N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

UNE MÉTHODE AMÉLIORÉE D'ESTIMATION DES COMPOSANTES DE LA VARIANCE

G. Hussain Choudhry, Gary Shapiro, Varma Nadimpalli et Joseph Croos[1]

RÉSUMÉ

Dans les plans d'échantillonnage à plusieurs degrés, les composantes de la variance facilitent la répartition de l'échantillon aux diverses étapes de l'échantillonnage. Le présent article discute d'un modèle de régression pour estimer les composantes de la variance pour un plan d'échantillonnage à deux degrés. Nous réalisons une étude de simulations pour comparer le biais et la racine carrée de l'erreur quadratique moyenne des composantes estimées de la variance obtenues selon la méthode classique et selon la méthode de régression proposée, et nous concluons que la méthode classique donne effectivement de meilleurs résultats. La méthode classique consiste à estimer la composante de variance à l'intérieur des unités d'échantillonnage de premier degré, puis à la soustraire de la variance totale afin d'obtenir la composante de variance entre les unités d'échantillonnage de premier degré, ce qui produit parfois des estimations négatives de cette composante. La méthode de régression peut aussi produire des estimations négatives de cette composante. Dans les deux cas, les valeurs négatives des composantes de la variance estimées sont ramenées à zéro. Ou bien, pour la méthode de régression, on peut estimer le modèle en imposant la contrainte que les paramètres du modèle ne soient pas négatifs.

MOTS CLÉS : Modèle de coût, nombre optimal d'UPE, plan d'échantillonnage à deux degrés, variance entre les UPE.

1. INTRODUCTION

Dans les enquêtes à plan de sondage à plusieurs degrés, il faut souvent estimer les composantes de la variance en plus d'estimer simplement la variance sur l'ensemble des degrés d'échantillonnage. Cette nécessité se manifeste surtout au moment où il faut décider du nombre d'unités d'échantillonnage de premier degré et de degrés ultérieurs qu'il convient de sélectionner. Le moyen habituel d'estimer la composante de premier degré de la variance consiste à estimer la variance totale et la variance à l'intérieur des unités de premier degré indépendamment, puis à calculer la variance entre les unités d'échantillonnage de premier degré par soustraction. Cette méthode donne habituellement des estimations instables des composantes de la variance et des estimations négatives assez fréquentes de la composante de la variance entre les unités d'échantillonnage de premier degré. Nous examinons dans le présent article une méthode de modélisation par régression linéaire pour obtenir les composantes de la variance. Nous pensions que cette approche produirait des estimations plus stables de la variance. Après la présentation de la communication au Symposium 2003, Wayne Fuller a mis cette possibilité en doute, puisque la méthode est fondée sur l'utilisation d'un sous-ensemble des données complètes et qu'elle ne permet pas d'introduire des ensembles plus grands de données. Suite aux questions soulevées par M. Fuller, nous avons comparé le biais et la racine carrée de l'erreur quadratique moyenne des estimations des composantes de la variance en nous appuyant sur une simulation à grande échelle. Cette simulation montre que la méthode classique est supérieure à la méthode de régression proposée.

[1] G. Hussain Choudhry, Gary Shapiro, Varma Nadimpalli et Joseph Croos, Westat, 1650 Research Boulevard, Rockville, Maryland 20850, USA.

2. DESCRIPTION DES MÉTHODES D'ESTIMATION DES COMPOSANTES DE LA VARIANCE

L'estimation du total pour un échantillon stratifié à deux degrés est donnée par

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{hij}$$

où Y_{hij} est la valeur de Y pour l'unité d'échantillonnage de second degré j sélectionnée à partir de l'unité d'échantillonnage de premier degré i à partir de la strate h et W_{hij} est le poids d'échantillonnage correspondant. La variance totale de l'estimation $\hat{Y}$ peut s'écrire sous la forme :

$$V = (1 - f_1)\frac{S_1^2}{n} + (1 - f_2)\frac{S_2^2}{nm}$$

où

$(1 - f_1)\frac{S_1^2}{n}$ = variance entre les UPE, et $(1 - f_2)\frac{S_2^2}{nm}$ = variance à l'intérieur des UPE;

n = nombre moyen d'UPE échantillonnées par strate;

m = nombre moyen de ménages échantillonnés par UPE;

f_1 = fraction d'échantillonnage au premier degré;

f_2 = fraction d'échantillonnage au deuxième degré.

