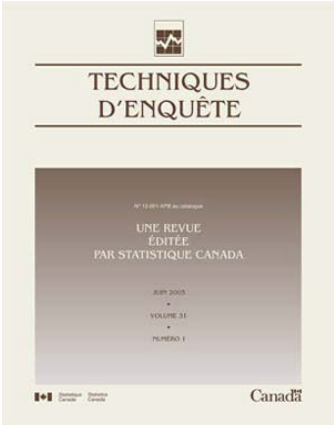
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Nouvel estimateur de la variance de l’estimateur par le ratio avec corrections de petit échantillon



par Paul Knottnerus et Sander Scholtus

Date de diffusion : le 17 décembre 2019

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. 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 sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Techniques d'enquête, décembre 2019
Vol. 45, nº 3, pp. 607-616
Statistique Canada, nº 12-001-X au catalogue

Nouvel estimateur de la variance de l'estimateur par le ratio avec corrections de petit échantillon

Paul Knottnerus et Sander Scholtus[1]

Résumé

Les formules largement utilisées pour la variance de l'estimateur par le ratio peuvent mener à une sérieuse sous-estimation quand l'échantillon est de petite taille; voir Sukhatme (1954), Koop (1968), Rao (1969) et Cochran (1977, pages 163 et 164). Nous proposons ici comme solution à ce problème classique de nouveaux estimateurs de la variance et de l'erreur quadratique moyenne de l'estimateur par le ratio qui ne sont pas entachés d'un important biais négatif. Des formules d'estimation semblables peuvent s'obtenir pour d'autres estimateurs par le ratio, comme il en est question dans Tin (1965). Nous comparons trois estimateurs de l'erreur quadratique moyenne de l'estimateur par le ratio dans une étude par simulation.

Mots-clés : Biais; moments-produits; variance d'échantillon; développement en série de Taylor.

1 Introduction

Considérons une population de N unités distinctes avec les valeurs (x_i, y_i) $(i = 1, \ldots, N)$ des variables x et y $(x_i > 0)$. Notons les moyennes correspondantes de population par $\bar{X}$ et $\bar{Y}$ dans $\bar{X} = \sum_{i=1}^{N} x_i \big/ N$ et $\bar{Y} = \sum_{i=1}^{N} y_i \big/ N$. Définissons R comme $R = \bar{Y} \big/ \bar{X}$. Posons qu'un échantillon aléatoire simple de taille n est prélevé sur cette population. Si $\bar{X}$ est connu, $\bar{Y}$ peut être estimé par l'estimateur par le ratio

$$\hat{\bar{Y}}_R = \hat{R}\bar{X}, \tag{1.1}$$

où $\hat{R} = \bar{y}_s \big/ \bar{x}_s$ avec $\bar{y}_s = \sum_{i=1}^{n} y_i \big/ n$ et $\bar{x}_s = \sum_{i=1}^{n} x_i \big/ n$; voir Cochran (1977, page 151). Pour un n grand, l'approximation bien connue de la variance de $\hat{\bar{Y}}_R$ est

$$\operatorname{var}\left(\hat{\bar{Y}}_R\right) \approx \frac{1-f}{n} S_e^2, \tag{1.2}$$

où $f = n/N$, $S_e^2 = \sum_{i=1}^{N} e_i^2 \big/ (N-1)$ et $e_i = y_i - Rx_i$ $(i = 1, \ldots, N)$; à noter que $\bar{E} = \sum_{i=1}^{N} e_i \big/ N = 0$. Quand n est petit, l'erreur d'approximation de (1.2) peut être considérable; voir Koop (1968). Qui plus est, l'erreur peut s'accroître quand, dans la pratique, S_e^2 en (1.2) est remplacé par son estimateur standard $s_{\hat{e}}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \hat{e}_i^2$, où $\hat{e}_i = y_i - \hat{R}x_i$ $(i = 1, \ldots, n)$; voir Cochran (1977, page 163). Comme le dit Koop (1968), la cause de l'écart par rapport à la variance vraie réside dans l'absence de prise en compte de termes en $1/n^2$ et $1/n^3$, et peut-être aussi de termes des ordres supérieurs.

Nous avons trois buts principaux dans cet exposé, à savoir (*i*) améliorer l'approximation (1.2) pour les petites valeurs de n par un développement en série de Taylor du deuxième ordre de $1/\bar{x}_s$; (*ii*) dériver un nouvel estimateur de S_e^2 qui est moins biaisé que $s_{\hat{e}}^2$; (*iii*) dégager un nouvel estimateur de variance pour l'estimateur par le ratio. Bien qu'une approximation distributionnelle normale puisse se révéler imprécise aux tailles d'échantillon dont il est question ici, un estimateur de variance de plus grande exactitude comme

1. Paul Knottnerus, Statistics Netherlands, C.P. 24500, 2490 HA, La Haye, Pays-Bas. Courriel : pkts@cbs.nl; Sander Scholtus, Statistics Netherlands, C.P. 24500, 2490 HA, La Haye, Pays-Bas. Courriel : sshs@cbs.nl.

celui qui est envisagé nous aide à mieux juger de la précision de l'estimateur par le ratio au regard de celle d'autres estimateurs. Dans le cas de petits échantillons de petites strates par exemple, l'estimé par le ratio combiné pour $\bar{Y}$ est à recommander sûrement plutôt que l'estimé par le ratio séparé quand les ratios (R_h, disons) sont constants de strate en strate; voir Cochran (1977, page 167).

