Properties and Modifications of a Probability Proportional to Size Sampling Procedure

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

Abstract

There are numerous procedures for selecting a sample PPS without replacement. Brewer and Hanif (1983) discuss the properties of 50 such procedures. One of the more interesting of these procedures originally appeared in Jessen (1969). This procedure possesses several desirable properties including: relative simplicity of sample selection, fixed sample size of any desired size, selection strictly PPS without replacement, unbiased estimators of variance, and simple calculation of joint inclusion probabilities. In this paper we study possible modifications of Jessen's procedure to obtain two other desirable properties, namely: (1) how to obtain a sample expansion with the expanded sample remaining PPS without replacement; and (2) how to adjust the joint selection probabilities in order to exclude the possibility of negative variance estimates when using the Yates-Grundy variance estimator.

Keywords: PPS sampling, Unbiased estimators of variance, Sample expansion, Jessen's method

1. Introduction

A summary of the paper is given in the Abstract above. In Section 2 we present some notation and terminology. In Section 3 we describe Jessen's method. In Section 4 we briefly discuss how to subsample a sample selected using Jessen's method, with the resulting subsample yielding a PPS subsample of the original universe. There is nothing really new here since the same type of subsampling would apply to the subsampling of a sample selected with any PPS sampling procedure. In Sections 5-7 we discuss (1) in the Abstract. Section 5 points out the difficulties in selecting a sample PPS in a way that allows for an expansion of the sample with the expanded sample selected PPS from the universe. Section 6 describes how to modify Jessen's method so that a PPS sample expansion is possible and Section 7 describes how to actually do the sample expansion. Finally in Section 8 we consider (2) in the Abstract.

2. Notation and Terminology

Consider a universe of N units, with an associated nondecreasing sequence of numbers M_i, $i=1,...,N$, where M_i denotes the measure of size for unit i. Then

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

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, where $n < N$, denoted $\pi_i(n)$, $i=1,...,N$, are obtained as follows. Initially let

$$\pi_i(n) = np_i, \ i=1,...,N \qquad (2.2)$$

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 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. Let $c(n)$ denote the number of certainty units.

For any PPS sampling procedure, any pair of units in the universe i,j, and any possible sample s of n units, we let $\pi_{ij}(n), \pi_s(n)$ denote, respectively, the probability of selection of the pair i,j and the sample s.

3. Jessen's Method

Jessen (1969) actually presents four methods for PPS sampling without replacement. (It is understood that all sampling procedures discussed are without replacement.) We consider here only his third method, which we will refer to as simply Jessen's method throughout this paper. The general idea of this method is as follows. First select a random number r with $0 \leq r < 1$ and calculate conditional probabilities $\pi_i(n,r), i=1,...,N$, that unit i is in the sample and $\pi_s(n,r)$ that s is the set of n sample units, given that random number r has been selected. To obtain these conditional probabilities, begin by selecting for $r=0$ all certainty units and selecting the remaining sample units simple random sample (SRS) from among all

noncertainty units. As r increases, the conditional selection probabilities at first remain the same, but are subject to change at various values of r denoted $r_1,...,r_f$. For any r, there are associated numbers $c(n,r)$, $z(n,r)$, where $c(n,r)$ is the number of units that are selected with certainty conditional on r and $z(n,r)$ is the number of units that have no chance of selection conditional on r. The set of conditional certainties at r is always the $c(n,r)$ largest units and the set that have no chance of selection is always the $z(n,r)$ smallest units. $c(n,r)$, $z(n,r)$ are nondecreasing functions of r. $c(n,r)$, $z(n,r)$, $\pi_i(n,r), \pi_s(n,r)$ do not change for $r_k \le r < r_{k+1}$ for any k. Also, the $c(n,0)$ largest units are the unconditional certainty units and $z(n,0) = 0$. The total probability assigned to a unit i for any value r or smaller, where $r_k \le r < r_{k+1,}$ $k = 0,1,..., f$, with $r_0 = 0$, $r_{f+1} = 1$, is denoted by $\pi_{Ti}(n,r)$, where

$$\pi_{Ti}(n,r) = (r - r_k)\pi_i(n,r_k) + \sum_{\ell=0}^{k-1}(r_{\ell+1} - r_\ell)\pi_i(n,r_\ell) \quad (3.1)$$

The conditional selection probabilities are defined recursively as follow. The conditional selection probabilities for $r = r_k$ are used for $r > r_k$ until r reaches a value where either (3.2) or (3.3) below occurs, that is:

$\pi_{Ti}(n,r) = \pi_i(n)$ and consequently $\pi_i(n,r) = 0$ for this and any greater r for a unit i with the smallest measure of size not among the $z(n,r_k)$ smallest units because the probability of selection of this unit been all used up. (3.2)

$\pi_{Ti}(n,r) = \pi_i(n) - 1 + r$ and consequently $\pi_i(n,r) = 1$ for this and any greater r for a unit i with the largest measure of size not among the $c(n,r_k)$ largest units, because this unit is a certainty unit for this and any greater r. (3.3)

If (3.2) occurs first or for the same r for which (3.3) does, then r_{k+1} is the smallest $r > r_k$ for which (3.2) holds and $z(n,r_{k+1}) = z(n,r_k)$ plus the number of units j for which $\pi_j(n) = \pi_i(n)$, where i is as in (3.2).

