11-617F
no.88-21F
ex.2

Statistics Canada | Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-88-021F

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-88-021F

DIRECTION DE LA MÉTHODOLOGIE

STATISTICS CANADA STATISTIQUE CANADA
JAN 9 1995
LIBRARY BIBLIOTHÈQUE

DÉTERMINATION, ALLOCATION ET SÉLECTION DE L'ÉCHANTILLON

par

M. Latouche

Mai 1988

DÉTERMINATION, ALLOCATION ET SÉLECTION DE L'ÉCHANTILLON

Présentation faite au Colloque sur
les méthodes et domaines d'application
de la statistique
dans le cadre du
cinqante-sixième congrès de
l'A.C.F.A.S.

le 10 mai 1988
à l'université de Moncton
Nouveau-Brunswick

Michel Latouche
Division des méthodes d'enquêtes
auprès des entreprises
Statistique Canada

ABSTRACT

A sample size determination, allocation and selection system has been created which minimizes the sample size for a given level of precision. The precision is expressed as the coefficient of variation (cv) defined by the ratio of the standard error to the estimate itself. Prior to the allocation, an optional module can distribute the desired global cv of the survey among the selection strata. The system can also allocate sample of a given size. The stratification can be done according to the size of one or two variables. A selection and panel rotation module completes the system. The overall system has been generalized in order to be usable by several surveys.

DÉTERMINATION, ALLOCATION ET SÉLECTION DE L'ÉCHANTILLON

Michel Latouche [1], Statistique Canada

Un système de détermination, d'allocation et de sélection de l'échantillon a été conçu de façon à minimiser la taille de l'échantillon pour un niveau de précision donné. Cette précision est exprimée par le coefficient de variation (c.v.) qui est le rapport entre l'écart-type et l'estimation elle-même. Avant de faire l'allocation, un module facultatif permet de répartir le c.v. global de l'enquête aux différentes strates de sélection. Le système peut également allouer un échantillon de taille déterminée. La stratification peut se faire selon la taille d'une ou de plusieurs variables. Un module de sélection et de renouvellement par panel complète ce système. Le système a été généralisé de façon à pouvoir être utilisé par plusieurs enquêtes.

INTRODUCTION

L'un des objectifs du remaniement des enquêtes sur les entreprises entrepris par Statistique Canada est de développer des systèmes qui peuvent être utilisés par plusieurs enquêtes [1]. Dans cet optique, un système de détermination, d'allocation et de sélection de l'échantillon (D.A.S.É.) a été récemment mis en place. Ce système a été construit par module, ce qui lui donne une flexibilité d'utilisation et permet de le modifier facilement au gré des développements méthodologiques. Le système D.A.S.É. a tout d'abord été conçu pour satisfaire aux exigences des enquêtes sur les commerces de gros et de détail. Plusieurs paramètres et options ont été ajoutés pour faciliter l'intégration d'autres enquêtes. Les prochaines sections décrivent la structure de l'univers des enquêtes sur les commerces de gros et de détail, pour ensuite présenter l'utilisation de D.A.S.É. faite par ces deux enquêtes. Finalement, la dernière section donne un aperçu des autres options disponibles ou en cours de développement.

1. STRUCTURE

L'univers des commerces de gros et de détail est formé de deux dimensions: industrielle et géographique. Chacune de ces dimensions est divisée en des éléments constituant un niveau de publication. À leur tour, ces éléments sont

---

1. Michel Latouche, division des méthodes d'enquêtes-entreprises, Statistique Canada, immeuble R.H. Coats, section 11-M, Ottawa, K1A 0T6.

divisés en d'autres éléments pour constituer un niveau de sélection tel que montré par les figures 1.a) et 1.b). En combinant les niveaux de sélection des deux dimensions, on obtient des cellules disjointes qui forment la structure de l'univers (figure 1.c)).

Figure 1. structure de l'univers

1.a) industrie

X		Y			Z	publication
1	2	3	4	5	6	sélection

1.b) géographie

A	1
B	2
	3
C	4
	5

pub. sél.

1.c) cellules

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56

L'utilisation d'une telle structure donne aux utilisatèurs l'opportunité de connaitre et de définir le degré de précision pour deux niveaux différents, et ce pour chaque dimension.

