Techniques d'enquête, juin 1979
Vol. 5, nº 1
Statistique Canada, nº 12-001-X au catalogue

Estimation sans biais des proportions dans un échantillonnage séquentiel

M.D. Bankier[1]

Résumé

Dans un plan de sondage séquentiel, la proportion défectueuse de l'échantillon est en général un estimateur biaisé de la valeur de la population. L'auteur de l'article propose un estimateur sans biais, dont un estimateur sans biais de la variance est également défini. Les résultats sont appliqués à un problème d'estimation tiré du recensement de 1976.

Mots-clés : Contrôle de la qualité; plan de sondage séquentiel; estimateur sans biais; recensement canadien.

1 Introduction

Dans le cadre d'un contrôle de la qualité, un échantillon de m unités d'un certain lot est examiné. Le nombre x d'unités défectueuses dans l'échantillon détermine si le lot est accepté ou rejeté. L'utilisation d'un plan d'échantillonnage séquentiel est courante. Une définition de tels plans est donnée à la section 2. Pour ces plans d'échantillonnage, le nombre d'unités échantillonnées est une variable aléatoire x/m ; il s'agit habituellement d'un estimateur avec biais de la proportion d'unités défectueuses dans le lot. À la section 3, il est question d'un estimateur p(m, x) sans biais selon l'échantillonnage séquentiel avec remise (a.r.) et sans remise (s.r.). Cet estimateur a d'abord été proposé par Girshick, Mosteller et Savage [3] dans le cadre de l'échantillonnage a.r. Un résultat donné par DeGroot [1] est également abordé. Il indique à quels plans d'échantillonnage p(m, x) correspond l'estimateur sans biais à variance minimale uniforme selon l'échantillonnage a.r. d'une population infinie. À la section 4, les estimateurs sans biais de V(p(m, x)) sont dérivés à la fois en fonction de l'échantillonnage a.r. et s.r. Ces estimateurs sont élaborés à l'aide d'une approche adoptée dans l'échantillonnage a.r. dans [3]. À la section 5, les résultats sont appliqués à un problème d'estimation tiré du recensement canadien de 1976.

1. M.-D. Bankier, Recensement et Division des méthodes d'enquête auprès des ménages, Statistique Canada.

2 Définition d'un plan d'échantillonnage

Si (m, x) représente un cas où m unités ont été échantillonnées et que de ces unités, x sont défectueuses. Un plan d'échantillonnage est défini comme une fonction S, de sorte que si :

$S((m, x)) = 1$, alors une autre unité est échantillonnée et inspectée. (2.1)

Si

$S((m, x)) = 0$, alors aucune autre unité n'est échantillonnée. (2.2)

L'événement (m, x) est appelé un point limite si $S((m, x)) = 0$ et $Pr((m, x)) > 0$. L'ensemble de tous les points limites d'un plan d'échantillonnage sera étiqueté B. Cet ensemble contient les résultats possibles du plan d'échantillonnage. On dit qu'un plan d'échantillonnage est fermé si :

$$\sum_{(m,x)\in B} Pr((m,x)) = 1 \tag{2.3}$$

Un plan d'échantillonnage borné est un plan où il existe un nombre entier positif b de sorte que

$$Pr(m \leq b) = 1 \tag{2.4}$$

La taille d'un plan d'échantillonnage est la plus petite valeur b dans l'équation (2.4). Seuls les plans d'échantillonnage fermés et bornés seront pris en compte dans le présent document. Les plans d'échantillonnage limitatifs simples, doubles, multiples et séquentiels utilisés pour le contrôle de la qualité satisfont tous aux critères susmentionnés.

