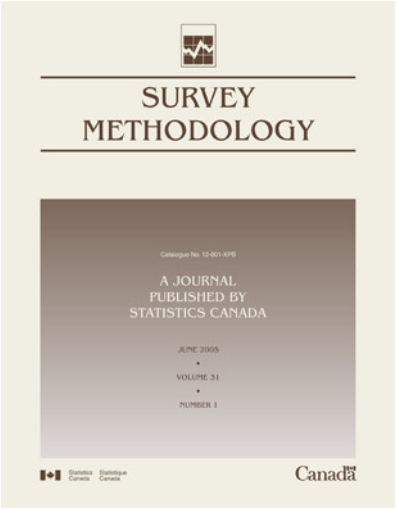
Catalogue no. 12-001-XIE

Survey Methodology

December 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 2, pp. 233-238
Statistics Canada, Catalogue No. 12-001

On the Correlation Structure of Sample Units

Alfredo Bustos [1]

Abstract

In this paper we make explicit some distributional properties of sample units, not usually found in the literature; in particular, their correlation structure and the fact that it does not depend on arbitrarily assigned population indices. Such properties are relevant to a number of estimation procedures, whose efficiency would benefit from making explicit reference to them.

Key Words: Census; Survey; Sampling; Sample units; Probability function; Mean; Covariance.

1. Introduction

In recent times, population and household censuses, as we know them, have become more difficult to perform for a number of reasons. Alternative ways of securing more frequent information for the production of local, state and national statistical results have been proposed. Continuous large national surveys, among them those known as rolling censuses, with large sample sizes and complex designs, are being considered.

However, in order to produce results at the local authority level the way a census does, different techniques for estimation as well as for validation and, in some cases, for imputation have to be developed and their efficiency improved. One way of achieving greater efficiency consists of taking into account all relevant information available. Of course, this includes the stochastic properties of sample units.

In what follows, beginning from basic principles, we derive a general explicit form for the probability function of an ordered sample. We also show how that function, as well as the inclusion probabilities, can be computed. Finally, we give a general form for the correlation matrix of sample units, which depends solely on inclusion probabilities, so that linear and maximum-likelihood estimation procedures can benefit from it.

2. The Basic Model

The basic model we start from represents the sequential random drawing of n units from a population U formed by N such units, and may be stated as follows. Let N and n be two positive constants such that $n \le N$, and let V represent an $N \times n$ matrix, whose components are each distributed as Bernoulli random variables with, possibly, different parameters. Then,

$$V_{N\times n} = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \vartheta_{13} & \cdots & \vartheta_{1n} \\ \vartheta_{21} & \vartheta_{22} & \vartheta_{23} & \cdots & \vartheta_{2n} \\ \vartheta_{31} & \vartheta_{32} & \vartheta_{33} & \cdots & \vartheta_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{N1} & \vartheta_{N2} & \vartheta_{N3} & \cdots & \vartheta_{Nn} \end{bmatrix}. \qquad (1.1)$$

Also part of the model is the restriction imposed on each column of V to add to one. In other words, we require that

$$\sum_{I=1}^{N} \vartheta_{Ik} = 1, \text{ for } k = 1, ..., n \qquad (1.2)$$

be satisfied.

This is required because if the j^{th} draw results in population unit I being selected, then entry (I, j) takes the value of one while all other entries of column j are equal to zero. Note that this is equivalent to imposing a non-stochastic constraint on the behavior of all components of the i^{th} column of V, regardless of the sampling scheme. Therefore, entries belonging to the same column do not behave independently.

When sampling takes place with replacement (WR), the sum of the elements of the I^{th} row of the above matrix is distributed as a Binomial (n, p_I) since each column is distributed independently of other columns. On the other hand, when sampling takes place without replacement (WOR), the total of row I can take only two values: one, if the I^{th} unit is drawn at some stage, or zero, otherwise, bringing us back to the Bernoulli case.

Disjoint subsets of rows may be formed according to different criteria. For instance, when rows are grouped with regard to their spatial vicinity, one could speak about clusters or primary sampling units. When one or more statistical indicators form the basis for the groupings, the term strata is usually used.