If (3.3) occurs first or for the same r for which (3.2) does, then r_{k+1} is the smallest $r > r_k$ for which (3.3) holds and $c(n,r_{k+1}) = c(n,r_k)$ plus the number of units j for which $\pi_j(n) = \pi_i(n)$, where i is as in (3.3).

In either case, for $r = r_{k+1}$ the $c(n,r_{k+1})$ largest units are selected with certainty and $n - c(n,r_{k+1})$ units are selected SRS from among all units that are neither among the $c(n,r_{k+1})$ largest units nor among the $z(n,r_{k+1})$ smallest units.

This recursive process continues until a k is reached for which there is no r for which $r_k < r < 1$ and either (3.2) or (3.3) holds. Then $k = f$ and $r_{k+1} = 1$.

One important property of Jessen's method is that generally

$$\pi_{ij}(n) > 0 \text{ for all distinct units } i,\ j \quad (3.4)$$

where $\pi_{ij}(n)$ denotes the selection probability for the pair i,j. This probability is generally positive since the sampling among all the noncertainty units is SRS for $r < r_1$. The only exception to this occurs when $\pi_i(n) = 1$ for the $n-1$ largest units, in which case $\pi_{ij}(n) = 0$ if neither i nor j are among the $n-1$ largest units. In this special case, that is when all the sample units but 1 are selected with certainty, it is not possible for (3.4) to be satisfied for any PPS sampling procedure.

4. Subsampling

Suppose a sample of m units is selected PPS using Jessen's method and it is desired to subsample down to n units, $n < m$, with the subsampling process yielding a PPS sample of the original universe. The most straightforward approach to this problem is to select the subsample based on the unconditional selection probabilities of the units being in the sample of m units and the original measures of size, rather than subsampling based on selection probabilities conditional on r. That is, the set of units i in the sample of m units is treated as a frame, with $M_i / \pi_i(m)$ the measure of size for each such unit in the sample of m units. Note that this measure of size is the same for all units that are noncertainty for a sample of m units and, in particular, the subsampling is with equal probability if there are no certainty units in the sample of m units. Any PPS sampling procedure can be used to perform the subsampling, including Jessen's method.

5. Sample Expansion

An expansion of a PPS sample of n units to a PPS sample of m units, where $m > n$, is not always possible. We illustrate by two examples.

Example 1. Consider a universe for which $N = 5$, $n = 3$, $m = 4$, and the p_i are 0.025, 0.075, 0.20, 0.20, 0.50, respectively. If the original sample of 3 units was chosen using any PPS procedure for which

$\pi_{ij}(3) > 0$ for all distinct i, j, then it is not possible to expand the sample to a sample of 4 units selected PPS. This is because units 3,4, and 5 are certainty units for a sample of 4 units. However, whenever units 1 and 2, in addition to certainty unit 5, are the selected units in the sample of 3 units, it is not possible for both units 3 and 4 to be in the expanded sample of 4 units and thus a PPS expansion is not possible unless $\pi_{12}(3) = 0$.

In particular, this problem would arise for this example with Jessen's method, since for this method $\pi_{ij}(3) > 0$ for all distinct i,j. Any modification of Jessen's method for this example that would allow a PPS sample expansion would have to remove the restriction that $\pi_{ij}(3) > 0$ for all distinct i, j.

Example 2. The expansion problem can arise even if there are no certainty units in the expanded sample. To illustrate, consider a second example for which $N = 4$, $n = 2$, $m = 3$ and the p_i are now 0.170, 0.170. 0.330, 0.330, respectively. Then there are no certainty units for a sample of 3 units, since

$$\pi_3(3) = \pi_4(3) = 0.990 \quad (5.1)$$

Furthermore, it follows from (5.1) that $\pi_{34}(3) \geq 0.980$ for any PPS sampling procedure of 3 units and hence

$$\pi_{12}(3) \leq 0.020 \quad (5.2)$$

Jessen's method in unmodified form does not lead to a sample expansion that is PPS in example 2. To see this, observe that we have $r_1 = r_f = 0.680$ for this example and that

$$\pi_i(2, r) = 0.500,\ i = 1,...,4,\ \text{for}\ r < 0.680 \quad (5.3)$$

$$\pi_i(2, r) = 1,\ i = 3,4\ \text{for}\ r \geq 0.680 \quad (5.4)$$

It follows from (5.3), (5,4) that

$$\pi_{12}(2) = 0.680/6 = 0.113 \quad (5.5)$$

which combined with (5.2) establishes that a PPS sample expansion to a sample of 3 units is not possible when a sample of 2 units had been selected originally using Jessen's method in unmodified form.

However if the original sample of *n* units had been selected using a specific modified form of Jessen's procedure, then a PPS sample expansion to a sample of *m* units is always possible. This modification is detailed in Section 6.

