MATHÉMATIQUES ET SCIENCES HUMAINES

M. C. WEISS

Variations pédagogiques sur le thème des échantillons systématiques

Mathématiques et sciences humaines, tome 102 (1988), p. 39-45

<http://www.numdam.org/item?id=MSH_1988__102__39_0>

Math. Inf. Sci. hum. (26e année, n°102, *1988*, p. 39-45).

VARIATIONS PEDAGOGIQUES SUR LE THEME DES ECHANTILLONS SYSTEMATIQUES

M.C. WEISS [1]

INTRODUCTION

Les sondages sont maintenant passés dans les moeurs... même pour exploiter les recensements de la population. Lorsqu'on dispose d'un gros fichier manuel ordonné (par exemple les pages d'annuaires) on est tenté de faire un sondage systématique, c'est à dire qu'après avoir déterminé une première page (unité statistique) on prendra aussi toutes les suivantes obtenues en ajoutant autant de fois qu'il faut un nombre entier q qui est donc la raison de la progression arithmétique donnant les numéros des pages. Les sondages systématiques sont aussi utilisés pour explorer de gros fichiers informatiques, voire dans des sondages à plusieurs degrés pour désigner des unités primaires (par exemple des cantons) lorsqu'elles sont de natures diverses, et que leur rangement ne présente pas de périodicité. Lorsque les unités statistiques proches ne sont pas trop différentes du point de vue de la variable étudiée, on obtient même souvent des estimations de la moyenne moins dispersées autour de la vraie valeur que pour les sondages équiprobables usuels. Ce ne sera pas le cas pour le fichier utilisé.

Cependant les sondages systématiques ont un inconvénient : une fois fixé le mode de tirage, le nombre d'échantillons possibles est limité, alors qu'il est très grand lorsque chaque unité est tirée selon une certaine loi de probabilité fixée à l'avance. (En général on utilise un tirage équiprobable exhaustif. Dans cet article ce type de sondage sera appelé élémentaire.) On sait calculer alors non seulement la variance théorique des estimateurs usuels à partir du fichier, mais aussi un estimateur sans biais de cette variance à partir de l'échantillon, ce qui n'est pas vrai pour les échantillons systématiques. (On utilise parfois la variance de l'échantillon ou la somme des carrés des différences. Voir pour cela les ouvrages cités en référence.) Dans le cadre de cet article, on ne pourra connaître la dispersion des estimateurs que pour le fichier et la variable considérés, par exemple à partir de l'ensemble des échantillons.

POSITION DU PROBLEME ET NOTATIONS

Nous étudierons pour un fichier de taille M des échantillons systématiques de taille m, ou à la rigueur m+1, pour estimer la moyenne y d'une variable numérique Y. Nous nous placerons dans le cas, non traité explicitement dans les ouvrages cités, où l'effectif M n'est pas un multiple de m. La division entière de M par m donne le résultat M=qm+r. On utilisera la valeur du quotient comme raison de la progression arithmétique. Toutes les notations sont empruntées au livre de J. Desabie, en particulier S pour une sommation sur les unités de l'échantillon. C'est la présence du reste r positif qui constitue une difficulté pour prendre en compte toutes les unités

1 Groupe Math et Psycho de l'Université René Descartes.

du fichier. Il y a de plus un biais statistique pour $\bar{y}$ quand on tire à probabilités égales les premières unités. Pour les estimateurs biaisés on définira comme à l'habitude le biais pour l'estimateur t de la moyenne μ par $B=E(t)-\mu$ et l'erreur totale E.T. par son carré : $(E.T.)^2=E(t-\mu)^2$.

Après l'exposé précis de deux modes de détermination des échantillons systématiques permettant de tirer toutes les unités du fichier, on effectuera pour chacun d'eux un tirage à probabilités égales (I) puis on cherchera un estimateur sans biais en tenant compte des probabilités d'apparition de chaque unité du fichier (II.1). On montrera de façon générale pourquoi les estimateurs usuels de la variance pour la méthode des probabilités inégales ne sont pas utilisables avec les tirages systématiques, l'un d'eux donnant d'ailleurs des variances toujours négatives, en développant la démonstration pour le deuxième mode de détermination des échantillons (II.2). On cherchera ensuite des probabilités inégales de tirages entre les échantillons systématiques donnant la meilleure variance théorique (pour des estimateurs sans biais)(III).