3 Un estimateur sans biais de la proportion défectueuse

Supposons :

$$x_i = \begin{cases} 1 \text{ si la } i^{\text{ème}} \text{ unité échantillonnée est défectueuse} \\ \\ 0 \text{ autrement.} \end{cases} \tag{3.1}$$

Un chemin vers (m, x) sera défini comme une séquence de points :

$$(j, \sum_{i=1}^{j} x_i) \qquad j = 1, \ldots, m \quad \sum_{i=1}^{m} x_i = x \tag{3.2}$$

de sorte que

$$s((j, \sum_{i=1}^{j} x_i)) = 1 \quad \text{pour} \quad j = 1, \ldots, m-1. \tag{3.3}$$

Supposons que les unités sont échantillonnées a.r., P étant la proportion défectueuse. Selon ces hypothèses, on peut voir que

$$Pr((m,x)) = c(m,x)P^x(1-P)^{m-x} \tag{3.4}$$

où

$$c(m,x) = \text{le nombre de chemins distincts vers } (m, x). \tag{3.5}$$

Supposons

$$\bar{y}_i = \frac{\text{nombre d'unités défectueuses dans le premier échantillon}}{i} \tag{3.6}$$

La moyenne de l'échantillon $\bar{y}_m = x/m$ est généralement un estimateur biaisé de P, sauf lorsque la taille de l'échantillon m est établie. Cependant, comme au moins une unité est toujours échantillonnée, $\bar{y}_1$ n'est pas biaisé. L'estimateur suivant :

$$\begin{aligned} p(m,x) &= E(\bar{y}_1 \mid (m,x)) \\ &= \frac{Pr(\bar{y}_1 = 1 \text{ and } (m,x))}{Pr((m,x))} \\ &= \frac{d(m,x)P^x(1-P)^{m-x}}{c(m,x)P^x(1-P)^{m-x}} \\ &= \frac{d(m,x)}{c(m,x)} \end{aligned} \tag{3.7}$$

est également un estimateur sans biais de P comportant une variance identique ou inférieure à $\bar{y}_1$ selon le théorème de Rao-Blackwell. Dans l'expression ci-dessus

$$d(m,x) = \text{le nombre de chemins distincts vers } (m, x) \text{ qui traversent le point } (1, 1) \qquad (3,8)$$

Selon l'échantillonnage s.r. d'une population finie, l'expression (3.7) tient toujours. Le $P^x(1 - P)^{m-x}$ est remplacé par :

$$G(m,x;M,X) = \begin{cases} \dfrac{\binom{X}{x}\binom{M-X}{m-x}}{\binom{m}{x}\binom{M}{m}} & \text{pour } m \leqq M,\ x \leqq X,\ X \leqq M,\ x \leqq m \\ 0 & \text{autrement} \end{cases} \qquad (3.9)$$

où M est le nombre d'unités dans la population et X correspond au nombre d'unités défectueuses. Dans ce cas, $P = X/M$.

Un plan d'échantillonnage est complet si le seul estimateur f qui fait que $E(f) = 0$ pour toutes les valeurs B, en est un défini par $f(m, x) = 0$ pour toutes les valeurs $(m, x) \in B$. Dans DeGroot [1], il est démontré que selon l'échantillonnage a.r., un plan d'échantillonnage de taille b est complet si et seulement si la limite B contient exactement b + 1 points. Si le plan d'échantillonnage est complet et que la taille de la population est infinie, alors selon le théorème de l'unicité de Lehmann-Scheffé (Roussas [4], p. 216), p(m, x) est l'estimateur sans biais à variance minimale de P.

4 Un estimateur sans biais de v(p(m, x))

Pour tout plan d'échantillonnage fermé, complet et borné ayant une taille $b \geq 2$, il est possible d'établir un estimateur sans biais v(p(m, x)) de V(p(m, x)) selon l'échantillonnage a.r. et s.r. Les expressions de v(p(m, x)) selon l'échantillonnage s.r. d'une population finie sont données ci-dessous pour trois types d'ensembles de points bornés B. Ces expressions se réduisent à celles de l'échantillonnage a.r. si $M - 1/M$ est remplacé par 1 et $1/M$ par 0.

Cas (1) : Nous supposons que (2, 2) ∈ B. Alors,

$$v(p(m,x)) = \begin{cases} 0 \text{ pour } (m,x) = (2,2) \\ p(m,x)\left(p(m,x) - \frac{1}{M}\right) \text{ sinon pour } (m,x) \in B. \end{cases} \tag{4.1}$$

Cas (2) : Nous supposons que (2, 2) ∉ B, mais (1, 1) ∈ B. Alors,

$$v(p(m,x)) = \begin{cases} 0 \text{ pour } (m,x) = (1,1) \\ \frac{M-1}{M} \frac{e(m,x)}{c(m,x)} \text{ sinon pour } (m,x) \in B \end{cases} \tag{4.2}$$

où e(m, x) = le nombre de chemins distincts passant par (1, 1) et (m, x) selon le plan d'échantillonnage avec le point limite établi B*. B* est identique à B sauf que le point (1, 1) a été retiré et le point (2, 2) a été ajouté. (4.3)

