BUREAU OF THE CENSUS
MANUFACTURING AND CONSTRUCTION DIVISION

MCD Working Paper Number: Census/MCD/WP-97/02

A MODIFIED APPROACH TO SAMPLE SELECTION AND VARIANCE ESTIMATION WITH PROBABILITY PROPORTIONAL TO SIZE AND FIXED SAMPLE SIZE

by
John G. Slanta and James T. Fagan
October 1997

Manufacturing and Construction Division
Bureau of the Census

This series contains a working paper, written by or in cooperation with staff members of the Manufacturing and Construction Division. The views reflected in this working paper are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries my be addressed to the authors or to the MCD Working Paper Coordinator, Manufacturing and Construction Division, Bureau of the Census, Washington, DC 20233.

A Modified Approach to Sample Selection and Variance Estimation with Probability Proportional to Size and Fixed Sample Size

Abstract

This method of sample selection is a modification of Tillé's (1996) method. This modification guarantees a nonnegative unbiased sample variance of the Horvitz-Thompson estimator of the total. An alternative to the unbiased sample variance is given which greatly reduces the number of calculations with negligible increase in bias. The modified method with the alternative variance estimator could be easily applied to large populations, making PPS fixed sample size selection more feasible.

Key Words: Yates-Grundy-Sen sample variance, Joint probability, Inclusion probability, Rejection probability.

I. Introduction

Probability Proportional to Size (PPS) sampling uses auxiliary information in the sample selection process so that unbiased simple linear estimators will be more efficient than when simple random sampling is used. The biased ratio and regression estimators use this information in the estimator and not in the sample selection process which is why their estimators are more complex. With PPS the probability of selecting a sampling unit is proportional to the auxiliary variable with the exception of certainty sampling units. The estimator of the total we will be referring to is the Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \qquad \textbf{(1)}$$

where y_i is the variable of interest for the *i*th sampling unit, π_i is the probability of selecting the *i*th sampling unit, and n is the sample size.

PPS sampling has advantages in manufacturing surveys because of the skewness of the population. When the population is skewed and the auxiliary variable is moderately to strongly proportional to the variable of interest, the variance will be smaller for PPS than for simple random sampling.

Many manufacturing surveys at the Census Bureau currently use Poisson PPS sampling. Poisson sampling means that the selection of any one sampling unit is independent of any other sampling unit.

The advantages of Poisson PPS sampling are:

1. Smaller variance than simple random sampling for skewed populations
2. Easy selection procedure
3. Easy variance calculation
4. Allows to easily rotate out, in unbiased fashion, selected units

A major disadvantage is variable sample size, a major contributor to the variance. The Annual Survey of Manufactures (ASM) incorporates the Horvitz-Thompson estimator into a difference estimator which reduces the variance due to the variability in sample size while retaining an unbiased estimator. But other manufacturing surveys do not use a difference estimator with Poisson sampling, so variable sample size is still a problem for them.

The unbiased sample variance of the population total with Poisson sampling is:

$$\hat{\sigma}^2_{Poisson} \quad = \quad \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} \, y_i^2 \qquad \textbf{(2)}$$

If sample selection were done in a way that guaranteed fixed sample size as opposed to using Poisson sampling then the appropriate sample variance would be the one developed by Yates and Grundy and by Sen (Cochran 1977, pp. 260-261):

$$\hat{\sigma}^2_{YGS} \quad = \quad \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\beta_{ij} - 1\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad \textbf{(3)}$$

where $\beta_{ij} = \pi_i\pi_j / \pi_{ij}$, and π_{ij} is the probability of selecting both the *i*th and the *j*th sampling units.

The advantage of this is that if y_i is proportional to π_i then the sample variance is zero. Also, certainty units and units from different strata will not contribute to the variance because in that case $\pi_{ij} = \pi_i\pi_j$ so $\beta_{ij} = 1$. Therefore PPS fixed sample size will be a major improvement over PPS Poisson sampling.

This approach has several disadvantages:

1. The values of π_{ij} are hard to calculate for many sampling schemes, because the formulas for π_{ij} may not be in closed form. These sampling schemes may also involve a large number of calculations for computing π_{ij} , which could introduce precision errors.

2. To guarantee that the sample variance be nonnegative, the following must hold true: $\pi_{ij} \leq \pi_i\pi_j$ if $i \neq j$.

3. The following should hold true for the sample variance to remain unbiased: $\pi_{ij} > 0$. If this is not so then the bias will be negative and the variance will be underestimated on the average.

4. All combinations of π_{ij} from the sample must be stored in a data file and be capable of being linked to the *i*th and *j*th sampling units. The number of different π_{ij} that must be stored is n (n−1) / 2. For example, if the sample size is 5000, then 12,497,500 different π_{ij} must be stored.

The literature on unequal probability sampling without replacement is quite extensive. Hanif and Brewer (1983) reviewed about 50 methods, but none addresses all of the above disadvantages. Some of the schemes are only for sample sizes of two, the largest n sampling units have the same probability, or the sample size is not fixed (so the Yates-Grundy-Sen sample variance is inappropriate). Sampford’s (1967) method addresses disadvantages (2) and (3). But it samples with replacement and resamples if duplicates are present. This means that the expected numbers of samples drawn to obtain a sample that is kept could be quite large and unacceptable. Hartley and Rao’s (1962) method addresses all the disadvantages and even has a variance formula which does not include π_{ij} , but their method is not strictly PPS. Other well known references to sampling with unequal probabilities are by Hanurav (1966, 1967), Vijayan (1968), and Sunter (1977, 1986).

Tillé (1996) addresses disadvantages (1) and (2). We extended his work to address disadvantages (3) and (4). His approach to (3) is impractical for large population sizes, but our approach to (3) is easy and practical for large population sizes. Our modification of Tillé’s method makes PPS fixed sample size sampling a reality for large population sizes that is very easy to implement in production.

In Section II we briefly explain Tillé’s sampling procedure. In Section III we present a modification which guarantees all $\pi_{ij} > 0$ while ensuring $\pi_{ij} \leq \pi_i\pi_j$ if $i \neq j$. In Section IV we exploit many properties of Tillé’s sampling scheme to derive a new variance formula which eliminates storing n (n − 1) / 2 unique π_{ij} values. Section V presents conclusions.