Dans une annexe numérique on donnera des résultats, estimateurs, $(E.T.)^2$ ou variance effective de ceux-ci pour deux variables corrélées d'un même fichier pour lequel M=334, m=30, donc q=11 et r=4. On commentera les résultats au fur et à mesure pour les différentes méthodes, par rapport à la méthode élémentaire. Dans la conclusion, on essayera de voir les résultats qui pourraient s'étendre à d'autres fichiers ayant des variables pour lesquelles l'échantillonnage systématique donne des résultats corrects.

I. MODES DE DETERMINATION DES ECHANTILLONS. TIRAGE EQUIPROBABLE.

1. *Les deux modes.*

On considérera deux modes de détermination des échantillons systématiques à partir d'une première unité qui seront repris tout le long de cet article.

Mode 1: On tire un nombre qui désigne une unité dans (1,q), on prend ensuite les unités de q en q et on va jusqu'au bout du fichier. On a alors q échantillons disjoints (ou grappes), ceux dont la première unité vaut de 1 à r sont constitués de m+1 unités, les q-r autres sont de taille m.

Mode 2: On connait la valeur de M, comment obtenir des échantillons systématiques exactement de taille m ? On tire un nombre entier compris entre 1 et q+r et on prend les m-1 unités suivantes désignées par la progression arithmétique. Parmi les q+r échantillons les unités de la fin des r premiers échantillons se retrouvent dans les r derniers. (Cette manière de procéder n'est pas habituelle. La procédure avec un pas fractionnaire n'est pas considérée ici. Dans ce dernier cas il y a aussi des unités qui se retrouvent dans deux échantillons.) Dans la suite on appellera A l'ensemble des unités qui appartiennent à deux échantillons, B l'ensemble des r premières et des r dernières unités du fichier, C celui des unités des q-r échantillons à m unités du premier mode.

2. *Tirage équiprobable*

On supposera d'abord que la première unité de chaque échantillon est tirée avec probabilités égales. L'estimateur usuel $\bar{y}$, moyenne arithmétique des observations de la variable Y est statistiquement *biaisé* pour les deux modes, beaucoup plus pour le deuxième, puisque les unités de type A sont tirées avec une probabilité double des autres.Il est donc préférable d'utiliser le premier mode, ce que montrent les résultats de l'annexe. On a de plus calculé pour chaque variable la variance de $\bar{y}$ correspondant au tirage élémentaire d'un échantillon de taille 30. Le carré de l'erreur totale est, pour les deux variables et dans les deux modes, du même ordre de grandeur, mais supérieur à cette variance.

II. RECHERCHE D'UN ESTIMATEUR SANS BIAIS DE LA MOYENNE.

On utilisera les probabilités d'inclusion, π_α, probabilité que l'unité U_α appartienne à l'échantillon, $\pi_{\alpha\beta}$ pour le couple U_α, U_β .

Voici le rappel des formules usuelles:
Estimateurs sans biais de la somme (respectivement de la moyenne) du caractère Y : $y'=S(y_i/\pi_i)$, $\bar{y}_{in}=y'/M$.
Variances théoriques : $V(\bar{y}_{in})=V(y') / (M)^2$.
Il existe deux formules pour calculer la variance de y' : celle d'Horvitz-Thomson:
$V1(y')=\sum \pi_\alpha(1-\pi_\alpha)(Y_\alpha/\pi_\alpha)^2+\sum\sum(\pi_{\alpha\beta} - \pi_\alpha\pi_\beta)Y_\alpha Y_\beta /\pi_\alpha\pi_\beta)$
et $V2(y')=1/2(\sum\sum(\pi_\alpha\pi_\beta-\pi_{\alpha\beta})((Y_\alpha/\pi_\alpha)-(Y_\beta /\pi_\beta))^2)$,
formule de Yates et Grundy.
(N.B. pour les sommes doubles les unités sont toujours supposées différentes).

Les deux estimateurs, obtenus en faisant les calculs sur les unités de l'échantillon après avoir divisé les termes des sommes sur un indice (respectivement des sommes à deux indices) par les probabilités d'inclusion correspondantes, ne sont pas utilisables, puisque, pour les échantillons systématiques, certains couples sont toujours dans des échantillons différents ce qui rend les probabilités correspondantes nulles. C'est pourquoi, seules les variances théoriques seront utilisées dans l'annexe numérique.

1. *Premiers résultats*

Le calcul des estimateurs sans biais des moyennes a été fait dans les deux modes pour le tirage de la première unité à probabilités égales.

