasible to increase the sample size. The original sample within each sampling cell had been selected by ordered, systematic PPS sampling. The sample expansion was accomplished by selecting a second sample using the same procedure, independently from the first. This approach resulted in a less than optimal design and processing complications that were due in part to the overlap of the units in the two samples. This could have been avoided if the original sample had been selected using Tillé's method.

2. Notation and Terminology

Consider a universe of N units, with T_i, $i = 1,...,N$, denoting the measure of size for unit i. Then

$$p_i = T_i / \sum_{j=1}^{N} T_j, \ i = 1,...,N \qquad (2.1)$$

is the probability of selection of unit i for a sample of 1

unit.

The probabilities of selection for a PPS sample of n units, denoted $\pi_i(n),\ i=1,...,N,$ are obtained as follows. Initially let

$$\pi_i(n) = np_i,\ i=1,...,N \tag{2.2}$$

Then for any i for which $np_i \geq 1$, redefine $\pi_i(n)=1$. For the remaining units redefine $p_i, \pi_i(n)$ using (2.1), (2.2), with i and j restricted to the remaining units rather than all N units in the universe, and with n replaced on the right hand side of (2.2) by n minus the number of units for which $\pi_i(n)=1$. Repeat this process, each time increasing the number of units i for which $\pi_i(n)=1$, until (2.2) yields no additional such units for which $\pi_i(n)>1$. The units for which the final value of $\pi_i(n)$ is 1 are the certainty units for a sample of n units and the other units are the noncertainty units.

As an example, the values of $\pi_i(k)$ for $k=11,...,4$ are presented in Table 1 at the end of the paper for a universe for which $N=12$ and the T_i are given in this table. We will return to this example throughout the paper.

The joint probability of selection for any pair of units $i, j,\ i \neq j,$ for a sample of n units is denoted $\pi_{ij}(n)$.

In general, we will let n, m, denote the number of units in the original and expanded samples, respectively.

3. Examples for Which a PPS Sample Expansion Is Not Possible

Consider a universe for which $N=4$, $n=2$, $m=3$, and the p_i are 0.4, 0.4, 0.1, 0.1, respectively. If the original sample of 2 units was chosen using the Brewer-Durbin procedure, or any PPS, without replacement procedure for which $\pi_{ij}(2)>0$ for all distinct i, j, then it is not possible to expand the sample to a sample of 3 units selected PPS, without replacement. This is because both units 1 and 2 must be certainty units for a sample of 3 units, since $3p_1 = 3p_2 > 1$. However, whenever 3 and 4 are the selected pair in the sample of 2 units, it is not possible for both units 1 and 2 to be in the expanded sample of 3 units.

The expansion problem can arise even if there are no certainty units in the expanded sample. To illustrate, consider a second example in which the only change is that the p_i are now 0.330, 0.330, 0.170, 0.170, respectively. Then there are no certainty units for a sample of 3 units, since $\pi_1(3)=\pi_2(3)=0.990$. However these probabilities cannot both be attained if the sample of 2 units is chosen with the Brewer-Durbin procedure. This is because $\pi_{34}(2)=0.051$ for that procedure. Consequently, since either unit 1 or 2 is not in the expanded sample of 3 units whenever units 3 and 4 are the selected pair for the sample of 2 units, we have that $\pi_i(3) \leq 1-0.051/2 < 0.98$ for either $i=1$ or $i=2$.

This expansion problem can never arise with equal probability, without replacement sampling, since to expand a sample of n units selected this way to a sample of m units, simply select a sample of $m-n$ units, equal probability, without replacement, from the remaining $N-n$ units.

