Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2005

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Décembre 2005

Publication autorisée par le ministre responsable de Statistique Canada

Mai 2006

Nº 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Vol. 31, N° 2, pp. 255-260
Statistique Canada, n° 12-001 au catalogue

Structure de corrélation des unités d'échantillonnage

Alfredo Bustos [1]

Résumé

Nous explicitons dans cet article certaines propriétés distributionnelles des unités d'échantillonnage qui ne sont habituellement pas décrites dans la documentation, notamment leur structure de corrélation et le fait que celle-ci ne dépend pas d'indices de population attribués arbitrairement. Ces propriétés importent pour plusieurs méthodes d'estimation, dont l'efficacité serait améliorée si on les mentionnait explicitement.

Mots clés : Recensement; enquête; échantillonnage; unités d'échantillonnage; fonction de probabilité; moyenne; covariance.

1. Introduction

Ces dernières années, la réalisation des recensements de la population et des ménages tels que nous les connaissons est devenue plus ardue pour plusieurs raisons. Par conséquent, d'autres moyens de recueillir plus fréquemment l'information requise pour la production de statistiques aux niveaux local, provincial et national ont été proposés. De grandes enquêtes nationales continues, notamment celles appelées recensement continu, réalisées auprès d'échantillons de grande taille selon des plans d'enquête complexes, sont envisagées.

Cependant, afin de produire des résultats au niveau local comparables à ceux d'un recensement, il faut mettre au point diverses méthodes d'estimation et de validation, ainsi que, dans certains cas, d'imputation et améliorer leur efficacité. Un moyen d'accroître l'efficacité consiste à tenir compte de toute l'information pertinente disponible. Naturellement, cela englobe les propriétés stochastiques des unités d'échantillonnage.

Dans la suite de l'exposé, en partant de principes fondamentaux, nous dérivons une forme générale explicite de la fonction de probabilité d'un échantillon ordonné. Nous montrons aussi comment on peut calculer cette fonction, ainsi que les probabilités d'inclusion. Enfin, nous donnons une forme générale de la matrice des corrélations des unités d'échantillonnage qui ne dépend que des probabilités d'inclusion, de sorte qu'il soit possible d'améliorer les méthodes d'estimation linéaires et du maximum de vraisemblance.

2. Le modèle de base

Le modèle de base dont nous partons représente le tirage séquentiel aléatoire de n unités à partir d'une population U formée de N de ces unités et peut être énoncé comme suit. Soit N et n deux constantes positives telles que $n \le N$, et soit V une matrice de dimensions $N \times n$, dont les composantes sont distribuées chacune comme des variables aléatoires de Bernoulli avec, éventuellement, des paramètres différents. Alors,

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

Fait aussi partie du modèle la contrainte voulant que la somme des éléments de chaque colonne de V soit égale à l'unité. Autrement dit, nous exigeons que la condition

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

soit satisfaite.

Cette condition est nécessaire, parce que si le j^{e} tirage donne lieu à la sélection de l'unité de population I, alors l'élément (I, j) prend la valeur de un, tandis que tous les autres éléments de la colonne j sont nuls. Notons que cela équivaut à imposer une contrainte non stochastique au comportement de toutes les composantes de la i^{e} colonne de V, indépendamment du plan d'échantillonnage. Par conséquent, les éléments appartenant à une même colonne ne se comportent pas de façon indépendante.

Lorsque l'échantillonnage a lieu avec remise (WR pour *with replacement*), la somme des éléments de la I^{e} ligne de la matrice susmentionnée suit une loi binomiale (n, p_I), puisque la distribution de chaque colonne est indépendante de celle des autres. Par ailleurs, si l'échantillonnage se fait sans remise (WOR pour *without replacement*), le total de la

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

ligne I ne peut prendre que deux valeurs : un, si la I^{e} unité est tirée à un certain degré, ou zéro, autrement, ce qui nous ramène au cas de Bernoulli.

Nous pouvons former des sous-ensembles disjoints de lignes conformément à divers critères. Par exemple, si nous regroupons les lignes en fonction de leur voisinage spatial, nous pourrions parler de grappes ou d'unités primaires d'échantillonnage. Si nous fondons le groupement sur un ou plusieurs indicateurs statistiques, nous utilisons habituellement le terme de strate.

Définissons maintenant les probabilités d'inclusion comme étant

$$\begin{aligned}\pi_I^{(k)} &= P(\text{unité de population } I \text{ dans l'échantillon} \\ &\qquad \text{de taille } k) \\ &= 0 \text{ si } k = 0.\end{aligned} \tag{2}$$

Notons que $\pi_I^{(n)} = \pi_I$, habituellement nommée probabilité d'inclusion de l'unité I.