Cas (3) : Nous supposons que (2, 2) ∉ B, mais (1, 1) ∉ B. Alors,

$$v(p(m,x)) = p(m,x)\left(p(m,x) - \frac{1}{M}\right) - \frac{M-1}{M} \frac{f(m,x)}{c(m,x)} \tag{4.4}$$

où f(m, x) = le nombre de chemins distincts passant par (2, 2) et (m, x) selon le plan d'échantillonnage avec le point limite établi B. (4.5)

Preuve : Cela n'est donné que pour le cas 1. Les preuves pour les deux autres cas sont semblables.

$$V(p(m,x)) = E(p^2(m,x)) - P^2$$

$$= \sum_{(m,x) \in B} p^2(m,x) Pr((m,x)) - P^2 \tag{4.6}$$

Nous pouvons voir voir que d(2, 2) = 1 et c(2, 2) = 1. Par conséquent, sous l'échantillonnage avec remise :

$$p^2(2,2)\Pr((2,2))-P^2 = P^2 - P^2 = 0. \tag{4.7}$$

Par conséquent,

$$V(p(m,x)) = \sum_{\substack{(m,x)\in B \\ (m,x) \neq (2,2)}} p^2(m,x)\Pr((m,x)) \ . \tag{4.8}$$

Donc,

$$v(p(m,x)) = \begin{cases} 0 \text{ pour } (m,x) = (2,2) \\ p^2(m,x) \text{ sinon pour } (m,x)\in B \end{cases} \tag{4.9}$$

est un estimateur sans biais de V(p(m, x)) selon l'échantillonnage a.r. Selon l'échantillonnage s.r., l'équation (4.7) n'est pas ramenée à 0. En fait,

$$\begin{aligned} & p^2(2,2)\Pr((2,2))-P^2 \\ & = G(2,2;M,X)-P^2 \\ & = -\frac{P(1-P)}{M-1} \ . \end{aligned} \tag{4.10}$$

Cela signifie que l'estimateur (4.9) selon l'échantillonnage s.r. a un biais de $\frac{P(1-P)}{M-1}$. Toutefois, si (2, 1) ∈ B alors,

$$q(m,x) = \begin{cases} 0 \text{ pour } (m,x) = (2,2) \\ \dfrac{d(m,x)}{c(m,x)} \text{ sinon pour } (m,x)\in B \end{cases} \tag{4.11}$$

est un estimateur sans biais de $P(1-P)\frac{M}{M-1}$ puisque

$$d(m,x) = \begin{cases} 1 \text{ pour } (m,x) = (2,1) \text{ or } (2,2) \\ 0 \text{ sinon pour } \quad (m,x) \varepsilon B \end{cases} \tag{4.12}$$

et

$$G(2,1;M,X) = P(1-P)\,\frac{M}{M-1}\,. \tag{4.13}$$

Si (2, 1) $\notin$ B, supposons que B′ est l'ensemble des points limites B avec (2, 1) ajouté. Aussi, supposons que

$$K(m, x) = \text{le nombre de chemins distincts vers } (m, x) \text{ selon le plan d'échantillonnage avec un point limite établi } B' \tag{4.14}$$

et

$$g(m,x) = \begin{cases} 0 \text{ si } (m,x) = (2,1) \\ \text{le nombre de chemins passant par (2, 1) et (m, x) selon le plan d'échantillonnage avec le point limite établi B si } (m, x) \neq (2, 1). \end{cases} \tag{4.15}$$

