REVUE DE STATISTIQUE APPLIQUÉE

A. ZINGER

Estimations de variances avec échantillonnage systématique

Revue de statistique appliquée, tome 11, n° 2 (1963), p. 89-97

<http://www.numdam.org/item?id=RSA_1963__11_2_89_0>

ESTIMATIONS DE VARIANCES AVEC ÉCHANTILLONNAGE SYSTÉMATIQUE

A. ZINGER [1]

Université de Montréal, Canada

1 - INTRODUCTION -

L'échantillonnage systématique possède un avantage pratique incontesté sur les autres méthodes d'échantillonnage. C'est celui de la simplicité du choix de l'échantillon. Par exemple, dans le cas de l'estimation du volume de bois dans une forêt tropicale, une des difficultés réside dans la localisation de l'échantillon. L'emploi d'échantillon systématique rend cette localisation relativement aisée. Un autre cas, plus fréquent, est celui du choix d'un échantillon de dossiers. Si la dimension de l'échantillon est considérable, un échantillonnage aléatoire simple peut facilement devenir fastidieux. Il est certainement plus facile de prélever chaque $n^{ième}$ dossier, disons chaque $20^{ième}$, que de choisir les dossiers en utilisant une liste de 500 nombres aléatoires compris entre 1 et 10 000.

Le seul inconvénient sérieux de l'échantillonnage systématique est qu'une estimation sans distorsion ou non biaisée de la variance semble impossible sans faire d'hypothèses sur la structure interne de la population étudiée. Ceci, si l'on se restreint à un seul échantillon systématique et si l'on n'utilise qu'un seul stage d'échantillonnage.

Nous verrons comment l'on peut rendre une telle estimation possible à un prix additionnel minime.

La méthode proposée consiste à choisir au hasard une ou plusieurs observations additionnelles, sans remise, parmi les unités non choisies. Cet échantillon combiné permet de résoudre deux problèmes.

1/ Comment obtenir une estimation sans distorsion de la variance totale entre les unités.

2/ Comment obtenir une estimation sans distorsion de la variance d'une estimation (d'un total) obtenue par échantillonnage systématique.

Le premier problème peut présenter un intérêt lorsque l'on s'attend à ce que l'échantillonnage systématique soit équivalent à un échantillonnage aléatoire simple. C'est le cas d'une population en "ordre aléatoire", Cochran (1953). Le cas des dossiers peut dans certains cas s'y rattacher.

Le second problème est la situation normale ou aucune hypothèse n'est faite sur l'ordre, la périodicité ou la tendance.

---

(1) Rédigé pendant un séjour de l'auteur à l'Institut de Statistique de l'Université de Paris, grâce à une bourse scientifique de l'OTAN décernée par le Conseil National des Recherches du Canada.

Nous verrons par des exemples artificiels que certaines estimations de variances peuvent être négatives. Un exemple d'échantillonnage par bande (strip sampling) en forêt est présenté et les résultats obtenus sont discutés.

2 - ESTIMATION DE VARIANCES -

Soit une population de $n = kr$ unités y_j, $j = 1, \ldots, n$. Les r échantillons systématiques consistent des unités :

$$y_j, y_{j+r}, \ldots, y_{j+(k-1)r}, \qquad j = 1, \ldots, r.$$

Soit un échantillon aléatoire simple de dimension b qui consiste des unités : $y_{n_1}, y_{n_2}, \ldots, y_{n_b}$, avec la condition que $n_1, n_2, \ldots, n_b$ sont tous différents entre eux et différents de $j, j+r, \ldots, j+(k-1)r$.

Posons :

$$\begin{aligned} Y &= y_j + y_{j+r} + \ldots + y_{j+(k-1)r} \\ X &= y_{n_1} + y_{n_2} + \ldots + y_{n_b} \end{aligned} \tag{1}$$

et définissons la moyenne pondérée :

$$\bar{y} = \frac{a\,Y + X}{a\,k + b}. \tag{2}$$

Définissons la somme de carrés S par :

$$S = (y_j - \bar{y})^2 + \ldots + (y_{j+(k-1)r} - \bar{y})^2 + (y_{n_1} - \bar{y})^2 + \ldots + (y_{n_b} - \bar{y})^2. \tag{3}$$

Pour calculer E(S), cherchons les contributions de y_j^2, $y_j y_{j+mr}$ et $y_j y_{n_i}$ pour j, m et n_i fixes, $j = 1, \ldots, r$; $m = 1, \ldots, k-1$ et $n_i \neq j, j+r, \ldots, j+(k-1)r$.