Représentons maintenant par $\underline{\vartheta}_{\circ j}$ la j^{e} colonne et par $\underline{\vartheta}_{I\circ}$ la I^{e} ligne de la matrice V. Par conséquent, en nous basant sur l'expression suivante,

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

nous pouvons écrire la fonction de probabilité conjointe des éléments de V sous la forme :

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

sachant que

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

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

et ici $p_I^{(k)}$, définie comme étant $p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)})$, représente la probabilité que l'unité de population I soit incluse dans l'échantillon lors du k^{e} tirage. La fonction susmentionnée est utile pour le calcul de la probabilité de tout échantillon ordonné de taille n. Manifestement, si l'on peut ignorer l'ordre d'inclusion, on obtiendra la probabilité d'un échantillon donné en ajoutant les n! valeurs obtenues au moyen de (4).

3. Les incidences de l'échantillonnage sur les propriétés stochastiques des unités de population

Conséquemment,

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

et, donc, nous pouvons écrire

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

À partir de là, nous pouvons calculer récursivement les probabilités d'inclusion étape par étape, dans les situations d'échantillonnage sans remise, comme le montre l'équation (7) qui suit.

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

Il convient de souligner que (7) nous permet de calculer les probabilités souhaitées à deux moments distincts : en premier lieu, quand aucun tirage n'a effectivement eu lieu, ce qui explique pourquoi nous calculons la moyenne sur l'ensemble de la population et, en deuxième lieu, quand on connaît le résultat du tirage précédent, moment auquel la probabilité que la J^{e} unité de population, disons, entre dans l'échantillon est égale à 1 et toutes les autres probabilités pour ce tirage sont nulles. Par conséquent, du moins en théorie, nous pouvons calculer l'inverse des facteurs dits d'expansion ou poids pour l'échantillonnage à un degré, ou étape par étape pour l'échantillonnage à plusieurs degrés. De toute évidence,

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

Si nous définissons les probabilités d'inclusion conjointes comme étant

$$\pi_{IJ}^{(k)} = P\left(\begin{array}{c}\text{unités de population } I \text{ et} \\ J \text{ dans l'échantillon de taille } k\end{array}\right), \tag{9}$$

alors nous savons qu'elles peuvent également être calculées comme suit :

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

Par exemple, dans le cas de l'échantillonnage aléatoire simple avec remise (EAS/WR), les expressions (7), (8) et (10) donnent lieu à (7.1), (8.1) et (10.1),

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

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

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

Dans le cas de l'EAS/WOR, nous obtenons, à la place, les expressions (7.2), (8.2) et (10.2).

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

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

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \text{ où } J \neq I$$
$$= \sum_{j=1}^{n-1}\left(\frac{n-j}{N(N-1)} + \frac{n-j}{N(N-1)} \right) = \frac{n(n-1)}{N(N-1)}. \qquad (10.2)$$

Considérons maintenant les vecteurs de ligne $\underline{\vartheta}_{I\circ}$. Alors, pour la matrice des covariances entre diverses lignes, nous obtenons

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

dans tous les cas où I est différent de J.

En cas d'échantillonnage avec remise où, par conséquent, $p_I^{(j)} = p_I \ \forall\ j = 1, ..., n,$ la matrice des covariances pour le I^{e} vecteur de ligne est donnée par

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

Dans le cas de l'échantillonnage sans remise, la matrice des covariances susmentionnée devient

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

Soit $\underline{\vartheta}$ le vecteur de dimension N qui résulte de l'addition des colonnes de V. De toute évidence, les composantes de ce vecteur peuvent être exprimées sous forme du produit de $\underline{\vartheta}_{I\circ}$ par un vecteur dont les composantes sont toutes égales à un. Autrement dit,

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

Certaines propriétés distributionnelles de ces sommes peuvent alors être obtenues directement d'après celles des lignes ou des colonnes de la matrice V.