4. Tillé's Method

To explain Tillé's method, we first review how to select a sample, S_n, of n units, PPS, without replacement, as a subsample of a sample of m units, S_m, selected PPS, without replacement. For any m, we view S_m as a random set and s_m as a specific realization of S_m. Let c_m, a_m denote the set of units in s_m that are certainty and noncertainty units, respectively, in S_m and let t_m denote the number of units in a_m. For $i \in s_m$, let $\pi_i(n|S_m = s_m)$ denote the conditional probability that unit i is in S_n given that $S_m = s_m$. Then

$$\pi_i(n|S_m = s_m) = \begin{cases} \pi_i(n), & i \in c_m \\ \dfrac{\left(n - \sum_{j \in c_m} \pi_j(n)\right)}{t_m}, & i \in a_m \end{cases} \tag{4.1}$$

It can be shown that for each $i=1,...,N$ and each s_m with $i \in s_m$,

$$\Pr(i \in S_n) = \pi_i(m)\pi_i(n|S_m = s_m) = \pi_i(n) \tag{4.2}$$

and hence the result of these two sampling steps is a sample of n units selected PPS, without replacement.

Tillé's procedure for selecting a sample of n units PPS is an $N-n$ step process where, corresponding to step $N-k,\ k=N-1,...,n,$ (that is, with k decreasing 1 in each succeeding step) a sample s_k of k units is subsampled from a sample s_{k+1} of $k+1$ units using

the subsampling procedure described above with n, m replaced by $k, k+1$. Since all but 1 unit from the preceding sample is retained at each step in the process, this is equivalent to selecting 1 unit to be eliminated at each step. Let $r_i(k)$ denote the conditional probability that unit i is eliminated at step $N-k$ given that it has not been eliminated at any of the preceding steps. Then from (4.1) it follows that

$$r_i(k) = \begin{cases} 1-\pi_i(k), & i \in c_{k+1} \\ 1 - \dfrac{\left(k - \sum_{j \in c_{k+1}} \pi_j(k)\right)}{t_{k+1}}, & i \in a_{k+1} \end{cases} \tag{4.3}$$

It is understood that in calculating $r_i(N-1)$ with (4.3), we use $c_N = s_N = \{1,...,N\}$ and $a_N = \varnothing$.

Then s_k, which is the set of sample units at the end of step $N-k$, is a sample of k units selected PPS. In particular, s_n is a sample of n units selected PPS and, for any $m > n$, s_m is an expanded sample of s_n selected PPS. If a record is kept of the order in which the units are eliminated in obtaining s_n, then s_m is known for each m. In the next section we consider the case when such information is not kept.

To illustrate this method, consider the universe for which N and T_i are as in as in Table 1, $n = 4$, and the $r_i(k)$ are given in Table 2. The unit eliminated at step $N-k$ for $k = 11,...,4$, is indicated by the row in bold in Table 2 for that value of k; for example unit 4 for $k = 10$. s_k is the set of units that are not among the first $N-k$ eliminated, for example $s_7 = \{1,3,6,9,10,11,12\}$; $s_4 = \{1,3,11,12\}$ is the final sample. If it were desired to expand s_4 to a sample of 7 units and the order of elimination of the units were known, then s_7 would be the expanded sample.

5. Expanding the Sample When the Order of the Elimination of Units in Obtaining the Original Sample Is Unknown

As in the previous section, we view S_k, $k = N-1,....,n$, as a random set of k units denoting the set of units at the end of step $N-k$ of Tillé's method, with s_k, $k = N-1,....,n$, denoting a specific realization for S_k. s_n, the final sample, is known, but s_k, $k = N-1,....,n+1$, which specifies the order of elimination of the units not in s_n, is unknown.

The general idea in obtaining an expanded sample of m units is to use a modification of Tillé's method that also selects the sample of m units in a $N-m$ step process in which a unit is eliminated at each step. The modification is that the eliminated unit can never be a unit in s_n and hence the conditional probabilities $r_i(k)$ of (4.3) must be changed, with the new conditional probabilities, denoted as $r_i^*(k)$, obtained as follows. For each k, the probability that unit i is the unit eliminated at step $N-k$ conditional on this unit having not been eliminated at any preceding step is

$$r_i^*(k) = \begin{cases} \dfrac{r_i(k)}{f(k)}, & i \notin s_n \\ 0, & i \in s_n \end{cases} \tag{5.1}$$

where

$$f(k) = \left(1 - \sum_{j \in s_n} r_j(k)\right), \quad k = N-1,...,n \tag{5.2}$$

