Techniques d'enquête, juin 1978
Vol. 4, nº 1
Statistique Canada, nº 12-001-X au catalogue


# Une comparaison des estimations de la variance des réponses corrélées obtenues dans les recensements de 1961, de 1971 et de 1976

**K.P. Krótki et C.J. Hill**[1]

Résumé

La variance totale d'un estimateur dans une enquête comprend la variance due à l'échantillonnage, la variance due aux réponses simples et la variance due aux réponses corrélées. Ce dernier composant reflète la partie de la variance totale causée par une influence commune sur un groupe de répondants. Dans le cas du recensement canadien, on a adopté l'auto-énumération comme méthode générale d'énumération pour le Recensement de 1971. Un facteur en faveur de l'introduction de cette méthode était l'évidence, dans le Recensement de 1961, que la variance due aux réponses corrélées apportait une contribution importante à la variance totale des estimations du recensement. Le présent article, basé sur une étude faite en utilisant l'interpénétration des intervieweurs, compare les variances dues aux réponses corrélées des recensements de 1961, 1971 et 1976. Les résultats démontrent que, bien que la méthode d'auto-énumération adoptée pour le Recensement de 1971 n'ait pas enlevé complètement la variance due aux réponses corrélées, cette approche a considérablement réduit l'importance de cette composante de la variance pour presque toutes les caractéristiques examinées.

**Mots-clés :** Recensement; autodénombrement; variance due aux réponses corrélées.

## 1 Introduction

La variance totale d'une estimation d'enquête incorpore à la fois la variance d'échantillonnage et la variance de réponse. Pour le voir de manière formelle, nous pouvons décomposer l'erreur quadratique moyenne totale en variance totale et en biais. À son tour, la variance totale peut être décomposée en variance d'échantillonnage et en variance de réponse. Enfin, la variance de réponse peut être exprimée comme la somme de la variance de réponse simple et de la variance de réponse corrélée. La première composante mesure la variabilité d'un essai à l'autre de la réponse d'un répondant donné. C'est la partie de la variance de réponse qui est produite par les tendances des répondants individuels à commettre des erreurs de réponse indépendamment des autres répondants. La variance de réponse corrélée, quant à elle, reflète la part de la variance totale attribuable à une influence commune sur un groupe de répondants.

1. K.P. Krótki et C.J. Hill, Division des méthodes d'enquête – recensement et enquêtes-ménages, Statistique Canada.

La décomposition de la variance totale est exposée en détail dans l'ouvrage de référence de Hansen, Hurwitz et Bershad [9]. C'est ce que l'on appelle souvent le modèle de recensement ([1], [2], [3]). La décomposition de l'erreur quadratique moyenne peut être exprimée algébriquement comme suit :

$$\text{E.Q.M.}(\bar{x}) = \frac{1}{n}\sigma_r^2 + \frac{n-1}{n}\rho_r\sigma_r^2 + \frac{N-n}{N-1}\frac{\sigma_s^2}{n} + \frac{2(n-1)}{n}\sigma_{rs} + B^2$$

où $\bar{x}$ est la variable moyenne X, n est la taille de l'échantillon, N est la taille de la population, $\sigma_s^2$ est la variance d'échantillonnage, $\sigma_r^2$ est la variance de la réponse simple, $\rho_r\sigma_r^2$ est la variance de la réponse corrélée, $\sigma_{rs}$ est la covariance échantillon-réponse et B est le biais. Des résultats similaires peuvent être obtenus par une approche de modèle linéaire (voir [4]) dans laquelle la valeur observée est exprimée comme une combinaison linéaire de la valeur réelle, d'un terme reflétant le biais constant de chaque agent recenseur et d'une composante aléatoire indépendante. Un examen général de la variance totale peut être trouvé dans [12].

Il a été démontré empiriquement que dans le cas du recensement par interview, la variance corrélée des réponses est la composante dominante de la variance totale des réponses ([5], p. 1035). En fait, dans le cas d'un recensement, pour des données recueillies auprès de 100 % de la population, il n'y a pas de variance d'échantillonnage et la variance de la réponse corrélée devient la composante dominante de la variance totale. Cette composante est attribuable à un effet sur les répondants qui fait en sorte que les répondants dans la même zone d'affectation des enquêteurs commettent des erreurs similaires. Il est postulé que cette homogénéité des erreurs est due à l'effet des intervieweurs sur les répondants.