II. Tillé’s sampling procedure

Tillé’s sampling procedure is a sample rejection procedure. One starts with the entire population and then rejects sampling units one by one from the sample with each iteration. At the end of each iteration the inclusion probabilities, $\pi(i|k)$, are the probabilities of selecting unit *i* given that the sample size equals k. The $\pi(i|k)$ are proportional to an auxiliary variable, x_i , with the exception of the inclusion probabilities greater than one. In that case the inclusion probabilities equal one. The rejection probability r_{ki} is the probability of rejecting the *i*th sampling unit when k sampling units will be left after this current iteration:

$$r_{ki} = 1 - \frac{\pi(i|k)}{\pi(i|k+1)} \quad , \quad n \leq k \leq N-1 \qquad \textbf{(4)}$$

This process continues until k = n. For his procedure $\pi(\mathrm{i}|\mathrm{N}) = 1$ and $\pi(\mathrm{i}|\mathrm{n}) = \pi_\mathrm{i}$ for all *i*.

Tillé also provides the formula for the probability of selecting together a cluster of m units $\mathrm{i}_1, \ldots, \mathrm{i}_\mathrm{m}$:

$$\pi(i_1,\ldots,i_m|n) \quad = \quad \prod_{k=n}^{N-1}\left(1 \; - \; \sum_{L=1}^{m} r_{ki_L}\right) \qquad \textbf{(5)}$$

III. New method to guarantee $\pi_{ij} > 0$

In our modification of Tillé's method, we begin by drawing a PPS fixed size sample according to his method, but not using the complementary procedure in section 6 of Tillé's paper. The population size is N, and the sample size is n. Our modification requires that the sampling units be sorted in ***ascending*** order of their probability of selection, π_i. Let us now define a variable q which means that no pairs of the first q sampling units could be selected pairwise with each other. There are q (q−1) / 2 unique π_{ij} which equal zero for $\mathrm{j} < i \leq \mathrm{q}$. When q = 0 then all $\pi_{\mathrm{ij}} > 0$ and no modification needs to be done.

We will give an example where q = 3, which means that no pairs of the first three sampling units could be selected together. The following probabilities from a population of size 8 would yield a sample size of 4: .05, .10, .15, .70, .72, .74, .76, .78. Since we sorted the probabilities of selection in ascending order, the zero joint probabilities in the joint probability matrix form two triangles in the upper left hand corner. We change the values of some π_{ij} , identified in bold italics, so that no zero joint probability exists. Then we can come up with a sampling scheme that will account for these probabilities, and be practical for large populations. The probability matrix is:

0.0500	***0.0000***	***0.0000***	***0.0250***	***0.0275***	0.0300	0.0325	0.0350
0.0000	0.1000	***0.0000***	***0.0500***	***0.0550***	0.0600	0.0650	0.0700
0.0000	***0.0000***	0.1500	***0.0750***	***0.0825***	0.0900	0.0975	0.1050
0.0250	***0.0500***	***0.0750***	0.7000	***0.4650***	0.4800	0.4950	0.5100
0.0275	***0.0550***	***0.0825***	***0.4650***	0.7200	0.4950	0.5100	0.5250
0.0300	0.0600	0.0900	0.4800	0.4950	0.7400	0.5250	0.5400
0.0325	0.0650	0.0975	0.4950	0.5100	0.5250	0.7600	0.5550
0.0350	0.0700	0.1050	0.5100	0.5250	0.5400	0.5550	0.7800

Let α be the value by which we will increase $\pi_{q+1,\, q+2}$. In this example, α = .009375. Decrease each π_{ij} (if q+1 $\leq i \leq$ q+2 and $j \leq$ q) by α / q =.003125. Finally, increase all the zero joint probabilities by 2α / (q (q−1)) =.003125. Section III B describes how α can be chosen. The rest of the joint probabilities will stay the same. Because of the symmetry of the matrix, $\pi_{ij} = \pi_{ji}$. The final probabilities are:

0.050000	0.003125	0.003125	0.021875	0.024375	0.030000	0.032500	0.035000
0.003125	0.100000	0.003125	0.046875	0.051875	0.060000	0.065000	0.070000
0.003125	0.003125	0.150000	0.071875	0.079375	0.090000	0.097500	0.105000
0.021875	0.046875	0.071875	0.700000	0.474375	0.480000	0.495000	0.510000
0.024375	0.051875	0.079375	0.474375	0.720000	0.495000	0.510000	0.525000
0.030000	0.060000	0.090000	0.480000	0.495000	0.740000	0.525000	0.540000
0.032500	0.065000	0.097500	0.495000	0.510000	0.525000	0.760000	0.555000
0.035000	0.070000	0.105000	0.510000	0.525000	0.540000	0.555000	0.780000

The following conditions hold which guarantee fixed sample size of integer n: $\sum_i \pi_i = n$, $\sum_j \pi_{ij} = n\pi_i$, and $\sum_i \sum_j \pi_{ij} = n^2$. Also, $\pi_{ij} \leq \pi_i\pi_j$ and $\pi_{ij} > 0$ if $i \neq j$, which guarantee nonnegative unbiased sample variances. It is easy to show with Tillé's method that there can be a range of α which guarantee that these prior conditions hold.

The following summarizes the changes in joint probabilities:

Increase π_{ij} by	α	if i = q + 2 and j = q + 1
Increase π_{ij} by	2α / (q (q−1))	if $j < i \leq$ q
Decrease π_{ij} by	α / q	if q+1 $\leq i \leq$ q+2 and $j \leq$ q
π_{ij} remains the same		otherwise

A. Sample modification

In the above example Tillé's sampling scheme would yield 35 possible outcomes. We will increase this by 9 more outcomes to a total of 44 possible outcomes. This is done systematically in a way that could easily be applied to large populations.

The first step is to sample according to Tillé's method. The sample drawn will be retained without alteration unless one of the following conditions occurs:

1. Sampling units q + 1, not q + 2, and any of the first q are selected, or
2. Sampling units q + 2, not q + 1, and any of the first q are selected.

If either condition occurs, proceed as follows.
Let a be the index of one of the first q sampling units selected. Let $\pi_{a,\, q+1,\, \overline{q+2}}$ be the probability

of selecting unit q + 1, not selecting unit q + 2, and selecting unit *a*. Then $\pi_{a,\, q+1,\, \overline{q+2}} \;=\; \pi_{a,\, q+1} - \pi_{a,\, q+1,\, q+2}$. $\pi_{a,\, q+1,\, q+2}$ is easy to calculate using Tillé's method.