Notre propos se résume ainsi. À l'aide de certains résultats de Nath (1968), nous dérivons à la section 2 une nouvelle formule d'approximation de la variance de $\hat{R}$ avec une erreur de l'ordre $1/n^3$. De plus, nous dégageons une nouvelle formule d'approximation pour le biais de la variance résiduelle d'échantillonnage $s_{\hat{e}}^2$ de l'ordre $1/n$. Enfin, nous proposons deux nouveaux estimateurs de l'erreur quadratique moyenne (EQM) de $\hat{\bar{Y}}_R$. À la section 3, nous faisons une étude de simulation permettant de comparer l'estimateur de variance standard aux nouveaux estimateurs proposés à la section 2. À la section 4, nous résumons nos principales conclusions.

2 Nouvel estimateur de variance

Nous posons $\hat{R} - R = \bar{e}_s / \bar{x}_s$, où $\bar{e}_s$ est la moyenne d'échantillon de e_i et, dans un développement en série de Taylor du deuxième ordre

$$\frac{1}{\bar{x}_s} = \frac{1}{\bar{X}} - \frac{1}{\bar{X}^2}(\bar{x}_s - \bar{X}) + \frac{1}{\bar{X}^3}(\bar{x}_s - \bar{X})^2 + O_p\left(\frac{1}{n^{1,5}}\right),$$

nous voyons que le développement en série de Taylor du troisième ordre de $\hat{R} - R$ est

$$\hat{R} - R = \frac{\bar{e}_s}{\bar{X}} - \frac{1}{\bar{X}^2}(\bar{x}_s - \bar{X})\bar{e}_s + \frac{1}{\bar{X}^3}(\bar{x}_s - \bar{X})^2 \bar{e}_s + O_p\left(\frac{1}{n^2}\right). \tag{2.1}$$

Par $\hat{\bar{Y}}_R = \bar{X}\hat{R}$, nous obtenons donc

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_R\right) &= \operatorname{var}(\bar{e}_s) + \frac{1}{\bar{X}^2}\operatorname{var}\{(\bar{x}_s - \bar{X})\bar{e}_s\} - \frac{2}{\bar{X}}\operatorname{cov}\{\bar{e}_s, (\bar{x}_s - \bar{X})\bar{e}_s\} \\ &\quad + \frac{2}{\bar{X}^2}\operatorname{cov}\{\bar{e}_s, (\bar{x}_s - \bar{X})^2 \bar{e}_s\} + O\left(\frac{1}{n^3}\right).\end{aligned} \tag{2.2}$$

En (2.2), nous avons omis une variance et une covariance, parce que les cinquième et sixième moments sous-jacents sont de l'ordre $1/n^3$; voir David et Sukhatme (1974). Il est possible d'évaluer toutes les (co)variances en (2.2) en utilisant les résultats suivants sur les moments-produits de quatre moyennes arbitraires d'échantillon, disons $\bar{x}_{sa}$, $\bar{x}_{sb}$, $\bar{x}_{sc}$ et $\bar{x}_{sd}$,

$$E(\bar{x}_{sa}\bar{x}_{sb}\bar{x}_{sc}) = (1 - f)(1 - 2f)S_{abc}/n^2 + O(n^{-3}) \tag{2.3}$$

$$E(\bar{x}_{sa}\bar{x}_{sb}\bar{x}_{sc}\bar{x}_{sd}) = \gamma(S_{ab}S_{cd} + S_{ac}S_{bd} + S_{ad}S_{bc}) + O(n^{-3}) \tag{2.4}$$

$$\operatorname{cov}(\bar{x}_{sa}\bar{x}_{sb}, \bar{x}_{sc}\bar{x}_{sd}) = \gamma(S_{ac}S_{bd} + S_{ad}S_{bc}) + O(n^{-3}) \tag{2.5}$$

$$E(\bar{x}_{sa}^2\bar{x}_{sb}^2) = \gamma(S_{aa}S_{bb} + 2S_{ab}^2) + O(n^{-3}), \tag{2.6}$$

où $\gamma = (1-f)^2/n^2$, $S_{ab} = \sum_{i=1}^N x_{ia}x_{ib}/(N-1)$ et $S_{abc} = \sum_{i=1}^N x_{ia}x_{ib}x_{ic}/(N-1)$. Nous supposons par commodité et sans perte de généralité que les moyennes de population sont nulles, c'est-à-dire que $\bar{X}_a = \bar{X}_b = \bar{X}_c = \bar{X}_d = 0$. Les formules (2.3) et (2.4) découlent des théorèmes 1 et 2 de Nath (1968) et les formules (2.5) et (2.6), de (2.4). Il s'ensuit de (2.2) à (2.6) que :