Mode 1: $\pi_\alpha=1/q$; l'estimateur sans biais de la moyenne est la moyenne de l'échantillon multipliée par q(m+1)/M pour les échantillons de taille m+1 et par q.m/M pour les échantillons de taille m. Pour ce fichier et pour les deux variables les estimateurs sans biais de la moyenne sont moins dispersés que la moyenne arithmétique des échantillons. Pour la première variable l'estimateur sans biais du tirage systématique est meilleur que pour le sondage élémentaire, pour la deuxième variable il est un peu moins bon.

Mode 2: On appellera A l'ensemble des unités qui se retrouvent dans deux échantillons possibles. Leur probabilité d'inclusion est 2/q+r. Celle des autres est de 1/q+r.

Les variances d'échantillonnage valent environ 250 fois celles de l'échantillon aléatoire de même taille et cet estimateur n'est donc pas intéressant pour ce mode de tirage.

A titre de démonstration nous allons expliciter la variance pour ce mode et voir pourquoi elle n'est pas estimable par les procédés usuels

2. *Analyse plus détaillée de la variance dans le mode 2.*

Il suffit d'étudier la variance V(y') de la somme. Soit B l'ensemble des r premières et des r dernières unités du fichier et C celui des unités des échantillons de taille m du mode 1. Posons Ech(u) pour l'ensemble des autres unités de l'échantillon auquel appartient l'unité u. On obtient alors pour deux unités différentes a et k de A avec $k\in Ech(a)$: $\pi_a=\pi_k=\pi_{ak}=2/(q+r)$. On a alors $\pi_a\pi_k-\pi_{ak}=(2/(q+r)^2)(2-q-r)<0$. Pour $b\in B$ et $a\in Ech(b)$, a étant une unité de A, on a $\pi_b=\pi_{ab}=1/(q+r)$, on obtient donc: $\pi_a\pi_b-\pi_{ab}=(1/(q+r)^2)(2-q-r)<0$. Ces résultats concernent les couples possibles pour les r premiers (resp. les r derniers échantillons) du mode 2. On appellera échantillons de type 1 (Ech1) ces 2r échantillons. (On aurait également des valeurs négatives avec ceux à m+1 unités du mode 1, appelés dans la suite échantillons de type 3 (Ech3)).

Appelons maintenant échantillons de type 2 (Ech2) les autres, communs d'ailleurs aux deux

modes, formés par définition d'unités de type C. Soient c et g deux unités du même échantillon de ce dernier type. On obtient alors $\pi_c=\pi_g=\pi_{cg}=1/(q+r)$, d'où, $\pi_c\pi_g-\pi_{cg}=(1/(q+r)^2)(1-q-r)<0$ aussi. Tous ces termes interviennent dans le calcul théorique de la variance. Pour les couples n'appartenant pas à un même échantillon on a $\pi_{\alpha\beta}=0$ et $\pi_\alpha\pi_\beta-\pi_{\alpha\beta}>0$.

Si on calculait l'estimateur de Yates et Grundy $V2(y')=1/2(\pi_i\pi_j-\pi_{ij})((y_i/\pi_i)-(y_j/\pi_j))^2/\pi_{ij}$, ce qui est toujours possible pour un échantillon, on trouverait, une valeur négative, estimant les termes négatifs de la variance théorique. Leur somme est en valeur absolue inférieure à celle des termes positifs puisqu'une variance est positive. Si la variable dont on estime la moyenne est toujours positive, l'estimateur d'Horvitz-Thomson, qui utilise ces termes croisés avec le signe opposé, ne prendrait en compte que les termes positifs de la variance. Ce deuxième estimateur, toujours incorrect mais calculable, surestimerait alors statistiquement la variance.

Pourrait-on utiliser l'estimateur de la variance correspondant à un tirage à probabilités inégales indépendant qui donnerait les mêmes probabilités d'inclusion, avec à chacun des m tirages $A_\alpha=\pi_\alpha/m$? Les formules aussi bien que le calcul sur les variables du fichier montrent que l'on obtient une valeur théorique parfois supérieure, parfois inférieure à V(y').

III. RECHERCHE DES MEILLEURES PROBABILITES INEGALES POUR LES PREMIERES UNITES DES ECHANTILLONS SYSTEMATIQUES.

On peut réécrire la variance théorique:

$V(y')=\Sigma((Y_\alpha)^2/\pi_\alpha)+\Sigma\Sigma(\pi_{\alpha\beta}\,Y_\alpha Y_\beta/\pi_\alpha\pi_\beta)-Y^2$. Pour la minimiser il suffit de dériver par les différentes probabilités en tenant compte des contraintes dues aux deux modes de tirage. Nous nous bornerons à distinguer les types d'échantillons qui interviennent dans chaque mode.