2. UTILISATION

Les enquêtes sur les commerces de gros et de détail utilisent les quatre modules du système D.A.S.É. Ces modules sont:

2.1 Détermination du degré de précision
2.2 Stratification
2.3 Allocation
2.4 Sélection et renouvellement

2.1 Détermination du degré de précision

Le degré de précision est exprimé par le coefficient de variation qui est donné par:

$$cv(\hat{Y}) = \frac{\hat{\sigma}(\hat{Y})}{\hat{Y}} \qquad (1)$$

où $\hat{Y}$ est l'estimation du total de la principale variable d'intérêt de l'enquête (ventes, recettes, nombre d'employés, etc) et $\hat{\sigma}(\hat{Y})$ est l'estimation de la variance de $\hat{Y}$.

Plusieurs enquêtes ont comme objectif d'obtenir de l'information avec un degré de précision égal pour tous les éléments d'un même niveau. Ce principe d'équité peut être respecté seulement si l'allocation est faite à partir d'un degré de précision pré-déterminé plutôt que d'allouer à l'aide d'une taille d'échantillon fixée d'avance. Le module se sert du fait que les éléments sont disjoints pour exprimer la variance d'un élément donné en une somme de variances d'éléments définis au niveau inférieur. En exprimant la variance en terme de $cv(\hat{Y})$ et de $\hat{Y}$, cette relation devient:

$$cv^2(\hat{Y}) \cdot \hat{Y}^2 = \sum_i cv^2(\hat{Y}_i) \cdot \hat{Y}_i^2 \qquad (2)$$

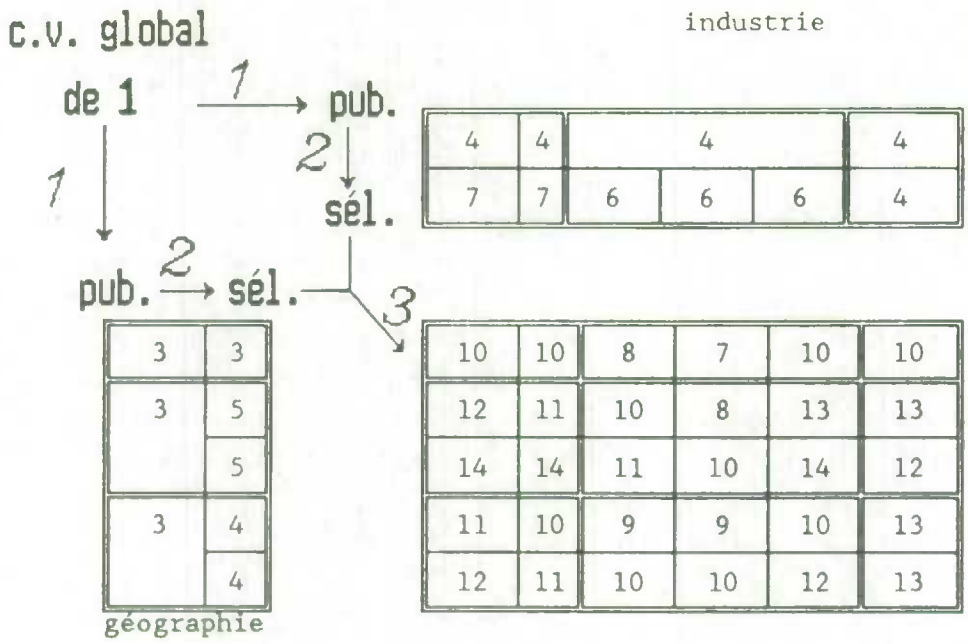
où "i" représente les éléments du niveau inférieur qui entrent dans la composition de l'élément majeur. Par exemple, l'élément majeur peut être la province du Nouveau-Brunswick (niveau de publication) et les éléments inférieurs, les régions métropolitaines de recensement (niveau de sélection à l'intérieur de la province).

La figure 2. montre les étapes du calcul des c.v. À l'aide d'un c.v. spécifié par les utilisateurs pour l'ensemble de l'enquête et d'estimations du total pour les différents éléments [2], l'équation (2) permet de calculer séparément pour chaque dimension un c.v. pour le niveau de publication (étape 1). À l'intérieur de chaque élément du niveau de publication, cette équation sert aussi à calculer un c.v. pour le niveau de sélection (étape 2). Une fois que les c.v. ont été calculés pour chaque élément du niveau sélection des deux dimensions, les c.v. sont calculés au niveau des cellules (étape 3) à l'aide d'un processus itératif similaire au processus de la méthode itérative des quotients [2].Cette méthode a pour but de distribuer les fréquences dans un tableau de fréquences à N dimensions de façon à respecter les fréquences marginales pré-établies de chaque dimension. Cette méthode a été adaptée pour que l'algorithme trouve un c.v. pour chaque cellule de manière à ce que les c.v. au niveau de la sélection soient respectés pour les deux dimensions.