Denote the random sample of k units obtained at step $N-k$ of this modified procedure by S_k^* and hence the final expanded sample by S_m^*. Let s_k^* denote a specific realization of S_k^*. If $s_m^* = s_m$ always, then the selection process would clearly be PPS. However, generally $s_m^* \neq s_m$. In fact, s_k^*, $k = N-1,...,m$, can be any sequence of subsets of $\{1,...,N\}$ of k units, such that for each k, $s_{k+1}^* \supset s_k^* \supset (c_k \cup s_n)$. We must demonstrate that the expanded sample of m units is selected PPS. To do this we will show that

$$\begin{aligned} &\Pr(S_k = s_k,\ k = N-1,...,m \,|\, S_n = s_n) \\ &= \Pr(S_k^* = s_k,\ k = N-1,...,m \,|\, S_n = s_n) \end{aligned} \tag{5.3}$$

and hence that

$$\Pr(S_m = s_m \,|\, S_n = s_n) = P(S_m^* = s_m \,|\, S_n = s_n) \tag{5.4}$$

Then, by (5.4), the distributions of the expanded sample of m units conditioned on $S_n = s_n$ obtained using the two methods are identical. Consequently, since the selection of S_m is PPS, so is the selection of S_m^*.

To establish (5.3), note that

$$\Pr(S_k = s_k,\ k = N-1,...,m \mid S_n = s_n)$$
$$= \frac{\Pr(S_k = s_k,\ k = N-1,...,m,\ \text{and } S_n = s_n)}{\Pr(S_n = s_n)}$$
$$= \frac{\left(\prod_{k=m}^{N-1} \Pr(S_k = s_k \mid S_{k+1} = s_{k+1})\right) \Pr(S_n = s_n \mid S_m = s_m)}{\Pr(S_n = s_n)} \quad (5.5)$$

Now let

$$d_k,\ k = N-1,...,n,\ \text{denote the unit in } s_{k+1} - s_k \quad (5.6)$$

that is, the unit eliminated in step $N-k$ of Tillé's method. Then by (5.6), (5.2), and the definition of $r_i(k)$,

$$\Pr(S_k = s_k \mid S_{k+1} = s_{k+1}) = r_{d_k}(k),\ k = N-1,...,m \quad (5.7)$$

$$\Pr(S_n = s_n \mid S_m = s_m) = \prod_{k=n}^{m-1} f(k) \quad (5.8)$$

$$\Pr(S_n = s_n) = \prod_{k=n}^{N-1} f(k) \quad (5.9)$$

Finally, we combine (5.5), (5.7)-(5.9), (5.1) to conclude that

$$\Pr(S_k = s_k,\ k = N-1,...,m \mid S_n = s_n) = \prod_{k=m}^{N-1} r^*_{d_k}(k) \quad (5.10)$$
$$= \Pr(S^*_k = s_k,\ k = N-1,...,m \mid S_n = s_n)$$

To illustrate, consider the Example of Table 3. Here, expanded samples of 11-5 units of the original s_4 in Table 2 are presented. They were obtained by successive elimination of a single unit in each step, using the probabilities $r^*_i(k)$ given in Table 3. The eliminated unit at each step, which was chosen arbitrarily from the eligible units, is indicated in bold, and s^*_k for each *k* consists of the remaining units in the column. For example $s^*_7 = \{1,3,8,9,10,11,12\}$, which is not the same as the original sample of 7 units, s_7, in Table 2.

6. Modification of Tillé's Method to Introduce a Quasi "Implicit Stratification"

One of the more commonly used PPS sampling methods is systematic PPS sampling, with units in the universe ordered by the measure of size. In addition to being a simple method to implement operationally, it yields an "implicit stratification" in the sense that if we divide the universe into size classes, using the size measure used to select the sample units, then the number of sample units in each size class consisting of only noncertainty units will be proportional to the total measure of size of the size class, ignoring rounding.