Dans le recensement canadien, l'autodénombrement été adopté comme méthode standard de dénombrement dans le Recensement de 1971. L'un des facteurs en faveur de l'introduction de cette méthode était la preuve, tirée du Recensement de 1961, que la variance des réponses corrélées contribuait de manière importante à la variance totale des estimations du recensement [5]. En principe, l'autodénombrement éliminerait cette composante de la variance. En pratique, même en utilisant cette méthode, il reste un

certain contact entre l'agent recenseur et le répondant. Dans ce cas, il faut s'attendre à ce que l'autodénombrement réduise plutôt qu'il ne supprime complètement la variance des réponses corrélées. Le présent article démontre que c'est effectivement le cas en offrant une comparaison des estimations obtenues pour le Recensement de 1961 et de celles obtenues en 1971 et en 1976.

## 2 Le modèle et le plan expérimental

Alors que la variance d'échantillonnage peut être calculée directement à partir des éléments de l'échantillon, le calcul de la variance des réponses corrélées suppose un plan expérimental pour fournir de multiples observations sur chaque réponse. Le rééchantillonnage de l'enquête est un moyen d'y parvenir. Cependant, des problèmes de contamination (manque d'indépendance entre la première et la seconde interview) rendent cette approche critiquable. Le plan utilisé dans les recensements de 1961, de 1971 et de 1976 pour mesurer la variance totale et la variance de réponse corrélée est fondé sur l'interpénétration des intervieweurs et des répondants. L'utilisation de l'interpénétration de cette manière est attribuable à Mahalanobis (1949). La théorie et les plans expérimentaux pour l'étude des erreurs de réponse et les méthodes particulières appliquées sont traités en détail ailleurs ([5], [8]). Un bref résumé est fourni ici.

Pour les recensements de 1961, de 1971 et de 1976, un modèle de conception d'interpénétration a été établi en sélectionnant des paires de secteurs de dénombrement (SD) adjacents dans lesquels les deux agents recenseurs partageaient le même superviseur, puis en divisant les SD au hasard en deux parties à peu près égales. Chaque agent recenseur de la paire est ensuite responsable de la réalisation de l'affectation en binôme. En 1961, l'étude a été réalisée dans 96 SD dans la région de Cornwall, alors qu'en 1971 et en 1976, un échantillon aléatoire stratifié de 376 paires de SD provenant de partout au Canada a été utilisé. Seul un petit nombre de SD de recensement par interview dans le Nord, de SD collectifs et de logements collectifs (établissements comme les hôpitaux, les prisons, etc.) au sein des SD ont été exclus du champ d'enquête de la présente étude.

Une autre caractéristique de l'étude de 1961 était l'utilisation d'un nouveau dénombrement. Un nouveau dénombrement permet l'estimation de paramètres supplémentaires, y compris l'estimation de la variance de la réponse simple, mais présente des coûts considérablement accrus. Un inconvénient d'un nouveau dénombrement, cependant, est qu'il est impossible (dans une étude sur l'erreur de réponse) de garantir l'indépendance entre les mesures répétées en raison de l'effet du rappel. Pour cette raison, et compte tenu de la preuve que, dans le cadre du recensement par interview, la variance des réponses simples contribue moins à la variance totale que la variance des réponses corrélées ([5], p. 1035), les études de 1971 et de 1976 se sont limitées à l'interpénétration.

# 3 Formules

Une dérivation détaillée des formules pour la variance de réponse totale et corrélée est donnée dans [5] pour le cas où l'interpénétration et le rééchantillonnage sont appliqués. À supposer que certains facteurs soient négligeables, une dérivation un peu plus courte applicable à la conception interpénétrée de 1971 et de 1976 est présentée dans [3]. Les éléments fondamentaux de ce travail sont présentés ci-dessous.

La notation suivante sera utilisée dans les formules.

$P$ est le nombre de SD au Canada.

$k$ est l'indice désignant le SD.

$h$ est l'indice désignant le ménage au sein d'un SD.

$k(i)$ désigne la $i^e$ moitié du SD $k$.

