REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Sur l'estimation de variance dans le cas d'échantillonnage systématique

Revue de statistique appliquée, tome 13, nº 4 (1965), p. 51-60
<http://www.numdam.org/item?id=RSA_1965__13_4_51_0>

SUR L'ESTIMATION DE VARIANCE DANS LE CAS D'ÉCHANTILLONNAGE SYSTÉMATIQUE

P. THIONET

Professeur à la Faculté des Sciences de Poitiers

1 - INTRODUCTION

On sait que l'estimation de la variance d'échantillonnage n'est pas possible en toute rigueur avec un échantillon systématique [1].

Observons que l'estimation de la variance d'échantillonnage constitue une information dont l'utilité est excessivement variable suivant la nature de l'enquête et de ses utilisateurs. Poussant les choses à l'extrême, cette information aura un prix infini aux yeux du spécialiste des sondages, mais n'en aura plus aucun pour bien des lecteurs des résultats de l'Enquête.

En France la majorité des utilisateurs s'estimera déjà assez satisfaite si on lui fournit sur la précision du sondage une information approximative. On se contente parfois de calculer (et publier) la précision qu'aurait eu le sondage s'il avait été obtenu par tirage au sort des unités-échantillons (à la façon des boules d'une urne), -bien qu'on ait employé en fait un autre procédé pour l'obtenir. Lorsqu'on est plus exigeant, on calcule et publie la variance d'un sondage stratifié à n/2 strates dont on aurait tiré au sort 2 unités par strate, en faisant comme si ces unités étaient 2 unités consécutives de l'échantillon systématique. On admet généralement que pareil calcul majore la variance réelle, mais modérément.

Un autre procédé pourra consister à introduire dans la formule correcte de variance quelque évaluation exogène de la corrélation intraclasse.

Tous ces procédés ont leurs mérites, le but cherché étant moins de connaître les variances d'échantillonnage que d'avoir une idée des ordres de grandeur qu'elles peuvent atteindre en mettant les choses au pire.

2 - QUELQUES TRAVAUX RECENTS

Nous nous intéresserons à deux essais récents (de Gautschi [2] et de Zinger [3]) destinés à fournir une solution théorique satisfaisante au problème qui vient d'être évoqué.

Nous devrons éviter d'ailleurs de confondre deux questions bien distinctes : Comment, avec un échantillon systématique :

- estimer la variance de population σ^2 ?
- estimer la variance du sondage systématique $\mathcal{V}\bar{x}_s$?

a) <u>Méthode des sous-échantillons systématiques</u> :

Divers auteurs (dont en dernier lieu Gautschi) ont réalisé ces estimations avec des sous-échantillons systématiques de même taille résultant de tirages aléatoires indépendants. C'est dire qu'on renonce à l'échantillon systématique banal.

Au minimum on prend 2 sous-échantillons ; la variance est alors estimée avec un seul degré de liberté, c'est-à-dire médiocrement, tandis que la qualité de l'estimation initiale est légèrement altérée ; mais il apparaît qu'on aurait plus à perdre qu'à gagner à augmenter le nombre de sous-échantillons.

Le seul critère employé dans ces études, pour apprécier la qualité de l'estimation, semble être la variance d'échantillonnage pour une taille donnée de l'échantillon. Cette façon de faire ne nous paraît pas satisfaisante.

En effet, le sondage systématique est employé par tous les praticiens parce qu'il est très commode, ce qu'on peut traduire ainsi :

parce qu'*il* est économique ;

nous devons donc admettre que k sous-échantillons, chacun de taille n/k, ont un coût supérieur à un échantillon de taille n, l'excédent de coût étant une fonction (croissante) de k dont la forme ne nous est pas connue.