Selon la méthode classique, on estime la variance totale, ainsi que la composante de variance à l'intérieur des UPE pour un plan d'échantillonnage stratifié à deux degrés à partir de l'échantillon. Puis, on estime la variance entre les UPE à partir de la variance totale estimée par soustraction. Le problème de l'estimation de la variance entre les UPE obtenue par soustraction est qu'elle est habituellement instable et peut donner des estimations négatives, surtout si la valeur de cette composante est petite.

Baskin (1992, 1993) a utilisé l'échantillonnage hiérarchique de Bayes et de Gibbs pour estimer certaines composantes de la variance pour l'indice des prix à la consommation (IPC) des États-Unis. Baskin et Johnson (1995) et Shoemaker (2001 et 2002) ont estimé les composantes de la variance du loyer et des biens et services compris dans l'IPC américain par la méthode du maximum de vraisemblance restreint (REML) et les estimateurs de type ANOVA habituels. Ces auteurs ont constaté que les méthodes REML produisent des estimations non négatives des composantes de la variance, tandis que les estimateurs ANOVA produisent des estimations négatives de certaines composantes. Leaver et Larson (à paraître en 2003) ont estimé les composantes de la variance pour l'IPC par la méthode du REML pondéré et ont comparé les résultats à ceux obtenus au moyen de deux modèles à effets aléatoires distincts. Choudhry et coll. (1985) se sont assurés d'obtenir des estimations non négatives des composantes de la variance pour l'Enquête sur la population active du Canada, dont le plan d'échantillonnage est à plusieurs degrés, en estimant ces composantes au moyen de données du recensement.

Dans le présent article, nous proposons un modèle de régression linéaire pour estimer les composantes de la variance pour un plan d'échantillonnage stratifié à deux degrés. Bien que la méthode de régression proposée puisse produire des estimations négatives des composantes de la variance, elle peut être modifiée par programmation non linéaire de sorte que les estimations soient contraintes d'être non négatives. Le ratio S_1^2 / S_2^2 ne peut être calculé directement. Nous utilisons un modèle des composantes de la variance pour estimer les deux composantes de la variance, puis nous estimons le ratio à partir des composantes de la variance ainsi estimées.

Le modèle des composantes de la variance pour l'estimation du total est :

$$V\left(\hat{Y}\right)=\left(1-f_1\right)\frac{S_1^2}{n}+\left(1-f_2\right)\frac{S_2^2}{nm}+\varepsilon,$$

et nous nous intéressons à l'estimation des paramètres S_1^2 et S_2^2. Nous utilisons l'échantillon complet pour créer 55 mini échantillons, tels que décrits à la section 3.3, pour estimer les paramètres S_1^2 et S_2^2 dans le modèle de régression susmentionné. Nous fixons $X_1=\left(1-f_1\right)/n$ et $X_2=\left(1-f_2\right)/nm$, et faisons la régression de $V\left(\hat{Y}_1\right)$ sur X_1 et X_2 pour estimer les paramètres S_1^2 et S_2^2 dans le modèle de régression. Les composantes de la variance peuvent être construites d'après les valeurs estimées des paramètres S_1^2 et S_2^2.

3. MÉTHODE DE SIMULATION

Nous avons utilisé quatre populations finies synthétiques pour sélectionner des échantillons répétés conformément à un plan d'échantillonnage stratifié à deux degrés avec échantillonnage aléatoire simple à chaque degré. Nous estimons les composantes de la variance pour chaque échantillon selon les méthodes classiques et de régression, et comparons les deux méthodes en ce qui concerne le biais et la racine carrée de l'erreur quadratique moyenne.