Par exemple, leurs valeurs attendues sont données par

$$E(\vartheta_I) = E(\underline{\vartheta}_{I\circ}^T \underline{1}) = E\left(\sum_{k=1}^{n} \vartheta_{Ik} \right)$$
$$= \sum_{k=1}^{n} p_I^{(k)} = \pi_I^{(1)} + \sum_{k=2}^{n} (\pi_I^{(k)} - \pi_I^{(k-1)}) = \pi_I^{(n)}. \qquad (14)$$

De (1.2), nous tirons la restriction non stochastique :

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

De (14) et (15) découlent directement les propositions bien connues (16) et (17),

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

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

Pour les moments de deuxième ordre, nous obtenons

$$\mathrm{Cov}(\vartheta_I, \vartheta_J) = \mathrm{Cov}(\underline{1}'\,\underline{\vartheta}_{I\circ}, \underline{1}'\,\underline{\vartheta}_{J\circ})$$
$$= \underline{1}'\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ})\,\underline{1} = -\sum_{k=1}^{n} p_I^{(k)} p_J^{(k)}$$
$$= \begin{cases} -np_I p_J & \text{WR} \\ (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}) & \text{WOR}, \end{cases} \qquad (18)$$

qui indique clairement que la covariance n'est jamais positive. À leur tour, les variances sont données par

$$\mathrm{Var}(\vartheta_I) = \mathrm{Var}(\underline{1}'\,\underline{\vartheta}_{I\circ}) = \underline{1}'\,\mathrm{Cov}(\underline{\vartheta}_{I\circ})\,\underline{1}$$
$$= \begin{cases} np_I q_I & \text{WR} \\ \pi_I^{(n)}(1-\pi_I^{(n)}) & \text{WOR}. \end{cases} \qquad (19)$$

Une autre conséquence importante de (15) concerne les moments de deuxième ordre du vecteur stochastique $\underline{\vartheta}$.

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

De toute évidence, les éléments diagonaux de la matrice C, la matrice des covariances de $\underline{\vartheta}$, ne sont pas tous nuls. Par conséquent, le tirage aléatoire d'un échantillon de taille fixe introduit dans les unités de population une dépendance qui donne lieu à des covariances non nulles sous-entendant

que la matrice C est singulière. Sinon, il est impossible que l'équation (20) soit satisfaite.

En fait, il est possible de prouver que la somme des éléments de toute ligne (ou colonne) de C doit être nulle, ce qui est un énoncé plus ferme. Sachant que la covariance entre une variable aléatoire et une constante est nulle, nous obtenons

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

Nous avons donc prouvé que, dans le cas de l'échantillonnage sans remise, (22.1) est vérifiée.

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

Le même énoncé peut être prouvé algébriquement en notant que

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

ce qui est évident si l'on se rend compte que la probabilité conditionnelle concernée représente la probabilité que l'unité de population J entre dans un échantillon de taille $n-1$ pour lequel s'applique aussi l'expression (19). En outre, en utilisant de nouveau (19), notons que

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

et, par conséquent,

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

Pour l'échantillonnage avec remise, (21) implique que :

$$\begin{aligned} 0 &= np_I q_I + \sum_{J \neq I} (n(n-1) p_I p_J - n^2 p_I p_J) \\ &= np_I q_I - np_I \sum_{J \neq I} p_J \end{aligned} \tag{22.2}$$

condition qui, on le voit directement, s'applique.

En tout cas, l'incidence la plus importante des résultats susmentionnés est que, indépendamment du plan d'échantillonnage, la matrice de corrélation des variables aléatoires de population $\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_N$ est singulière. En ce qui concerne les situations pratiques décrites dans l'introduction, la conséquence la plus importante tient principalement au fait que l'inverse de la matrice des covariances est utilisée dans de nombreuses méthodes d'ajustement et d'estimation de modèles.

4. Les deux premiers moments des unités d'échantillonnage

Après avoir établi les moments de premier et de deuxième ordre du vecteur $\underline{\vartheta}$, il nous est possible de déterminer les moments correspondants de sous-vecteurs de diverses tailles et dont les composantes sont sélectionnées aléatoirement, c'est-à-dire l'échantillon. À cette fin, définissons les variables aléatoires $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r}$, où r représente le nombre d'unités de population différentes dans l'échantillon et dont les indices $I_k, 1 \leq k \leq r \leq n,$ peuvent prendre la valeur I avec la probabilité $\pi_I^{(n)}$. Autrement dit, sous les conditions susmentionnées, nous sommes en présence d'un jeu de variables aléatoires dont les indices sont eux-mêmes aléatoires.