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_R\right) &= \frac{1-f}{n}S_e^2 + \left(\frac{1-f}{n\bar{X}}\right)^2 (S_x^2 S_e^2 + S_{xe}^2) - \frac{2}{n^2\bar{X}}(1-f)(1-2f)S_{xee} \\ &\quad + 2\left(\frac{1-f}{n\bar{X}}\right)^2 (S_x^2 S_e^2 + 2S_{xe}^2) + O(n^{-3}) \\ &= \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n\bar{X}^2}\right)S_x^2\right\} + 5\left(\frac{1-f}{n\bar{X}}\right)^2 S_{xe}^2 - 2\frac{(1-f)(1-2f)}{n^2\bar{X}}S_{xee} + O(n^{-3}), \quad (2.7)\end{aligned}$$

où

$$\begin{aligned}S_x^2 &= \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{X})^2, \quad S_{xe} = \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{X})e_i \\ S_{xee} &= S_{ge} = \frac{1}{N-1}\sum_{i=1}^N e_i^2 (x_i - \bar{X}), \quad g_i = (x_i - \bar{X})e_i.\end{aligned}$$

Des formules semblables en cumulants sont dégagées par Tin (1965) à l'aide de certains résultats de Kendall et Stuart (1958). Malheureusement, les nombreux cumulants dans les formules de Tin ne nous éclairent guère sur la structure de $\operatorname{var}(\hat{\bar{Y}}_R)$ et, par conséquent, les corrections de petit échantillon pour l'estimateur de variance nous imposent des calculs quelque peu fastidieux. En revanche, on peut voir par (2.7) que, pour un n suffisamment grand, l'approximation (1.2) mène à une sous-estimation à moins que $S_{xee}(= S_{ge})$ ne soit très positif. Ajoutons que Tin parle de trois autres estimateurs par le ratio, mais sans tenir compte des corrections de petit échantillon dans l'estimation des diverses variances.

Il s'ensuit de (2.1) et (2.3) que

$$\operatorname{biais}\left(\hat{\bar{Y}}_R\right) = -\frac{1-f}{n\bar{X}}S_{xe} + O\left(\frac{1}{n^2}\right); \quad (2.8)$$

voir aussi Cochran (1977, page 161). Par $S_{xee} = S_{ge}$, on voit ensuite, à partir de (2.7) et (2.8), que l'erreur quadratique moyenne de $\hat{\bar{Y}}_R$ est

$$\operatorname{EQM}\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n\bar{X}^2}\right)S_x^2\right\} + 6\left(\frac{1-f}{n\bar{X}}\right)^2 S_{xe}^2 - 2\frac{(1-f)(1-2f)}{n^2\bar{X}}S_{ge} + O(n^{-3}). \quad (2.9)$$

Si le coefficient de variation $C_x\ (\equiv S_x/\bar{X})$ est connu, il est bon d'écrire (2.9) sous la forme suivante :

$$\operatorname{EQM}\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n}\right)C_x^2(1 + 2\rho_{xe}^2) - 2\frac{(1-2f)}{n}C_x\rho_{ge}\,S_g/(S_x S_e)\right\}, \quad (2.10)$$

où $\rho_{xe} = S_{xe}/S_x S_e$, $\rho_{ge} = S_{ge}/S_g S_e$ et $S_g^2 = \frac{1}{N-1}\sum_{i=1}^N (g_i - \bar{G})^2$. Dans la pratique, nous pouvons estimer $\operatorname{EQM}(\hat{\bar{Y}}_R)$ en (2.10) par

$$\widehat{\mathrm{EQM}}_1\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n} s_{\hat{e}}^2 \left\{1 + 3\left(\frac{1-f}{n}\right) C_x^2 \left(1 + 2\hat{\rho}_{x\hat{e}}^2\right) - 2\frac{(1-2f)}{n} C_x \hat{\rho}_{\hat{g}\hat{e}}\, s_{\hat{g}} \big/ (s_x s_{\hat{e}})\right\}, \quad (2.11)$$

où $\hat{\rho}_{x\hat{e}} = s_{x\hat{e}} \big/ s_x s_{\hat{e}}$, $\hat{\rho}_{\hat{g}\hat{e}} = s_{\hat{g}\hat{e}} \big/ s_{\hat{g}} s_{\hat{e}}$ et

$$s_{\hat{g}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{g}_i - \bar{\hat{g}}_s\right)^2 \quad [\text{nota : } \hat{g}_i = (x_i - \bar{x}_s)\hat{e}_i]$$

$$s_{x\hat{e}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)\hat{e}_i, \; s_{\hat{g}\hat{e}} = s_{x\hat{e}\hat{e}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)\hat{e}_i^2.$$

Il reste que l'estimateur en (2.11) ne tient pas compte du biais de $s_{\hat{e}}^2$ déjà défini.