3.1 Populations finies

Nous avons créé quatre populations synthétiques à partir desquelles nous avons tiré des échantillons répétés. Chaque population comprend 10 strates, 40 UPE dans chaque strate et 100 unités de second degré (comme les ménages) dans chaque UPE, ce qui donne en tout 40 000 unités d'échantillonnage de second degré. Le tableau 1 donne les moyennes pour chaque strate dans les quatre populations.

Tableau 1. Moyenne par strate

Strate	1	2	3	4	5	6	7	8	9	10
Moyenne	820	860	900	940	980	1 020	1 060	1 100	1 140	1 180

Pour la première population, les UPE ont été sélectionnées à partir d'une loi normale d'écart-type 8,28. Donc, dans la strate 1, par exemple, 40 UPE ont été sélectionnées à partir de la loi $N\left(820, 68{,}56\right)$, où $68{,}56=\left(8{,}28\right)^2$. Puisque l'ensemble des 40 UPE a été sélectionné au hasard, la valeur moyenne pour les UPE de la strate 1 n'est pas exactement 820, quoique la valeur prévue de la moyenne soit 820.

Les quatre populations diffèrent en ce sens que les écarts-types diffèrent pour les UPE. Le tableau 2 donne les écarts-types pour les populations.

Tableau 2. Écart-type (prévu) entre les UPE selon la population

Population	1	2	3	4
Écart-type	8,28	13,10	18,57	23,89

Après avoir déterminé la population d'UPE pour chacune des dix strates dans chacune des quatre populations, nous avons pu sélectionner les unités d'échantillonnage de second degré. Pour chaque UPE, nous avons sélectionné

100 unités d'échantillonnage de second degré à partir d'une loi normale de moyenne égale à la moyenne de l'UPE et d'écart-type égal à 100. (Nous avons utilisé le même écart-type égal à 100 pour toutes les UPE dans chaque population.)

Le tableau 3 donne la variance entre les UPE, exprimée en proportion du total de la variance, pour les populations obtenues.

Tableau 3. Variance entre les UPE, exprimée en proportion du total de la variance

Population	1	2	3	4
Variance entre les UPE en proportion du total	12,43 %	19,73 %	29,45 %	39,02 %

3.2 Échantillons de simulation

Nous avons tiré 2 000 échantillons simulés conformément à un plan d'échantillonnage stratifié à deux degrés avec échantillonnage aléatoire simple à chaque degré. Il convient de souligner que chaque échantillon simulé a été sélectionné à partir de quatre populations finies similaires, mais différentes. Les quatre populations finies ont été tirées de quatre superpopulations tel que décrit plus haut. Aux fins de notre étude de simulations, nous traitons ces populations finies comme étant fixes.

Pour chaque exécution des simulations, nous avons sélectionné 4 UPE par strate et 10 ménages par UPE, par échantillonnage aléatoire simple (EAS) à chaque degré d'échantillonnage. Donc, chaque échantillon simulé comprend 400 ménages — 10 ménages par UPE, 4 UPE par strate et 10 strates. Nous avons sélectionné indépendamment 2 000 échantillons simulés.

3.3 Mini échantillons pour la méthode de régression

Nous avons créé 55 mini échantillons à partir de chaque échantillon simulé et les avons utilisés pour l'estimation des composantes de la variance à l'intérieur des UPE et entre les UPE au moyen du modèle de régression décrit à la section 2. Soit P1, P2, P3 et P4, la notation générique des quatre UPE dans n'importe quelle strate de l'échantillon principal. Par exemple, lors d'une simulation donnée, l'UPE P1 dans une strate pourrait être « 02 », tandis que la même itération de l'UPE P1 dans une autre strate pourrait être « 06 ». Pour chaque strate, il existe 11 sous-ensembles possibles d'UPE contenant les quatre UPE échantillonnées, trois UPE échantillonnées sur quatre et deux UPE échantillonnées sur quatre. Pour la simulation, nous avons créé des sous-échantillons à partir de l'échantillon original de telle façon que le nombre d'UPE soit le même dans chaque strate. Pour chacun des 11 sous-ensembles d'UPE échantillonnées définis plus haut, nous avons créé cinq sous-ensembles de ménages, comme suit :