The problems with the sample expansion just illustrated can never arise with SRS, 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 SRS from the $N - n$ units not selected in the sample of *n* units.

Also in Ernst (2003) it was observed that for the PPS sampling method of Tillé (1996) a sample expansion is always possible using a different approach. However, Tillé's method also does not always satisfy the condition that $\pi_{ij}(n) > 0$ for all *i,j*.

6. Modification of Jessen's Method to Allow for a PPS Sample Expansion

In this section we present a modification of Jessen's method for selecting a PPS sample of *n* units that does always allow for an expansion to a sample of *m* units, where the expanded sample is chosen using a different modified form of Jessen's method described in Section 7.

In presenting this modified form of Jessen's method for a sample of *n* units, we will use the following additional notation. For this modification $\pi_i'(n,r)$, $\pi_{ij}'(n)$, $\pi_s'(n,r)$, $c'(n,r)$, $z'(n,r)$ are analogous to the same functions without the primes, except they apply to the modified instead of the original Jessen's method. Likewise $r_1',..., r_f'$ are the values of r for which the $\pi_i'(n,r)$ are subject to change under the modified method and the $\pi_{Ti}'(n,r)$ are as given by (3.1) except π, r_k, r_ℓ are replaced by π', r_k', r_ℓ'.

For the modification to be presented, we will have that $\pi_{ij}'(n) > 0$ for all *i,j*, except if for some $m > n$

$$c(m) = m - 1 \quad (6.1)$$

where $c(m)$ is the number of certainty units in a PPS design with *m* sample units; if (6.1) holds then

$$\pi_{ij}'(n) = 0 \ \text{if neither unit } i \text{ nor unit } j \text{ are among the } m-1 \text{ largest units} \quad (6.2)$$

In particular, example 1 satisfies (6.1) and hence (6.2) for $m = 4$. The reason that (6.1) yields (6.2) is that if $c(m) = m - 1$, then $n - 1$ of the $m - 1$ largest units must be selected for any PPS sample scheme of *n* units in order to allow for a sample expansion to *m* units that always includes the $m - 1$ largest units and hence (6.2) must hold.

Instead of giving a formal algorithm for this expansion, we will illustrate it by examples. First consider an example for which $N = 20$, $n = 7$, and: $c(7) = 2$; $c(8) = 4$; $c(m) = 7, m = 9,10$; $c(11) = 8$; $c(m) = 10, m = 12,...,19$. For Jessen's method in unmodified form for $n = 7$ we would select for $r = 0$ the 2 largest units with certainty and select 5 units SRS from among the remaining 18 units in the population. The first problem with that approach is that since $c(8) = 4$ either the third or the fourth largest unit must be in the sample of 7 units to allow for a PPS expansion to $m = 8$ units. Therefore, for $n = 7$, $r = 0$, we must, in addition to selecting the largest 2 units with certainty, also select 1 of the next 2 largest units SRS. However, this change is not enough. Since $c(9) = 7$, 5 of the 7 largest units must be in the sample of 7 units in order to be able to expand to a sample of 9

units with 7 certainties at $r = 0$. Since so far we only have guaranteed that the 2 largest units and one of the next 2 largest units are in the sample of 7 units, we select 2 units SRS among the 4 units out of the 7 largest not already selected. Finally, no $m > 9$ requires any additional restrictions on the selection for $n = 7$ at $r = 0$. That is, we can select the final 2 units SRS from among all units not already selected. This is because we have already selected 5 of the 7 largest units for $n = 7$ and can therefore expand to a sample of 9 or 10 units with 7 certainties or 11 units with 8 certainties or a sample of 12 or more units with 10 certainties since $c(m) \le m - 2$ for all *m*.

Now consider one change in this example, namely that

$$c(m) = 11,\ m = 12,...,19$$

In that case, for the sample of 7 units, in addition to requiring that the 2 largest units in the frame are in sample for $r = 0$, that 1 of the third and fourth largest units be in sample and 2 of the remaining units be among the largest 7, it is also required that 1 of the 6 units not already chosen at this point among the largest 11 units be selected SRS and then that the final unit is chosen SRS among all units except the 6 already selected. The additional restriction is due to the fact that now $c(m) = m - 1$ for $m = 12$.

For $r > 0$, the selection procedure remains the same as for $r = 0$ until we reach the smallest value of *r*, denoted r_1', for which one of (6.3)-(6.5) below happens.

