JOINT PROBABILITIES OF SELECTION OF UNITS IN SYSTEMATIC SAMPLES[1]

G.B. Gray, Statistics Canada

Introduction

Many schemes for sampling n units out of N with probability proportional to size (pps) have been devised. Examples of pps methods of selection are given by I.P. Fellegi [1], H.O. Hartley and J.N.K. Rao[3], Horvitz and Thompson [4] and many others. In each case, a Horvitz-Thompson estimate of the population total from the sample may be obtained. Upon calculation of joint probabilities of pairs of units in the sample, variance estimates may be calculated by either the Horvitz-Thompson [4] or Yates-Grundy [6] formulas.

The simplest of all sampling schemes is systematic pps sampling as described in [2] and is particularly useful for rotating samples in recurring surveys. The units may be in a fixed pre-determined order or in random order prior to selection but in either case, unbiased estimates of the total may be obtained. In the case of rotating samples, the units may be randomly ordered prior to the first survey and the order may be maintained for the first and subsequent surveys with a partial or complete rotation of units.

In the case of a fixed pre-determined order of listing, most of the joint probabilities of selection are zero so that unbiased estimates of the variance are not possible. However, if the order of listing has been undertaken in such a manner to ensure a negative serial correlation for sampled units, then an over-estimate of the variance may be obtained by treating the units as though they had been sampled with pps with replacement.

In the case of randomly ordered units, non-zero joint probabilities nearly always exist for every pair of units and these may be easily calculable by the algorithm described below for small populations[2]. H.O. Hartley and J.N.K. Rao [2] derived an asymptotic formula in 1962, which holds approximately true only for pairs of units selected from large populations. Alternatively, the algorithm could be applied to a large sample of possible arrangements of units and the joint probability estimated by averaging over the large sample of units.

Systematic Sample Procedure, Notation, and Assumptions

p_i = relative size of i^{th} unit where $i = 1,2, \ldots, N$ (relative number of persons, relative Census sales, for example). In a systematic pps selection procedure, probability of selection of unit no. $i = np_i$ where $n \geq 2$.

It will be assumed here that $np_i \leq 1$. In some universes or strata, the relative size of a unit, say i, may be so large that, by the procedure of systematic sampling, np_i would lie in the range (1,2) so that the unit could be selected once with probability $1/np_i$ or twice with probability $(np_i - 1)/np_i$. In most cases, it would be preferable from cost/variance benefit and operational point of view simply to include the unit with certainty and adjust the relative sizes of the remaining units. In practice, the procedures could be adapted to include units in a certainty stratum if $np_i > a$ for some arbitrary a less than one such as 0.5 for example. In small populations, however, it is difficult to achieve such a condition while maintaining a fixed sample size and ensuring a sampling stratum containing at least two selected units. If $n = 1$, then no joint probabilities exist and no variance estimate is possible.

Considering a particular order of listing, for the moment, say, in serial number order, we have the units 1,2,3, ..., N together with their relative sizes, p_i, probabilities of selection, np_i and accumulated probabilities $t_i = \sum_{j=1}^{i} np_j$, in Table 1:

Table 1: Set-up of Units for Systematic pps Selection of n Units

Unit No.	Relative Size	Probability of Selection	Accumulated Probabilities
1	p_1	np_1	$t_1 = np_1 = 0 + d_1$
2	p_2	np_2	$t_2 = t_1 + np_2 = I_2 + d_2$
3	p_3	np_3	$t_3 = t_2 + np_3 = I_3 + d_3$
4	p_4	np_4	$t_4 = t_3 + np_4 = I_4 + d_4$
.	.	.	.
.	.	.	.
.	.	.	.
i	p_i	np_i	$t_i = t_{i-1} + np_i = I_i + d_i$
.	.	.	.
.	.	.	.
.	.	.	.
N	p_N	np_N	$t_N = n = I_N + 0$

[1] A similar study was undertaken by W.S. Connor [1] in 1966.

[2] Zero joint probabilities however do occur in an example stated by J.N.K. Rao in his thesis [5] and in W.S. Connor's article [1] even when 5 units are randomly ordered and 2 units are selected. (Relative sizes .1, .1, .25, .275, .275).