Pour examiner le biais de $s_{\hat{e}}^2 = \sum_{i=1}^{n} \hat{e}_i^2 \big/ (n-1)$, nous employons de nouveaux symboles

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)^2, \; s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e}_s)^2$$

$$s_{xe} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)e_i, \; q_i = (x_i - \bar{X})^2 \; (i = 1, \ldots, N).$$

Nous pouvons maintenant formuler $E(s_{\hat{e}}^2)$ ainsi :

$$\begin{aligned} E(s_{\hat{e}}^2) &= E\frac{1}{n-1}\sum_{i=1}^{n}\left\{y_i - \bar{y}_s - R(x_i - \bar{x}_s) - \left(\hat{R} - R\right)(x_i - \bar{x}_s)\right\}^2 \\ &= E(s_e^2) + E\left\{\left(\hat{R} - R\right)^2 s_x^2\right\} - 2E\left\{\left(\hat{R} - R\right) s_{xe}\right\} \\ &= S_e^2 + E\left\{\left(\hat{R} - R\right)^2 \bar{q}_s\right\} - 2E\left\{\left(\hat{R} - R\right)\bar{g}_s\right\} + O(n^{-2}), \end{aligned} \quad (2.12)$$

où $\bar{q}_s$ et $\bar{g}_s$ sont respectivement les moyennes d'échantillon de q_i et g_i. En (2.12), nous avons utilisé

$$s_x^2 = \left\{\bar{q}_s - \left(\bar{x}_s - \bar{X}\right)^2\right\}\left(1 + \frac{1}{n-1}\right), \; s_{xe} = \left\{\bar{g}_s - \left(\bar{x}_s - \bar{X}\right)\bar{e}_s\right\}\left(1 + \frac{1}{n-1}\right)$$

et, par conséquent, nous obtenons ce qui suit par (2.1), (2.3) et (2.4) :

$$\begin{aligned} E\left\{\left(\hat{R} - R\right)^2 (\bar{q}_s - s_x^2)\right\} &= E\left\{\bar{e}_s^2 \left(\bar{x}_s - \bar{X}\right)^2 \big/ \bar{X}^2\right\}\{1 + o(1)\} = O(n^{-2}) \\ E\left\{\left(\hat{R} - R\right)(\bar{g}_s - s_{xe})\right\} &= E\left\{\bar{e}_s \left(\bar{x}_s - \bar{X}\right)\bar{e}_s \big/ \bar{X}\right\}\{1 + o(1)\} = O(n^{-2}). \end{aligned}$$

À partir de (2.1) et (2.12), nous pouvons voir que

$$\begin{aligned} \text{biais}(s_{\hat{e}}^2) &= E\left\{\left(\hat{R} - R\right)^2 \left(\bar{q}_s - \bar{Q} + \bar{Q}\right)\right\} - 2E\left\{\left(\hat{R} - R\right)\left(\bar{g}_s - \bar{G} + \bar{G}\right)\right\} + O(n^{-2}) \\ &= \frac{1-f}{n\bar{X}^2} S_e^2 \bar{Q} - 2E\left[\left\{\frac{\bar{e}_s}{\bar{X}} - \frac{1}{\bar{X}^2}\left(\bar{x}_s - \bar{X}\right)\bar{e}_s\right\}\left(\bar{g}_s - \bar{G} + \bar{G}\right)\right] + O(n^{-2}) \\ &= \frac{1-f}{n\bar{X}^2} S_e^2 S_x^2 - 2\frac{1-f}{n}\left(\frac{S_{ge}}{\bar{X}} - \frac{S_{xe}^2}{\bar{X}^2}\right) + O(n^{-2}), \end{aligned} \quad (2.13)$$

où nous avons employé $\bar{Q} = S_x^2 (1 - N^{-1})$ et $\bar{G} = S_{xe}(1 - N^{-1})$. À noter qu'il découle de (2.13) que, pour un n suffisamment grand, la quantité $s_{\hat{e}}^2$ donne lieu à une sous-estimation de $S_{\hat{e}}^2$ sauf quand S_{ge} est très

positif. Autant que nous sachions, la formule en (2.13) ne figure nulle part ailleurs dans les études spécialisées.

En nous fondant sur (2.13), nous obtenons un nouvel estimateur de $\mathrm{EQM}(\hat{\bar{Y}}_R)$ qui prend en compte le biais de $s_{\hat{e}}^2$:

$$\widehat{\mathrm{EQM}}_2\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}\hat{S}_e^2\left\{1+3\left(\frac{1-f}{n}\right)C_x^2\left(1+2\hat{\rho}_{x\hat{e}}^2\right)-2\frac{(1-2f)}{n}C_x\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}\big/(s_x s_{\hat{e}})\right\}, \quad (2.14)$$

où $\hat{S}_e^2$ est corrigé en fonction du biais relatif de $s_{\hat{e}}^2$ qui découle de (2.13). En d'autres termes,

$$\hat{S}_e^2 = s_{\hat{e}}^2\left[1-\frac{1-f}{n}C_x^2\left(1+2\hat{\rho}_{x\hat{e}}^2\right)+2\frac{1-f}{n}C_x\frac{\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}}{s_x s_{\hat{e}}}\right].$$

Il convient de noter que nous avons employé ici $S_{xe}^2\big/\bar{X}^2S_e^2 = \rho_{xe}^2C_x^2$ et $S_{ge}\big/\bar{X}S_e^2 = \rho_{ge}S_gC_x\big/S_eS_x$. Signalons enfin que les autres estimateurs $\hat{\rho}_{x\hat{e}}^2$ et $\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}\big/(s_x s_{\hat{e}})$ en (2.11) sont aussi entachés d'un biais, mais il est moins simple de dégager ce genre de biais. Il est à espérer que leur biais sera modeste si on prend toutes les (co)variances de l'échantillon, dont s_x^2. Précisons que, dans les simulations de la section 3, nous avons constaté que le remplacement de C_x par $s_x\big/\bar{x}_s$ n'améliorait pas les résultats.