1. Victor Alfredo Bustos y de la Tijera, Instituto Nacional de Estadística, Geografia e Informática, H. de Nacozari 2301, 20270, Aguascalientes, Ags., México. E-mail: alfredo.bustos@inegi.gob.mx.

Let us now define the inclusion probabilities as

$$\pi_I^{(k)} = P(\text{population unit } I \text{ in sample of size } k) \\ = 0 \text{ if } k = 0. \tag{2}$$

Note that $\pi_I^{(n)} = \pi_I$, commonly refered to as the inclusion probability for unit I.

Now let $\underline{\vartheta}_{\circ j}$ represent the j^{th} column and $\underline{\vartheta}_{I\circ}$ the I^{th} row of matrix V. Therefore, based on the following expression,

$$f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) = f(\underline{\vartheta}_{\circ 1}) f(\underline{\vartheta}_{\circ 2} \mid \underline{\vartheta}_{\circ 1}) \\ f(\underline{\vartheta}_{\circ 3} \mid \underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}) ... f(\underline{\vartheta}_{\circ n} \mid \underline{\vartheta}_{\circ 1}, ..., \underline{\vartheta}_{\circ n-1}) \tag{3}$$

we can write the joint probability function of the elements of V as:

$$f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) = \prod_{k=1}^{n} \left[\prod_{I=1}^{N} (\pi_I^{(k)} - \pi_I^{(k-1)})^{\vartheta_{Ik}} \right] \\ = \prod_{k=1}^{n} \left[\prod_{I=1}^{N} (p_I^{(k)})^{\vartheta_{Ik}} \right] \tag{4}$$

subject to

$$\sum_{I=1}^{N} \vartheta_{Ik} = 1, k = 1, \ldots, n \text{ and}$$

$$\sum_{k=1}^{N} \vartheta_{Ik} \leq \begin{cases} 1, \text{ WOR} \\ n, \text{ WR} \end{cases} I = 1, \ldots, N;$$

and here $p_I^{(k)}$, defined as $p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)})$, stands for the probability that population unit I is included in the sample at the k^{th} draw. The above function is useful for calculating the probability of any ordered sample of size n. Clearly, when the order of inclusion can be ignored, the probability of a given sample would be obtained by adding the n! values obtained through (4).

3. The Implications of Sampling on the Stochastic Properties of Population Units

Consequently,

$$E(\vartheta_{Ik}) = p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)}) \tag{5}$$

and therefore, we can write

$$E[V] = \begin{bmatrix} p_1^{(1)} & p_1^{(2)} & p_1^{(3)} & \cdots & p_1^{(n)} \\ p_2^{(1)} & p_2^{(2)} & p_2^{(3)} & \cdots & p_2^{(n)} \\ p_3^{(1)} & p_3^{(2)} & p_3^{(3)} & \cdots & p_3^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_N^{(1)} & p_N^{(2)} & p_N^{(3)} & \cdots & p_N^{(n)} \end{bmatrix}. \tag{6}$$

From here, step-by-step inclusion probabilities, in WOR sampling situations, may be recursively computed, as is shown in (7), below.

$$p_I^{(k)} = \begin{cases} p_I & \text{if } k = 1 \\ p_I^{(k-1)} \sum_{J \neq I}^{N} \dfrac{p_J^{(k-1)}}{1 - p_J^{(k-1)}} & \text{if } k > 1. \end{cases} \tag{7}$$

Note that (7) enables us to compute the desired probabilities at two different moments: first, when no draw has actually occurred, which explains why we average over the whole population, and secondly, when the result of the previous draw is known, at which time the probability of the J^{th} population unit, say, entering the sample equals one and all other probabilities for that draw are equal to zero. Hence, at least in theory, we can compute the inverse of the so called expansion factors or weights for one stage sampling, or stage by stage in multistage sampling. Clearly,

$$\pi_I^{(n)} = \sum_{k=1}^{n} p_I^{(k)}. \tag{8}$$

If we define the joint inclusion probabilities as

$$\pi_{IJ}^{(k)} = P\left(\begin{matrix} \text{population units } I \text{ and} \\ J \text{ in sample of size } k \end{matrix} \right), \tag{9}$$

then we have that they can also be computed as follows:

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right). \tag{10}$$

For example, in simple random sampling WR (SRS/WR), expressions (7), (8) and (10) result in (7.1), (8.1) and (10.1),

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \geq 1 \tag{7.1}$$

$$\pi_I^{(n)} = \frac{n}{N} \tag{8.1}$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \\ = \sum_{j=1}^{n-1} \left(\frac{n-j}{N^2} + \frac{n-j}{N^2} \right) = \frac{n(n-1)}{N^2}. \tag{10.1}$$

While in SRS/WOR we get expressions (7.2), (8.2) and (10.2), instead.