$c'(n,r)$ increases because an additional unit *i* must be certainty for any greater *r* in order that $\pi_{Ti}'(n,1) = \pi_i(n)$, that is for the *r* for which $c'(n,r)$ increases, $\pi_i(n) - \pi_{Ti}'(n,r) = 1 - r$. (6.3)

$z'(n,r)$ increases because an additional unit *i* (actually unit 1 for r_1') must have no chance of selection for any greater *r* in order that $\pi_{Ti}'(n,1) = \pi_i(n)$, that is for the *r* for which $z'(n,r)$ increases $\pi_{Ti}'(n,r) = \pi_i(n)$. (6.4)

$c(m,r)$ increases for some $m > n$, which may increase the number of larger units required to be selected in the sample of *n* in order to be able to able to meet the requirement of additional conditional certainties in the sample of *m* units. (6.5)

We then recalculate the selection probabilities $\pi_i'(n,r),\ i = 1,...,N$, and $\pi_s'(n,r)$ for all set of *n* units for $r \ge r_1'$, using the new values of $c'(n,r_1')$, $z'(n,r_1')$, $c(m,r_1')$. The sampling procedure is basically the same as it is for $r < r_1$ except since $z'(n,r) = 0$ for $r < r_1'$, $z'(n,r)$ did not enter the sampling procedure for $r < r_1'$. Now, if $z'(n,r_1') > 0$, we have the additional constraint that none of the $z'(n,r)$ smallest units can be selected for $r \ge r_1'$.

We then repeat the process as we did for the unmodified version of Jessen's method described in Section 3, with the conditional probabilities changing at r_k', $k = 1,...., f'$.

We illustrate the modified procedure first by example 1 of Section 5. In this example, we have $n = 3$, and

$$c(3) = 1,\ c(4) = 3 \qquad (6.6)$$

Since $N - 1 = 4$, we need not calculate $c(m)$ for any $m > 4$, that is the only possible expansion is from a sample of 3 units to a sample of 4 units. Also since $\pi_5(3) = 1$, we must have $\pi_5'(3,r) = 1$ for all *r*.

It follows from (6.6) that 1 of the units 3,4 must also be selected for a sample of 3 units. Therefore, 1 of these 2 units is selected SRS and 1 additional unit is selected from the remaining 3 unselected units SRS, yielding the probabilities $\pi_s'(3,0)$ given in the first numerical column of Table 1.

To determine r_1' note that (6.3) first occurs for $r = 0.60$ since for the 2 largest noncertainty units *i*, that is units 3 and 4, we have $\pi_i'(3,0) = 0.67$ and $\pi_i(3) = 0.80$ and for these units

$$\pi_i(3) - \pi_{Ti}'(3,0.60) = 0.80 - 0.40 = 1 - 0.60$$

As for (6.4), this occurs first for $r = 0.30$ since for the smallest unit, that is unit 1, we have $\pi_{T1}'(3,0.30) = 0.10 = \pi_1(3)$.

Finally, for (6.5), $c(4,r)$ increases to 4 for $r = 0.50$ because unit 2 becomes certainty for this *r* since $\pi_{T2}(4,0.50) = 0.25 = \pi_2(4) - 1 + 0.50$.

Thus the minimum of the smallest *r*'s needed to satisfy (6.3)-(6.5) is 0.30 and this is the value of r_1'. Consequently, $\pi_1'(3,r) = 0$ for $r \ge 0.30$. Therefore, at r_1' select 1 of the units 3 and 4 SRS and then select from among the nonselected unit and unit 2 SRS, that is $\pi_i'(3,r_1') = 0.75,\ i = 3,4,$ and $\pi_2'(3,r_1) = 0.5$.

To obtain r_2', first note that (6.3) occurs next for $r = 0.70$ since

$$\pi_i(3) - \pi_{Ti}'(3,0.70) = 0.80 - (0.20 + 0.30)$$
$$= 1 - 0.70,\ i = 3,4$$

Also (6.4) occurs next for the same *r*, since $\pi_2'(3, 0.70) = 0.10 + 0.20 = 0.30 = \pi_2(3)$. Finally, as noted above, $c(4,r)$ increases to 4 for $r = 0.50$, with unit 2 becoming certainty. However, the conditional probabilities $\pi_s'(3,r)$ need not change at $r = 0.50$, since the only requirement imposed by $c(4,r) = 4$ is

that 3 of the 4 largest units must be selected for a sample of 3 units, a condition that is already met at $r_1' = 0.30$. Furthermore, $c(4,r)$ cannot increase for any $r > 0.50$ and hence (6.5) does not need to determine any r_k'.

Therefore $r_2' = 0.70$ and hence for $r \geq 0.70$, we have that units 3,4,5 are selected with certainty. The set of conditional probabilities is given in Table 1. (Note in example 1, 0.33 is actually 1/3, while in example 2, 0.333 is actually 1/3 and 0.330 is an exact decimal.)

Table 1. $\pi_s'(3,r)$ *for Example* 1 *Using Modified Form of Jessen's Method*

s	r		
	[0,3)	[.3,0.7)	[0.70,1)
{1,2,5}	0	0	0
{1,3,5}	0.17	0	0
{1,4,5}	0.17	0	0
{2,3,5}	0.17	0.25	0
{2,4,5}	0.17	0.25	0
{3,4,5}	0.33	0.50	1

As will be shown in the next section, this modification allows for a sample expansion for example 1 but, as we have shown, the condition that $\pi_{ij}'(3) > 0$ for all distinct *i,j* is not met since $\pi_{12}'(3) = 0$.