Il est évident que

$$k(m,x) = c(m,x) - g(m,x). \tag{4.16}$$

De plus,

$$\sum_{(m,x)\varepsilon B'} k(m,x)\; G(m,x;M,X) = 1 \quad , \tag{4.17}$$

$$k(2,1)G(2,1;M,X) + \sum_{(m,x)\varepsilon B} k(m,x)\; G(m,x;M,X) = 1 \, , \tag{4.18}$$

$$k(2,1)G(2,1;M,X) + \sum_{(m,x)\in B} c(m,x)\ G(m,x;M,X)$$

$$- \sum_{(m,x)\in B} g(m,x)\ G(m,x;M,X) = 1 \qquad (4.19)$$

et

$$\sum_{(m,x)\in B} g(m,x)\ G(m,x;M,X) = k(2,1)\ G(2,1;M,X)\ . \qquad (4.20)$$

Maintenant si $(1, 0) \in B$, alors $k(2, 1) = 1$ et

$$g(m,x) = \begin{cases} 0 \text{ pour } (m,x) = (2,2) \\ d(m,x) \text{ sinon pour } (m,x)\in B. \end{cases} \qquad (4.21)$$

Si $(1, 0) \notin B$, alors $k(2, 1) = 2$ et

$$g(m,x) = \begin{cases} 0 \text{ pour } (m,x) = (2,2) \\ 2d(m,x) \text{ sinon pour } (m,x)\in B. \end{cases} \qquad (4.22)$$

Donc, à partir des équations (4.21) et (4.22), l'équation (4.20) peut être réécrite comme suit :

$$\sum_{\substack{(m,x)\in B \\ (m,x)\neq(2,2)}} d(m,x)\ G(m,x;M,X) = G(2,1;M,X)\ . \qquad (4.23)$$

Donc $q(m, x)$ étant donné que dans l'équation (4.11) figure aussi un estimateur sans biais de $G(2, 1; M, X)$ si $(2, 1) \notin B$. Par conséquent, un estimateur sans biais de $v(p(m, x))$ selon l'échantillonnage s.r. d'une population finie est donné par l'équation (4.1).

5 Estimation de la proportion de questionnaires défectueux dans le recensement de 1976

Dans le recensement canadien de 1976, on a fait appel à des techniciens du contrôle de la qualité pour examiner et rejeter les secteurs de

dénombrement (SD) où les données étaient de piètre qualité, sur la base d'un échantillon. Un échantillon de questionnaires sur la population et le logement (formulaires 2B) a été échantillonné sans remise un à la fois. Si un formulaire 2B échantillonné ne respectait pas certaines normes de qualité, il était rejeté. Après avoir examiné un formulaire 2B, le technicien examinait le tableau 1 pour déterminer s'il fallait accepter le SD, le rejeter ou continuer l'échantillonnage des formulaires 2 B. Par exemple, après l'examen de 17 ménages, le SD était accepté, rejeté ou l'échantillonnage se poursuivait si 2, 4 ou 3 ménages avaient été rejetés respectivement. Au moins deux formulaires 2B et un maximum de 24 formulaires 2B ont été échantillonnés. Un plan d'échantillonnage séquentiel exige un plus petit échantillon en moyenne qu'un plan dont la taille de l'échantillon est établie pour distinguer les SD de bonne et de mauvaise qualité avec une exactitude raisonnable (voir Duncan [2]), p. 178).

Aux fins de planification future, il fallait une estimation pour le Canada de la proportion de formulaires 2B défectueux (ceux qui ne répondaient pas aux normes de qualité) au début de l'opération de contrôle de la qualité. Un échantillon stratifié de SD a été sélectionné avec une répartition proportionnelle. Les SD ont été stratifiés selon le bureau régional et la méthodologie relative aux SD. Les renseignements consignés dans les formules de contrôle de la qualité comprenaient le nombre de formulaires 2B échantillonnés et le nombre de formulaires rejetés. La proportion de formulaires 2B défectueux au niveau canadien a été estimée comme suit :

$$\hat{p} = \frac{\sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} M_{ij} p_{ij}}{\hat{M}} \qquad (5.1)$$

où

N_i = Le nombre de SD dans la $i^{ème}$ strate, (5.2)

n_i = Le nombre de SD dans le $i^{ème}$ échantillon de la strate, (5.3)

M_{ij} = Le nombre de formulaires 2B dans le $j^{ième}$ SD échantillonné de la $i^{ème}$ strate, (5.4)

p_{ij} = L'estimateur de la proportion de formulaires 2B défectueux dans le $j^{ième}$ SD échantillonné de la $i^{ème}$ strate. (5.5)