Il conviendrait finalement de se demander si l'information de plus en plus précise obtenue sur les variances σ^2 et $\mathcal{V}\bar{x}_s$ compense ou non cet accroissement du coût d'enquête associé à une légère perte de précision sur $\bar{x}_s$ (un gonflement de $\mathcal{V}\bar{x}_s$). Un tel problème ne peut guère se traiter que par le biais d'une courbe d'indifférence traduisant les utilités que nous attachons respectivement à mieux connaître soit $\bar{X}$ soit $\mathcal{V}\bar{x}_s$.

b) <u>Méthode de l'échantillon additionnel non systématique</u> :

M. A. Zinger (de l'Université de Montréal) a imaginé un procédé nouveau d'estimation de la variance, en ajoutant à l'échantillon systématique un tout petit échantillon non systématique (au minimum une unité de sondage supplémentaire). Le coût additionnel est infime mais on peut l'évaluer. On aura donc à mettre en balance (avant de prendre la décision de suivre M. Zinger) :

- le coût de cette donnée supplémentaire (et même le coût des calculs d'erreurs s'ils sont notables),

- contre l'utilité qu'on attribue à la connaissance de la variance.

L'étude de M. Zinger a été présentée au Séminaire de l'I. S. U. P. en 1963 et son résumé a été publié dans la Revue de Statistique Appliquée [3]. Son idée pour sortir de l'impasse actuelle paraît très intéressante.

Malheureusement l'estimation des variances est apparue comme étant d'assez médiocre qualité, vu qu'il lui arrive d'être négative. Comme il s'agit d'une estimation sans distorsion (non biaisée), il est clair que les écarts par excès doivent être aussi grands en moyenne que les écarts par défaut, et quand elle n'est pas négative l'estimation est peut-être beaucoup trop grande.

3 - DESCRIPTION DE NOUVELLES

Nous allons maintenant décrire deux méthodes voisines, qui ressemblent l'une à la méthode de M. Gautschi, l'autre à la méthode de M. Zinger. Nous en ferons ensuite la théorie.

a') Méthode de l'échantillon systématique complété d'un échantillon en grappes.

Supposons la population disposée suivant les cases d'un tableau à double entrée, de telle sorte qu'une ligne du tableau constitue un échantillon systématique, alors qu'une colonne du tableau en constitue une grappe. Nous admettrons que l'une des lignes et l'une des colonnes sont tirées au sort par deux tirages indépendants. Ces deux échantillons sont soumis à l'enquête.

Pour que cette technique présente un intérêt pratique, il convient de supposer le tableau beaucoup plus large que haut, donc l'échantillon en grappe beaucoup plus petit que l'échantillon systématique. En outre, un tel échantillon en grappe de h éléments dont $(h - 1)$ ne figurent pas dans l'échantillon systématique, est en pratique beaucoup moins coûteux que $(h - 1)$ éléments tirés au sort par des tirages indépendants : c'est du moins ce qu'on supposera.

On est donc en possession de deux estimations de la moyenne $\overline{X}$ du tableau, à savoir $\bar{x}_g$ (moyenne de la colonne) et $\bar{x}_s$ (moyenne de la ligne), estimations l'une et l'autre sans biais et indépendantes entre elles ; toutefois $\bar{x}_g$ est (selon toute vraisemblance) beaucoup moins précise que $\bar{x}_s$. On peut à la fois les combiner linéairement pour constituer l'estimateur sans biais "le meilleur" de $\overline{X}$, et en déduire une estimation de la variance de la forme $K(\bar{x}_s - \bar{x}_g)^2$, estimation médiocre mais sûrement positive.

En effet, on a :

$$\mathcal{E}(\bar{x}_s - \bar{x}_g)^2 = \mathcal{V}\bar{x}_s + \mathcal{V}\bar{x}_g$$

Soit V_s et V_g les estimateurs sans biais que nous désirons pour $\mathcal{V}\bar{x}_s$ et $\mathcal{V}\bar{x}_g$, on devra donc avoir :

$$(\bar{x}_s - \bar{x}_g)^2 = V_s + V_g$$

Dès lors, si l'on adopte pour $\overline{X}$ l'estimateur $(\bar{x}_s + \bar{x}_g)/2 = \bar{\bar{x}}$, on a :

$$\mathcal{V}\bar{\bar{x}} = \frac{1}{4}(\mathcal{V}\bar{x}_s + \mathcal{V}\bar{x}_g) \quad ,$$

donc : estimateur sans biais de $\mathcal{V}\bar{\bar{x}} = (\bar{x}_s - \bar{x}_g)^2/4$ ·

Si l'on adopte l'estimateur linéaire plus général $\alpha\bar{x}_s + (1-\alpha)\bar{x}_g = \bar{\bar{x}}$, on voit aisément (et c'est bien connu) que le meilleur couple α, $1-\alpha$ (c'est-à-dire celui qui rend $\mathcal{V}\bar{\bar{x}}$ minimum) est proportionnel aux inverses des variances.