$N_k$, $n_k$ représente le nombre total de ménages dans la population et dans l'échantillon respectivement pour le $k^e$ SD.

$n_{ki}$ est le nombre de ménages dans l'échantillon pour la $i^e$ moitié du $k^e$ SD.

$X_{kh}$ désigne la caractéristique observée pour le ménage $h$ dans le SD $k$.

$\sigma_k^2$ est une mesure de la variance de réponse.

$\rho_k\sigma_k^2$ est une mesure de la variance de réponse corrélée.

$S_{xk}^2$ est la variance d'échantillonnage.

L'estimation du chiffre de population pour une caractéristique de l'échantillon du recensement peut être écrite comme suit :

$$\hat{X} = \sum_{k=1}^{P} \frac{N_k}{n_k} \sum_{h \varepsilon S_k} x_{kh}$$

où $S_k$ est l'ensemble des ménages de l'échantillon dans le SD k,

et sa variance totale est alors donnée par :

$$V(\hat{X}) = \sum_{k=1}^{P} [N_k^2 \frac{\sigma_k^2}{n_k} [1+(n_k-1)\rho_k] + (1-\frac{n_k}{N_k}) \frac{S_{xk}^2}{n_k}].$$

Deux estimateurs peuvent être obtenus à partir du plan expérimental. La première est la variance entre agents recenseurs, $C_k$, qui est une mesure de la variance pour le SD k.

$$C_k = \frac{1}{2} [\bar{x}_{k(1)} - \bar{x}_{k(2)}]^2$$

et

$$E(C_k) = \frac{2\sigma_k^2}{n_k} [(1+(\frac{n_k}{2} - 1)\rho_k] + \frac{2S_{xk}^2}{n_k} .$$

La deuxième est $D_k$, une variance au sein de l'agent recenseur pour le SD k.

$$D_k = \frac{\sum_{i=1}^{2} \sum_{h \varepsilon S_{ki}} (x_{kh}-\bar{x}_{k(i)})^2}{n_{k1} + n_{k2} - 2}$$

et

$$E(D_k) = \frac{2}{n_k} [\sigma_k^2 (1-\rho_k) + S_{xk}^2].$$

Enfin, on peut montrer que, avec un certain biais,

$E(C_k-D_k) = \rho_k\sigma_k^2$ une mesure de la variance des réponses corrélées pour le SD k.

Afin de rendre les résultats de 1971 et de 1976 comparables à ceux de 1961, nous avons calculé une moyenne pondérée de $\rho_k \sigma_k^2$ sur tous les SD du projet. Les poids reflétaient la taille du SD pour ce qui est du nombre de personnes.

Il convient de souligner que, pour plusieurs raisons, les résultats exposés dans le présent document ne correspondent pas aux résultats de la variance totale des publications de 1971 et de 1976. L'une des raisons est que les publications contiennent des estimations de la variance totale qui diffèrent des estimations de la variance des réponses corrélées. Deuxièmement, les publications présentent des résultats exagérés au niveau du Canada. Les formules utilisées pour calculer ces résultats de variance totale aux échelons national et infranational peuvent être trouvées dans [3]. Enfin, aux fins de publication, pour tout SD dans lequel $C_k - D_k$ était négative, cette quantité a été mise à zéro. En regardant les résultats, pour plusieurs caractéristiques, il y a un nombre suffisant de valeurs négatives de $C_k - D_k$ pour que la valeur moyenne globale de l'estimation de la variance de la réponse corrélée soit négative. Une analyse de ce problème peut être trouvée dans [13] dans lequel on donne une explication statistique pour les estimations négatives de la variance de la réponse corrélée.

# 4 Limites de la comparaison de la variance des réponses corrélées de 1961, de 1971 et de 1976

Les principales limites de la comparaison des estimations de la variance des réponses corrélées de 1961, de 1971 et de 1976 sont les différences de portée et de conception des projets. Comme nous l'avons indiqué ci-dessus, l'étude de 1961 comprenait à la fois le rééchantillonnage et l'interpénétration dans sa conception, mais elle n'a été appliquée que dans la région de Cornwall, tandis que les études de 1971 et de 1976, qui ne comprenaient que l'interpénétration, ont été appliquées à un échantillon pancanadien. Ces différences reflètent des objets différents pour les études. Dans l'étude de 1961, on a tenté de réaliser un examen détaillé des facteurs contribuant à la variance totale. L'étude de 1971 a tenu pour acquis le résultat de 1961 selon lequel la variance de réponses corrélées domine la variance de réponses simples. Ainsi, plutôt que de fournir de plus amples précisions sur les composantes de la

variance totale, l'étude de 1971 visait à donner une mesure fiable de la variance à l'échelle du Canada pour accompagner les estimations publiées du recensement.