As a second illustration we consider example 2 of Section 5. This example, unlike example 1, illustrates that (6.5) can impact the conditional selection probabilities. In this example $n = 2$, we need only consider $m = 3$, and we have $c(2) = c(3) = 0$; this imposes no restrictions on the selection probabilities for $r = 0$ and consequently 2 of the 4 units are selected SRS for $r = 0$, yielding the $\pi_s'(2,0)$ in the first numerical column of Table 2.

To determine r_1', note that (6.3) first occurs for $r = 0.680$ since for $i = 3,4$,

$$\pi_i(2) - \pi_{Ti}'(2,0.680) = 0.660 - 0.340 = 1 - 0.680$$

Now (6.4) also occurs first for $r = 0.680$ since

$$\pi_{Ti}'(2,0.680) = \pi_i(2) = 0.340 \text{ for } i = 1,2.$$

Finally for (6.5), $c(3,r)$ increases from 0 to 2 at 0.040 since for $i = 3,4$ we have

$$\pi_i(3) - \pi_{Ti}(3,0.040) = .990 - .030 = 1 - 0.040.$$

So the smallest *r* needed to satisfy (6.3)-(6.5) is 0.040 and this is the value of r_1'. Therefore, in order to meet the requirement that $c(3,0.040) = 2$, we must first select 1 of the units 3 and 4 SRS and then select from among the nonselected unit and units 1 and 2 SRS. Consequently, the $\pi_s'(2,r_1')$ are as given in the last column of Table 2.

As for r_2', (6.3) does not occur for any *r* for which $0.040 < r < 1$, since for $i = 3,4$, and such *r*

$$\pi_i(2) - \pi_{Ti}'(2,r) = 0.660 - (0.020 + 0.667(r - 0.040))$$
$$= 0.667(1 - r) < 1 - r$$

Similarly (6.4) does not occur for any $r > 0.40$, since for $i = 1,2$, and such *r*,

$$\pi_{Ti}'(2,r) = 0.020 + 0.333(r - 0.040) < 0.340 = \pi_i(2)$$

Finally $c(3,r)$ cannot increase for any $r > 0.040$ because any increase would require that $c(3,r) = 4$ since $\pi_1(3) = \pi_2(3)$.

Therefore, $r_2' = 1$ and the conditional selection probabilities do not change for any $r > 0.040$.

Table 2. $\pi_s'(2,r)$ *for Example* 2 Using *Modified Form of Jessen's Method*

s	r	
	[0,0.040)	[0.040,1)
{1,2}	0.167	0
{1,3}	0.167	0.167
{1,4}	0.167	0.167
{2,3}	0.167	0.167
{2,4}	0.167	0.167
{3,4}	0.167	0.333

7. Expansion of Modified Version of Jessen's Method

We have shown in the previous section how to obtain a modified version of Jessen's method that assigns selection probabilities, denoted $\pi_s'(n,r)$, for the sets of all samples *s* of *n* units, corresponding to any random number *r* with $0 \leq r < 1$, with the resulting sampling being PPS. In this section we explain how to expand the sample from a sample *s* of *n* units to a sample *t* of *m* units for any $m > n$, with the selection probabilities resulting in a PPS sampling procedure of *m* units. Here $\pi_t''(m,r,n,s)$ denotes the conditional probability that the set of units *t* is selected for a sample of *m* units given that the set *s* was selected for the sample of *n* units using the modified form of Jessen's method and that *r* was the random number used in choosing the sample of *n* units. (Similarly, for each unit *i*, $\pi_i''(m,r,n,s)$ denotes the conditional probability that unit *i* is selected for a sample of *m* units conditional on the same *s* and *r*.) In order for this sampling of *m* units to be PPS these conditional selection probabilities $\pi_t''(m,r,n,s)$ must yield the unconditional selection probability of $\pi_i(m)$ for unit *i*

for a sample of *m* units.

Before describing the steps in the expansion, we present additional notation. $\pi_i''(m,r,n)$ is the probability of selecting unit *i* in the expanded sample of *m* units using the modified version of Jessen's method to be described, conditional on choosing random number *r* and with the original sample of *n* units having been selected using the modified version of Jessen's method described in the previous section. $r_k'', k = 1,..., f''$ are the values of *r* where $\pi_i''(m,r,n)$ are subject to change under this expansion method and $\pi_{Ti}''(m,r,n)$ is obtained by replacing (3.1) with

$$\pi_{Ti}''(m,r,n) = (r - r_k'')\pi_i''(m,r_k'',n) + \sum_{\ell=0}^{k-1}(r_{\ell+1}'' - r_\ell'')\pi_i''(m,r_\ell'',n) \quad (7.1)$$

that is $\pi_{Ti}''(m,r,n)$ is the total probability assigned to unit *i* under this expansion procedure for random number *r* or smaller.

We now describe the steps in the expansion.