---

2. Ces estimations proviennent de données antérieures ou sont fournies par d'autres sources.

Figure 2. calcul des c.v. (%)



4	4	4			4
7	7	6	6	6	4

3	3
3	5
	5
3	4
	4

10	10	8	7	10	10
12	11	10	8	13	13
14	14	11	10	14	12
11	10	9	9	10	13
12	11	10	10	12	13

Soit $cv_r(\hat{Y}_{ij})$ le c.v. calculé à la $r^{ième}$ itération pour la cellule formée par la combinaison de l'élément "i" du niveau géographique de sélection et de l'élément "j" du niveau industriel de sélection.

Les $cv_r(\hat{Y}_{ij})$ sont donnés par:

$$cv_r(\hat{Y}_{ij}) = \frac{cv_{r-1}(\hat{Y}_{ij}).cv(\hat{Y}_{i.}).cv(\hat{Y}_{.j}).\hat{Y}_{i.}.\hat{Y}_{.j}}{\sqrt{\sum_i cv^2_{r-1}(\hat{Y}_{ij}).\hat{Y}^2_{ij}} \times \sqrt{\sum_j cv^2_{r-1}(\hat{Y}_{ij}).\hat{Y}^2_{ij}}} \qquad (3)$$

où $cv(\hat{Y}_{i.})$ est le c.v. pour le $i^{ième}$ élément du niveau géographique de sélection; $cv(\hat{Y}_{.j})$ est le c.v. pour le $j^{ième}$ élément du niveau industriel de sélection; $\hat{Y}_{ij}, \hat{Y}_{.j}$ et $\hat{Y}_{i.}$ sont les estimations du total pour les éléments correspondants; $cv_0(\hat{Y}_{ij})$ est la moyenne entre le c.v. de la cellule calculé en répartissant à l'aide de l'équation (2) le $cv(\hat{Y}_{i.})$ aux cellules de l'élément "i" et celui obtenu en répartissant $cv(\hat{Y}_{.j})$ aux cellules de l'élément "j".

L'algorithme ne converge pas de façon absolue; cependant après un grand nombre d'itérations, les $cv_r(\hat{Y}_{ij})$ oscillent. L'algorithme arrête lorsque la différence entre $cv_r(\hat{Y}ij)$ et $cv_{r-1}(\hat{Y}_{ij})$ est jugée négligeable pour toutes les cellules. Si,

après un nombre déterminé d'itérations, une seule de ces différences est trop élevée, alors le système prend pour valeur finale la moyenne des deux dernières itérations.

2.2 Stratification

Le système permet de stratifier selon la variable d'intérêt utilisée précédemment. Une telle stratification permet de minimiser la taille de l'échantillon et de réduire le biais des estimations particulièrement lorsque la variable d'intérêt possède une distribution asymétrique [3]. Hidiroglou (1986), a développé une méthode permettant de diviser chaque cellule en une strate auto-représentative où toutes les unités sont recensées et une strate non auto-représentative où les unités sont échantillonnées. La borne délimitant les deux strates est calculée à l'aide d'un processus itératif. À la $r^{ième}$ itération, la borne (B_r) est donnée par:

$$B_r = \bar{Y}_{r-1} + \sqrt{\frac{N_{r-1} - 1}{N_{r-1}^2} \cdot cv^2(\hat{Y}_{ij}) \cdot \hat{Y}_{ij}^2 + S_{r-1}^2} \qquad (4)$$

où N_r, $\bar{Y}_r$ et S_r^2 sont respectivement le nombre d'unités, la moyenne et la variance de la strate non auto-représentative délimitée par la $r^{ième}$ borne. Les $cv(\hat{Y}_{ij})$ et $\hat{Y}_{ij}$ sont les c.v. et les totaux des cellules (ij). Initialement, N_0, $\bar{Y}_0$ et S_0^2 sont calculés sur l'ensemble de la cellule. Ce processus est appliqué séparément à chaque cellule. La borne choisie est celle parmi toutes les itérations qui minimise la taille de l'échantillon [3].

Le système améliore cette méthode en permettant de rediviser la strate non auto-représentative en plusieurs strates ayant des fractions de sondage différentes. Le nombre de strates est déterminé par l'utilisateur. La méthode employée est une modification de celle présentée par Dalenius et Hodges (1959): elle consiste à répartir également le cumulé de la racine carrée du total de la variable d'intérêt entre chaque strate. Cochran (1977) donne un exemple de la méthode de Dalenius qui répartit également le cumulé de la racine carrée de la fonction de densité de la variable d'intérêt entre les strates.