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \geq 1 \tag{7.2}$$

$$\pi_I^{(n)} = \frac{n}{N} \tag{8.2}$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \text{ where } J \neq I$$

$$= \sum_{j=1}^{n-1}\left(\frac{n-j}{N(N-1)} + \frac{n-j}{N(N-1)} \right) = \frac{n(n-1)}{N(N-1)}. \tag{10.2}$$

Let us now consider the row vectors $\underline{\vartheta}_{I\circ}$. Then, for the covariance matrix between different rows, we get

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ}) = \begin{bmatrix} -p_I^{(1)}p_J^{(1)} & 0 & \cdots & 0 \\ 0 & -p_I^{(2)}p_J^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -p_I^{(n)}p_J^{(n)} \end{bmatrix}_{n\times n} \tag{11}$$

whenever I is different from J.

When sampling takes place WR, and therefore, $p_I^{(j)} = p_I \ \forall\ j = 1, ..., n,$ the covariance matrix for the I^{th} row vector is given by

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I q_I & 0 & 0 & \cdots & 0 \\ 0 & p_I q_I & 0 & \cdots & 0 \\ 0 & 0 & p_I q_I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_I q_I \end{bmatrix}_{n\times n}. \tag{12.1}$$

In a WOR setting the above covariance matrix becomes

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I^{(1)}(1-p_I^{(1)}) & -p_I^{(1)}p_I^{(2)} & \cdots & -p_I^{(1)}p_I^{(n)} \\ -p_I^{(1)}p_I^{(2)} & p_I^{(2)}(1-p_I^{(2)}) & \cdots & -p_I^{(2)}p_I^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_I^{(1)}p_I^{(n)} & -p_I^{(2)}p_I^{(n)} & \cdots & p_I^{(n)}(1-p_I^{(n)}) \end{bmatrix}_{n\times n} \cdot \tag{12.2}$$

Let $\underline{\vartheta}$ represent the N–dimensional vector which results from adding the columns of V. Clearly, the components of this vector may be expressed as the product of $\underline{\vartheta}_{I\circ}$ by a vector whose components are all equal to one. In other words,

$$\underline{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vdots \\ \vartheta_N \end{pmatrix} = \begin{pmatrix} \underline{\vartheta}_{1\circ}^T \underline{1} \\ \underline{\vartheta}_{2\circ}^T \underline{1} \\ \underline{\vartheta}_{3\circ}^T \underline{1} \\ \vdots \\ \underline{\vartheta}_{N\circ}^T \underline{1} \end{pmatrix}. \tag{13}$$

Some distributional properties of these sums may be then obtained directly from those of the rows or the columns of matrix V.

For instance, their expected values are given as

$$E(\vartheta_I) = E(\underline{\vartheta}_{I\circ}^T \underline{1}) = E\left(\sum_{k=1}^{n} \vartheta_{Ik} \right)$$

$$= \sum_{k=1}^{n} p_I^{(k)} = \pi_I^{(1)} + \sum_{k=2}^{n} (\pi_I^{(k)} - \pi_I^{(k-1)}) = \pi_I^{(n)}. \tag{14}$$

From (1.2), we get the non-stochastic restriction:

$$\underline{1}'\,\underline{\vartheta} = \vartheta_1 + \vartheta_2 + \vartheta_3 + ... + \vartheta_N = n. \tag{15}$$