If condition 1 occurs then retain the sample with probability $1 - \alpha / (q\ \pi_{a,\, q+1,\, \overline{q+2}})$. Replace unit *a* with unit q + 2 with probability $\alpha / (2\ q\ \pi_{a,\, q+1,\, \overline{q+2}})$. Replace unit q + 1 with one of the remaining first q − 1 units not selected (from the first q sampling units in the population) with probability $\alpha / (2\ q\ \pi_{a,\, q+1,\, \overline{q+2}})$. The selection of one of the remaining q − 1 sampling units is with equal probability.

If condition 2 occurs then retain the sample with probability 1 − ξ, where

$$\xi \quad = \quad \frac{\alpha}{q\ \pi_{a,q+1,\overline{q+2}}} \cdot \frac{\pi_{a,q+1,m_3,\ldots,m_n}}{\pi_{a,q+2,m_3,\ldots,m_n}} \qquad \textbf{(6)}$$

where m_3 through m_n are the third through n^{th} sampling units that are selected. In order to maintain precision the following method should be used to calculate a portion of the term:

$$\frac{\pi_{a,q+1,m_3,\ldots,m_n}}{\pi_{a,q+2,m_3,\ldots,m_n}} \quad = \quad \prod_{k=n}^{N-1} \left(\frac{1 \;-\; \sum\limits_{i \in a,q+1,m_3,\ldots,m_n} r_{ki}}{1 \;-\; \sum\limits_{i \in a,q+2,m_3,\ldots,m_n} r_{ki}} \right) \qquad \textbf{(7)}$$

This eliminates having to calculate the probability of selecting a particular sample which could be extremely small and will inevitably lead to loss of precision. Replace unit *a* with unit q + 1 with probability ξ / 2. Replace unit q + 2 with one of the remaining first q − 1 units not selected (from the first q sampling units in the population) with probability ξ / 2. The selection of one of the remaining q − 1 sampling units is with equal probability.

Table 1 displays the probability space when this method is used. The columns labeled S_1, S_2, S_3, and S_4 state what sampling units were selected from the population. For example, in the first row $S_1 = 1$, $S_2 = 4$, $S_3 = 5$, and $S_4 = 6$. This means that the first, fourth, fifth, and sixth sampling units were selected, expressed as {1,4,5,6}. π_S is the probability of selecting this sample using Tillé's method. The column labeled π_S *modified* is the probability of selecting the sample using the sampling modification mentioned above. The column labeled *difference* is (π_S *modified*) − π_S.

Notice the pattern in the column labeled *difference*. This is by design, not by accident, because the modification exploits another unique property of Tillé's method,

$$\frac{\pi_{1,m_2,m_3,\ldots,m_n}}{\pi_1} \quad = \quad \frac{\pi_{a,m_2,m_3,\ldots,m_n}}{\pi_a} \;, \qquad a \leq q \qquad \textbf{(8)}$$

Let M be the set of all subsets of $\{q + 3, \dots , N\}$ of cardinality $n - 2$. Let **m** be a set $\{m_3, m_4, \dots, m_n\}$ where $m_i \geq q + 3$. When the *difference* is negative then it may be expressed as $\Delta_{a,\, q+1,\, \mathbf{m}} = \pi_{a,\, q+1,\, \mathbf{m}}\, [\alpha / (q\, \pi_{a,\, q+1,\, \overline{q+2}})]$ if condition 1 happens, or $\Delta_{a,\, q+2,\, \mathbf{m}} = \pi_{a,\, q+2,\, \mathbf{m}}\, [\alpha / (q\, \pi_{a,\, q+1,\, \overline{q+2}})\,]\, [\pi_{a,\, q+1,\, \mathbf{m}} / \pi_{a,\, q+2,\, \mathbf{m}}]$ if condition 2 happens. Then $\Delta_{a,\, q+1,\, \mathbf{m}} = \Delta_{a,\, q+2,\, \mathbf{m}}$. It can also be seen from equation **(8)** that $\Delta_{a,\, q+1,\, \mathbf{m}} = \Delta_{b,\, q+1,\, \mathbf{m}}$ when both *a* and *b* are less than or equal to q , because $\pi_{a,\, q+1,\, \mathbf{m}}$ is proportional to π_a. Another identity with this method is: $\sum_{\mathbf{m} \in M} \Delta_{a,\, q+1,\, \mathbf{m}} = \alpha / q$, because $\sum_{\mathbf{m} \in M} \pi_{a,\, q+1,\, \mathbf{m}} = \pi_{a,\, q+1,\, \overline{q+2}}$. These identities explain the changes in the revised joint probability matrix. $\pi_{q+1,\, q+2}$ increases by α, π_{ij} increases by $2\alpha / [q\, (q - 1)]$ if *i* and $j \leq q$, π_{ij} decreases by α / q if $q + 1 \leq i \leq q + 2$ and $j \leq q$, and all other probabilities stay the same. See the Appendix for the proof.

B. Choosing α

This section discusses the acceptable range for α. Keep in mind we want to guarantee that $\pi_{ij} \leq \pi_i \pi_j$ and $\pi_{ij} > 0$ if $i \neq j$.

1. To guarantee new $\pi_{q+1,\, q+2} \leq \pi_{q+1}\, \pi_{q+2}$ then $\alpha \leq \pi_{q+1}\, \pi_{q+2} - \pi_{q+1,\, q+2}$

2. To guarantee new $\pi_{ij} \leq \pi_i \pi_j$, $j < i \leq q$, then $\alpha \leq \pi_1 \pi_2\, q\, (q - 1) / 2$

3. To guarantee new $\pi_{q+1,1} > 0$ then $\alpha < q\, \pi_{q+1,\, 1}$

4. To guarantee any new $\pi_S \geq 0$ then $\alpha \leq q\, \pi_{1,\, q+1,\, \overline{q+2}}$

An acceptable range of α would always be a value greater than zero and a value less than the minimum value of all the above conditions. If the minimum value of all the above conditions did not come from condition 3 then α may be less than or equal to the minimum value.

We discuss below three alternative schemes of choosing α where α is in this acceptable range. It is a judgement call by the user for choosing α, but α must be chosen before sample selection, in the acceptable range.