Le nombre total d'échantillons différents est $r\binom{n-k}{b}$. Une classification de ces échantillons suivant leurs types nous permet de dresser le tableau suivant :

Nombre d'échantillons	Type : Partie systématique	Type : Partie aléatoire
$\binom{n-k-1}{b-1}$	$j, j+mr, \ldots$	$n_i, \ldots$
$\binom{n-k-1}{b}$	$j, j+mr, \ldots$	non $n_i, \ldots$
$\binom{n-k-2}{b-2}$	$n_i, \ldots$	$j, j+mr, \ldots$
$\binom{n-k-2}{b-1}$	$n_i, \ldots$	j, non $j+mr, \ldots$
$(r-2)\binom{n-k-3}{b-1}$	non $n_i, \ldots$	j, non $j+mr$, non $n_i, \ldots$
$(r-2)\binom{n-k-3}{b-2}$	non $n_i, \ldots$	$j, j+mr$, non $n_i, \ldots$
$(r-2)\binom{n-k-3}{b-2}$	non $n_i, \ldots$	j, non $j+mr$, $n_i, \ldots$
$(r-2)\binom{n-k-3}{b-3}$	non $n_i, \ldots$	$j, j+mr$, $n_i, \ldots$
$(r-1)\binom{n-k-1}{b}$	non j	non j

Utilisant ces résultats, nous trouvons que les contributions de $y_j^2, y_j y_{j+mr}$ et $y_j y_{n_i}$ dans E(S) sont respectivement :

$$\frac{(ka+b)^2\,(k+b-2) + (k+b)\,(ka^2+b)}{n(ka+b)^2} \tag{4}$$

$$\frac{2k(n-k-1)\,(a^2b-ka^2-2ab) + 2b(b-1)\,(k-2ka-b)}{n(n-k-1)\,(ka+b)^2} \tag{5}$$

$$\frac{2b(b-1)\,(r-2)\,(k-2ka-b) - 4b(ka^2+b)\,(n-k-1)}{n(n-k-1)\,(r-1)\,(ka+b)^2} \tag{6}$$

3 - CAS DE LA VARIANCE TOTALE ENTRE LES UNITES -

Considérons le premier problème qui consiste à trouver une estimation sans distorsion de :

$$V = n \sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2. \tag{7}$$

Puisque les contributions respectives de y_j^2, $y_j y_{j+mr}$ et $y_j y_{n_i}$ sont n-1, -2 et -2, on peut imposer deux conditions sur les équations (4), (5) et (6) pour satisfaire :

$$A\ E(S) = V, \quad \text{où A est une constante.}$$

Il est facile de voir que ces deux conditions se réduisent à une seule : a doit être solution de :

$$a^2k(n-k-1)\,(n-k-b-br) + 2akb(nr-n+k-r+b) + b(kb-2bn+b+b^2+k) = 0. \tag{8}$$

Quelques calculs montrent que :

$$A = \frac{n(brn-br-k-b)}{bn+n-2k-2b+br(b-1)}. \tag{9}$$

Le discriminant de l'équation (8) étant :

$$D = kb(n-k-b)\,(brn-br-k-b)\,(n-k-b-1), \tag{10}$$

nous voyons qu'il est positif si $b \geqslant 1$ et $r \geqslant 2$, conditions qui sont toujours réalisées.

Posons $a = \dfrac{\alpha \pm \sqrt{D}}{\beta}$, où D est donné par (10).

Ecrivons S sous la forme :

$$S = y_j^2 + \ldots + y_{j+(k-1)r}^2 + y_{n_1}^2 + \ldots + y_{n_b}^2 + a_1Y^2 + a_2\,XY + a_3X^2, \tag{11}$$

où X et Y sont donnés par (1).

Des calculs montrent que tous les signes ± disparaissent et que a_1, a_2 et a_3 sont donnés par :

$$a_1 = -\frac{b(2n-k-b-1)-k}{k(brn-k-b-br)},$$

$$a_2 = -\frac{2(nr-2n+k-r+b)}{brn-k-b-br}, \qquad (12)$$

$$a_3 = \frac{(n-k-1)(n-k-b-br)}{b(brn-k-b-br)},$$

Remarquons que la formule (11) se prête mieux aux calculs que la formule (3).

Nous pouvons résumer en disant que la méthode proposée nous permet d'obtenir une estimation sans distorsion AS de V, où V, A et S sont donnés par (7), (9), (11) et (12).