Tableau 1
Table de décision pour le plan d'échantillonnage séquentiel des formulaires 2B du recensement de 1976

Nombre de formulaires 2B échantillonnés	Accepter le secteur de dénombrement si le nombre de formulaires 2B suivant sont rejetés	Rejeter le secteur de dénombrement si le nombre de formulaires 2B suivant sont rejetés
2	*	2
3	*	2
4	*	2
5	*	2
6	*	2
7	0	3
8	0	3
9	0	3
10	0	3
11	0	3
12	1	4
13	1	4
14	1	4
15	1	4
16	1	4
17	2	4
18	2	5
19	2	5
20	2	5
21	2	5
22	2	5
23	3	5
24	4	5

* indique que le secteur de dénombrement ne peut pas être accepté avec cette taille d'échantillon.

et

$$\hat{M} = \sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} M_{ij}. \qquad (5.6)$$

Les estimateurs de la variance et du biais de p sont donnés par :

$$v(\hat{p}) = \frac{1}{\hat{M}^2} \sum_i \left(s(r_{ij}, r_{ij}) + \left(\frac{N_i}{n_i}\right)^2 \sum_{j=1}^{n_i} M_{ij}\, v(p_{ij}) \right) \qquad (5.7)$$

et

$$b(\hat{p}) = -\frac{1}{\hat{M}^2} \sum_i s(M_{ij}, r_{ij}) \qquad (5.8)$$

où

$$r_{ij} = M_{ij}(p_{ij} - \hat{p}) \qquad (5.9)$$

et

$$s(w_{ij},v_{ij}) = N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\frac{1}{n_i-1}\left(\sum_{j=1}^{n_i} w_{ij}v_{ij} - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} w_{ij}\right)\left(\sum_{j=1}^{n_i} v_{ij}\right)\right). \qquad (5.10)$$

Les formules standard ont été utilisées pour linéariser l'estimateur par le ratio $\hat{p}$ dans v(p) et b($\hat{p}$). Des espérances conditionnelles ont été appliquées aux dérivations en raison de l'échantillonnage à deux degrés. Les quantités d'échantillons ont été remplacées par des valeurs de population au besoin.

Pour calculer p_{ij} il faut trouver c(m, x) et d(m, x) dans l'expression (3.7). Cela a été fait aux figures 1 et 2. Le nombre dans une cellule de la figure 1, par exemple, est la valeur de c(m, x) pour ce m et ce x. Les cellules entourées de lignes continues sont des points où un SD pourrait être accepté ou rejeté dans le tableau 1. Les calculs ont commencé à la figure 1 en plaçant le chiffre 1 dans la première colonne. Le nombre dans une cellule à droite a été calculé en additionnant les nombres dans toutes les cellules à gauche ayant des flèches pointant vers cette cellule. La figure 1 montre que ce plan d'échantillonnage comporte 25 points limites. Le point (8, 0), par exemple, n'est pas un point limite puisque Pr((8, 0)) = 0. Cela indique (voir la section 3) que si les questionnaires avaient été échantillonnés à partir d'une population infinie, alors p(m, x) aurait été l'estimateur sans biais à variance minimale de P. L'estimateur sans biais de V(p_{ij}) est donné par l'équation (4.1).

Figure 1 : Calcul de c(m, x) pour le tableau 1, plan d'échantillonnage.