Scheme 1

Scheme 1 ensures that if $\pi_j < \pi_i$ then $\pi_{jk} \leq \pi_{ik}$, $i \neq j$, $i \neq k$, $j \neq k$. This is accomplished by requiring that in addition to the first four conditions the following must be true as well.

5. $\alpha \leq \pi_{q+1,\, 1}\, q\, (q - 1) / (q + 1)$ and

6. $\alpha \leq \pi_{q+1,\, k} - \pi_{q+1,\, q+2}$ where *k* is the smallest *k* that meets the criteria $k > q+2$ and $\pi_{k,\, q+1} - \pi_{q+1,\, q+2} > 0$. If *k* is not defined under these criteria then we may disregard condition 6.

Condition 5 also guarantees condition 3, therefore α is the minimum value of conditions 1, 2, 4, and 5 (and 6 if needed).

Scheme 2

Scheme 2 ensures that $\beta_{ij} \geq \beta_{i+1,\,j}$ with the exception of $\beta_{q+1,\,q+2}$ for all values of β_{ij} from a sample. This is accomplished by requiring that in addition to conditions 1 and 4 the following must be true:

5. $\alpha \leq \delta\, q\, (q - 1) / 2$

where

$$\delta = \frac{\pi_{q+1,1}}{\left(\dfrac{\pi_{q+1}}{\pi_2} + \dfrac{q-1}{2}\right)} \qquad \textbf{(9)}$$

Condition 5 also guarantees conditions 2 and 3. Therefore α in this case is the minimum value of conditions 1, 4 and 5. It was this scheme that was used to determine α in the example.

Scheme 3

To minimize the effects that large values of β_{ij} can have on the sample variance, the minimax rule may be employed. This method chooses the α that minimizes the maximum β_{ij} over all α subject to $\beta_{ij} \geq 1$ for every *i* and *j*. The α that minimizes the maximum β_{ij} is:

5. $$\alpha = \frac{\pi_{q-1}\ \pi_q\ \pi_{1,\,q+1}\ q\ (q-1)}{2\ \pi_1\ \pi_{q+1} + \pi_{q-1}\ \pi_q\ (q-1)}$$

When we include the restriction $\beta_{ij} \geq 1$ then we must also include conditions 1 and 2 since condition 5 only guarantees condition 3. Condition 4 must also be guaranteed regardless of the restriction $\beta_{ij} \geq 1$. Therefore, α in this case is the minimum of conditions 1, 2, 4, and 5.

C. Variance Considerations

We will now discuss how much the variance changes from modifying the sample procedure. The change in variance is $\sigma^2(\hat{Y}_{\text{modified}}) - \sigma^2(\hat{Y}_{\text{Tillé}}) =$

$$\sum_{i=q+1}^{q+2} \sum_{j=1}^{q} \frac{\alpha}{q} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 - \sum_{i=2}^{q} \sum_{j=1}^{i-1} \frac{2\alpha}{q\ (q-1)} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 - \alpha \left(\frac{y_{q+2}}{\pi_{q+2}} - \frac{y_{q+1}}{\pi_{q+1}} \right)^2$$

Whether or not the actual variance increases or decreases will depend not only on α but on the values of y_i . It is the first double sum that increases the variance while the rest of the terms decrease the variance which means that there is some effect of canceling out. It should also be noted that α is very small because of the restrictions on its bounds, therefore any change would be small relative to the overall variance.

The bias in the sample variance when Tillé's method contains some $\pi_{ij} = 0$ is:

$$- \sum_{i=2}^{q} \sum_{j=1}^{i-1} \pi_i\ \pi_j \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

The amount of bias depends on the values of y_i .

We will now address how much change there is in the variance of the variance because of the sample modification. With Tillé's method as well as the sample modification, there are at most 3 sampling units that can be selected from the first q+2 sampling units no matter how large q is. This is because not all possible samples of size n are possible with this method as with other methods such as Sampford's. With the modification the change in the sample variance will come from at most 3 terms out of n(n−1)/2 terms. This should not significantly affect the variance of the variance when n is large.

The Yates-Grundy-Sen sample variance can be expressed in the following form:

$$\hat{\sigma}^2_{YGS} = \sum_{u=1}^{\tilde{N}} t_u \frac{x_u}{\pi_u}$$

where

$u = (i-1)\ (i-2)\ /\ 2 \ + \ j$,	$j < i$ and $i \geq 2$
$\tilde{N} = N\ (N-1)\ /\ 2$	
$\pi_u = \pi_{ij}$	
$t_u = 1$	if i and j sampling units selected, 0 otherwise

$$x_u = (\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Therefore the variance of the Yates-Grundy-Sen sample variance is:

$$\sigma^2(\hat{\sigma}^2_{YGS}) = \sum_{u=2}^{\tilde{N}}\sum_{v=1}^{u-1}\left(\pi_u\pi_v - \pi_{uv}\right)\left(\frac{x_u}{\pi_u} - \frac{x_v}{\pi_v}\right)^2$$

Note that:

$$\sum_{u=1}^{\tilde{N}} \pi_u = \frac{n(n-1)}{2} = \tilde{n} \quad ,$$

$$\sum_{u=2}^{\tilde{N}}\sum_{v=1}^{u-1}\pi_{uv} = \frac{\tilde{n}(\tilde{n}-1)}{2} \qquad and \qquad \sum_{v=1}^{\tilde{N}}\pi_{uv} = \tilde{n}\pi_u$$

With the sample modification some π_u will decrease thus decreasing the variance of the variance while other π_u will increase thus increasing the variance of the variance. Also some π_{uv} will increase thus decreasing the variance of the variance while other π_{uv} will decrease thus increasing the variance of the variance. Again we are left with a canceling out effect. Whether or not the variance of the variance will actually increase or decrease will depend not only on α but on the values of y_i as well.

IV. New Variance formula

It can be proven that when Tillé's method is used then $\beta_{ij} \geq \beta_{i\,(j+1)}$ and $\beta_{ij} \geq \beta_{(i+1)\,j}$ given *j* < *i*. We also discovered when we reviewed Tillé's method that for many values $\beta_{ij} = \beta_{i1}$ for *j* < *i*. We were later able to prove why this is the case when the population is large and the π_i are skewed. In the Appendix we will show when $\beta_{ij} = \beta_{i1}$ as well as when $\beta_{ij} \neq \beta_{i1}$ given *j* < *i*. The proof that $\beta_{ij} \geq \beta_{i\,(j+1)}$ and $\beta_{ij} \geq \beta_{(i+1)\,j}$ is similar to identifying when $\beta_{ij} = \beta_{i1}$.