3 Une étude par simulation

3.1 Configuration et principaux résultats

Dans cette section, nous appliquons les résultats qui précèdent à 11 populations. Les populations 1 à 5 viennent de Cochran (1977, pages 152, 182, 203 et 325), les populations 6 et 7 de Sukhatme (1954, pages 183 et 184), la population 8 de Kish (1995, page 42) et les populations 9 à 11 de Koop (1968). Les tailles de ces populations varient de 10 à 49. Les coefficients de corrélation entre y et x varient de 0,32 à 0,98 et les coefficients de variation de x, de 0,14 à 1,19. Pour plus de détails, voir le tableau 3.1.

Nous avons considéré des échantillons aléatoires simples sans remise de tailles $n = 4, 6, \ldots, 14$ prélevés sur ces populations (en excluant les cas où $n \geq N$). Pour chaque population, nous avons simulé tous les $\binom{N}{n}$ échantillons possibles de taille n en autant qu'il n'y en ait pas plus de un million. Quand $\binom{N}{n} > 10^6$, nous nous sommes bornés à prélever un million d'échantillons aléatoires de taille n sur la population. Avec ces échantillons simulés, nous avons calculé (un estimé précis de) l'erreur quadratique moyenne vraie de $\hat{\bar{Y}}_R$ pour une population et une taille d'échantillon donnée devant servir d'étalon.

Pour chaque échantillon, nous avons calculé l'estimateur de variance standard pour $\hat{\bar{Y}}_R$, disons $\widehat{\mathrm{var}}(\hat{\bar{Y}}_R)$, en nous fondant sur (1.2) avec remplacement de S_e^2 par $s_{\hat{e}}^2$. Cet estimateur est aussi l'estimateur standard de l'erreur quadratique moyenne de $\hat{\bar{Y}}_R$, disons $\widehat{\mathrm{EQM}}_0(\hat{\bar{Y}}_R)$, avec une erreur de l'ordre $1/n^2$. Nous avons en outre calculé les nouveaux estimateurs $\widehat{\mathrm{EQM}}_1(\hat{\bar{Y}}_R)$ et $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ pour l'erreur quadratique moyenne de $\hat{\bar{Y}}_R$ à partir de (2.11) et (2.14). Nous prévoyons que ces estimateurs seront plus exacts que l'estimateur standard, leur erreur étant de l'ordre $1/n^3$.

Tableau 3.1
Principales caractéristiques des 11 populations de l'étude par simulation

	Source	N	$\bar{Y}$	$\bar{X}$	R	S_e^2	C_x	ρ_{xy}	ρ_{xe}	ρ_{ge}
1	Cochran, page 152	49	128	103	1,24	621	1,01	0,98	-0,34	0,02
2	Cochran, page 182	34	2,91	8,37	0,35	5,72	1,03	0,72	-0,24	0,56
3	Cochran, page 182	34	2,59	4,92	0,53	4,81	1,02	0,73	-0,14	0,38
4	Cochran, page 203	10	54,3	56,9	0,95	6,71	0,17	0,97	0,38	-0,01
5	Cochran, page 325	10	101	58,8	1,72	150	0,14	0,65	-0,29	-0,29
6	Sukhatme, pages 183 et 184	34	201	218	0,92	3 304	0,77	0,93	-0,23	0,93
7	Sukhatme, pages 183 et 184	34	218	765	0,29	8 735	0,62	0,83	0,05	0,44
8	Kish, page 42	20	12,8	21,8	0,59	17,8	1,19	0,97	0,23	0,75
9	Koop, population 1	20	4,40	6,30	0,70	0,41	0,67	0,98	-0,06	0,50
10	Koop, population 2	20	4,50	51,2	0,09	4,87	0,44	0,42	-0,50	-0,85
11	Koop, population 3	20	15,6	30,0	0,52	36,3	0,40	0,32	-0,88	0,11

Pour une comparaison d'exactitude de ces trois estimateurs, nous avons évalué leur biais relatif par rapport à la valeur étalon pour l'erreur quadratique moyenne vraie de $\hat{\bar{Y}}_R$:

$$\mathrm{BR}_k = \frac{E\left\{\widehat{\mathrm{EQM}}_k\left(\hat{\bar{Y}}_R\right)\right\} - \mathrm{EQM}\left(\hat{\bar{Y}}_R\right)}{\mathrm{EQM}\left(\hat{\bar{Y}}_R\right)} \times 100\ \%, \quad k \in \{0, 1, 2\}.$$

L'erreur quadratique moyenne $\mathrm{EQM}(\hat{\bar{Y}}_R)$ se compose du biais^2 $(\hat{\bar{Y}}_R)$ et de la variance $\mathrm{var}(\hat{\bar{Y}}_R)$. Pour toutes les populations de notre étude, nous avons constaté que, malgré les petites tailles d'échantillon, le biais de $\hat{\bar{Y}}_R$ comme estimateur de $\bar{Y}$ était plus ou moins négligeable. En fait, le biais relatif le plus grand de $\hat{\bar{Y}}_R$ se présentait toujours pour $n = 4$ et variait entre -4 % et +4 %. En d'autres termes, les erreurs quadratiques moyennes vraies et estimées dans cette étude étaient dominées par leurs composantes de variance.