Tillé's method, like most other PPS sampling methods, lacks this implicit stratification property to a large extent. However, it does attain this property to at least some extent in comparison with some other PPS sampling methods. For example, for Sampford's method (Cochran 1977, Sec. 9A.8) it is possible for the selected sample of *n* units to consist of any *n* distinct units that includes all the certainty units. To illustrate, it is possible with Sampford's method for $s_4 = \{1,2,3,12\}$ for the example we have been discussing, that is, for the 3 smallest units to all be in sample. This is not possible for Tillé's method since, by Table 2, we must have $d_{11} \in \{1,2,3\}$, where d_k is as in (5.6). However, it is possible with Tillé's method for 2 of the 3 smallest units, to be in this sample. In fact, $s_4 = \{1,3,11,12\}$ for Table 2.

In this section we discuss a modification of Tillé's method which allows it to attain the advantages of implicit stratification in the selection of S_n to a greater extent. To begin this method, we assume that the *N* units in the universe are arranged in order of increasing measure of size. First, select a preliminary sample of *n* units using Tillé's method in the standard way, with the order of elimination of the units recorded. We then replace subsets of the eliminated units by new sets of eliminated units chosen using ordered, systematic, equal probability sampling. It is this replacement that allows us to gain the advantages of implicit stratification to a greater extent. To do this replacement, we first specify an integer $g \geq 2$ and a multiplier $z > 1$. Each ordered, systematic, equal probability sample used in this replacement consists of *g* units chosen from a set of at least zg units. To obtain the first subset to which the replacement of eliminated units applies, let k_1 be the largest integer *k* for which $t_k \geq zg$, where t_k is as in Section 4. Chose an ordered, systematic, equal probability sample of *g* units from a_{k_1} denoted $u_{11}, u_{12},...,u_{1g}$. Then let $k_{1i},\ i = 1,...,g,$ be in descending order the first *g* integers for which $k_{1i} < k_1$ and $d_{k_{1i}} \in a_{k_1}$. Then we replace the original values of these $d_{k_{1i}}$ by

$$d_{k_{1i}} = u_{1i},\ i = 1,...,g \quad (6.1)$$

thus yielding an ordered, systematic, equal probability,

sample of g units eliminated from a_{k_1}. Note that equal probability sampling is used since, by (4.3), the units in a_{k_1} are eliminated with equal probability.

The process can be repeated. To obtain a second set of eliminated units to be replaced, let k_2 be the largest integer k for which $k \le k_{1g}$ and $t_k \ge zg$. Then obtain $u_{21}, u_{22}, ..., u_{2g}$, k_{2i}, $i = 1,...,g$, $d_{k_{2i}} = u_{2i}$, $i = 1,...,g$, analogously to the way they were obtained for the first set of eliminated units to be replaced.

The process continues this way until either: (1) there is a j for which there is no k with $k \le k_{(j-1)g}$ and $t_k \ge zg$, or (2) there are only g' integers, k_{ji}, $i = 1,...,g'$, where $g' < g$, for which $k_{ji} < k_j$ and $d_{k_{ji}} \in a_{k_j}$. If (1) holds, or if (2) holds with $g' = 1$, then the process of redefining sets of eliminated units stops after set $j-1$. If (2) holds with $g' > 1$, then there are j sets, but set j consists of g' rather than g units and is obtained using an ordered, systematic sample of g' units from a_{k_j} as the replacement units to be eliminated from a_{k_j} in obtaining s_n.

To illustrate this procedure, consider the example of Table 2 and let $g = 2$, $z = 2$. Then $k_1 = 9$ since 9 is largest value of k for which $t_k \ge 4$. We then choose a systematic sample of 2 units from $a_9 = \{1,3,6,7\}$, for example $u_{11} = 6$, $u_{12} = 1$. Now $k_{11} = 8$, $k_{12} = 6$ and hence, by (6.1), we replace the original values of d_8, d_6 by $d_8 = 6$, $d_6 = 1$. Continuing, we have $k_2 = 6$, and now $a_6 = \{3,7,9,10\}$ due to the changes in d_8, d_6. We then choose a systematic subsample of 2 units from a_6, for example $u_{21} = 7$, $u_{22} = 10$. Then $k_{21} = 5$, $k_{22} = 4$ and $d_5 = 7$, $d_4 = 10$. Thus the final sample of 4 units is now $s_4 = \{3,9,11,12\}$.