In **Table 2** are listed 62 π_i and y_i , where n = 20, which were randomly generated with a skewed distribution. **Table 3** is a frequency table that outputs the number of times there was a unique value of β_{ij} for a given *i*. For example, when *i* = 62 then $\beta_{ij} = \beta_{i1} = 1.0020947368$ for all *j* < *i*.

We will use these properties of the β_{ij} along with the following statistical identity in the derivation of the new variance formula. See the Appendix for the proof of this identity.

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = n\sum_{j=1}^{n}\left(\frac{y_j}{\pi_j}\right)^2-\left(\sum_{j=1}^{n}\frac{y_j}{\pi_j}\right)^2 \qquad \textbf{(10)}$$

Since most $\beta_{ij} = \beta_{i1}$ and the variability of β_{ij} and is very small for a fixed *i* and *j* < *i*, and since $\beta_{ij} \geq \beta_{i\,(j+1)}$, if we substitute β_{i1} for all β_{ij}, the Yates-Grundy-Sen sample variance will be slightly overestimated. If we allow the subscript 1 to indicate the lowest index from the sample selected as opposed to the lowest index of the population then the sample variance may be closer yet to the Yates-Grundy-Sen sample variance. With this substitution the Yates-Grundy-Sen variance formula reduces to:

$$\hat{\sigma}^2_{YGS} \leq \sum_{i=1}^{n}\gamma_i\left(i\sum_{j=1}^{i}\left(\frac{y_j}{\pi_j}\right)^2-\left(\sum_{j=1}^{i}\frac{y_j}{\pi_j}\right)^2\right) \qquad \textbf{(11)}$$

where

$$\begin{aligned} \gamma_i &= 0 && \text{if} \quad i = 1 \\ &= \beta_{i1} - \beta_{(i+1)\,1} && \quad 1 < i < n \\ &= \beta_{i1} - 1 && \quad i = n \end{aligned}$$

A proof is in the Appendix.

If the sample modification took place then it is possible for some γ_i to be negative, but the overall sample variance will be nonnegative because this variance is greater than or equal to the Yates-Grundy-Sen sample variance which is nonnegative under this sampling scheme.

The variance from the data set in **Table 2** is 1,597,337.92. The bias of **(11)** is 800.11, .05% of the variance.

We wrote a computer program that calculates the β_{ij} that are different from β_{i1} from the sample. This can be done without having to calculate all π_{ij} from the sample. It is important to know that there are a small number of such cases. One may get the same result as the Yates-Grundy-Sen sample variance using the following formula which has fewer calculations than equation **(3)**.

$$\hat{\sigma}^2_{YGS} = \sum_{i=1}^{n}\gamma_i\left(i\sum_{j=1}^{i}\left(\frac{y_j}{\pi_j}\right)^2-\left(\sum_{j=1}^{i}\frac{y_j}{\pi_j}\right)^2\right) - \sum_{\beta_{ij}\neq\beta_{i1}}\left(\beta_{i1}-\beta_{ij}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \qquad \textbf{(12)}$$

We believe that for large populations where π_i are skewed the second term is very small and therefore **(12)** will be slightly less than **(11)**.

Advantages of **(11)**:

1. The pairing of the joint terms for indexes *i* and *j* is eliminated.
2. Calculating all π_{ij} from the sample is eliminated.
3. The number of mathematical computations are greatly reduced.

Disadvantage:

> If one is interested in publishing the standard error as opposed to the relative standard error, RSE, then this formula could result in a loss of precision. If RSEs are published as opposed to the standard errors then this should not be a problem since the Census Bureau publishes at most one decimal of a percent. If too much precision is lost in calculating the variance using **(11)** then the RSEs calculated from both **(3)** and **(11)** would be the same when rounded to the first decimal of a percent, i.e. 0.0.

V. Conclusions

We believe that when sampling is PPS then this method is beneficial, and we recommend it for consideration because of the following reasons:

1. Is easy to implement.
2. Makes it easy to calculate the terms π_{i1} used in variance estimation with high precision.
3. Has lower variance than Poisson PPS sampling.

The new sample variance formula:

1. Does not require deriving $n(n-1)/2$ different π_{ij} .
2. Does not require having to pair π_{ij} with the ith and jth sampling units.
3. Has fewer computations than the Yates-Grundy-Sen sample variance with little increase in bias.

Table 1 Probability Space for Example on page 5

				Probability of Selecting Sample		
Index of sample units selected				Tillé procedure	with modification	π_S modified − π_S
S_1	S_2	S_3	S_4	π_S	π_S modified	difference
1	4	5	6	0.00250	0.00250	0.00000
1	4	5	7	0.00333	0.00333	0.00000
1	4	5	8	0.00417	0.00417	0.00000
1	4	6	7	0.00417	0.00330	-0.00087
1	4	6	8	0.00500	0.00396	-0.00104
1	4	7	8	0.00583	0.00462	-0.00122
1	5	6	7	0.00500	0.00413	-0.00087
1	5	6	8	0.00583	0.00479	-0.00104
1	5	7	8	0.00667	0.00545	-0.00122
1	6	7	8	0.00750	0.00750	0.00000
2	4	5	6	0.00500	0.00500	0.00000
2	4	5	7	0.00667	0.00667	0.00000
2	4	5	8	0.00833	0.00833	0.00000
2	4	6	7	0.00833	0.00747	-0.00087
2	4	6	8	0.01000	0.00896	-0.00104
2	4	7	8	0.01167	0.01045	-0.00122
2	5	6	7	0.01000	0.00913	-0.00087
2	5	6	8	0.01167	0.01063	-0.00104
2	5	7	8	0.01333	0.01212	-0.00122
2	6	7	8	0.01500	0.01500	0.00000
3	4	5	6	0.00750	0.00750	0.00000
3	4	5	7	0.01000	0.01000	0.00000
3	4	5	8	0.01250	0.01250	0.00000
3	4	6	7	0.01250	0.01163	-0.00087
3	4	6	8	0.01500	0.01396	-0.00104
3	4	7	8	0.01750	0.01628	-0.00122
3	5	6	7	0.01500	0.01413	-0.00087
3	5	6	8	0.01750	0.01646	-0.00104
3	5	7	8	0.02000	0.01878	-0.00122
3	6	7	8	0.02250	0.02250	0.00000
4	5	6	7	0.13000	0.13260	+0.00260
4	5	6	8	0.13500	0.13813	+0.00313
4	5	7	8	0.14000	0.14365	+0.00365
4	6	7	8	0.14500	0.14500	0.00000
5	6	7	8	0.15000	0.15000	0.00000
1	2	6	7	0.00000	0.00087	+0.00087
1	2	6	8	0.00000	0.00104	+0.00104
1	2	7	8	0.00000	0.00122	+0.00122
1	3	6	7	0.00000	0.00087	+0.00087
1	3	6	8	0.00000	0.00104	+0.00104
1	3	7	8	0.00000	0.00122	+0.00122
2	3	6	7	0.00000	0.00087	+0.00087
2	3	6	8	0.00000	0.00104	+0.00104
2	3	7	8	0.00000	0.00122	+0.00122