Le tableau 3.2 en présente les résultats. On peut voir d'abord que l'estimateur standard $\widehat{\mathrm{EQM}}_0(\hat{\bar{Y}}_R)$ se trouve ordinairement à sous-estimer l'erreur quadratique moyenne vraie. Le biais négatif de cet estimateur peut être très grand (jusqu'à dépasser -60 % dans le cas de la population 8). Ensuite, il est frappant que, pour les trois populations (9 à 11) de l'étude de Koop, $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ estime toujours l'EQM vraie de $\hat{\bar{Y}}_R$ avec un biais relatif de moins de 5 %. Dans le cas des autres populations, le biais relatif est toujours de moins de 7 % sauf pour les populations 1, 6 et 8 avec $n = 4$ et $n = 6$. Pour $n \geq 10$, $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ est toujours plus exact que $\widehat{\mathrm{EQM}}_0(\hat{\bar{Y}}_R)$, et, en réalité, cette constatation vaut pour la plupart des cas avec $n < 10$. Pour $n \geq 8$, $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ donne presque toujours un meilleur résultat que $\widehat{\mathrm{EQM}}_1(\hat{\bar{Y}}_R)$, d'où l'utilité d'une correction en fonction du biais dans $s_{\hat{e}}^2$. On peut également voir au tableau 3.2 que, en règle générale, $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ est bien moins entaché d'un biais négatif que $\widehat{\mathrm{EQM}}_0(\hat{\bar{Y}}_R)$, alors que $\widehat{\mathrm{EQM}}_1(\hat{\bar{Y}}_R)$ accuse un biais positif.

Tableau 3.2
Biais relatif BR_k pour les trois estimateurs de $EQM(\widehat{\bar{Y}}_R)$

population	estimateur	$n = 4$	$n = 6$	$n = 8$	$n = 10$	$n = 12$	$n = 14$
1	$\widehat{EQM}_0$	-48,2 %	-35,6 %	-27,1 %	-21,6 %	-17,2 %	-14,2 %
	$\widehat{EQM}_1$	27,4 %	15,8 %	10,9 %	7,7 %	6,3 %	5,1 %
	$\widehat{EQM}_2$	-30,9 %	-11,7 %	-5,6 %	-3,5 %	-2,1 %	-1,4 %
2	$\widehat{EQM}_0$	-34,9 %	-27,7 %	-22,3 %	-18,7 %	-16,1 %	-13,6 %
	$\widehat{EQM}_1$	32,6 %	10,1 %	3,3 %	0,5 %	-0,9 %	-0,9 %
	$\widehat{EQM}_2$	2,8 %	3,4 %	1,7 %	0,4 %	-0,5 %	-0,5 %
3	$\widehat{EQM}_0$	-37,2 %	-28,4 %	-22,4 %	-17,9 %	-14,4 %	-11,6 %
	$\widehat{EQM}_1$	26,1 %	7,7 %	2,6 %	1,0 %	0,6 %	0,7 %
	$\widehat{EQM}_2$	-2,8 %	-0,6 %	-1,3 %	-1,3 %	-1,1 %	-0,6 %
4	$\widehat{EQM}_0$	-1,0 %	-0,4 %	-0,1 %			
	$\widehat{EQM}_1$	1,4 %	0,5 %	0,2 %			
	$\widehat{EQM}_2$	0,7 %	0,3 %	0,1 %			
5	$\widehat{EQM}_0$	0,4 %	0,7 %	0,8 %			
	$\widehat{EQM}_1$	2,0 %	1,0 %	0,5 %			
	$\widehat{EQM}_2$	0,8 %	0,4 %	0,2 %			
6	$\widehat{EQM}_0$	-19,2 %	-17,3 %	-15,8 %	-14,7 %	-14,1 %	-13,5 %
	$\widehat{EQM}_1$	21,1 %	0,8 %	-5,4 %	-7,4 %	-7,9 %	-7,8 %
	$\widehat{EQM}_2$	20,6 %	10,2 %	4,9 %	2,3 %	0,7 %	-0,3 %
7	$\widehat{EQM}_0$	-17,8 %	-12,0 %	-8,7 %	-6,7 %	-5,3 %	-4,3 %
	$\widehat{EQM}_1$	4,9 %	0,3 %	-0,1 %	0,0 %	0,0 %	0,0 %
	$\widehat{EQM}_2$	0,0 %	-0,6 %	-0,5 %	-0,3 %	-0,3 %	-0,2 %
8	$\widehat{EQM}_0$	-62,3 %	-45,8 %	-34,9 %	-28,0 %	-23,4 %	-20,3 %
	$\widehat{EQM}_1$	-11,1 %	-8,2 %	-6,5 %	-5,7 %	-5,3 %	-4,8 %
	$\widehat{EQM}_2$	-34,4 %	-13,3 %	-6,4 %	-4,0 %	-3,3 %	-3,2 %
9	$\widehat{EQM}_0$	-20,1 %	-13,2 %	-9,7 %	-7,6 %	-6,2 %	-5,2 %
	$\widehat{EQM}_1$	7,4 %	1,0 %	-0,5 %	-0,8 %	-0,8 %	-0,7 %
	$\widehat{EQM}_2$	0,4 %	0,1 %	-0,2 %	-0,3 %	-0,4 %	-0,4 %
10	$\widehat{EQM}_0$	-8,9 %	-2,0 %	0,9 %	2,5 %	3,5 %	4,2 %
	$\widehat{EQM}_1$	21,1 %	15,4 %	10,9 %	7,7 %	5,4 %	3,7 %
	$\widehat{EQM}_2$	0,9 %	2,1 %	2,0 %	1,7 %	1,4 %	1,1 %
11	$\widehat{EQM}_0$	-17,5 %	-10,1 %	-6,5 %	-4,4 %	-3,0 %	-2,1 %
	$\widehat{EQM}_1$	3,4 %	3,0 %	2,3 %	1,7 %	1,2 %	0,8 %
	$\widehat{EQM}_2$	-4,3 %	-1,2 %	-0,3 %	0,0 %	0,0 %	0,1 %
moyenne	$\widehat{EQM}_0$	-24,2 %	-17,4 %	-13,3 %	-13,0 %	-10,7 %	-8,9 %
	$\widehat{EQM}_1$	12,4 %	4,3 %	1,7 %	0,5 %	-0,1 %	-0,4 %
	$\widehat{EQM}_2$	-4,2 %	-1,0 %	-0,5 %	-0,6 %	-0,6 %	-0,6 %