4 - CAS DE LA VARIANCE AVEC ECHANTILLONNAGE SYSTEMATIQUE -

Considérons la variance W d'une estimation (d'un total) obtenue par échantillonnage systématique. Nous avons :

$$W = r\sum_{j=1}^{r}(y_j + \ldots + y_{j+(k-1)r})^2 - \Big(\sum_{j=1}^{n} y_j\Big)^2 \qquad (13)$$

Dans ce cas les contributions de y_j, $y_j y_{j+mr}$ et $y_j y_{n_i}$ sont respectivement r-1, 2(r-1) et -2. Nous pouvons imposer deux conditions sur les équations (4), (5) et (6) pour obtenir :

$$B\,E(S) = W, \quad \text{où } B \text{ est une constante.}$$

De même que dans la section 3, les deux conditions sont identiques et nous trouvons que a doit être solution de :

$$a^2k(k+b)(n-k-1) + 2abk(n-k-b-r+br) + b(bn-kb-b+n-k+rb^2-b^2-br) = 0 \qquad (14)$$

Quelques calculs donnent :

$$B = \frac{n(r-1)(n-k-b)}{2k + 2b-n+br(1-b)-nb}. \qquad (15)$$

Cette fois le discriminant de l'équation (14) est négatif et a est imaginaire.

Posons $a = p \pm iq$ et écrivons S sous la forme :

$$S = y_j^2 + \ldots + y_{j+(k-1)r}^2 + y_{n_1}^2 + \ldots + y_{n_b}^2 + b_1Y^2 + b_2XY + b_3X^2, \qquad (16)$$

où X et Y sont donnés par (1).

Des calculs montrent que tous les signes $\pm$ disparaissent et que b_1, b_2 et b_3 sont réels et donnés par :

$$b_1 = -\frac{(k+b)(b+1)(r-1)-2br}{k(n-k-b)}, \qquad (17)$$

$$b_2 = \frac{2(k+b)(r-1)-2r}{n-k-b},$$

$$b_3 = -\frac{(k+b)(n-k-1)}{b(n-k-b)}. \qquad (17)$$

Nous pouvons résumer en disant que la méthode proposée nous permet d'obtenir une estimation sans distorsion BS de W, où W, B et S sont donnés par (13), (15), (16) et (17).

5 - EXEMPLES ARTIFICIELS -

Soit une population de six valeurs : 3,2,5,5,4,2. Considérons le cas k = 3, r = 2 et b = 1. V = 57. Les six valeurs de AS sont 66, -150, 66, 174, 66 et 120. Nous remarquons que AS est bien une estimation sans distorsion, mais une de ses valeurs est négative. De plus, la dispersion de ces valeurs est considérable.

Utilisant la même population dans le cas où k = 3, r = 2 et b = 2, nous trouvons les six valeurs 51,51,51,51,69 et 69 pour AS. Puisque la vraie dimension échantillonnale est 3+2 = 5, nous avons intérêt à comparer les valeurs obtenues aux estimations sans distorsion provenant d'échantillons aléatoires simples de dimension 5. Les estimations sont les mêmes.

Considérons maintenant la population suivante : 3,2,5,5,4,2,0,-1 et le cas k = 4, r = 2 et b = 2. V = 272. Les douze estimations de V sont comprises entre AS = 170 et AS = 368 et leur variance est 3 390. Dans le cas des 28 estimations obtenues par échantillonnage aléatoire simple de dimension 6 nous trouvons que les estimations extrêmes sont 106,4 et 375,2 et que leur variance est 4611. Ainsi, non seulement les estimations obtenues par la méthode proposée sont sans distorsion, mais dans ce cas elles sont plus stables que celles obtenues par échantillonnage aléatoire simple de même dimension.

En utilisant les mêmes populations et les mêmes valeurs de k, r et b nous trouvons :

1/ W = 9 BS = 30, 3, 30, -18, 18 et -9 ;

2/ W = 9 BS = -1, -1, 35, -1,5 et 17 ;

3/ W = 16 BS = -14,5, 2, 81,5, -14,5, -16, 81,5, 44, 17,5, -18,5, 81,5, -26,5, -26.

Nous remarquons que dans chacun de ces cas certaines estimations sont négatives.

6 - APPLICATION A L'ECHANTILLONNAGE SYSTEMATIQUE EN FORET -

Nous allons appliquer les méthodes décrites dans les sections 3 et 4 à une population qui consiste en une énumération complète de Celtis soyanxii sur 156 bandes de la forêt de recherches Mpanga en Uganda. Ces données ont été mises gracieusement à la disposition de l'auteur par Mr. H. C. Dawkins du "Commonwealth Forestry Institute" de Oxford. Les mesures utilisées sont les sommes par bande des carrés des périmètres des arbres mesurés (en pieds) à hauteur de poitrine.