Table 2 Randomly Generated Data Set

π_i	y_i	π_i	y_i	π_i	y_i
0.0101	7	0.1696	122	0.4409	206
0.0191	12	0.1738	272	0.4442	641
0.0245	36	0.1787	145	0.4547	419
0.0253	8	0.2156	212	0.4547	349
0.0262	33	0.2207	379	0.4859	319
0.0288	46	0.2340	313	0.4966	687
0.0394	11	0.2395	282	0.5060	208
0.0431	36	0.2481	255	0.5381	412
0.0499	31	0.2645	95	0.5436	324
0.0720	61	0.2704	301	0.5446	462
0.0783	65	0.2740	155	0.6102	841
0.0826	121	0.2840	204	0.6377	906
0.0827	90	0.2974	519	0.7191	1112
0.0889	122	0.2995	231	0.7242	383
0.0983	60	0.3030	210	0.7458	1166
0.1090	102	0.3066	424	0.7475	371
0.1190	205	0.3197	165	0.7615	659
0.1230	91	0.3616	141	0.8161	828
0.1232	119	0.4000	534	0.9080	1028
0.1530	85	0.4194	547	0.9602	1587
0.1536	186	0.4303	266		

Table 3 Frequency Count of Unique β_{ij} by *i*

i	π_i	β_{ij}	Frequency Count
2	0.0191	4.4754551738	1
3	0.0245	1.8198425576	1
3	0.0245	2.0058101943	1
4	0.0253	1.7745948291	1
4	0.0253	1.7962045080	1
4	0.0253	1.8892131989	1
5	0.0262	1.7697085136	1
5	0.0262	1.7699378544	1
5	0.0262	1.7719910584	1
5	0.0262	1.7803407545	1
6	0.0288	1.7133755781	5
7	0.0394	1.5413843457	6
8	0.0431	1.4974386940	1
8	0.0431	1.4987854552	6
9	0.0499	1.4397153396	8
10	0.0720	1.2965254581	9
11	0.0783	1.2587337055	1
11	0.0783	1.2661367368	9
12	0.0826	1.2471920370	1
12	0.0826	1.2479543829	1
12	0.0826	1.2487235589	9
13	0.0827	1.2467289480	1
13	0.0827	1.2470981174	1
13	0.0827	1.2477191152	1
13	0.0827	1.2483455343	9
14	0.0889	1.2322974972	13
15	0.0983	1.2124422016	14
16	0.1090	1.1928582820	1
16	0.1090	1.1931013533	14
17	0.1190	1.1778995311	16
18	0.1230	1.1714591303	1
18	0.1230	1.1725813602	16
19	0.1232	1.1710044845	1
19	0.1232	1.1713728271	1
19	0.1232	1.1723257321	16
20	0.1530	1.1410588275	19
21	0.1536	1.1393759878	1
21	0.1536	1.1405076651	19
22	0.1696	1.1273170058	21
23	0.1738	1.1243445857	22
24	0.1787	1.1204317467	1
24	0.1787	1.1211387443	22
25	0.2156	1.0994125956	24
26	0.2207	1.0951385228	1
26	0.2207	1.0967027780	24
27	0.2340	1.0903049069	26
28	0.2395	1.0867826258	1
28	0.2395	1.0879156211	26
29	0.2481	1.0843286003	1
29	0.2481	1.0843671329	1
29	0.2481	1.0844110173	26
30	0.2645	1.0784369119	29
31	0.2704	1.0765059541	30
32	0.2740	1.0750618382	1
32	0.2740	1.0754147411	30
33	0.2840	1.0722173942	1
33	0.2840	1.0724393033	1
33	0.2840	1.0725393814	30
34	0.2974	1.0690619995	33
35	0.2995	1.0676448300	1
35	0.2995	1.0685507964	33
36	0.3030	1.0671214011	1
36	0.3030	1.0672282955	1
36	0.3030	1.0677155911	33
37	0.3066	1.0667709912	1
37	0.3066	1.0667952678	1
37	0.3066	1.0668101085	1
37	0.3066	1.0668777303	33
38	0.3197	1.0641870164	37
39	0.3616	1.0561229523	38
40	0.4000	1.0491893863	39
41	0.4194	1.0459349464	40
42	0.4303	1.0438822071	1
42	0.4303	1.0442098236	40
43	0.4409	1.0426411669	42
44	0.4442	1.0419833750	1
44	0.4442	1.0421974483	42
45	0.4547	1.0405966646	1
45	0.4547	1.0407403463	1
45	0.4547	1.0408308418	42
46	0.4547	1.0401536204	1
46	0.4547	1.0405966646	1
46	0.4547	1.0407403463	1
46	0.4547	1.0408308418	42
47	0.4859	1.0370895202	46
48	0.4966	1.0359031822	47
49	0.5060	1.0346720112	1
49	0.5060	1.0349032986	47
50	0.5381	1.0316537644	49
51	0.5436	1.0310794130	1
51	0.5436	1.0311435406	49
52	0.5446	1.0306375415	1
52	0.5446	1.0309913646	1
52	0.5446	1.0310519333	49
53	0.6102	1.0253104581	52
54	0.6377	1.0230310142	53
55	0.7191	1.0166263162	54
56	0.7242	1.0161136218	1
56	0.7242	1.0162519312	54
57	0.7458	1.0146275206	1
57	0.7458	1.0146943525	1
57	0.7458	1.0147259252	54
58	0.7475	1.0143338956	1
58	0.7475	1.0145366362	1
58	0.7475	1.0145862952	1
58	0.7475	1.0146097545	54
59	0.7615	1.0136936933	58
60	0.8161	1.0103502713	59
61	0.9080	1.0049880169	60
62	0.9602	1.0020947368	61