D'où l'estimateur :

$$\hat{x} = \left[\frac{\bar{x}_s}{\mathcal{V}\bar{x}_s} + \frac{\bar{x}_g}{\mathcal{V}\bar{x}_g}\right] \cdot \left[\frac{1}{\mathcal{V}\bar{x}_s} + \frac{1}{\mathcal{V}\bar{x}_g}\right]^{-1}$$

et sa variance :

$$\mathcal{V}\hat{x} = \frac{\mathcal{V}\bar{x}_s \cdot \mathcal{V}\bar{x}_g}{\mathcal{V}\bar{x}_s + \mathcal{V}\bar{x}_g}$$

Posons :

$$\mathcal{V}\bar{x}_s = \rho\,\mathcal{E}(\bar{x}_s - \bar{x}_g)^2 \quad \text{et} \quad V_s = \rho\,(\bar{x}_s - \bar{x}_g)^2,$$

il vient

$$\text{estimation de : } \mathcal{V}\hat{x} = (\bar{x}_s - \bar{x}_g)^2\,\rho(1-\rho)$$

donc

$$\text{estimation de : } \mathcal{V}\hat{x} \leqslant (\bar{x}_s - \bar{x}_g)^2/4$$

Bien entendu cette estimation est médiocre car elle a le même défaut que celle de Gautschi ; en revanche on notera que pour $\rho = 0,33 = (1-\rho)/2$, on a encore :

$$\rho(1-\rho) = 2/9 \quad \text{au lieu de} \quad 1/4 = 2/8$$

Pour $\rho = 0,2 = (1-\rho)/4$, on a encore $\rho(1-\rho) = 0,16$ au lieu de 0,25.

On n'a donc guère de raisons de descendre beaucoup en dessous de $(\bar{x}_s - \bar{x}_g)^2/4$; mais on risque d'énormes fluctuations (en revanche) sur la différence $|\bar{x}_s - \bar{x}_g|$ proprement dite.

b') Méthode de l'échantillon systématique complété par un petit sous-échantillon systématique.

Considérons toujours la population disposée suivant un tableau à double entrée ; mais plaçons-nous dans le cas où ce tableau est beaucoup plus haut que large. L'échantillon en grappe du cas (a) coûterait plus cher que l'échantillon systématique et il faut y renoncer (puisqu'il fournit par hypothèse une estimation médiocre de $\overline{X}$). On pourrait se contenter d'en extraire un petit échantillon au hasard : mais on perd ainsi l'avantage du faible coût du sondage en grappe. Nous supposerons plutôt que l'échantillon systématique formant une ligne du tableau est complété par un échantillon (aussi réduit que possible) d'une seconde ligne du même tableau. Le coût additionnel d'enquête et le coût du calcal de variance doivent être mis en balance avec l'information (très médiocre et éventuellement négative) qu'on peut ainsi obtenir sur les variances.

4 - THEORIE DE L'ANALYSE DE VARIANCE (RAPPEL)

Soit σ_e^2 la variance des moyennes de ligne du tableau à double entrée (de h lignes et n colonnes) ; soit σ_i^2 la moyenne des variances des éléments de chaque ligne du tableau. La variance σ^2 des éléments du tableau est liée à σ_e^2 et σ_i^2 par :

$$\sigma^2 = \sigma_e^2 + \sigma_i^2$$

La variance d'échantillonnage $\mathcal{V}\bar{x}_s$ n'est autre que σ_e^2 ; tandis que la variance des éléments de l'échantillon (constitué par la ligne a) est σ_a^2, avec :

$$\sigma_i^2 = \frac{1}{h} \sum_a \sigma_a^2 \quad ;$$

ainsi σ_a^2 est estimateur sans biais de σ_i^2.