1. Mode 2

On tire les 2r échantillons de type 1 avec une probabilité π_1 et les q-r échantillons de type 2 avec la probabilité π_2. On a l'égalité : $2r\pi_1+(q-r)\pi_2=1$ ce qui entraine $0<\pi_1<1/2r$.

Appelons f_A(resp. f_B, f_C) la somme des carrés de la variable pour les unités correspondant à A (resp. B, C). Posons aussi des notations pour la somme des produits des variables pour les divers types d'échantillons; on désigne par β les unités de B et $\beta'\in Ech(\beta)$ et par α et α' les couples d'unités de A appartenant à $Ech(\beta)$, alors $f_{Ech1}=\Sigma\Sigma Y_\alpha Y_{\alpha'}+2\Sigma\Sigma Y_\beta Y_{\beta'}$. De même avec les unités α de C et $\alpha'\in Ech(\alpha)$ on note $f_{Ech2}=\Sigma\Sigma Y_\alpha Y_{\alpha'}$. Les probabilités d'inclusion des couples sont nulles en dehors des échantillons. En mettant les valeurs des probabilités d'inclusion des couples en fonction de π_1 et π_2 on obtient :

$$V(y')=(f_A/2\pi_1)+(f_B/\pi_1)+(f_C/\pi_2)+(f_{Ech1}/2\pi_1)+(f_{Ech2}/\pi 2)-Y^2.$$

En remplaçant, par exemple, π_2 par sa valeur en fonction de π_1 et en dérivant par ce dernier on obtient la solution optimale qui est unique.

Pour le fichier de référence et pour la première variable, V(y'/M) est de l'ordre de 4.1 alors que pour la méthode aléatoire élémentaire exhaustive on obtiendrait 5.1. Les deux méthodes sont sans biais. Pour le fichier considéré on a $\pi_2\cong\pi_1$, ce qui donne à peu près la même pondération à toutes les unités.

2. Mode 1

On peut aussi trouver une meilleure répartition entre les échantillons des deux types concernés, Ech2 et Ech3. En donnant l'indice correspondant aux probabilités de tirage des premières unités de chaque échantillon on a l'égalité $r\pi_3+(q-r)\pi_2=1$. Appelons D la réunion des ensembles A et B et posons $f_D=f_A+f_B$. Si α est une unité de D et $\beta\in Ech(\alpha)$ alors $f_{Ech3}=\Sigma\Sigma Y_\alpha Y_\beta$. On a cette fois-ci $V(y')=(f_D/\pi_3)+(f_C/\pi_2)+(f_{Ech3}/\pi_3)+(f_{Ech2}/\pi_2)-Y^2$. On obtient là aussi une solution unique pour les probabilités optimales π_2 et π_3.

Cette méthode est a priori supérieure à la précédente puisqu'il existe des échantillons de taille m+1 plus "informatifs" que ceux de taille m. Ceci est vérifié pour les deux variables du fichier considéré. Les valeurs optimales pour π_2 sont à peu près les mêmes que pour le mode 2.

On obtient pour les deux variables et pour les deux modes de meilleurs résultats qu'avec la moyenne arithmétique des échantillons systématiques équiprobables.

CONCLUSION

Il faut être prudent pour généraliser ce qui est observé sur le fichier (structuré, forcément petit pour pouvoir effectuer aisément tous les calculs) analysé dans l'annexe.

Dans la pratique les sondages systématiques utilisés à bon escient peuvent apporter un gain relativement important pour la précision des observations, même si on ne peut pas la mesurer en une seule enquête.

Voici cependant en termes généraux ce qui a été observé. Parmi les deux modes proposés, il vaut mieux utiliser des échantillons sans recouvrement possible (mode 1). Ceci est peut-être dû aussi à la taille plus élevée de certains échantillons. La connaissance de M n'est pas nécessaire pour le tirage équiprobable dans le mode 1. Si $r<<q$, pour la plupart des échantillons la moyenne arithmétique des observations est peu différente de l'estimateur sans biais, ce qui doit rendre celui-ci négligeable. Pour notre fichier l'estimateur sans biais est de toute façon meilleur dans ce mode. Son utilisation et le calcul des probabilités inégales optimales pour les deux modes nécessite de connaître M, la taille de la population.