In the above table, t_i is partitioned into an integer I_i and a decimal d_i component. As will be seen later, this device will facilitate the computation of joint probabilities. Thus $I_2=0$ or 1 in the above table according as $np_1 + np_2 < 1$ or ≥ 1.

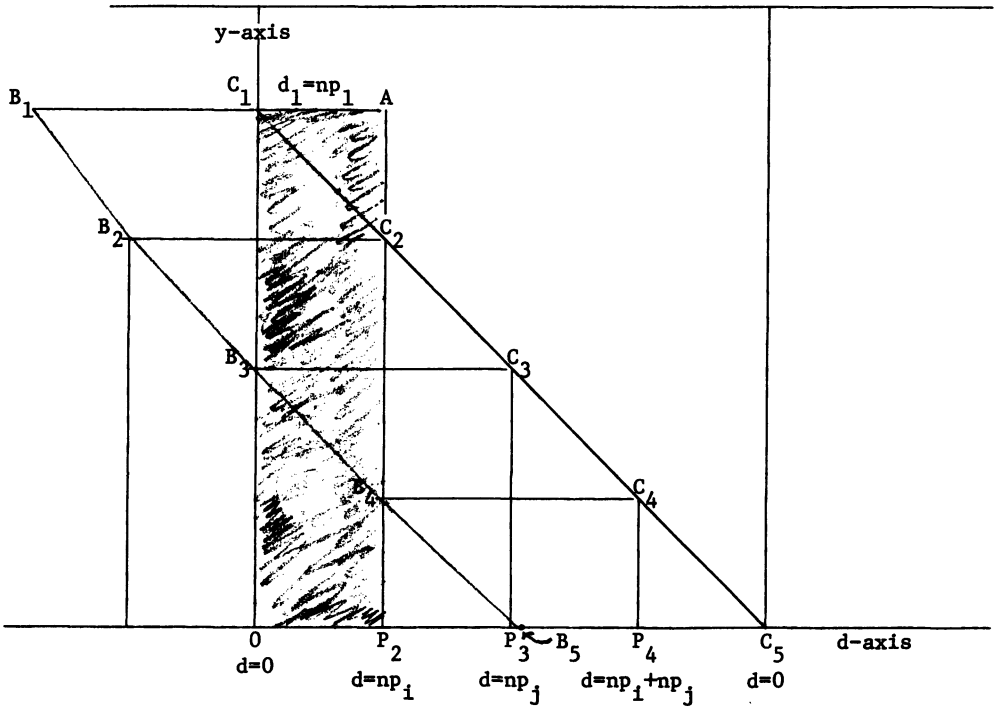
The procedure of systematic sampling is then to select a random number r in the range [0,1) and select n units $i_1, i_2, \ldots, i_n$ such that $t_{i_1} \leq r < t_{i_1+1}$; $t_{i_2} \leq r+1 < t_{i_2+1}$; $t_{i_3} \leq r+2 < t_{i_3+1}$, $\ldots$ $t_{i_m} \leq r+m-1 < t_{i_m+1}$, $\ldots$, and finally $t_{i_n} \leq r+n-1 < t_{i_n+1}$. If $n = N$, t_{i_n+1} may be taken as any arbitrary number greater than n, say $t_{i_N} + np_i$.

In practice, the probabilities of selection and their accumulations are usually replaced by actual sizes $S_1, S_2, \ldots S_N$ and $T_i = \sum_{j=1}^{i} S_j$ and the systematic pps selection is undertaken by selecting a random integer $R \varepsilon [1, T_N/n]$ and selecting units $i_1, i_2, \ldots i_m, \ldots i_n$ so that $T_{i_m} \leq R + (m-1)\, T_N/n < T_{i_m+1}$. The resultant sample is exactly the same as before with $r = R/(T_N/n)$ and $p_i = S_i/T_N$.