4.1 Moyenne et variance pour l'échantillonnage avec remise

Dans ce cas, la fonction de probabilité de ϑ_{I_j} est donnée par

$$\begin{aligned} P(\vartheta_{I_j} = x) &= \sum_{I=1}^{N} p_I P(\vartheta_I = x) \\ &= \sum_{I=1}^{N} p_I \binom{n}{x} p_I^x (1 - p_I)^{n-x}. \end{aligned} \tag{23}$$

Les deux premiers moments peuvent aussi être obtenus par la voie d'un argument conditionnel. La moyenne de sa distribution est donnée par

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

À son tour, sa variance est calculée à l'aide de la formule bien connue

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

Dans ce cas, nous avons

$$\begin{aligned} E(\vartheta_{I_j} \mid I_j = I) &= np_I \\ \text{et } V(\vartheta_{I_j} \mid I_j = I) &= np_I(1 - p_I). \end{aligned} \tag{26}$$

Donc,

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

et, par conséquent

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

Pour le cas de l'EAS, l'équation (24) qui précède donne

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

Tandis que (28) donne

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

4.2 Moyenne et variance pour l'échantillonnage sans remise

Dans ce cas, la fonction de probabilité de ϑ_{I_i} est donnée par

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

et, par conséquent,

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

En utilisant de nouveau (25), nous commençons par noter que

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

dont nous tirons

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

et

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

Donc, la variance est donnée par

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

Une fois de plus, afin d'illustrer ces résultats au moyen d'un exemple, considérons l'EAS. L'expression (30) devient

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

Tandis que (31) donne

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

4.3 La covariance entre les unités d'échantillonnage

Afin d'établir la covariance entre les diverses unités d'échantillonnage, nous recourons à une simple extension de (25),

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

Dans ce cas, nous savons que

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

et

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

tandis qu'il est facile de voir que la covariance entre crochets dans le deuxième membre de (34) est égale à

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

À partir de (35) et (36), nous obtenons

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

tandis que, de (37), nous obtenons

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

Enfin, en additionnant ces deux dernières expressions, nous obtenons la covariance souhaitée

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

Dans le cas de l'EAS/WR, (40) donne

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

tandis que dans le cas sans remise, on peut voir que la covariance est égale à

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

Il convient de souligner que, dans le cas de l'EAS, que le tirage se fasse avec ou sans remise, les coefficients de corrélation sont donnés par

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

indépendamment de la taille de l'échantillon.

De surcroît, nous savons que, à mesure que la valeur de n s'approche de celle de N dans l'échantillonnage sans remise, $\pi_I^{(n)}$ et $\pi_{IJ}^{(n)}$ s'approchent l'une et l'autre de l'unité. En particulier, quand $n = N$, les valeurs des expressions (31) et (40) deviennent nulles.

5. La matrice des corrélations des unités d'échantillonnage

Dès que l'on se rend compte qu'aucune des expressions (28), (31) et (40) ne dépend d'aucun des indices arbitraires utilisés pour différencier les unités de population, il devient clair que la matrice $r \times r$ des corrélations pour le vecteur aléatoire $\underline{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r})$, où $r \leq n$, peut s'écrire :

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

Il convient de souligner que les éléments de $R_r(\rho)$ dans (44) dépendent uniquement des probabilités d'inclusion qui, pour toute taille d'échantillon, peuvent être calculées entièrement d'après la récursion (7) et les expressions (8) et (10). Autrement dit, elles ne dépendent d'aucun paramètre de population inconnu qu'il faut estimer ni des valeurs des variables qui doivent être mesurées sur les unités d'échantillonnage.

6. Remarques finales

En théorie, l'efficacité de toute méthode d'estimation s'accroîtrait dans une certaine mesure si l'on tenait compte explicitement de la corrélation entre les unités d'échantillonnage. Il en serait certainement ainsi pour l'estimation linéaire et, dans certains cas, pour l'estimation du maximum de vraisemblance.

Par ailleurs, il convient d'insister sur le fait que $R_n(\rho)$ peut devenir singulière à mesure que la taille de l'échantillon n s'approche de la taille de la population N; il en est ainsi pour l'EAS ($R_N(-1/(N-1))$, ainsi que pour l'échantillonnage sans remise en général. Par conséquent, numériquement, nombre de méthodes d'estimation qui s'appuient sur l'inverse ou le déterminant de R plutôt que sur la matrice des corrélation proprement dite pourraient également bénéficier du remplacement de l'hypothèse simplifiante d'indépendance entre les observations par une hypothèse plus réaliste d'observations corrélées quand la taille de l'échantillon est grande relativement à taille de la population. Les cas où cela est possible se produisent à certaines étapes dans l'échantillonnage à plusieurs degrés (par exemple nombre de ménages dans un îlot) et dans de grandes enquêtes à l'échelle du pays.