5 - ETUDE D'UN TABLEAU (2 x 2) (Méthode a')

a	b
a'	b'

Pour commencer, supposons $n = h = 2$. On tire au sort 1 ligne et 1 colonne. On se trouve donc dans l'un des 4 cas dont les schémas sont figurés ici :

(1)

a	b
a'	

(2)

a	b
	b'

(3)

	b
a'	b'

(4)

a	
a'	b'

On veut estimer σ^2 (variance des 4 éléments) et σ_e^2 variance entre lignes. Rappelons l'expression de σ^2, σ_i^2 et σ_e^2.

$$16\,\sigma^2 = (a - b)^2 + (a' - b')^2 + (a - a')^2 + (b - b')^2 + (a - b')^2 + (a' - b)^2$$

$$8\,\sigma_i^2 = (a - b)^2 + (a' - b')^2 \quad ; \quad \sigma_e^2 = \sigma^2 - \sigma_i^2 \Longrightarrow$$

$$16\,\sigma_e^2 = (a - a')^2 + (b - b')^2 + (a - b')^2 + (a' - b)^2 - (a - b)^2 - (a' - b')^2$$

Estimation de σ^2 : Les données fournissant toujours 3 des 6 carrés. Pour estimer $16\sigma^2$ on multipliera par 2 la somme des 3 carrés connus.

Estimation de σ_i^2 : on connaît toujours l'un des 2 carrés qui composent $8\sigma_i^2$. Il suffira de le multiplier par 2 pour estimer $8\sigma_i^2$.

Estimation de σ_e^2 : par différence : est σ_e^2 = est σ^2 - est σ_i^2. On obtient les estimations de σ_e^2 qui suivent :

Schéma (1)	est $8\sigma_e^2 = (a - a')^2 + (a' - b)^2 - (a - b)$
	$= 2\,(a' - a)\,(a' - b)$
(2)	...... $= 2\,(a - b')\,(b - b')$
(3)	...... $= 2\,(b - a')\,(b - b')$
(4)	...... $= 2\,(a - a')\,(a - b')$

Ainsi l'estimation de σ_e^2 par différence est négative dans les cas suivants :

Schéma (1)	$b < a' < a$	ou	$a < a' < b$
Schéma (2)	$a < b' < b$	ou	$b < b' < a$
Schéma (3)	$a' < b < b'$	ou	$b' < b < a'$
Schéma (4)	$a' < a < b'$	ou	$b' < a < a'$

Pourtant l'espérance mathématique de cet estimateur est bien positive :

$$\frac{1}{4}\ 2\,[(a'-a)(a'-b) + (a-b')(b-b') + (b-a')(b-b') + (a-a')(a-b')]$$

$$\equiv \frac{1}{2}(a - a' + b - b')^2 \equiv \frac{1}{2}(a + b - a' - b')^2$$

Changement d'estimateur.

Au lieu d'estimer la moyenne $\overline{X}$ du tableau par la moyenne de la ligne échantillon (ou de la colonne échantillon), procédons comme suit : soit $\bar{x}$ l'estimateur de $\overline{X}$, estimons $\bar{x} - \overline{X}$ dans les 4 cas possibles :

Schéma :

$$\begin{array}{cccc}
(1) & (2) & (3) & (4) \\
\boxed{\begin{array}{cc} a & b \\ a' & (a) \end{array}} & \boxed{\begin{array}{cc} a & b \\ (b) & b' \end{array}} & \boxed{\begin{array}{cc} (b') & b \\ a' & b' \end{array}} & \boxed{\begin{array}{cc} a & (a') \\ a' & b' \end{array}} \\
\downarrow & \downarrow & \downarrow & \downarrow \\
4(\bar{x} - \overline{X}) = (a - b')\ ; & (b - a')\ ; & (b' - a)\ ; & (a' - b)
\end{array}$$

d'où : $\mathcal{V}\bar{x} = \mathcal{E}(\bar{x} - \overline{X})^2 = [(a - b')^2 + (b - a')^2]/32$

Dans chacun des 4 cas, on dispose soit de $(a - b')^2$, soit de $(b - a')^2$ et par conséquent on sait écrire une estimation toujours positive de $\mathcal{V}\bar{x}$.

Autre façon de retrouver $\bar{x}$.