Now if X_i is a characteristic value for the i^{th} unit of some item whose total $X = \sum_{i=1}^{N} X_i$ is to be estimated, an estimate of the total is given by:

$$\hat{X} = \sum_{m=1}^{n} X_{i_m} / np_{i_m}, \qquad (1)$$

FIG. I Diagram showing joint probability of units i and j when $np_i + np_j \leq 1$



Explanation of Above Table

C_1 A of length $d_1 = np_i$ is projected upon the d-axis, represented by OP_2 or any parallel line in the shaded area. The abscissa of $AP_2 = d_1$. The distances $B_i C_i$ between the two diagonal lines $B_1 B_2 \ldots B_5$ and $C_1 C_2 \ldots C_5$ is of length np_j and the abscissae of all points on $C_1 C_2 \ldots C_5$ represent all possible values of d_{k+2} while those of $B_1 B_2 \ldots B_5$ represent those of d_{k+1}. If B and C represent the positions on $B_1 B_5$ and $C_1 C_5$ for a given set of k units between units i and j, then the portion of BC covered in the shaded area represents the joint probability of selection of units i and j.

Here, we are assuming $np_i \leq np_j$ and $np_i + np_j \leq 1$.

and its true and estimated variance are given by:

$$V(\hat{X}) = \sum_{i=1}^{N} \sum_{j>i} (np_i \cdot np_j - \pi_{ij}) \left(\frac{X_i}{np_i} - \frac{X_j}{np_j}\right)^2 \quad (2)$$

and

$$\hat{V}(\hat{X}) = \sum_{m=1}^{n} \sum_{q>m} \left(\frac{np_{i_m} np_{i_q}}{\pi_{i_m i_q}} - 1\right) \left(\frac{X_{i_m}}{np_{i_m}} - \frac{X_{i_q}}{np_{i_q}}\right)^2 \quad (3)$$

respectively, applying Yates-Grundy formulas [6].

Here, π_{ij} is the joint inclusion probability of units i and j in the sample.

(3) is an unbiased estimate of (2) only if the joint probabilities of all $\binom{N}{2}$ pairs of units in the population are non-zero. Otherwise, by direct substitution of the individual and joint probabilities and the statistics of the selected units in (3), (2) will be under-estimated by the contribution by the double summation $\sum_{i=1}^{N} \sum_{j>i}$ of (2), in which $\pi_{ij} = 0$.

(2) may be written also as:

$$V(\hat{X}) = \sum_{i=1}^{N} np_i \left(\frac{X_i}{np_i} - \frac{X}{n}\right)^2 + \sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} \left(\frac{X_i}{np_i} - \frac{X}{n}\right)\left(\frac{X_j}{np_j} - \frac{X}{n}\right) \quad (4)$$

and be defining

$$\sigma^2 = \sum_{i=1}^{N} p_i \left(\frac{X_i}{Np_i} - \frac{X}{N}\right)^2 \quad (5)$$

(6)

and

$$r_{FP} = \frac{1}{n(n-1)\sigma^2} \sum_{i \neq j}^{N} \pi_{ij} \left(\frac{X_i}{Np_i} - \frac{X}{N}\right)\left(\frac{X_j}{Np_j} - \frac{X}{N}\right)$$

and $(\hat{X})$ may be simply written as:

$$V(X) = N^2 \frac{\sigma^2}{n} [1 + (n-1)\, r_{FP}]. \quad (7)$$

The above formulas hold true, for all pps sampling procedures where Horvitz-Thompson estimators are employed. When the sizes are equal, σ^2 reduces to the classical formula $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - X/N)^2$ and in the case of simple random sampling without replacement, $r_{FP} = -1/(N-1)$, the classical finite population correlation. When sampling with pps with replacement, $r_{FP} = 0$ in (7).

In the case of systematic sampling with an equal step interval of N/n for units in a predetermined order, r_{FP} becomes the serial correlation whereby $\pi_{ij} = n/N$ for units N/n or a multiple of N/n apart and =0 for all other pairs. In the case of systematic sampling with pps among units in a fixed order of listing, r_{FP} may be defined as the serial correlation generalized to pps systematic whereby $\pi_{ij} = 0$ for those pairs of units that have no chance of entering the sample together.

A general formula for the joint probability of pairs of units being selected by systematic pps in a pre-determined order of listing and then the algorithm to calculate π_{ij} over all random arrangements of units will be described.