Trente couples (k, b) ont été choisis et cent échantillons par couple ont été tirés dans la population. Chaque échantillon comprenait un échantillon systématique de dimension k et un échantillon aléatoire simple sans remise de dimension b pris parmi les unités non choisies systématiquement.

Les calculs ont été effectués par le Centre de Calcul de l'Institut Blaise Pascal, Paris.

a) Estimation de V.

Remarquons d'abord que les 3 000 estimations de V étaient positives. Comme nous l'avons vu dans la section 5, la théorie admet des estimations négatives.

De plus, puisque les estimations AS de V sont basées sur k+b mesures, il est assez naturel de calculer les quantités :

$$M = \frac{n^2}{b-1}\left[\sum_{i=1}^{b} y_{n_i}^2 - \left(\sum_{i=1}^{b} y_{n_i}\right)^2/b\right],$$

qui sont aussi des estimations sans distorsion de V, mais basées sur b observations. Nous pouvons alors définir l'efficacité de la méthode proposée en calculant :

$$\text{Eff.} = \frac{(b-1)\,\text{var}(M)}{(k+b-1)\,\text{var}(AS)}, \quad b > 1.$$

Si au lieu de prendre les degrés de liberté associés à M et à AS, nous utilisons b et k+b, l'efficacité augmente. Posons $\overline{AS}$ = moy(AS) et $\overline{M}$ = moy(M), la moyenne étant prise sur l'ensemble des 100 échantillons par couple (k, b). La vraie valeur à estimer est $V = 10{,}50 \times 10^9$. Le tableau 1 donne les résultats obtenus.

Tableau 1

Estimation de V

k	b	$\overline{AS}\times10^{-9}$	$\overline{M}\times10^{-9}$	var(AS)$\times10^{-18}$	var(M)$\times10^{-18}$	Eff. %
4	1	10,01	-----	56,13	------	---
3	2	11,24	10,30	46,94	215,10	115
2	3	8,89	9,66	57,06	129,15	113
4	2	10,37	10,69	39,12	280,96	144
3	3	10,07	10,43	36,74	119,10	130
2	4	11,96	10,34	53,14	81,01	91
6	1	10,22	-----	49,84	------	---
4	3	10,96	10,98	39,59	119,54	101
3	4	9,76	9,09	35,25	70,15	100
6	2	10,66	10,95	36,18	250,03	99
4	4	11,02	12,01	35,50	106,06	128
3	5	11,08	11,36	21,68	51,99	137
12	1	10,01	-----	47,20	------	---
6	7	10,87	10,82	14,00	41,18	147
4	9	10,30	10,21	16,71	30,55	122
2	11	11,02	10,81	19,42	23,54	101
13	1	9,19	-----	37,17	------	---
12	2	10,04	11,63	27,01	204,04	58

Tableau 1 (Suite)

Estimation de V

k	b	$\overline{AS} \times 10^{-9}$	$\overline{M} \times 10^{-9}$	$var(AS) \times 10^{-18}$	$var(M) \times 10^{-18}$	Eff. %
6	8	11,31	11,68	13,25	38,44	156
3	11	10,80	10,90	17,67	26,45	115
13	2	11,22	10,71	47,37	258,66	39
12	3	10,32	9,90	14,86	125,19	120
6	9	10,64	10,25	13,20	35,66	154
4	11	10,33	10,34	16,28	24,75	109
26	1	11,35	-----	61,27	------	---
26	2	10,38	12,82	10,95	315,37	107
26	3	10,89	10,84	17,01	116,24	49
78	1	10,58	-----	20,85	------	---
78	2	10,96	13,40	5,25	292,55	71
78	3	10,58	9,98	3,04	95,82	79

Nous voyons qu'une estimation de V est parfaitement possible même en ne prenant qu'une observation additionnelle.

Considérons maintenant le problème qui consiste à trouver des valeurs optimum de k et de b, au sens de minimum de la variance des estimations de V pour un coût total donné. Etant donné la difficulté du calcul théorique de var(AS), nous utiliserons une méthode numérique qui nous donnera quelques indications sur la situation optimum.