1. The first step in the selection of units to be added in the expansion is to add to *s* all units that are among the $c(m)$ largest units and not in *s*, that is all units that are certainty for a sample of *m* units This is always possible since under the modified version of Jessen's method for *n* sample units we always select in the sample of *n* enough of the units among the $c(m)$ largest to ensure we can include all of the $c(m)$ largest units in the expansion. That is, in the sample of *n* units the number of units selected among the $c(m)$ largest for any *m* is always at least $c(m)-(m-n)$, so it is always possible to add all of the $c(m)$ largest units not in *s* without exceeding the sample size of *m* units. Now we must show that at this point, with this partially expanded sample, the total assigned probability to each unit *i* does not exceed $\pi_i(m)$. This is clearly the case for the $c(m)$ largest units, since for these units $\pi_i(m) = 1$, while for the other units, the total assigned probability is $\pi_i(n) \le \pi_i(m)$.

2. Next we show how to add the remaining units needed to bring the sample up to a sample of *m* units. To obtain $\pi_i''(m,0,n,s)$ for $r = 0$, simply select SRS among all units which are neither in *s* nor are among the $c(m)$ largest units.

3. For $r > 0$ the conditional probabilities $\pi_i''(m,r,n,s)$ remain the same as a function of *r* and *s* as for $r = 0$ until we reach the smallest *r*, denoted r_1'', for which at least one of the following among (7.1)-(7.3) occurs.

The first possibility is

$$r_1'' = r_1' \quad (7.1)$$

where r_1' is as in Section 6. If (7.1) occurs first, then the set of conditional selection probabilities for the possible samples *s* of *n* units change for this *r* as they did in Section 6. For each sample *s*, the set of units that are in $c(m)$ but not in *s* are added to the sample as part of the expansion as they are in step 1, and the remaining units are selected SRS as they are in step 2, although these probabilities change since the set of possible *s* and their selection probabilities change in the sample of *n* units, as described in Section 6.

The second possibility is that r_1'' is the smallest *r* such that for some unit *i*, the unit becomes certainty for that *r* and any greater *r*, that is

$$\pi_i(m) - \pi_{Ti}''(m,r,n) = 1 - r \quad (7.2)$$

in which case for that *r* and any greater *r* and for any *s*,

$$\pi_i''(m,r,n,s) = \pi_i''(m,r,n) = 1$$

The final possibility is that r_1'' is the smallest *r* such that for some *i* (actually $i = 1$)

$$\pi_{Ti}''(m,r,n) = \pi_i(m) - \pi_i(n) + \pi_{Ti}'(n,r) \quad (7.3)$$

in which case for that *r* and any greater *r*, unit *i* will be in the sample of *m* units if and only if it is in the sample *s* of *n* units selected under the modification of Jessen's method of Section 6.

If (7.2) or (7.3) occurs first or both occur for the same *r*, then the selection is SRS among the units which do not satisfy (7.2) or (7.3) and are not in *s*.

As with the description of the original Jessen's method in Section 3 and the modified form of Jessen's method for the original sample of *n* units in Section 6, this process repeats until an f'' is reached for which $r_{f''+1} = 1$.

To demonstrate this expansion procedure, let us consider example 1 again with an expansion from $n = 3$ to $m = 4$ beginning with $r = 0$. Since $\pi_i(4) = 1, i = 3,4,5,$ we have that $\pi_i''(4,0,3) = 1, i = 3,4,5$. If $s = \{3,4,5\}$ then by step 2, we have $\pi_i''(4,0,3,s) = 0.5, i = 1,2;$ while if *s* is any other possible triple in the first column of Table 1, that is all *s* in this table except $\{1,2,5\}$, then the expanded sample of 4 units consists of units 3,4,5 and whichever unit of 1,2 is in *s*. Consequently, $\pi_i''(4,0,3) = 0.5, i = 1,2$.

To obtain r_1'', we first observe that by Table 1, $r_1' = 0.30$. Furthermore, (7.2) does not occur for $i = 2$ for any $r \le 0.30$, since for any such *r*

$$\pi_2(4) - \pi_{T2}''(4,r,3) = 0.75 - 0.50r < 1 - r$$

and (7.3) does not occur for $i = 1$ for any $r \le 0.30$ since

$$0.50r = \pi_{T1}''(4,r,3) < \pi_1(4) - \pi_1(3) + \pi_{T1}'(3,r) = 0.25 - 0.10 + 0.33r$$

Consequently $r_1'' = 0.30$. Then for $r = 0.30$ we have that

$$\pi''_{Ti}(4,0.30,3) = 0.15,\ i = 1,2, \quad (7.4)$$

$\pi''_1(4,0.30,3,s) = 0.5$ for $s = \{3,4,5\}$ and $\pi''_1(4,0.30,3,s) = 0$ for all other possible s, that is $\{2,3,5\}$ and $\{2,4,5\}$ (7.5)

$\pi''_2(4,0.30,3,s) = 0.5\,0$ for $s = \{3,4,5\}$ and $\pi''_2(4,0.30,3,s) = 1$ for $s = \{2,3,5\}$ and $s = \{2,4,5\}$ (7.6)