La différence dans le domaine d'étude, c'est-à-dire Cornwall par opposition au Canada, peut ne pas avoir d'importance pour la majorité des variables. Il n'y a aucune raison de croire que les agents recenseurs de Cornwall étaient plus ou moins enclins à influencer les réponses que partout ailleurs au Canada. Toutefois, sous certains aspects cruciaux, Cornwall est atypique par rapport à l'ensemble du Canada. Il s'agit d'une zone frontalière entre les régions francophones et anglophones, qui contient une proportion importante des deux groupes linguistiques. Il est clair que si 100 % des personnes d'un secteur de dénombrement ont les mêmes caractéristiques, la probabilité qu'il y ait des erreurs de réponse pour une caractéristique particulière est très faible, alors que dans une zone hétérogène, il faut s'attendre à des erreurs, ne serait-ce que parce qu'il existe une proportion de personnes pour lesquelles la « vraie » réponse est ambiguë. Par conséquent, pour les estimations de la variance du français ou de l'anglais, la comparaison se fait entre une zone où les erreurs de réponse corrélées sont potentiellement élevées et l'ensemble du Canada qui comprend à la fois des zones de cette nature et d'autres zones où les erreurs de réponse seront faibles.

Il y a deux limitations supplémentaires à prendre en compte lors des comparaisons.

(1) Dans le Recensement de 1961, on a posé toutes les questions à l'ensemble de la population, alors qu'en 1971, certaines questions n'ont été posées qu'à un échantillon représentant le tiers de la population. Les comparaisons se font donc parfois entre 100 % de questions et à d'autres moments entre 100 % de questions et des questions types. Il n'est pas clair si cette différence est capitale ou non.

(2) La formulation des questions a souvent changé d'un recensement à l'autre, et dans certains cas, il y a même un léger changement dans le concept englobé par les questions.

# 5 Les résultats

Les résultats sont présentés dans deux tableaux. Le premier fournit les résultats de 1961, de 1971 et de 1976 pour les caractéristiques qui

étaient comprises dans [5]. Le deuxième tableau est une extension du tableau 1 dans la mesure où il donne les comparaisons de 1961, 1971 et 1976 pour les caractéristiques non publiées dans [5].

Deux estimations de la variance de réponses corrélées pour 1961 sont disponibles. La première est calculée comme $\delta_{21}\sigma_{r1}^2$, au moyen de la notation de [5]. En fait, l'estimation de cette quantité suppose plusieurs autres termes (voir les formules 39 et 41 dans [5]) qui sont censés être négligeables. De plus, l'estimation de cette version de la variance de réponse corrélée est basée à la fois sur les aspects interpénétrés et répétés de l'enquête de 1961. Cette estimation n'est donc pas directement comparable à celle utilisée en 1971 et en 1976. L'estimation de [5] qui est comparable à celle utilisée en 1971 et en 1976 est $\frac{1}{n}$ $(C_1 - F_1)$ où n est la taille moyenne de l'affectation des agents recenseurs, $C_1$ est la variance entre agents recenseurs pour la première enquête et $F_1$ est la variance au sein des agents recenseurs pour la première enquête ([5], p. 1033). Cette estimation est basée uniquement sur la partie interpénétrée de l'étude. C'est cette estimation qui est utilisée pour comparer les résultats de 1961, de 1971 et de 1976.

Si l'on compare les résultats de 1961 avec ceux de 1971 et de 1976 dans le tableau 1, il est évident que l'estimation de la variance de réponses corrélées est en général considérablement réduite. Toutes les caractéristiques ayant des valeurs positives en 1961 donnent des estimations plus faibles en 1971 et en 1976, y compris deux qui sont négatives. La seule estimation qui n'est pas inférieure en 1971 est celle de l'âge = 5 qui, de toute façon, est à peine différente de zéro. La différence la plus importante se produit pour le groupe ethnique = français.