Joint Probability of Selection of Units i and j when units in fixed order

For a specified order of listing of the N units with units i and j separated by k units, π_{ij} will remain unchanged if we renumber the units so that i is the first unit, j is the (k+2)th unit and (i-1) is the last unit. π_{ij} will also remain unchanged if we interchange units i and j, if necessary, so that $np_i \leq np_j$ since the order of listing could be reversed without affecting the joint probabilities. If this is undertaken, Table 1 would be altered thus:

Table 2: Data as in Table 1 with renumbering as above

Unit No.	Relative Size	Probability of Selection	Accumulated Probabilities
i	p_i	np_i	$np_i = d_1$
	p_{i_2}	np_{i_2}	.
	.	.	.
k units	.	np_{i_3}	.
$\equiv a_k$	.	.	.
	$p_{i_{k+1}}$	$np_{i_{k+1}}$	.
j	p_j	np_j	$I_{k+2} + d_{k+2}$
N-(k+2) remaining units	.	.	.
	.	.	.
	.	.	.
$\equiv a_{N-2-k}$	p_{i_n}	np_{i_N}	n

Originally the cumulative probabilities for unit i was $I_i + d_i$ and for unit j, $I_{j+k+1} + d_{j+k+1}$ before renumbering the units. By deducting $c = I_i + d_i - np_i$ from the cumulative probabilities for units i to N and by adding n-c to the cumulative probabilities for the remaining units, we arrive at the new cumulative probabilities in Table 2. The original random number r may be adjusted to $r' = r - c + A$ for some integer A to ensure that $r' \in [0,1)$ and the random number r' will yield exactly the same sample in the renumbered list as r in Table 1.

If the original listing of units is simply reversed, the random number r may be adjusted to $r' = 1 - r$ with $r' = 0$ if $r = 0$ to ensure that $r' \in [0,1)$ instead of $(0,1]$. By employing r' in the reverse order the same selection will result as with random number r.

Units i and j are both selected with a fixed selection of k units or its complement of N-k-2 units lying between i and j when the adjusted random number $r' \leq np_i$ and $r' + a\,\varepsilon(I_{k+1} + d_{k+1},\ I_{k+2} + d_{k+2}]$ for some integer a.

In FIG. I, the individual and joint probabilities of units i and j are illustrated by a series of lines parallel to the d-axis with C_1 A and its parallel lines in the shaded area representing $d_1 = np_i$ and BC representing np_j.

By observing the relative positions of B_iC_i (for different i) with respect to the shaded area, we can determine the value of π_{ij} for each position as follows: beginning with $d_{k+2} = 0$. If $\underline{a}_k$ represents a vector of k selected units between units i and j, $\pi_{ij|\underline{a}_k}$ represents the conditional joint probability for the given set.

Table 3: Value of $\pi_{ij|\underline{a}_k}$ according to range of values of d_{k+2} $(np_i + np_j \leq 1)$[2]

	Range of d_{k+2}	Value of $\pi_{ij\|\underline{a}_k}$
(3.1)	$[0,\ np_i)$	d_{k+2}
(3.2)	$[np_i,\ np_j)$	np_i
(3.3)	$[np_j,\ np_i + np_j)$	$np_i + np_j - d_{k+2}$
(3.4)	$[np_k + np_j,\ 1)$	0

By an argument similar to the above, using a figure such as the above for the case when $np_i + np_j > 1$, although each <1, we can obtain the probabilities $\pi_{ij|\underline{a}_k}$.

Table 4: Value of $\pi_{ij|\underline{a}_k}$ according to range of values of d_{k+2} $(np_i + np_j > 1)$[2]

	Range of d_{k+2}	Value of $\pi_{ij\|\underline{a}_k}$
(4.1)	$[0,\ np_i + np_j - 1)$	$np_i + np_j - 1$
(4.2)	$[np_i + np_j - 1,\ np_i)$	d_{k+2}
(4.3)	$[np_i,\ np_j)$	np_i
(4.4)	$[np_j,\ 1)$	$np_i + np_j - d_{k+2}$

[2] These results agree closely with those derived by W.S. Connor [1] in 1966.

Consider the two sets of units given below and the characteristic values.