From (14) and (15), well known propositions (16) and (17) follow immediately,

$$E[\underline{\vartheta}'] = (\pi_1^{(n)}, \pi_2^{(n)}, \pi_3^{(n)}, \ldots, \pi_N^{(n)}) \tag{16}$$

$$\pi_1^{(n)} + \pi_2^{(n)} + \pi_3^{(n)} + \ldots + \pi_N^{(n)} = n. \tag{17}$$

For the second order moments, we get

$$\mathrm{Cov}(\vartheta_I, \vartheta_J) = \mathrm{Cov}(\underline{1}'\,\underline{\vartheta}_{I\circ}, \underline{1}'\,\underline{\vartheta}_{J\circ})$$

$$= \underline{1}'\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ})\underline{1} = -\sum_{k=1}^{n} p_I^{(k)} p_J^{(k)}$$

$$= \begin{cases} -np_I p_J & \text{WR} \\ (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}) & \text{WOR}, \end{cases} \tag{18}$$

which clearly indicates that the covariance is never positive. In turn, the variances are given by

$$\mathrm{Var}(\vartheta_I) = \mathrm{Var}(\underline{1}'\,\underline{\vartheta}_{I\circ}) = \underline{1}'\,\mathrm{Cov}(\underline{\vartheta}_{I\circ})\,\underline{1}$$

$$= \begin{cases} np_I q_I & \text{WR} \\ \pi_I^{(n)}(1-\pi_I^{(n)}) & \text{WOR}. \end{cases} \tag{19}$$

Another important consequence of (15) has to do with the second order moments of the stochastic vector $\underline{\vartheta}$.

$$0 = \mathrm{Var}(n) = \mathrm{Var}(\underline{1}'\,\underline{\vartheta}) = \underline{1}'\mathrm{Cov}(\underline{\vartheta})\underline{1} = \underline{1}'C\underline{1}. \tag{20}$$

Clearly, the diagonal elements of matrix C, the covariance matrix of $\underline{\vartheta}$, are not all equal to zero. Therefore, randomly drawing a fixed-size simple introduces a dependency in the population units which results in non-null covariances implying that matrix C is singular. Otherwise, it is impossible for (20) to be satisfied.

As a matter of fact, it is possible to prove that the sum of any row (or column) of C must be equal to zero, which is a stronger statement. Given that the covariance between a random variable and a constant equals zero, we get

$$0 = \mathrm{Cov}(\vartheta_I, n) = \mathrm{Cov}(\vartheta_I, \vartheta_1 + \vartheta_2 + \vartheta_3 + \cdots + \vartheta_N)$$
$$= C_{I1} + C_{I2} + \cdots + C_{IN}$$
$$= \mathrm{Var}(\vartheta_I) + \sum_{J \neq I} \mathrm{Cov}(\vartheta_I, \vartheta_J). \quad (21)$$

We have thus proven that in WOR sampling (22.1) holds.

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}). \quad (22.1)$$

The same statement can be proven algebraically by noting that

$$\sum_{J \neq I} \pi_{IJ}^{(n)} = \pi_I^{(n)} \sum_{J \neq I} \pi_{J|I}^{(n)}$$
$$= (n-1)\pi_I^{(n)},$$

which is obvious once we realize that the conditional probability involved represents the probability that population unit J enters a sample of size $n-1$ for which (19) also applies. Additionally, using (19) again, note that

$$\sum_{J \neq I} \pi_J^{(n)} = (n - \pi_I^{(n)}),$$

and therefore,

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)})$$
$$= \pi_I^{(n)} - (\pi_I^{(n)})^2 + (n-1)\pi_I^{(n)} - \pi_I^{(n)}(n - \pi_I^{(n)}).$$

For WR sampling (21) implies:

$$0 = np_Iq_I + \sum_{J \neq I} (n(n-1)p_Ip_J - n^2 p_I p_J)$$
$$= np_Iq_I - np_I \sum_{J \neq I} p_J \quad (22.2)$$

which is immediately seen to apply.

In any case, the most important implication of the above results is that regardless of the sampling scheme, the correlation matrix of the population random variables $\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_N$ is singular. For the practical situations described in the introduction, the most important implication of this fact lies mainly in the use made by many model fitting and estimation procedures of the inverse of the covariance matrix.