Considérons le modèle :

$$var(AS) = c\, k^{\alpha} b^{\beta} (k+b)^{\gamma}.$$

Le tableau d'analyse de la variance nous donne :

Source	d. l.	S. C.	C. M.
Total	29	15,2504	
Régression sur k et b	2	12,3507	6,1753
Régression sur k+b	1	0,0816	0,0816
Erreur	26	2,8181	0,1084

Les meilleures estimations sont $\hat{\alpha} = -0,603$, $s_{\hat{\alpha}} = 0,063$ et $\hat{\beta} = -0,765$, $s_{\hat{\beta}} = 0,087$. Utilisant le modèle $var(AS) = c\, k^{\alpha} b^{\beta}$, les valeurs optimum de k et de b pour un coût donné $C = k + bc_o$ sont données par :

$$k = \frac{\alpha C}{\alpha + \beta} \quad \text{et} \quad b = \frac{\beta C}{(\alpha + \beta) c_o}.$$

En supposant $c_o = 1$, nous obtenons $\hat{k} = 0,44\ C$ et $\hat{b} = 0,56\ C$. Cette allocation correspond assez bien aux grandes valeurs de l'efficacité indiquées dans le tableau 1. Nous pouvons en déduire que, pour l'exemple considéré, la méthode proposée est plus efficace que l'échantillonnage aléatoire simple. Cette dernière conclusion peut ne pas être vraie en général.

b) Estimation des W.

Dans ce cas nous observons un nombre considérable d'estimations négatives. Posons $\overline{BS}$ = moy (BS). Le tableau 2 donne les résultats. La colonne marquée % indique le pourcentage observé d'estimations négatives pour chaque couple (k, b). Les valeurs extrêmes des BS sont aussi indiquées.

Tableau 2

Estimation des W

k	b	%	$\overline{BS} \times 10^{-9}$	extrêmes $\times 10^{-9}$	$W \times 10^{-9}$	$var(BS) \times 10^{-18}$
2	3	34	2,882	- 6,49 ; 24,25	3,436	39,47
	4	28	6,545	- 4,69 ; 51,25		103,62
	11	27	3,641	- 1,65 ; 21,00		25,09
3	2	44	1,336	- 8,13 ; 45,10	1,159	55,70
	3	54	0,125	- 7,19 ; 16,04		12,95
	4	50	1,373	- 3,91 ; 25,73		24,66
	5	44	1,714	- 3,80 ; 23,41		21,01
	11	41	1,613	- 1,41 ; 10,68		7,72
4	1	71	-0,202	-10,46 ; 27,30	0,706	40,67
	2	54	0,586	- 8,86 ; 36,08		40,87
	3	54	1,341	- 5,31 ; 42,64		45,28
	4	56	0,176	- 5,47 ; 10,22		7,71
	9	56	0,938	- 1,71 ; 11,75		6,98
	11	52	1,010	- 1,36 ; 9,39		5,71
6	1	74	-0,110	- 6,74 ; 26,73	0,269	43,00
	2	56	0,885	- 6,53 ; 39,30		34,36
	7	63	0,214	- 1,60 ; 9,06		3,93
	8	69	0,186	- 2,08 ; 9,13		4,55
	9	57	0,513	- 1,44 ; 8,76		3,59
12	1	76	-0,330	- 7,06 ; 34,47	0,126	47,59
	2	69	-0,406	- 4,77 ; 43,87		27,21
	3	60	0,117	- 3,70 ; 35,54		19,93
13	1	78	-0,951	- 6,66 ; 27,17	0,143	32,41
	2	58	1,301	- 4,53 ; 46,99		53,91
26	1	71	0,790	- 5,58 ; 29,66	0,016	58,38
	2	70	-0,580	- 3,59 ; 19,86		9,90
	3	67	0,502	- 2,65 ; 22,71		16,89
78	1	82	0,122	- 2,79 ; 18,06	0,007	20,84
	2	64	0,304	- 1,38 ; 11,44		5,22
	3	65	0,153	- 0,93 ; 11,61		2,57

Il ne semble pas y avoir de moyen simple pour sortir de cette impasse. Une solution peu satisfaisante consiste à prendre la valeur absolue. Dans ce cas on est certain d'avoir une surestimation de biais inconnu. En utilisant la même approche numérique qu'auparavant, nous trouvons que la quantité $var(BS)/W^2$ est minimum pour $k = 2$ et b maximum pour un coût donné. Si nous minimisons var(BS), nous trouvons $k = 0,35\,C$ et $b = 0,65\,C$ lorsque $c_o = 1$.

7 - CONCLUSION -

A en juger par l'exemple, l'estimation de la variance totale entre les unités, en se servant d'un échantillonnage systématique, ne présente aucune difficulté, même lorsque le nombre d'observations additionnelles est égal à un. Pour certaines valeurs de k et de b la méthode proposée semble plus efficace que l'échantillonnage aléatoire simple.

D'autre part, même si théoriquement le problème de l'estimation sans distorsion de la variance associée à un échantillonnage systématique est dans un sens résolu, l'utilisation de la méthode décrite demeure douteuse, à moins que l'exemple n'ait été mal choisi.

REFERENCE

COCHRAN W.G. (1953) - Sampling Techniques. New York : John Wiley and Sons, Inc.