Table 5: Set No. 1 (Horvitz-Thompson [4])

Unit No.	p_i	$2p_i$	$t_i = I_i + d_i$	X_i
1	18/394	18/197	18/197	19
2	9/394	9/197	27/197	9
3	14/394	14/197	41/197	17
4	12/394	12/197	53/197	14
5	24/394	24/197	77/197	21
6	25/394	25/197	102/197	22
7	23/394	23/197	125/197	27
8	24/394	24/197	149/197	35
9	17/394	17/197	166/197	20
10	14/394	14/197	180/197	15
11	18/394	18/197	1 + 1/197	18
12	40/394	40/197	1 + 41/197	37
13	12/394	12/197	1 + 53/197	12
14	30/394	30/197	1 + 83/197	47
15	27/394	27/197	1 110/197	27
16	26/394	26/197	1 + 136/197	25
17	21/394	21/197	1 + 157/197	13
19	19/394	19/197	1 + 185/197	19
20	12/394	12/197	2	12

p_i based on eye-estimated count of households in each of 20 blocks of Ames, Iowa X_i = actual count of households in the correcponding blocks.

Table 6: Set No. 2 (I.P. Fellegi[2])

Unit No.	p_i	$2p_i$	$t_i = I_i + d_i$	X_i
1	.10	.20	.20 = 0 + .20	.60
2	.14	.28	.48 = 0 + .48	.98
3	.17	.34	.82 = 0 + .82	1.53
4	.18	.36	1.18 = 1 + .18	2.16
5	.19	.38	1.56 = 1 + .56	2.85
6	.22	.44	2.00 = 2 + .0	4.18

To find the joint probability of units 2 and 5 being selected in Set No. 2, we note that unit 2 is the smaller so we renumber; thus:

Unit No.	p_i	$2p_i$	t_i	
2	.14	.28	.28	$d_1 = .28$
3	.17	.34	.62	
4	.18	.36	.98	
5	.19	.38	1.36	$d_1 = .36$
6	.22	.44	1.80	
1	.10	.20	2.00	

Here, k = 2, $d_1 = .28$, $d_4 = .36$, $2p_2 + 2p_5 = .66 < 1$ so we use Table 3 to find $\pi_{2,5|\underline{a}_2}$ where $\underline{a}_2 = (3,4)$.

Now $d_4 = .36$ lies between $2p_2 = .28$ and $2p_5 = .38$ and so by condition 3.2 of Table 3, $\pi_{2,5|\underline{a}_2} = .28$.

Proceeding in this manner for all pairs of units of Set 2 in the given order of listing, we find the joint probabilities, as follows:

Table 7: Value of π_{ij} (fixed order) as of Set 2

Unit i / Unit j	1	2	3	4	5	6
1	-	.0	.0	.18	.02	.0
2	.0	-	.0	.0	.28	.0
3	.0	.0	-	.0	.08	.26
4	.18	.0	.0	-	.0	.18
5	.02	.28	.08	.0	-	.0
6	.0	.0	.26	.18	.0	-
Sum	.20	.28	.34	.36	.38	.44

Turning now to the case of randomized order of units, we see that, for each given pair of units i and j, we need only find $\pi_{ij|\underline{a}_k}$ for $k = 0,1,2, \ldots, \frac{N-3}{2}$ if N odd and for $k = 0,1,2, \ldots, \frac{N-2}{2}$ if N even and only half of the sets for $k = \frac{N-2}{2}$ since each set of $\underline{a}_k$ yields a complement set with also N-2/2 units with the same joint probability. The arrangement of the k units in a given set $\underline{a}_k$ lying between i and j is immaterial since the cumulative probabilities remain unchanged for units i and j. Hence, we need only consider all possible selections of k units as k proceeds from k = 0 (when i and j are adjacent in the list) to k = N-3/2 or N-2/2. Or, we find $\pi_{ij|\underline{a}_k}$ for each set $\underline{a}_k$ and take an average over $\binom{N-2}{k}$ selections of k units and over (N-1)/2 distinct values of k, or for random ordering of units or