4. The First Two Moments of Sample Units

Once the first and second order moments of the vector $\underline{\vartheta}$ have been established, we are in a position to determine the corresponding moments for sub-vectors of different sizes and whose components are randomly chosen, *i.e.*, the sample. To this end, let us define the random variables $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r}$, where r represents the number of different population units in the sample, and whose indices $I_k, 1 \leq k \leq r \leq n,$ can take the value I with probability $\pi_I^{(n)}$. In other words, under the above conditions, we are in the presence of a set of random variables whose indices are random themselves.

4.1 Mean and Variance for WR Sampling

For this case, the probability function of ϑ_{I_i} is given by

$$P(\vartheta_{I_j} = x) = \sum_{I=1}^{N} p_I P(\vartheta_I = x)$$
$$= \sum_{I=1}^{N} p_I \binom{n}{x} p_I^x (1 - p_I)^{n-x}. \quad (23)$$

The first two moments may also be obtained via a conditional argument. The mean of its distribution is given by

$$E(\vartheta_{I_j}) = \sum_{I=1}^{N} p_I E(\vartheta_I) = \sum_{I=1}^{N} np_I\, p_I = n\sum_{I=1}^{N} p_I^2. \quad (24)$$

In turn, its variance is computed using the well known formula

$$V(\vartheta_{I_j}) = V_{I_j}[E(\vartheta_{I_j} \mid I_j)] + E_{I_j}[V(\vartheta_{I_j} \mid I_j)]. \quad (25)$$

In this case, we have

$$E(\vartheta_{I_j} \mid I_j = I) = np_I$$
$$\text{and } V(\vartheta_{I_j} \mid I_j = I) = np_I(1 - p_I). \quad (26)$$

Hence,

$$V_{I_j}[E(\vartheta_{I_j} \mid I_j)] = V_{I_j}(np_{I_j})$$
$$= n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})],$$
$$E_{I_j}[V(\vartheta_{I_j} \mid I_j)] = nE_{I_j}[p_{I_j}(1 - p_{I_j})]$$
$$= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] \quad (27)$$

and therefore

$$V(\vartheta_{I_j})$$
$$= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] + n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})]$$
$$= \sum_{I=1}^{N} np_I^2 \left(1 + (n-1)p_I - \sum_{J=1}^{N} np_J^2\right). \quad (28)$$

For the case of SRS, (24) above results in

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N} (\pi_I^{(n)})^2 = \frac{1}{n}\sum_{I=1}^{N} \left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2 \frac{N}{n} = \frac{n}{N}.$$

While (28) yields

$$V(\vartheta_{I_j}) = \sum_{I=1}^{N} n\frac{1}{N^2}\left(1 + (n-1)\frac{1}{N} - \sum_{J=1}^{N} n\frac{1}{N^2}\right) = n\frac{1}{N}\left(1 - \frac{1}{N}\right).$$

4.2 Mean and Variance for WOR Sampling

For this case, the probability function of ϑ_{I_i} is given by

$$P(\vartheta_{I_j} = x) = \frac{1}{n}\sum_{I=1}^{N} \pi_I^{(n)} \sum_{k=1}^{n} (p_I^{(k)})^x (1 - p_I^{(k)})^{1-x} \quad (29)$$

and therefore

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}E(\vartheta_I) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}\sum_{j=1}^{n}(p_I^{(j)}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2. \quad (30)$$

Using (25) again, we note firstly that

$$E(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)} \text{ and } V(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)}(1-\pi_{I_j}^{(n)})$$

from which we get

$$V[E(\vartheta_{I_j} \mid I_j)] = V(\pi_{I_j}^{(n)}) = E[(\pi_{I_j}^{(n)})^2] - [E(\pi_{I_j}^{(n)})]^2$$

and

$$E[V(\vartheta_{I_j} \mid I_j)] = E[\pi_{I_j}^{(n)}(1-\pi_{I_j}^{(n)})] = E[(\pi_{I_j}^{(n)})] - [E(\pi_{I_j}^{(n)})^2].$$

Hence, the variance is given by

$$V(\vartheta_{I_j}) = E(\pi_{I_j}^{(n)}) - E^2(\pi_{I_j}^{(n)}) = E(\pi_{I_j}^{(n)})[1-E(\pi_{I_j}^{(n)})] = \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\left[1-\left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\right]. \quad (31)$$

Once again, in order to exemplify these results, let us turn to SRS. Expression (30) becomes