2.3 Allocation

Une fois la stratification terminée, le système calcule un facteur d'allocation pour chaque strate non auto-représentative à l'intérieur de la cellule. Ce facteur est proportionnel au total de la variable d'intérêt dans la strate. Le facteur d'allocation est donné par:

$$a_h = \frac{\hat{Y}_h^e}{\sum_h \hat{Y}_h^e} \qquad 0 < e \leq 1 \qquad (5)$$

où $\hat{Y}_h$ est le total pour la $h^{ième}$ strate non auto-représentative.

La sommation est faite sur l'ensemble des strates non auto-représentatives à l'intérieur de la cellule. L'allocation peut être aussi faite de façon proportionnelle au nombre d'unités dans les strates. La puissance "e" permet de modifier l'allocation de façon à assurer une meilleure représentativité des petites strates. Une puissance de 1/2 est recommandé si on désire une meilleure précision au niveau des strates [4]. La taille de l'échantillon pour l'ensemble des strates non auto-représentatives de la cellule dépend des caractéristiques des strates, du facteur d'allocation et du c.v. de la cellule:

$$n = \frac{\sum_h P_h N_h^2 \left(S_h^2 + (1 - P_h) \cdot \bar{Y}_h^2\right) / a_h}{cv^2(\hat{Y}_{ij}) \cdot \sum_k (N_k \cdot P_k \cdot \bar{Y}_k)^2 + \sum_h P_h N_h \left(S_h^2 + (1 - P_h) \cdot \bar{Y}_h^2\right)} \qquad (6)$$

où "k" représente toutes les strates de la cellule, tandis que "h" représente uniquement les strates non auto-représentatives. P_h est la proportion d'unités vivantes dans la strate au moment de tirer l'échantillon. (Ce paramètre peut être utile s'il y a un délai entre la création de la base de sondage et la sélection de l'échantillon.) Par la suite, la taille de l'échantillon de chaque strate non auto-représentative est donnée par:

$$n_h = n \cdot a_h \qquad (7)$$

où n_h peut être fractionnaire.

Il se peut que pour certaines strates, la taille de l'échantillon soit supérieure à la taille de la population. Dans ce cas, la strate est déclarée auto-représentative et la taille de l'échantillon pour les autres strates est recalculée à l'aide de (5), (6) et (7) avec maintenant deux strates auto-représentatives dans cette cellule au lieu d'une seule. Ces étapes sont répétées tant et aussi longtemps qu'il y a de la sur-allocation. À la limite, toute la cellule est auto-représentative.

En dernier lieu, l'utilisateur peut modifier les tailles d'échantillon calculées de manière à ce qu'elles soient au moins égales à une taille minimale. Il peut également contraindre les tailles d'échantillon à respecter un poids de sélection maximum.

2.4 Sélection et renouvellement

Le dernier module du système consiste à tirer un échantillon dans chaque strate en accord avec les tailles calculées. Les responsables des enquêtes sur les commerces de gros et de détail désirent un système de sélection avec un renouvellement des unités échantillonnées. De cette façon, l'on obtient à la fois une plus grande régularité des estimations d'un mois à l'autre et de meilleures estimations des tendances sans trop alourdir le fardeau des répondants. Pour ce faire, D.A.S.É. utilise un système de renouvellement par panels. Les unités de la population à l'intérieur de la strate sont réparties en groupes. Ces groupes, appelés panels, ont le même nombre d'unités à une unité près. Ensuite, des panels sont sélectionnés pour faire partie de l'échantillon de façon à respecter la fraction de sondage (F_h). Lors du renouvellement de l'échantillon, un panel est retiré de l'échantillon et un nouveau y est ajouté par échantillonnage systématique. Le nombre de panels dans l'échantillon (P_{IN}) est calculé de manière à satisfaire les trois contraintes suivantes:

1- La fraction de sondage F_h doit être respectée c.à.d.

$$\frac{P_{IN}}{(P_{IN} + P_{EX})} = \frac{n_h}{N_h} = F_h \qquad (8)$$

où P_{EX} est le nombre de panels en dehors de l'échantillon.