Le tableau 2 qui offre une comparaison des estimations non publiées de 1961 avec les chiffres de 1971 et de 1976 donne essentiellement les mêmes résultats. La caractéristique « langue officielle parlée = français seulement » donne une estimation plus élevée, mais tous les autres résultats sont inférieurs ou négligeables. La principale réserve dans l'interprétation de ces estimations est la présence de résultats négatifs qui donnent des valeurs numériquement importantes.

Les explications possibles de ces résultats sont que (1) des erreurs sont introduites qui réduisent en fait la variance totale et (2) la variance

des réponses corrélées est élevée, en particulier pour les variables qui étaient des variables d'échantillon en 1971 et en 1976. La variance de la réponse corrélée peut être encore gonflée par le fait que ces variables sont groupées et ne s'appliquent qu'aux personnes de plus de 15 ans.

Dans certains cas (p. ex. pour la langue maternelle = anglais), une diminution de la variance de la réponse corrélée peut être observée sur les trois moments temporels. Le calcul des variances pour un plus grand nombre de caractéristiques pour lesquelles les résultats ne sont pas négatifs pourrait permettre de mieux comprendre la prévalence de cette situation. Il est également intéressant de mentionner que pour les trois recensements, les valeurs pour la langue maternelle = anglais sont plus importantes que celles pour la langue maternelle = français. On ne peut guère se prononcer sur les implications de ce résultat tant qu'on n'en saura pas plus sur les langues des intervieweurs et des répondants.

## 6 Conclusion

La comparaison des estimations de la variance des réponses corrélées de 1961, de 1971 et de 1976 donne des preuves empiriques d'une réduction de la variance entre 1961 et 1971. Cette réduction peut vraisemblablement être attribuée au passage du recensement par interview à l'autodénombrement pour ce qui est de la collecte des données. La réduction de la variance semble persister jusqu'au Recensement de 1976. En fait, dans certains cas, la variance est même réduite encore plus.

Ces résultats sont toutefois nuancés par un certain nombre de considérations concernant la variance et la précision des estimateurs et les problèmes de préparation des résultats obtenus dans des circonstances différentes. Les travaux de recherche concernant la variance et la précision des estimateurs sont maintenant menés sur deux fronts. Premièrement, une nouvelle méthode de calcul de la variance des réponses [6] est à l'étude avec les données des recensements de 1971 et de 1976.

Certains résultats préliminaires sont déjà disponibles ([9], [11]). Deuxièmement, l'ancien estimateur fait l'objet d'une étude approfondie afin de mettre en lumière ses propriétés théoriques et empiriques.

# Bibliographie

[1] Bailar, B. et Dalenius, T. (1969), "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", *Sankhya* B 31 (parts 3 & 4): 341-60, décembre.

[2] Bailey, L., Moore, T.F. et Bailar, B. (1978), "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample", *Journal of the American Statistical Association*, 73(361): 16-23, mars.

[3] Brackstone, G.J. et Hill, C.J. (1976), "The Estimation of Total Variance in the 1976 Census", *Techniques d'enquête*, 2(2): 195-208, décembre.

[4] Dodds, D.J. et Smith, T.M.F. (1973), "Estimation of Correlated Response Variance Under a Linear Additive Model", présenté à la conférence de l'IASS, août.

[5] Fellegi, I.P. (1964), "Response Variance and its Estimation", *Journal of the American Statistical Association*, 59(308): 1016-41, décembre.

[6] Fellegi, I.P. (1974), "An Improved Method of Estimating the Correlated Response Variance", *Journal of the American Statistical Association*, 69(346): 496-501, juin.

[7] Hansen, M.H., Hurwitz, W.N. et Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the ISI, 38(2): 359-74.

[8] Hill, C.J. (1976), "1971 Census Evaluation Programme, 1971 Response Variance Project, Methodology Report and Results", rapport interne, Census Survey Methods Division, Statistique Canada.

[9] MacLeod, A.D. (1978), "Investigation of Fellegi's 'Improved' Method of Estimating Correlated Response Variance", rapport interne, Census Survey Methods Division, Statistique Canada, avril.