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2 = \frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2\frac{N}{n} = \frac{n}{N}. \quad (32)$$

Whereas (31) results in

$$V(\vartheta_{I_j}) = \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\left[1-\left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\right] = \frac{n}{N}\left(1-\frac{n}{N}\right). \quad (33)$$

4.3 The Covariance Between Sample Units

In order to establish the covariance between different sample units we resort to a simple extension to (25),

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \mathrm{Cov}_{I_j, I_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)] + E_{I_j, I_j}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)]. \quad (34)$$

In this case, we have that

$$E(\vartheta_{I_j} \mid I_j = I) = \pi_I^{(n)} \quad (35)$$

and

$$E(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j = I, I_k = J) = \pi_{IJ}^{(n)} \quad (36)$$

while the covariance between brackets on the right-hand side of (34) is easily seen to equal

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}/I_j = I, I_k = J) = \pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_j^{(n)}. \quad (37)$$

From (35) and (36), we obtain

$$\mathrm{Cov}_{I_jI_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)] = E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}) - E_{I_j}(\pi_{I_j}^{(n)})E_{I_k}(\pi_{I_k}^{(n)}) \quad (38)$$

whereas from (37) we get

$$E_{I_jI_k}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)] = E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}). \quad (39)$$

Finally, adding these last two expressions we arrive at the desired covariance

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - [E_{I_j}(\pi_{I_j}^{(n)})][E_{I_k}(\pi_{I_k}^{(n)})] = \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{\substack{J=1\\J\neq I}}^{N}(\pi_{IJ}^{(n)})^2 - \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)^2. \quad (40)$$

In the SRS/WR (40) results in

$$\begin{aligned}\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{J=1}^{N}\left(\frac{n(n-1)}{N^2}\right)^2 - \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)^2 \\ &= \frac{n(n-1)}{N^2} - \frac{n^2}{N^2} \\ &= -\frac{n}{N^2}, \end{aligned} \quad (41)$$

while for the WOR case the covariance can be seen to equal

$$\begin{aligned}\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{\substack{J=1\\J\neq I}}^{N}\left(\frac{n(n-1)}{N(N-1)}\right)^2 - \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)^2 \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \\ &= -\left(\frac{n(N-n)}{N^2(N-1)}\right). \end{aligned} \quad (42)$$

It should be stressed that for SRS, regardless of whether it takes place with or without replacement, the correlation coefficients are given by

$$\text{Corr}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{-1}{(N-1)}, \qquad (43)$$

independently of the sample size.

Furthermore, we have that, as the value of n approaches that of N in WOR sampling, both $\pi_I^{(n)}$ and $\pi_{IJ}^{(n)}$ approach one. In particular, when $n = N$, the values of expressions (31) and (40) become zero.

5. The Correlation Matrix for Sample Units

Once we realize that none of the expressions in (28), (31) and (40) depend on any of the arbitrary indices used to differentiate population units, it should become clear that the $r \times r$ correlation matrix for the random vector $\underline{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r})$, where $r \leq n$, may be written as:

$$\text{Corr}(\underline{\theta}) = R_r(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}. \qquad (44)$$

It should be noted that the elements of $R_r(\rho)$ in (44) depend only on the inclusion probabilities which, for any sample size, may be fully computed from recursion (7), and expressions (8) and (10). In other words, they do not depend on any unknown population parameters to be estimated nor on the values of the variables to be measured on the sample units.

6. Final Remarks

In theory, the efficiency of every estimation procedure will experience some gain whenever explicit allowance for the correlation between sample units is made. This would certainly be the case for linear as well as for some instances of maximum-likelihood estimation.

On the other hand, it should be emphasized that $R_n(\rho)$ may become singular as the sample size n approaches the population size N; this is the case for SRS ($R_N(-1/(N-1))$ as well as for WOR sampling in general. Therefore, numerically, many estimation procedures which rely on the inverse or the determinant of R, rather than on the correlation matrix itself, may also benefit from replacing the simplifying assumption of independence between observations by a more realistic one of correlated observations whenever sample sizes are large relative to population sizes. Instances where this can happen are given by some stages in multi-stage sampling (*e.g.*, number of households in a block) and by large country-wide surveys.