Introduisons la variance à l'intérieur des colonnes, symétrique de la variance à l'intérieur des lignes :

$$8\,\sigma'^2 = (a - a')^2 + (b - b')^2$$

avec $\quad \sigma^2 = \sigma_i'^2 + \sigma_e'^2$

Posons $\quad 16\,\sigma^2 = 8\,\sigma_i^2 + 8\,\sigma_i'^2 + 8\,\sigma_o^2$

avec $\quad 8\,\sigma_o^2 = (a - b')^2 + (a' - b)^2$

Autrement dit :

$$2\,\sigma^2 = \sigma_i^2 + \sigma_i'^2 + \sigma_o^2$$

avec $\quad 2\,\sigma_e^2 = 2(\sigma^2 - \sigma_i^2) = \sigma_i'^2 + \sigma_o^2 - \sigma_i^2$

et par symétrie d'où

$$2\sigma_e'^2 = \sigma_i^2 + \sigma_o^2 - \sigma_i'^2$$

$$\sigma_e^2 + \sigma_e'^2 = \sigma_o^2$$

Ainsi la moyenne de l'estimation de $\overline{X}$ par les lignes et de celle par les colonnes

$$\bar{x} = \frac{1}{2}(\bar{x}_s + \bar{x}_g)$$

a pour variance

$$\mathcal{V}\bar{x} = \frac{1}{4}(\sigma_e^2 + \sigma_e'^2) = \frac{1}{4}\sigma_o^2$$

$$= \frac{1}{32}[(a - b')^2 + (b - a')^2]$$

expression trouvée plus haut.

Conclusion :

- Avec 1 ligne échantillon on ne sait pas estimer la variance σ_e^2 de la moyenne de ligne.

- Si on y ajoute 1 élément hors ligne, on sait estimer σ_e^2, mais on aura souvent une estimation négative.

- Si on procède à l'estimation de la moyenne générale, à l'aide de la ligne échantillon et de l'élément supplémentaire, on sait estimer la variance de cette estimation par une statistique toujours positive.

6 - THEORIE DE LA METHODE (a') AVEC UN TABLEAU $h \times n$ COMME POPULATION

L'estimateur :

$$\bar{x} = \alpha\,\bar{x}_s + \beta\,\bar{x}_g, \quad \alpha + \beta = 1, \quad \text{de } \overline{X}$$

a pour variance

$$\mathcal{V}\bar{x} = \alpha^2\,\mathcal{V}\,\bar{x}_s + \beta^2\,\mathcal{V}\,\bar{x}_g$$

et nous proposons d'estimer séparément $\mathcal{V}\bar{x}_s$ et $\mathcal{V}\bar{x}_g$ sur échantillon. Nous ferons usage de la formule

$$\sigma^2 \equiv \sum_{a=1}^{m} \frac{(x_a - \bar{x})^2}{m} = \sum_{ab}{}' \frac{(x_a - x_b)^2}{m^2}$$

$\sum'$ désignant la double sommation $\sum\sum$ étendue à tous les couples (a, b) $a < b$, c'est-à-dire que (ab) et (ba) ne figurent pas simultanément dans l'énumération. En conséquence, revenant au tableau rectangulaire, on a

$n^2 h^2 \sigma^2$	= Somme de	$n\,h\,(n\,h - 1)/2$	différences carrées
$n^2 h\, \sigma_i^2$	= Somme de	$n\,h\,(n - 1)/2$	différences carrées entre termes de la même ligne
$n\, h^2\, \sigma_i'^2$	= Somme de	$n\,h\,(h - 1)/2$	différences carrées entre termes de la même colonne

Enfin, les différences carrées entre termes ne figurant ni dans la même ligne ni dans la même colonne, sont au nombre de :

$$\frac{1}{2}\, n\, h\, (n\, h - 1) - \frac{1}{2}\, n\, h\, (n - 1) - \frac{1}{2}\, n\, h\, (h - 1) = \frac{1}{2}\, n\, h\, (n - 1)\, (h - 1)$$

Soit γ^2 leur moyenne. Divisant par $n^2 h^2$, il vient ainsi :

$$\sigma^2 = \frac{1}{h}\, \sigma_i^2 + \frac{1}{n}\, \sigma_i'^2 + \frac{(n - 1)\, (h - 1)}{n\, h}\, \frac{\gamma^2}{2}$$

Cependant l'analyse de variance classique donne

$$\mathcal{V}\, \bar{x}_s = \sigma_e^2 = \sigma^2 - \sigma_i^2, \quad \mathcal{V}\bar{x}_g = \sigma_e'^2 = \sigma^2 - \sigma_i'^2$$

Estimons sur l'échantillon σ_i^2 par la variance σ_s^2 de l'échantillon systématique ; $\sigma_i'^2$ par σ_g^2 variance de l'échantillon en grappe ; γ^2 par c^2, moyenne des $(n - 1)(h - 1)$ écarts carrés entre termes distincts des deux échantillons.