This modified sample, unlike the original sample, only includes 1 of the 3 smallest units. However, the modified procedure in certain circumstances does not produce the desired results. To illustrate, consider the example of Table 2, except now suppose for the original sample that $d_{11},....,d_7$ are as in Table 2, but that $d_6 = 9$, $d_5 = 10$, and $d_4 = 11$. The modification then breaks down since, although k_1 remains 9, now 8 is the only k for which $k < 9$ and $d_k \in a_9$. Thus it would not be possible in that situation to eliminate 2 units selected systematically from a_9 when using this modified procedure. Consequently, it would appear that further research would be appropriate to determine if a different modification of Tillé's procedure would come closer to attaining the characteristics of implicit stratification possessed by ordered, systematic PPS sampling.

If a sample of n units had been selected and it is desired to expand the sample to m units while applying the modification described in this section to attain the advantages of implicit stratification to a greater extent, then proceed as follows. First select a preliminary expanded sample of m units. If the order of elimination of the units in obtaining the final sample of n units had been recorded, then the units remaining in sample after step $N - m$ of the selection of the final sample of n units is the preliminary sample of m units. Otherwise, use the method of Section 5 to obtain this preliminary sample. In either case, the final sample of m units is obtained by applying the modification described earlier in this section to the preliminary sample of m units.

7. References

Brewer, K. R. W. and Hanif, M. (1983). Sampling with Unequal Probabilities. New York: Springer-Verlag.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

Tillé, Y. (1996). An Elimination Procedure for Unequal Probability Sampling Without Replacement. Biometrika, 83, 238-241.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

Table 1. Unconditional Selection Probabilities $\pi_i(k)$

i	T_i	p_i	k							
			11	10	9	8	7	6	5	4
1	20	0.01	0.44	0.29	0.21	0.16	0.11	0.08	0.06	0.05
2	30	0.02	0.67	0.43	0.32	0.24	0.17	0.13	0.10	0.07
3	40	0.02	0.89	0.57	0.42	0.32	0.23	0.17	0.13	0.10
4	50	0.03	1	0.71	0.53	0.39	0.28	0.21	0.16	0.12
5	70	0.04	1	1	0.74	0.55	0.40	0.29	0.22	0.17
6	80	0.04	1	1	0.84	0.63	0.45	0.34	0.26	0.19
7	90	0.05	1	1	0.95	0.71	0.51	0.38	0.29	0.22
8	150	0.08	1	1	1	1	0.85	0.63	0.48	0.36
9	200	0.10	1	1	1	1	1	0.84	0.64	0.48
10	220	0.11	1	1	1	1	1	0.93	0.70	0.53
11	300	0.15	1	1	1	1	1	1	0.96	0.72
12	750	0.38	1	1	1	1	1	1	1	1

Table 2. Conditional Probabilities $r_i(k)$ of Eliminating Units Using Tillé's Method

i	k							
	11	10	9	8	7	6	5	4
1	0.56	0.36	0.26	0.25	0.28	0.26	0.24	0.25
2	**0.33**							
3	0.11	0.36	0.26	0.25	0.28	0.26	0.24	0.25
4	0	**0.29**						
5	0	0	**0.26**					
6	0	0	0.16	0.25	0.28	**0.26**		
7	0	0	0.05	**0.25**				
8	0	0	0	0	**0.15**			
9	0	0	0	0	0	0.16	0.24	**0.25**
10	0	0	0	0	0	0.07	**0.24**	
11	0	0	0	0	0	0	0.04	0.25
12	0	0	0	0	0	0	0	0

Table 3. Conditional Probabilities $r_i^*(k)$ of Eliminating Units When Expanding the Sample and Original Order of Elimination Is Unknown

i	k						
	11	10	9	8	7	6	5
1	0	0	0	0	0	0	0
2	**1**						
3	0	0	0	0	0	0	0
4	0	**1**					
5	0	0	0.56	0.50	**0.65**		
6	0	0	**0.33**				
7	0	0	0.11	**0.50**			
8	0	0	0	0	0.35	**0.53**	
9	0	0	0	0	0	0.32	**0.50**
10	0	0	0	0	0	0.15	0.50
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0