3.2 Examen de deux résultats en particulier

Si on revient au tableau 3.1, on peut noter que, pour les deux populations 1 et 8 présentant les plus grandes erreurs négatives relatives pour $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$, il existe à la fois une corrélation forte ρ_{xy} et une valeur de C_x relativement grande si nous comparons ces populations aux autres populations de notre étude ($\rho_{xy} \geq 0{,}97$ et $C_x \geq 1{,}01$). Il est donc intéressant d'examiner de plus près l'effet de ces quantités sur l'exactitude de l'estimation de l'erreur quadratique moyenne.

Supposons d'abord que la transformation suivante est appliquée aux valeurs de x, e et y dans une population donnée :

$$x' := x, \quad e' := ae, \quad y' := Rx + e',$$

avec $a \neq 0$. Dans cette transformation, le ratio des deux variables ne change pas ($R' = \bar{Y}' / \bar{X}' = R$) contrairement à leur coefficient de corrélation ($\rho_{x'y'} \neq \rho_{xy}$ sauf si $a = 1$). Il est évident que $C_{x'} = C_x$ et $S_{e'}^2 = a^2 S_e^2$. Si nous employons maintenant les expressions (1.2), (2.8), (2.11) et (2.14), il n'est pas difficile de constater que $E\{\widehat{\mathrm{EQM}}_k(\hat{\bar{Y}}_R')\} = a^2 E\{\widehat{\mathrm{EQM}}_k(\hat{\bar{Y}}_R)\}$ pour tout $k \in \{0, 1, 2\}$. On peut aussi voir par (2.1) que l'erreur pour $\hat{R}$ est linéaire dans $\bar{e}_s$ et il s'ensuit que l'identité $\mathrm{EQM}(\hat{\bar{Y}}_R') = a^2 \mathrm{EQM}(\hat{\bar{Y}}_R)$ se vérifie exactement. Ainsi, cette transformation n'a aucun effet sur le biais relatif BR_k de quelque estimateur de l'erreur quadratique moyenne que ce soit dans cette étude. Il semblerait que le biais n'est pas touché par un changement de la corrélation ρ_{xy} quand les autres caractéristiques de la population sont toujours constantes. C'est aussi dire plus particulièrement que les grandes valeurs de ρ_{xy} dans les populations 1 à 8 n'expliquent pas à elles seules le manque d'exactitude de $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ dans ces populations.

Considérons ensuite la nouvelle transformation suivante :

$$x'' := \bar{X} + b(x - \bar{X}), \quad e'' := be, \quad y'' := \bar{Y} + b(y - \bar{Y}),$$

avec $0 < b \leq 1$. On peut voir dans ce cas que $R'' = R$, $\rho_{x''y''} = \rho_{xy}$ et $C_{x''} = bC_x \leq C_x$. Ainsi, une telle transformation peut servir à réduire le coefficient de variation de x dans une population donnée, tout en gardant fixes le ratio et la corrélation de y et x.

Nous avons appliqué cette transformation aux populations 1 et 8 pour $n = 4$ avec $b = 1{,}0; 0{,}9; \ldots, 0{,}2$. Le tableau 3.3 indique le biais relatif résultant de $\widehat{\mathrm{EQM}}_k(\hat{\bar{Y}}_R'')$ pour les populations transformées, ce qui s'obtient quand on simule l'ensemble des $\binom{49}{4} = 211\,876$ et $\binom{20}{4} = 4\,845$ échantillons possibles respectivement. On observe que les trois estimateurs de l'erreur quadratique moyenne deviennent moins biaisés avec la réduction du coefficient de variation de x. En particulier, $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R'')$ devient raisonnablement exact (si on considère que $n = 4$) une fois que le coefficient de variation de x tombe sous 0,8 pour la population 1 et sous 1 pour la population 8.