$$\pi_{ij} = \frac{1}{N-1} \sum_{k=0}^{N-2} \pi_{ij|k}$$

$$\pi_{ij} = \frac{2}{N-1} \sum_{k=0}^{(N-3)/2} \frac{1}{\binom{N-2}{k}} \sum_{\underline{a}_k} \pi_{ij|\underline{a}_k} \text{ for N odd}$$

$$= \frac{2}{N-1} \left[\sum_{k=0}^{(N-4)/2} \frac{1}{\binom{N-2}{k}} \sum_{\underline{a}_k} \pi_{ij|\underline{a}_k} + \frac{1}{\binom{N-2}{\frac{N-2}{2}}} \sum^{*}_{\underline{a}_{(N-2)/2}} \pi_{ij|\underline{a}_{(N-2)/2}} \right]^{3} \quad (8)$$

for N even.

[3] This formula was also derived by W.S. Connor [1] who noted the symmetry in the joint probabilities for k and N-2-k units apart but not for each half of all the sets when k=(N-2)/2.

Σ^{*} denotes summation over only 1/2 of the sets of N-2/2 units (easily accomplished by fixing a particular unit in each set when calculating the conditional joint probabilities).

In effecting the calculation of $\pi_{ij|k}$, the conditional joint probability of k units lie between i and j, the $\binom{N-2}{k}$ selections may be partitioned into four separate groups satisfying conditions 3.1, 3.2, 3.3, or 3.4 by examining the ranges of values of d_{k+2} over sub-sets of k selected units. In fact, $\pi_{ij|k}$ may be zero regardless of the selected units listed between i and j, a fact easily established by determining the range of values of $I_{k+2} + d_{k+2}$ between the smallest k units and the largest k units. Thus, the algorithm for calculating $\pi_{ij|k}$ for each value of k may be simplified and the calculations reduced by ordering the N-2 units in ascending order of size between i and j and examining the range of value of $I_{k+2} + d_{k+2}$.

Considering set 2 with 6 units, suppose now that $\pi_{2,5}$ is to be calculated over all possible random orderings of units.

Table 8: Re-ordering for calculation of $\pi_{2,5}$ (Set 2)

Unit	$2p_i$	Sum	Comments
2	.28		
5	.38	.66	Sum < 1, so use Table 3
remaining units ordered by size	$2p_i$		
1	.20	.20	
3	.34	.54	
4	.36	.90	
6	.44	1.34	

Calculations of $\pi_{2,5|k}$ for:

k=0 since $d_2 = .66$, condition 3.4 holds and so $\pi_{2,5|0} = 0$

k=1 Min. $I_3 + d_3 = .86$ and Max. $I_3 + d_3 = 1.10$ so for some sets $\underline{a}_1$, $\pi_{2,5|\underline{a}_1} \neq 0$ since $d_3 = .10$ for Max. $I_3 + d_3$, for which condition 3.1 holds. So proceeding thus; for

$\underline{a}_1=(6)$, $d_3=.10$, $\pi_{2,5|(6)}=.10$ by 3.1,

$\underline{a}_1=(4)$, $d_3=.02$, $\pi_{2,5|(5)}=.02$ by 3.1,

$\underline{a}_1=(3)$ or (1), $d_3 \in [.66,1)$, $\pi_{2,5|\underline{a}_1} = 0$

Sum .12

Average $.12/\binom{4}{1} = .12/4 = .03 = \pi_{ij|1}$

k=2=(N-2)/2 so that we need consider only one-half of the $\binom{4}{1}$ sets, by fixing a particular unit, say no. 1 in each set.

Then, with unit no. 1 included in the 3 possible sets.

Min $I_4 + d_4 = .66 + .20 + .34 = 1.20$ or $d_4 = .20$

Min $I_4 + d_4 = .66 + .20 + .44 = 1.30$ or $d_4 = .30$

both yielding non-zero joint probabilities so for k=2, when

$\underline{a}_2=(1,3)$, $d_4=.20$, $\pi_{2,5|(1,3)}=.20$ by 3.1,

$\underline{a}_2=(1,4)$, $d_4=.22$, $\pi_{2,5|(1,4)}=.22$ by 3.1, and

$\underline{a}_2=(1,6)$, $d_4=.30$, $\pi_{2,5|(1,6)}=.28$ by 3.2

Sum =.70

Average =.70/6=.116=$\pi_{2,5|2}$

Hence, $\pi_{2,5} = \frac{2}{5}$ (0+.03 + .116) =

.4 x .146 = .0586.