It follows from Table 1 and (7.5), (7.6) that

$$\pi''_1(4,0.30,3) = 0.25 \text{ and } \pi''_2(4,0.30,3) = 0.75 \quad (7.7)$$

To obtain r_2'', we first combine (7.4) and (7.7) to find that (7.2) occurs for $i = 2$ when $r = 0.70$ since

$$\pi_2(4) - \pi''_{T2}(4,0.70,3)$$
$$= 0.75 - (0.15 + 0.75 \times 0.40) = 1 - 0.70$$

and that (7.3) occurs for $i = 1$ for the same r since

$$\pi''_{T1}(4,0.70,3) = 0.25 = 0.25 - 0.10 + 0.33 \times 0.30$$
$$= \pi_1(4) - \pi_1(3) + \pi'_{T1}(3,0.70)$$

Since also $r_2' = 0.70$, we have that $r_2'' = 0.70$ and that $\pi''_i(4,r,3) = 1, i = 2,3,4,5$ for $r \geq 0.70$ since the only possible s for such r is $\{3,4,5\}$ by Table 1 and since (7.3) holds for $i = 1$ and $r = 0.70$.

Next let us consider example 2 with an expansion from $n = 2$ to $m = 3$ beginning with $r = 0$. Since $c(3) = 0$, for $r = 0$ we simply select in the expansion 1 unit SRS from the 2 units in the population not in s to expand the sample from 2 units to 3 units.

To obtain r_1'', we first observe that by Table 2, $r_1' = 0.040$. Furthermore (7.2) first occurs for $i = 3,4$ at $r = 0.040$ since for these i,

$$\pi_i(3) - \pi''_{Ti}(3,0.040,2) = 0.990 - 0.030 = 1 - 0.040$$

Finally (7.3) cannot occur for any $r \leq 0.040$, since for any such r and $i = 1,2$, we have

$$\pi''_{Ti}(3,r,2) = 0.750r < 0.510 - 0.340 + 0.50r$$
$$= \pi_i(3) - \pi_i(2) + \pi'_{Ti}(2,r)$$

Therefore $r_1'' = 0.040$, and

$\pi''_i(3,0.040,2,s) = 1, i = 3,4$ for all possible s, that is all of the 6 pairs of units except $\{1,2\}$ (7.8)

As for $i = 1,2$, we have

$$\pi''_i(3,0.040,2,s) = 1 \text{ if } i \in s \quad (7.9)$$

$$\pi''_i(3,0.040,2,s) = 0.50 \text{ if } s = \{3,4\} \quad (7.10)$$

$$\pi''_i(3,0.040,2,s) = 0 \text{ otherwise} \quad (7.11)$$

To obtain r_2'', we first note, as mentioned at the end of Section 6, that $r_2' = 1$. Also (7.2) cannot occur next for any $r < 1$ since this would require that both units 1 and 2 become additional certainty units for this r and $m = 3$. Similarly (7.3) cannot occur next for any $r < 1$ since this would require that units 1 and 2 be in the sample of 3 units if and only if they were in the sample of 2 units for the modification of Jessen's method presented in Section 6. However, this is not possible since if $s = \{3,4\}$ then either unit 1 or unit 2 must be in the sample of 3 units. Consequently, $r_2'' = 1$ and therefore the conditional selection probabilities do not change for any $r > 0.040$. That is, the conditional selection probabilities remain as given by (7.8)-(7.11).

8. Avoiding Negative Variance Estimates

In this section we will show how to change the unmodified version of Jessen's method described in Section 3 so that the Yates-Grundy estimator of variance is nonnegative or equivalently that

$$0 < \pi_{ij} \leq \pi_i \pi_j \quad (8.1)$$

for all distinct units i, j provided, as discussed in Section 3, the $n - 1$ largest units are not all certainty.

To demonstrate that Jessen's method in unmodified form does not satisfy (8.1), we present in Table 3 the unmodified version of Jessen's method for example 1. For this example,

$$\pi_{12}(3) = 0.033 > \pi_1(3)\pi_2(3) = 0.030$$

Table 3. $\pi_s(3,r)$ *for Example Using Original Form of Jessen's Method*

s	r		
	[0, 0.2)	[0.2, 0.5)	[0.5, 1)
{1,2,5}	0.17	0	0
{1,3,5}	0.17	0	0
{1,4,5}	0.17	0	0
{2,3,5}	0.17	0.33	0
{2,4,5}	0.17	0.33	0
{3,4,5}	0.17	0.33	1

To modify Jessen's method so that $\pi_{ij} \leq \pi_i \pi_j$, proceed as follows. We generally follow the approach used in obtaining Jessen's method described in Section 3. As in Section 3, we use here no primes in our notation, even though the values are different than in Section 3. Since, unlike in Sections 5-7, the sample size here is fixed, we also omit a parameter for sample size. That is, for example, the probability of selection for unit i for a sample of n units is now π_i rather than $\pi_i(n)$. The main substantive change is that for any k, instead of r_{k+1} being the smallest $r > r_k$ satisfying (3.2) or (3.3), it is now the smallest $r > r_k$ satisfying (3.2), (3.3), (8.2), or (8.3) where (8.2) and (8.3) are as described below.