Frequency

Appendix

Proof that:

π_{ij} increases by	α	if $i = q + 2$ and $j = q + 1$
π_{ij} increases by	$2\alpha / (q\,(q-1))$	if $j < i \leq q$
π_{ij} decreases by	α / q	if $q+1 \leq i \leq q+2$ and $j \leq q$
π_{ij} does not change		otherwise

The following identities will be used in the proof:

$$\Delta_{a,\, q+1,\, \mathbf{m}} = \Delta_{a,\, q+2,\, \mathbf{m}}$$
$$\Delta_{a,\, q+1,\, \mathbf{m}} = \Delta_{b,\, q+1,\, \mathbf{m}}$$
$$\sum_{\mathbf{m} \in M} \Delta_{a,\, q+1,\, \mathbf{m}} = \alpha / q$$

Let π_{ij} be the joint probability using Tillé's method, and π^{*}_{ij} be the joint probability after the sample modification.

$\pi^{*}_{ij} = \pi_{ij} -$ P{i & j selected and not retained in sample} + P{i & j not selected and in sample}

1) $i > q + 2$ and $j > q + 2$

$$\pi^{*}_{ij} = \pi_{ij} - 0 + 0 = \pi_{ij}$$

2) $i > q + 2$ and $q + 1 \leq j \leq q + 2$

$$\pi^{*}_{i,\, q+1} = \pi_{i,\, q+1} - \tfrac{1}{2} \sum_{a=1}^{q} \sum_{i \in \mathbf{m} \in M} \Delta_{a,\, q+1,\, \mathbf{m}} + \tfrac{1}{2} \sum_{a=1}^{q} \sum_{i \in \mathbf{m} \in M} \Delta_{a,\, q+2,\, \mathbf{m}}$$

$$= \pi_{i,\, q+1} - \tfrac{1}{2} \sum_{a=1}^{q} \sum_{i \in \mathbf{m} \in M} \Delta_{a,\, q+1,\, \mathbf{m}} + \tfrac{1}{2} \sum_{a=1}^{q} \sum_{i \in \mathbf{m} \in M} \Delta_{a,\, q+1,\, \mathbf{m}} = \pi_{i,\, q+1}$$

Same procedure for $\pi_{i,\, q+2}$

3) $i > q + 2$ and $j \leq q$

$$\pi^{*}_{ij} = \pi_{ij} - \tfrac{1}{2} \sum_{b=q+1}^{q+2} \sum_{i \in \mathbf{m} \in M} \Delta_{j,\, b,\, \mathbf{m}} + \tfrac{1}{2} \left[\sum_{b=q+1}^{q+2} \sum_{i \in \mathbf{m} \in M} \sum_{a \neq j}^{q} \Delta_{a,\, b,\, \mathbf{m}} \right] / (q-1)$$

$$= \pi_{ij} - \tfrac{1}{2} \sum_{b=q+1}^{q+2} \sum_{i \in \mathbf{m} \in M} \Delta_{j,\, b,\, \mathbf{m}} + \tfrac{1}{2} \left[\sum_{b=q+1}^{q+2} \sum_{i \in \mathbf{m} \in M} \sum_{a \neq j}^{q} \Delta_{j,\, b,\, \mathbf{m}} \right] / (q-1) = \pi_{ij}$$

4) i = q + 2 and j = q + 1

$$\pi^*_{q+2,\,q+1} = \pi_{q+2,\,q+1} - 0 + ½ \textstyle\sum_{a=1}^{q} \sum_{\mathbf{m} \in M} \sum_{b=q+1}^{q+2} \Delta_{a,\,b,\,\mathbf{m}} = \pi_{q+2,\,q+1} + ½ \sum_{a=1}^{q} \sum_{\mathbf{m} \in M} 2\, \Delta_{a,\,q+1,\,\mathbf{m}}$$

$$= \pi_{q+2,\,q+1} + \textstyle\sum_{a=1}^{q} \sum_{\mathbf{m} \in M} \Delta_{a,\,q+1,\,\mathbf{m}} = \pi_{q+2,\,q+1} + \sum_{a=1}^{q} \alpha / q = \pi_{q+2,\,q+1} + \alpha$$

5) q + 1 ≤ i ≤ q + 2 and j ≤ q

$$\pi^*_{q+1,\,j} = \pi_{q+1,\,j} - \textstyle\sum_{\mathbf{m} \in M} \Delta_{j,\,q+1,\,\mathbf{m}} + 0 = \pi_{q+1,\,j} - \alpha / q$$

6) i ≤ q and j ≤ q

$$\pi^*_{ij} = 0 - 0 + ½ \textstyle\sum_{\mathbf{m} \in M} \sum_{b=q+1}^{q+2} (\Delta_{i,\,b,\,\mathbf{m}} + \Delta_{j,\,b,\,\mathbf{m}}) / (q - 1)$$

$$= ½ \textstyle\sum_{\mathbf{m} \in M} 2\, (\Delta_{i,\,q+1,\,\mathbf{m}} + \Delta_{j,\,q+1,\,\mathbf{m}}) / (q - 1)$$

$$= \textstyle\sum_{\mathbf{m} \in M} (\Delta_{i,\,q+1,\,\mathbf{m}} + \Delta_{j,\,q+1,\,\mathbf{m}}) / (q - 1)$$

$$= 2 \textstyle\sum_{\mathbf{m} \in M} \Delta_{i,\,q+1,\,\mathbf{m}} / (q - 1)$$

$$= 2\, \alpha / [q\, (q - 1)]$$

++

Proof of equation **(10)**

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

The summations on the left are for $j < i$ while the summations on the right just add the values of $(y_i / \pi_i - y_j / \pi_j)^2$ where $j \geq i$. Since $(y_i / \pi_i - y_j / \pi_j)^2 = (y_j / \pi_j - y_i / \pi_i)^2$ and $(y_i / \pi_i - y_i / \pi_i)^2 = 0$ the quantity on the right doubles, therefore multiplying by one half brings us back to the original total on the left.