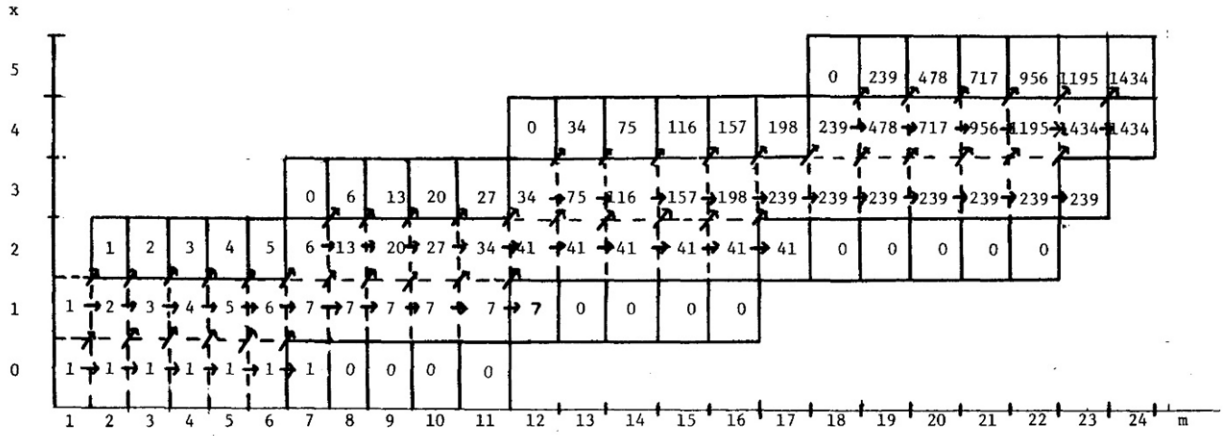
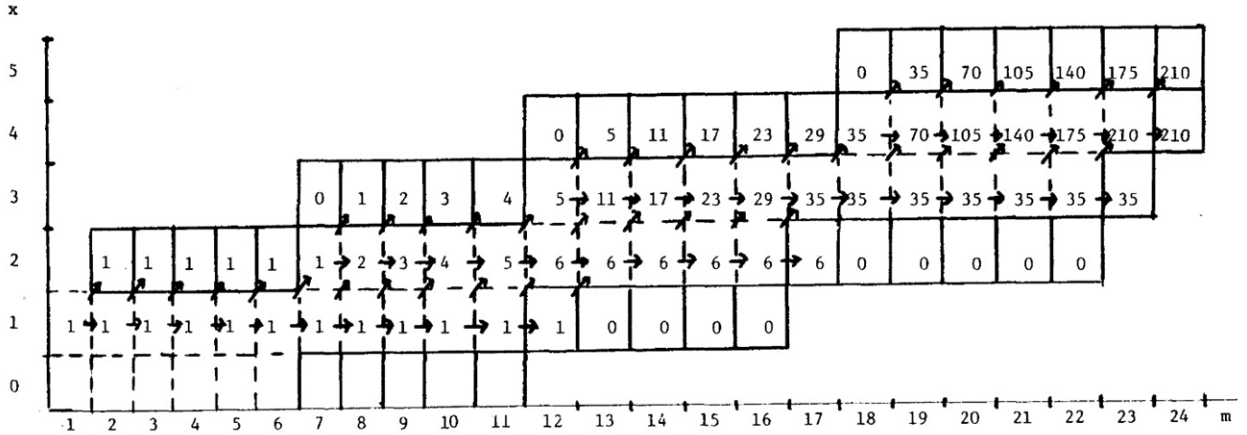


Figure 2 : Calcul de d(m, x) pour le tableau 1, plan d'échantillonnage.



Les résultats ci-dessous sont fondés sur un échantillon de 1 199 SD :

$$\hat{p} = 2{,}63 \times 10^{-2}, \tag{5.11}$$

$$b(\hat{p}) = 2{,}82 \times 10^{-8}, \tag{5.12}$$

$$\sqrt{v(\hat{p})} = 3{,}22 \times 10^{-3}, \tag{5.13}$$

$$\frac{\sqrt{v(\hat{p})}}{\hat{p}} \times 100\ \% = 12{,}2\ \%. \tag{5.14}$$

Nous pouvons voir que l'estimation de la proportion canadienne défectueuse a un coefficient de variation raisonnablement faible et un biais très faible.

Remerciements

Je tiens à remercier mon superviseur, R. Burgess, pour ses conseils, ainsi que G. Brackstone, pour sa suggestion d'étudier p(m, x).

Bibliographie

[1] DeGroot, M.H. (1959), "Unbiased Sequential Estimation for Binomial Populations", *Ann. Math. Stat.*, 30, 80-101.

[2] Duncan, A.J. (1965), "Quality Control and Industrial Statistics", Third Edition, Richard D. Irwin, Inc.

[3] Girshick, M.A., Mosteller, F. et Savage, L.J. (1946), "Unbiased Estimates for Certain Binomial Sampling Problems with Applications", *Ann. Math. Stat.*, 17, 13-23.

[4] Roussas, G.G. (1973), "A First Course in Mathematical Statistics", Addison-Wesley Publishing Company.