2- Les unités et du même coup les panels doivent être en dehors de l'échantillon au moins T_{EX} périodes de temps c.à.d.

$$P_{EX} \geq T_{EX}.$$

3- Les unités et du même coup les panels doivent rester dans l'échantillon T_{IN} périodes de temps c.à.d.

$$P_{IN} = T_{IN}.$$

Lorsque les trois contraintes peuvent être respectées, P_{EX} est calculé en posant P_{IN} égal à T_{IN} dans l'équation (8), et on obtient :

$$P_{EX} = \text{partie entière de } \{(T_{IN} * (1-F_h)/F_h) + 0.5\}$$

S'il est impossible de respecter les trois contraintes, alors on modifie la troisième contrainte en acceptant que P_{IN} soit supérieur à T_{IN}, on pose P_{EX} égal à T_{EX} et on résout l'équation (8). De cette façon, P_{IN} est donné par:

$$P_{IN} = \text{partie entière de } \{(T_{EX} * F_h/(1-F_h)) + 0.5\}$$

Des exemples de calculs de P_{IN} et P_{EX} sont présentés à l'annexe A.

3. AUTRES OPTIONS

Le système D.A.S.É. peut traiter également des univers ayant une seule dimension et/ou un seul niveau. De plus, il n'est pas nécessaire que les dimensions soient industrielles ou géographiques; l'utilisateur peut définir lui-même ses dimensions et ses niveaux.

La répartition du c.v. global peut être faite de façon à ce que les éléments du niveau de sélection aient le même c.v. peu importe l'élément de publication auquel ils appartiennent. Par contre, avec cette option, les éléments du niveau de publication vont avoir des c.v. différents. Le système peut utiliser des options différentes pour chaque dimension. Un module est en développement pour permettre de spécifier en plus du c.v. global, des c.v. pour les éléments du niveau de sélection. Un autre module est également à l'étude pour permettre la stratification selon plusieurs variables. On pense également à permettre à l'utilisateur de définir lui-même ses bornes de stratification.

Au lieu de travailler avec un niveau de précision, le système pourra sous peu allouer un échantillon de taille déterminée. Dans ce cas l'allocation pourra se faire séparément pour chaque cellule ou sera faite sur l'ensemble des strates et des cellules.

Il n'est pas nécessaire de faire un processus de renouvellement dans la sélection de l'échantillon. Des modules sont en cours de développement pour faire de la sélection systématique ou à l'aide de fonctions Hash [7] pour accommoder les enquêtes faisant des analyses longitudinales. Un autre module est à l'étude pour mettre à jour le répertoire des unités échantillonnales en identifiant les naissances, les morts et les changements de classification.

Finalement, il existe un module intéractif utilisable sur micro-ordinateur permettant à l'usager de spécifier les options et les paramètres. Une fois complétées, les spécifications sont transmises à l'ordinateur principal qui effectue tous les calculs.

4. BIBLIOGRAPHIE

[1] Colledge, M. et Lussier, R. (1987), "Une méthode d'enquête généralisée pour les enquêtes économiques", rapport interne de Statistique Canada.

[2] Deming, W. E. et Stephan, F. F. (1940),"On a Least Squares Adjustment of Sampled Frequency Table When the Expected Marginal Totals Are Known", Annals of Mathematical Statistics 11:427-444

[3] Hidiroglou, M. A. (1986), "The Construction of a Self-Representating Stratum of Large Units in Survey Design", The American Statistician, 40, 27-31.

[4] Bankier, M. (1981), "The Square Root Allocation", rapport interne de Statistique Canada

[5] Cochran, W. G. (1977), "Sampling Techniques", 3ème édition, John Wiley and sons, New-York,pp.129-130

[6] Dalenius, T., et Hodges, J.L.,Jr. (1959), "Minimum Variance Stratification.", Journal of the American Statistical Association, 54, pp. 88-101.

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010180962

003

[7] Sunter, Alan B. (1986), "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique", Journal of Official Statistics, Vol. 2. No 2, pp. 161-168.

ANNEXE A

CALCUL DU NOMBRE DE PANELS

Exemple 1: cas où les trois contraintes sont respectées.

Supposons que $N_h = 300$, $n_h = 30$, $T_{IN} = 24$ mois et $T_{EX} = 12$ mois

alors $P_{EX} = ENT((24*0.9/0.1) + 0.5) = 216$ panels $\geq T_{EX}$

et $P_{IN} = 24$ panels

Exemple 2: cas où seulement deux contraintes sont respectées.

Supposons que $N_h = 60$, $n_h = 50$, $T_{IN} = 24$ mois et $T_{EX} = 12$ mois

avec les trois contraintes on a:

$P_{EX} = ENT((24*0.17/0.83) + 0.5) = 5 < T_{EX}$

il faut donc omettre la troisième contrainte et

$P_{IN} = ENT((12*0.83/0.17) + 0.5) = 60$ panels

$P_{EX} = 12$ panels