[10] Mahalanobis, P.C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute", *Journal of the Royal Statistical Society*, 109: 325-70.

[11] Maurice, C., (1978), "Un nouvel estimateur de la variance corrélée de réponse", rapport interne, Census Survey Methods Division, Statistique Canada, août 1977, révisé en avril.

[12] Nisselson, H. et Bailar, B.A. (1976), "Measurement, Analysis and Reporting of Non-sampling Errors in Surveys", Proceedings of the 9th International Biometric Conference, papier invité, Vol. 2: 301-22, Boston, 22 au 27 août.

[13] United States Bureau of the Census (1968), "Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Interviewers and Crew Leaders", Series ER60, no. 7, U.S. Govt. Printing Office, Washington, D.C.

**Tableau 1**
**Une comparaison des estimations de la variance des réponses corrélées de 1961, de 1971 et de 1976 pour les caractéristiques publiées dans Fellegi (1964)**

| | Estimation pour 1961 × $10^4$ comparable à 1971 | Estimation pour 1971 × $10^4$ | Estimation pour 1976 × $10^4$ (4) |
|---|---|---|---|
| Sexe : homme | 13,0 | 0,2 | -3,4 |
| Âge : 5 ans | -0,9 | 0,4 | -0,6 |
| Groupe ethnique : français | 1 688,0 | 151,1 | - |
| Niveau d'études le plus élevé : secondaire, 5e année ou université (2) | 23,3 | 9,8 | - |
| Personnes à la recherche d'un emploi la semaine dernière | 40,8 | 5,6 | - |
| Personnes travaillant habituellement 40 heures par semaine (3) | 83,1 | 48,9 | - |
| Secteur d'activité : fabrication | 6,7 | -27,0 | - |
| Secteur d'activité : commerce | 18,5 | -3,4 | - |

Notes :

(1) L'âge, le sexe et le groupe ethnique étaient des variables provenant de 100 % de la population dans les trois recensements. Les autres variables provenaient de 100 % de la population en 1961, mais elles étaient tirées d'un échantillon en 1971 et en 1976.

(2) Libellé de 1961. En 1971 et en 1976, l'estimation concerne la 12e ou la 13e année ou l'université.

(3) Libellé de 1961. En 1971, l'estimation concerne les personnes qui travaillent habituellement de 40 à 44 heures par semaine.

(4) Les entrées marquées d'un « - » ne sont pas disponibles pour 1976.

**Tableau 2**
**Une comparaison des estimations de la variance des réponses corrélées de 1961, de 1971 et de 1976 pour les caractéristiques non publiées dans Fellegi (1964)**

| Caractéristique | Estimation pour 1961 $\times 10^4$ comparable à 1971 | Estimation pour 1971 $\times 10^4$ | Estimation pour 1976 $\times 10^4$ (1) |
|---|---|---|---|
| Langue officielle, anglais parlé uniquement | 129,2 | 112,5 | - |
| Langue officielle, français parlé uniquement | 193,8 | 681,5 | - |
| Langue maternelle anglaise | 228,2 | 12,0 | 9,8 |
| Langue maternelle française | 97,2 | -0,7 | 2,2 |
| ÉDUCATION (niveau de scolarité le plus élevé) | | | |
| École primaire seulement | 93,6 | -196,0 | - |
| École secondaire (première ou deuxième année) | 77,0 | -52,9 | - |
| École secondaire (troisième ou quatrième année) | 6,1 | -69,3 | - |
| ÂGE | | | |
| 4 ans | -0,6 | -0,4 | -0,2 |
| 5 ans | -0,9 | 0,4 | -0,6 |
| 6 ans | 0,8 | -0,3 | -0,7 |
| 64 ans | 0,2 | -0,1 | -0,7 |
| 65 ans | 0,5 | -0,2 | -0,6 |
| 66 ans | -0,5 | 0,2 | -0,5 |
| AUTRES CARACTÉRISTIQUES | | | |
| Relation avec le chef du ménage = fils | 16,1 | 7,4 | 0,7 |
| État matrimonial = marié(e) | 1,6 | 0,8 | -1,4 |

(1) Les entrées marquées d'un « - » ne sont pas disponibles.