Posons enfin $(n - 1)(h - 1)/n\, h = \lambda$; il vient :

estimation sans biais de $\sigma^2 = \sigma_s^2/h + \sigma_g^2/n + \lambda\, c^2/2$

" " " " $\mathcal{V}\bar{x}_s = \sigma_s^2/h + \sigma_g^2/n + \lambda\, c^2/2 - \sigma_s^2$

" " " " $\mathcal{V}\bar{x}_g = \sigma_s^2/h + \sigma_g^2/n + \lambda\, c^2/2 - \sigma_g^2$

En général h et n seront assez grands pour qu'on puisse se contenter des estimations suivantes (dont le biais est de l'ordre de h^{-1} et n^{-1}) :

estimation de $\sigma^2 = c^2/2$

$\mathcal{V}\bar{x}_s = c^2/2 - \sigma_s^2$

$\mathcal{V}\bar{x}_g = c^2/2 - \sigma_g^2$

On voit que les estimations de $\mathcal{V}\bar{x}_s$ et $\mathcal{V}\bar{x}_s$ sont exposées à devenir négatives.

Conditions dans lesquelles les estimations sont négatives.

Supposons les σ_s^2 et σ_g^2 constants, alors que c^2 varie. Quand aura-t-on c^2 petit ? Imaginons pour cela que les termes de la colonne échantillon soient fixes mais que ceux de la ligne échantillon subissent tous la même variation t (bien entendu nous négligeons le terme situé au croisement des ligne et colonne échantillons). Le terme c^2 était de la forme :

$$c^2 = \sum_j \sum_k z_{jk}^2/(n - 1)(h - 1), \quad j = 1\ 2 \ldots h\ ,\ k = 1\ 2 \ldots n$$

et $(z_{jk} - t)$ est substitué à z_{jk}. D'autre part on a :

$$h\, n\, \bar{x}_g - h\, n\, \bar{x}_s = \sum_j \sum_k z_{jk}$$

(z_{jk} désigne la différence entre le terme j de la colonne et le terme k de la ligne échantillon).

Ainsi (en confondant n et n-1, h et h-1) $c^2 - 2\sigma_l^2$ est remplacé par :

$$t^2 - 2t\,(\bar{x}_g - \bar{x}_s) + c^2 - 2\,\sigma_l^2 \qquad (l = g,\ s)$$

Ce trinôme a des racines si

$$(\bar{x}_g - \bar{x}_s)^2 > c^2 - 2\,\sigma_l^2$$

C'est donc bien (comme on pouvait le prévoir) un écart excessif entre les moyennes des 2 échantillons (systématique et en grappe), écart dû à un hasard malheureux, qui peut provoquer des estimations négatives pour $\mathcal{V}\bar{x}_g$ ou $\mathcal{V}\bar{x}_s$.

Remarque : On peut songer à ne tirer qu'un échantillon de la grappe colonne au lieu de la prendre en totalité. Tout le calcul subsiste, à condition de modifier l'estimation c^2 de γ^2 et d'estimer σ_g^2 par s_g^2. Mais on est ainsi conduit (pour la grappe) à un sondage à 2 degrés. On sait que celui-ci augmente la variance d'échantillonnage ; et il faudrait ajouter un terme correctif à $\mathcal{V}\bar{x}_g$. En outre, on accroît vraisemblablement les risques d'aboutir à des estimations de variance négatives.