Attention à ne pas utiliser les estimateurs habituels de la variance! Si l'on a quelques connaissances sur la répartition des unités du fichier entre les sous-ensembles A,B,C tels qu'ils sont définis dans le corps de l'article on peut évaluer les probabilités inégales optimales de tirage parmi les échantillons désirés (mode 1 ou 2). Le tirage à probabilités optimales pour les deux modes donne le meilleur résultat, non biaisé.

Pour estimer la variance d'échantillonnage, le mieux serait de doubler la raison de la progression arithmétique et de tirer deux demi-échantillons différents. Pour le mode 1 on retrouve alors le problème classique du tirage de deux grappes.

BIBLIOGRAPHIE TRES SUCCINTE

Livres de base en français.

[1] DEROO, M. et DUSSAIX, A.M., *Pratique et analyse des enquêtes par sondage*, Paris, Presses Universitaires de France, 1980.

[2] DESABIE, J.,*Théorie et pratique des sondages*, Paris, Dunod, 1966.

[3] GROSBRAS, J.M., *Méthodes statistiques des sondages*, Paris, Economica, 1987.

ANNEXE NUMERIQUE

Fichier de référence "INOP"

Variables Y=QIV, Z=QINV.

Les 334 individus sont rangés selon 4 groupes d'effectifs 86,80,86,82, dont les variances sont presque identiques et les moyennes dans l'ordre QIV, QINV, sont respectivement: 113.3,116.4 ; 108.0,109.6 ; 104.4,109.9 ; 103.3,104.3.

Moyennes pour le fichier : γ =107.29, ζ=110.11.

On cherche à estimer ces moyennes selon les différentes méthodes :
Un tirage aléatoire équiprobable exhaustif de 30 U.S. donnerait $V(\bar{y})$=5.1 , $V(\bar{z})$=5.7.

Notation: Moy(.) désigne la moyenne arithmétique d'un estimateur pour tous les échantillons concernés.

MODE 1:
Il y a 4 échantillons de taille 31, de type 3, et 7 échantillons de taille 30, de type 2.

A .*Tirage équiprobable.*

I. Moy($\bar{y}$)=107.31, $V(\bar{y})$=5.404, Biais=0.01, $(E.T.)^2$=5.405.
Moy($\bar{z}$)=110.13, $V(\bar{z})$=8.129, Biais=0.016, $(E.T.)^2$=8.1296.

II. Estimateur sans biais.
Chaque échantillon est tiré avec la probabilité π = 1/11=0.09091.
Moy($\bar{y}_{in}$) = 107.29, $V(\bar{y}_{in})$ =4.143.
Moy($\bar{z}_{in}$) = 110.11, $V(\bar{z}_{in})$ =7.640.

B. *Tirage à probabilités inégales entre les échantillons.*

III. Tirage avec les probabilités inégales optimales correspondant à chaque variable.
a. Pour Y : π_3=0.0914378, π_2 = 0.0906069
$V(\bar{y}_{in})$ = <u>3.924</u>, $V(\bar{z}_{in})$ = 7.181.
b. Pour Z : π_3=0.0916979, π_2=0.0904583
$V(\bar{y}_{in})$ = 3.977, $V(\bar{z}_{in})$ = <u>7.122</u>.

On constatera que les grappes ne sont pas efficaces pour Z, car le tirage élémentaire de 30 observations est meilleur.

MODE 2 :
Il y a 8 échantillons de type 1 et les 7 échantillons de type 2.

A.*Tirage équiprobable.*

I. Moy($\bar{y}$) = 106.92, V($\bar{y}$) = 5.308, Biais =-0.37, $(E.T.)^2$ = 5.447.
Moy($\bar{z}$) = 109.85, V($\bar{z}$) = 8.474, Biais =-0.26, $(E.T.)^2$ = 8.5436.

II. Estimateur sans biais. Chaque échantillon est tiré avec la probabilité π=1/15=0.0667. Les variances sont grandes.
Moy($\bar{y}_{in}$) = 107.29, V($\bar{y}_{in}$) = 1303.41.
Moy($\bar{z}_{in}$) = 110.11, V($\bar{z}_{in}$) = 1359.275.

B.*Tirage à probabilités inégales entre les échantillons.*

III. Tirage avec les probabilités inégales optimales correspondant à chaque variable.
a. Pour Y: π_1 = 0.0457, π_2 = 0.0906
V($\bar{y}_{in}$) = 4.053, V($\bar{z}_{in}$) =7.399.
b. Pour Z: π_1 = 0.0458, π_2 = 0.0905
V($\bar{y}_{in}$) = 4.107, V($\bar{z}_{in}$) = 7.342.