For any distinct pair of units *i,j*, and random number *r*, let

$$\pi_{Tij}(r) = (r - r_k)\pi_{ij}(r_k) + \sum_{\ell=0}^{k-1}(r_{\ell+1} - r_\ell)\pi_{ij}(r_\ell)$$

that is the total probability assigned to this pair for any value *r* or smaller and let

$$\pi_{Ri}(r) = \pi_i - \pi_{Ti}(r), \quad \pi_{Rij}(r) = \pi_i\pi_j - \pi_{Tij}(r)$$

That is $\pi_{Ri}(r)$ is the probability remaining to be assigned to unit *i* for any value *r* or greater and $\pi_{Rij}(r)$ is the maximum remaining probability that can be assigned to the pair *i,j* for any value *r* or greater. Then conditions (8.2) and (8.3) are as follows:

For a pair of units *i,j*, $\pi_{Rij}(r) = 0$ for some *r*, in which case $\pi_s(r) = 0$ for this and any greater *r* for any set *s* of *n* units that include both units *i* and *j*. (8.2)

For some pair of units *i,j*, and some *r*

$$(n-1)\pi_{Ri}(r) = \sum_{k\neq i}\pi_{Rik}(r) \text{ and } \pi_{Ri}(r) \le \pi_{Rij}(r)$$

in which case if $i \in s$ then $j \in s$ for this or any greater *r* (8.3)

To illustrate this procedure consider Table 4 below for example 1 with $n = 3$.

Table 4. $\pi_s(r)$ *for Example Using Original Jessen's Method Modified to Insure* $\pi_{ij} \le \pi_i\pi_j$ *for All i,j*

s	*r*			
	[0, 0.18)	[0.18, 0.205)	[0.205, 0.505)	[0.505, 1)
{1,2,5}	0.17	0	0	0
{1,3,5}	0.17	0.20	0	0
{1,4,5}	0.17	0.20	0	0
{2,3,5}	0.17	0.20	0.33	0
{2,4,5}	0.17	0.20	0.33	0
{3,4,5}	0.17	0.20	0.33	1

We begin by selecting unit 5 with certainty and 2 of the other 4 units SRS. (8.2) occurs first for $i, j = 1,2$ at $r = 0.18$ since $\pi_{T12}(0.18) = 0.03$, $\pi_1 = 0.10$, , $\pi_2 = 0.30$, while at $r = 0.18$, none of (3.2), (3.3), (8.3) has occurred. (3.2) has not occurred since

$$\pi_{Ti}(0.18) = 0.09 < \pi_i, \ i = 1,2,3,4.$$

(3.3) has not occurred since

$$\pi_{Ti}(0.18) = 0.09 > \pi_i - 1 + 0.18, \ i = 1,2,3,4$$

Finally, (8.3) has not occurred because for $i = 1$, for example, we have that for $r < 0.18$,

$$2\pi_{R1}(r) = 0.20 - r < \sum_{k=2}^{5}\pi_{R1k}(r) = 0.29 - r$$

with similar relationships for units 2,3,4. Unit 5 does not satisfy (8.3) either since $\pi_{R5}(r) = 1 - r > \pi_{5j}(r)$ for $j = 1,2,3,4$. Thus $r_1 = 0.18$ and the $\pi_s(r_1)$ are as given in the second column of probabilities in Table 4.

The first occurrence of (3.2), (3.3), (8.2), (8.3) for $r > 0.18$ is (3.2) for $i = 1$ at $r = 0.205$, that is $\pi_{T1}(0.205) = 0.10 = \pi_1$. None of (3.3), (8.2), (8.3) occur for $0.018 < r < 0.205$. (3.3) does not occur since

$$\pi_{Ti}(0.205) = 0.105 > \pi_i - 1 + 0.205, i = 3,4$$

(8.2) does not occur since $\pi_{Tij}(0.205) = 0.035 < \pi_i\pi_j$ for all *i,j* other than 1,2 for which $1 \le i < j \le 4$. Finally (8.3) does not occur for $i = 1$, for example, since for $0.018 < r < 0.205$, we have that

$$2\pi_{R1}(r) = 0.02 - 0.80(r - 0.18) < \sum_{k=2}^{5}\pi_{R1k}(r)$$
$$= 0.11 - 0.80(r - 0.18)$$

Thus $r_2 = 0.205$.

The first occurrence of (3.2), (3.3), (8.2), (8.3) for $r > 0.205$ is for $r = .505$. For this value of *r*, (3.2) holds for $i = 2$ and (3.3) holds for $i = 3,4$. (8.2) does not occur for $r < 0.505$ since

$$\pi_{Tij}(0.505) = 0.135 < \pi_i\pi_j, (i, j) = (2,3),(2,4),(3,4)$$

(8.3) can be shown not to occur for $r < 0.505$ similarly to the previous intervals. So $r_3 = 0.505$ and $f = 3$ since $s = \{3,4,5\}$ is selected with certainty for $r \ge 0.505$.

References

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

Jessen, R. J. (1969). Some Methods of Probability Non-Replacement Sampling. Journal of the American Statistical Association, 64, 175-193.

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

_____ (2005). Sampling Algorithms. New York: Springer.

.

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