Il semble que la valeur de C_x – qui est connue dans la pratique – constitue un important facteur pour le biais (négatif) de l'estimateur $\widehat{\mathrm{EQM}}_2(\hat{\bar{Y}}_R)$ que nous proposons. Si nous supposons que l'ensemble de populations naturelles dans cette étude par simulation est suffisamment varié pour représenter la plupart des populations susceptibles de se présenter dans la pratique, nous pouvons avancer une conclusion et dire que,

même pour $n = 4$, $\widehat{\text{EQM}}_2(\hat{\bar{Y}}_R)$ est un estimateur précis de l'erreur quadratique moyenne de l'estimateur par le ratio, et ce, sans grand biais négatif quand $C_x < 0{,}8$. Pour $C_x \geq 0{,}8$, ce ne sera pas nécessairement le cas.

Tableau 3.3
Biais relatif BR_k pour les versions transformées des populations 1 et 8 avec $n = 4$

		Population 1				Population 8		
		biais relatif				biais relatif		
b	$C_{x'}$	$\widehat{\text{EQM}}_0$	$\widehat{\text{EQM}}_1$	$\widehat{\text{EQM}}_2$	$C_{x'}$	$\widehat{\text{EQM}}_0$	$\widehat{\text{EQM}}_1$	$\widehat{\text{EQM}}_2$
1,0	1,01	-48,2 %	27,4 %	-30,9 %	1,19	-62,3 %	-11,1 %	-34,4 %
0,9	0,91	-39,1 %	32,0 %	-16,5 %	1,07	-48,4 %	7,6 %	-12,9 %
0,8	0,81	-31,0 %	31,8 %	-6,2 %	0,95	-38,3 %	14,2 %	-0,7 %
0,7	0,71	-24,0 %	28,5 %	0,3 %	0,83	-30,0 %	15,0 %	5,9 %
0,6	0,61	-17,8 %	23,4 %	3,6 %	0,72	-23,1 %	12,5 %	8,4 %
0,5	0,51	-12,5 %	17,6 %	4,6 %	0,60	-17,2 %	8,6 %	8,0 %
0,4	0,40	-8,2 %	11,9 %	4,1 %	0,48	-12,3 %	4,2 %	6,0 %
0,3	0,30	-4,7 %	6,8 %	2,8 %	0,36	-8,1 %	0,4 %	3,5 %
0,2	0,20	-2,1 %	3,0 %	1,4 %	0,24	-4,7 %	-1,9 %	1,2 %

4 Conclusion

Dans cet article, nous avons dégagé une nouvelle formule d'approximation de $\text{EQM}(\hat{\bar{Y}}_R)$ de l'ordre $1/n^2$, ainsi qu'une nouvelle formule pour le biais de $s_{\hat{e}}^2$ de l'ordre $1/n$. Le nouvel estimateur $\widehat{\text{EQM}}_2(\hat{\bar{Y}}_R)$, qui tient compte du biais de $s_{\hat{e}}^2$, paraît moins entaché d'un biais que $\widehat{\text{EQM}}_0(\hat{\bar{Y}}_R) = \widehat{\text{Var}}(\hat{\bar{Y}}_R)$ et $\widehat{\text{EQM}}_1(\hat{\bar{Y}}_R)$. Pour $n \geq 8$, le biais de $\widehat{\text{EQM}}_2(\hat{\bar{Y}}_R)$ était de moins de 7 % dans tous les cas de cette simulation, ce qui est bien mieux que pour l'estimateur de variance standard; le plus souvent, ce résultat s'obtient même pour $n \geq 4$. Pour un n très petit, $\widehat{\text{EQM}}_2(\hat{\bar{Y}}_R)$ peut être entaché d'un important biais négatif si la population présente un grand coefficient de variation C_x. À en juger par notre étude de simulation, ce problème semble improbable dans la mesure où $C_x < 0{,}8$.

Rappelons enfin que, pour les populations de cette étude, le biais de l'estimateur par le ratio même était considérablement petit même pour $n = 4$. En règle générale, ce même biais pourrait ne pas être négligeable pour d'autres populations. Cochran (1977, pages 174 et 175) examine plusieurs autres estimateurs par le ratio sans biais.

Bibliographie

Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

David, I.P., et Sukhatme, B.V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association*, 69, 464-466.

Kendall, M.G., et Stuart, A. (1958). *The Advanced Theory of Statistics, Volume I*. Londres : Charles Griffin and Company.

Kish, L. (1995). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Koop, J.C. (1968). An exercise in ratio estimation. *The American Statistician*, 22, 29-30.

Nath, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

Rao, J.N.K. (1969). Ratio and regression estimators. Dans *New Developments in Survey Sampling*, (Éds., N.L. Johnson et H. Smith), New York: John Wiley & Sons, Inc., 213-234.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications*, Iowa State College Press, Ames, IA.

Tin, M. (1965). Comparison of some ratio estimators. *Journal of the American Statistical Association*, 60, 294-307.