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\left(\frac{y_i}{\pi_i} \right)^2 - 2 \frac{y_i y_j}{\pi_i \pi_j} + \left(\frac{y_j}{\pi_j} \right)^2 \right)$$

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{y_i}{\pi_i}\right)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{y_iy_j}{\pi_i\pi_j} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{y_j}{\pi_j}\right)^2\right)$$

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = \frac{1}{2}n\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}\right)^2 - \left(\sum_{i=1}^{n}\frac{y_i}{\pi_i}\right)\left(\sum_{j=1}^{n}\frac{y_j}{\pi_j}\right) + \frac{1}{2}n\sum_{j=1}^{n}\left(\frac{y_j}{\pi_j}\right)^2$$

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = n\sum_{j=1}^{n}\left(\frac{y_j}{\pi_j}\right)^2 - \left(\sum_{j=1}^{n}\frac{y_j}{\pi_j}\right)^2 \qquad \textbf{(10)}$$

+++

Derivation of equation **(11)**

$$\hat{\sigma}^2_{YGS} = \sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = \sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\beta_{ij}-1\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

If we substitute β_{i1} for β_{ij} then we get:

$$\hat{\sigma}^2_{YGS} \le \sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\beta_{i,1}-1\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

$$= \left(\beta_{n,1}-1\right)\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\beta_{i,1}-\beta_{n,1}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

$$= \gamma_n\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 + \sum_{i=2}^{n-1}\sum_{j=1}^{i-1}\left(\beta_{i,1}-\beta_{n,1}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Note the index change on the last term. This is because when i = n then $(\beta_{i,1} - \beta_{n,1}) = 0$. The above expression can be rewritten as:

$$\gamma_n \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \left(\beta_{n-1,1} - \beta_{n,1}\right) \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \left(\beta_{i,1} - \beta_{n-1,1}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \gamma_n \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \gamma_{n-1} \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \sum_{i=2}^{n-2} \sum_{j=1}^{i-1} \left(\beta_{i,1} - \beta_{n-1,1}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \gamma_n \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \gamma_{n-1} \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \cdots + \gamma_2 \sum_{i=2}^{2} \sum_{j=1}^{i-1} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Let $\gamma_1 = 0$, and then by equation **(10)**,

$$\hat{\sigma}^2_{YGS} \leq \sum_{i=1}^{n} \gamma_i \left(i \sum_{j=1}^{i} \left(\frac{y_j}{\pi_j} \right)^2 - \left(\sum_{j=1}^{i} \frac{y_j}{\pi_j} \right)^2 \right) \qquad \textbf{(11)}$$

****

When does $\beta_{ij} = \beta_{i1}$ where $j < i$?

By definition:

$$\beta_{ij} = \frac{\pi_i \pi_j}{\pi_{ij}} = \frac{\prod_{k=n}^{N-1} \left(1 - r_{ki} \right) \prod_{k=n}^{N-1} \left(1 - r_{kj} \right)}{\prod_{k=n}^{N-1} \left(1 - r_{ki} - r_{kj} \right)} = \prod_{k=n}^{N-1} \frac{\left(1 - r_{ki} \right)\left(1 - r_{kj} \right)}{\left(1 - r_{ki} - r_{kj} \right)}$$

From Tillé’s paper we have:

$$\begin{aligned} r_{ki} &= 0 && \text{if } i \in A_k \\ &= 1 - \pi(i|k) && \text{if } i \in B_k \\ &= \frac{1 - \sum_{i \in B_k} \left(1 - \pi(i|k) \right)}{k + 1 - \#A_k - \#B_k} && if\ i \in C_k \end{aligned}$$

Where $A_k = \{i : \pi(i|k) = 1\}$
$B_k = \{i : \pi(i|k) < 1 \text{ and } \pi(i|k+1) = 1\}$

$C_k = \{i : \pi(i|k+1) < 1\}$

When $i \in A_k$ then

$$\frac{(1 - r_{ki})(1 - r_{kj})}{(1 - r_{ki} - r_{kj})} = \frac{(1 - r_{ki})(1 - r_{k1})}{(1 - r_{ki} - r_{k1})} = 1$$

because $r_{ki} = 0$.

When $j \in C_k$ then

$$\frac{(1 - r_{ki})(1 - r_{kj})}{(1 - r_{ki} - r_{kj})} = \frac{(1 - r_{ki})(1 - r_{k1})}{(1 - r_{ki} - r_{k1})}$$

because $r_{kj} = r_{k1}$. Therefore when $j < i$ and $i \in B_k$ when $j \in C_k$ then $\beta_{ij} = \beta_{i1}$. Otherwise if $i \in B_k$ and $j \in B_k$ then

$$\frac{(1 - r_{ki})(1 - r_{kj})}{(1 - r_{ki} - r_{kj})} < \frac{(1 - r_{ki})(1 - r_{k1})}{(1 - r_{ki} - r_{k1})}$$

and $\beta_{ij} < \beta_{i1}$.

Acknowledgment

The authors thank the Research and Methodology Branch of the Manufacturing and Construction Division of the Census Bureau for their review and comments on this paper.

Reference List

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

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

Hanurav, T. V. (1966) "Optimum πPS Sampling." *Annals of Mathematical Statistics*, **37**. 1859.

Hanurav, T. V. (1967) "Optimum Utilization of Auxiliary Information: πps Sampling of Two Units from a Stratum." *Journal of the Royal Statistical Society*, **B**, **29**. 374-391.

Hartley, H. O. and Rao J. N. K. (1962) "Sampling with Unequal Probabilities and without Replacement." *Annuals of Mathematical Statistics*, **33**, 350-374.

Sampford, M. R. (1967) "On Sampling Without Replacement with Unequal Probabilities of Selection." *Biometrika*, **54**, 3 and 4. 499-513.

Sunter, A. (1977) "List Sequential Sampling with Equal or Unequal Probabilities without Replacement." *Applied Statistics*, **26**, 3. 261-268.

Sunter, A. (1986) "Solutions to the Problem of Unequal Probability Sampling Without Replacement." *International Statistical Review*, **54**. 33-50.

Tillé, Y. (1996) "An Elimination Procedure for Unequal Probability Sampling Without Replacement." *Biometrika*, **83**, 1. 238-241.

Vijayan, K. (1968) "An Exact πps Sampling Scheme - Generalization of a Method of Hanurav." *Journal of the Royal Statistical Society*, **B**, **30**. 556-566.