Proceeding in like manner as above, for the remaining 14 pairs of units, we may derive the joint probability matrix in Table 9 and a similar matrix in Table 10 for the case of n=3 for the same population.

Table 9: Joint Probability for each pair of units of Set 2 (n=2) Unit i

Unit j	1	2	3	4	5	6
1	-	.0386	.0386	.0386	.0420	.0420
2	.0386	-	.0486	.0553	.0586	.0786
3	.0386	.0486	-	.0753	.0786	.0986
4	.0386	.0553	.0753	-	.0853	.1053
5	.0420	.0586	.0786	.0853	-	.1153
6	.0420	.0786	.0986	.1053	.1153	-
Sum	.2000	.2800	.3400	.3600	.3800	.4400

Table 10: Joint Probabilities between Pairs of Units (Population 2 when n=3)

Unit i

Unit j	1	2	3	4	5	6
1	-	.1080	.1163	.1213	.1263	.1280
2	.1080	-	.1630	.1680	.1730	.2280
3	.1163	.1630	-	.2096	.2346	.2963
4	.1213	.1680	.2096	-	.2596	.3213
5	.1263	.1730	.2346	.2596	-	.3463
6	.1280	.2280	.2963	.3213	.3463	-
Sum	.6000	.8400	1.0200	1.0800	1.1400	1.3200

Check = 2 x $3p_1$ = 2 x $3p_2$, etc.

and the calculations for say $\pi_{5,6}$ in Table 10:

i	$3p_i$	Sum					
5	.57						
6	.66	1.23 > 1 so $\pi_{5,6\|0}$ = .23 (by 4.1)					
-		k=1	$\pi_{5,6\|\underline{a}_1}$		$\underline{a}_2$	k=2	$\pi_{5,6\|\underline{a}_2}$
1	.30	1.53	.53 (by 4.2)		1,2	1.95	.28(by 4.4)
2	.42	1.65	.57 (by 4.3)		1,3	2.04	.23(by 4.1)
3	.51	1.74	.49 (by 4.4)		1,4	2.07	.23(by 4.1)
4	.54	1.77	.46 (by 4.4)				

$\pi_{5,6|1} = \frac{2.05}{4} = .5125$ $\pi_{5,6|2} = \frac{.74}{6} = .123\dot{3}$

Hence, π_{56} = .4 (.2300 + .5125 + .1233) =

.4 x .86583 = .3463.

For a population of 6 units, to calculate the joint probability of any pair of units requires only the calculation for 8 distinct arrangements of units and reading off the probabilities from Table 3 or 4 are required before a direct substitution in formula (8) can be made. However, for N units in general, 2^{N-3} distinct arrangements and readings from Table 3 or 4 must be obtained before substitution in formula (8). This is clearly impractical for manual calculations beyond say N=10 and perhaps even for a computer operation beyond say N=20 since 2^{17} = 131,072 and the running time on the computer could be extensive. The calculations may sometimes be considerably reduced by averaging over many possible selections of units between a given pair of units rather than by reading off the joint probability for each of the 2^{N-3} possible selections of units and then averaging the probabilities.

References

[1] Connor, W.S., "An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement", Journal of the American Statistical Association, Vol. 61, (1966), pp. 384-390.

[2] Fellegi, I.P., "Sampling with varying probabilities without replacement: rotating and non-rotating samples", Journal of the American Statistical Association, Vol. 58, (1963), pp. 183-201.

[3] Hartley, H.O. and Rao, J.N.K., "Sampling with unequal probabilities without replacement from a finite universe", Journal of the American Statistical Association, Vol 47, (1952), pp. 663-685.

[5] Rao, J.N.K., "Sampling procedures involving unequal probability selection", (1961), unpublished Ph.D Thesis, Iowa State University, Ames, Iowa.

[6] Yates, F. and Grundy, P.M., "Selection without replacement from within strata and with probability proportional to size", Journal of the Royal Statistical Society, Series B, Vol. 15 (1953), pp. 253-261.