7 - THEORIE DE LA METHODE (b') AVEC UN TABLEAU h × n

Supposons que, dans la méthode de Gautschi, on ait tiré 2 lignes j j' du tableau (tirages indépendants) mais qu'on n'ait conservé qu'un sous échantillon de la ligne j'. Si l'on combine encore $\bar{x}_s$ et $\bar{x}'_s$ linéairement pour estimer $\bar{X}$, il devient nécessaire d'ajouter à la variance $\mathcal{V}\bar{x}$ un terme correspondant à un sondage à 2 degrés. Nous écartons ce procédé et n'envisageons que l'estimateur $\bar{x}_s$ et sa variance $\mathcal{V}\bar{x}_s$, de sorte que le sous-échantillon extrait de la ligne j' n'intervient que pour estimer la variance et est supposé de très faible taille. Nous nous trouvons ainsi dans un cas voisin de celui étudié par Zinger. Quand on supposera ce sous-échantillon réduit à un seul élément, notre problème sera également le problème de ZINGER le plus réduit (mais non le plus simple).

Analyse de Variance :

$$\sigma^2 = \sigma_e^2 + \sigma_i^2$$

avec $\mathcal{V}\bar{x}_s = \sigma_e^2$; estimation sans biais de $\sigma_i^2 = \sigma_s^2 = \sigma^2\,(j)$.

Estimation directe de σ_e^2 : Ce terme désigne la variance des moyennes de ligne, autrement dit on a :

$$h^2\ \sigma_e^2 = \sum_{j\,j'} (\bar{x}_j - \bar{x}_{j'})$$

Cette somme aurait $h(h-1)/2$ termes si les lignes-échantillons étaient connues complètement ; on aurait donc :

estimation sans biais de $\sigma_e^2 = \dfrac{h(h-1)/2}{h^2}\,(\bar{x}_j - \bar{x}'_j)^2$

Mais $\bar{x}_j$ est seul connu ; et $\bar{x}_{j'}$ est estimé sans biais (sur échantillon) à ξ :

$$\mathcal{E}(\bar{x}_j - \xi)^2 = \mathcal{E}(\bar{x}_j - \bar{x}_{j'} + \bar{x}_{j'} - \xi)^2 = (\bar{x}_j - \bar{x}_{j'})^2 + \mathcal{E}(\bar{x}_{j'} - \xi)^2$$

Ce dernier terme est la variance (intérieure) de la ligne j' ; et on a

$$\text{estim. sans biais de } \sigma_e^2 = \frac{h-1}{2h}(\bar{x}_j - \xi)^2 - \sigma^2(j')$$

Nous pouvons estimer $\sigma_{j'}^2$ sur l'échantillon de la ligne j' si celui-ci comprend au moins 2 éléments. S'il est réduit à 1 élément (et même autrement, s'il est très petit) il peut paraître sensé d'estimer $\sigma_{j'}^2$ par $\sigma_{(j)}^2$ c'est-à-dire σ_s^2.

Condition pour avoir une estimation négative de σ_e^2 :

On voit que le risque est grand d'avoir $(\xi - \bar{x}_j)^2 < \frac{2h}{h-1}\,\hat{\sigma}^2(j')$

soit que ξ (sur la ligne j') s'écarte peu de la moyenne de la ligne j, soit que $\hat{\sigma}^2(j')$ soit estimée avec malchance. Il suffit de songer à la probabilité qu'on ait :

$$|\eta - \bar{x}_j| < \sqrt{2}\,\sigma(j), \quad \eta \text{ désignant un élément de la ligne j.}$$

Nota : en revanche l'estimation de $\sigma^2 = \sigma_e^2 + \sigma_1^2$, sera :

$$\frac{h-1}{2h}(\bar{x}_1 - \xi)^2 + \sigma^2(j) - \hat{\sigma}^2(j'), \text{ ou (même) } \frac{h-1}{2h}(\bar{x}_1 - \xi)^2 \# (\bar{x}_1 - \xi)^2/2$$

Résumé :

On examine comment estimer les variances d'échantillonnage quand l'échantillon systématique est complété soit d'un échantillon en grappe dual, soit d'un sous échantillon du même sondage systématique. On recherche notamment dans quelles conditions peuvent surgir des estimations négatives de variance.

REFERENCES

[1] MADOW (W. G) et MADOW (L.H) - On the theory of systematic samply Annals of Math. Stat. 15 (1944) p. 1-24.

[2] GAUTSCHI (Werner) - Some remarks on systematic sampling Annals of Math. Stat. 28 (1957) p. 385-394.

[3] ZINGER (A) - Estimations de variance avec échantillon systématique. Revue de Statistique Appliquée XI n° 2 (1963) p. 89-98.