1. ensemble des 10 ménages échantillonnés dans chaque UPE;
2. moitié des 10 ménages échantillonnés choisis aléatoirement (cinq ménages par UPE);
3. autre moitié des 10 ménages échantillonnés (les 5 ménages non sélectionnés à l'étape 2 ci-dessus);
4. sept des 10 ménages échantillonnés sélectionnés aléatoirement par UPE;
5. les trois ménages restants non sélectionnés à l'étape 4 ci-dessus.

Donc, nous avons généré 55 mini échantillons à partir de chaque échantillon simulé, l'un de ces mini échantillons étant l'échantillon simulé réel.

La qualité de l'ajustement des modèles était très bonne pour les quatre populations. Pour la première, la valeur du R-carré ajusté variait de 0,64 à 0,95, avec une moyenne de 0,87. Pour la deuxième population, le R-carré ajusté variait de 0,67 à 0,96, pour la troisième, il variait de 0,69 à 0,96 et pour la quatrième, il variait de 0,70 à 0,96. Les valeurs moyennes du R-carré pour les deuxième et troisième populations étaient identiques à celle obtenue pour la première population (0,87) et la moyenne pour la quatrième population était de 0,88.

4. RÉSULTATS DES SIMULATIONS

Nous nous concentrons ici sur le ratio entre la variance au niveau des UPE et la variance totale. Nous examinons aussi les estimations de la variance à l'intérieur des UPE et de la variance entre les UPE. Toutes les méthodes produisent des estimations de la variance comparables et de bonne qualité à l'intérieur des UPE. Le profil de la variance entre les UPE est, dans l'ensemble, comparable à celui du ratio de la variance entre les UPE à la variance totale.

Nous avons examiné trois versions de la méthode classique. La plus simple consiste à utiliser directement l'estimation de l'échantillon, y compris les valeurs négatives de la variance entre les UPE. Particulièrement pour la population 1, pour laquelle la variance réelle entre les UPE est assez faible, le nombre de valeurs négatives est élevé, soit 709 des 2 000 échantillons. Le tableau 4 donne le nombre de valeurs négatives pour chaque population, pour l'estimation par régression non ajustée, ainsi que pour la méthode classique non ajustée. La deuxième version de la méthode classique consiste à remplacer les valeurs négatives par une valeur nulle, c'est-à-dire que le ratio estimé est de 0 %. La troisième version consiste à apporter une correction pour le biais en plus de remplacer les valeurs négatives par zéro. Cochran (1977, p. 278) et d'autres auteurs ont montré que s_1^2/n , l'estimation de la variance échantillonale entre les unités d'échantillonnage de premier degré, n'est pas une estimation sans biais de $V(\hat{Y})$. Donc, la troisième version comprend un ajustement pour ce fait[2]. La racine carrée de l'erreur quadratique moyenne est beaucoup plus élevée pour la première version que pour les deux autres. Celle calculée pour la troisième version est à peine inférieure à celle pour la deuxième version. Nous basons nos comparaisons ici sur la deuxième version.

Tableau 4. Fréquence des estimations non négatives de la variance entre les UPE

Population	1	2	3	4
Classique non ajustée	709	471	240	90
Régression non ajustée	812	636	385	217

Comme nous en avons discuté à la section 2, nous avions aussi deux versions de la méthode de modélisation par régression (l'une avec des valeurs négatives ramenées à zéro et l'autre avec des contraintes empêchant l'obtention de valeurs négatives). Les deux versions produisent des résultats quasiment identiques. Le tableau 5 montre que la racine de l'erreur quadratique moyenne (REQM) est nettement plus élevée pour la méthode de régression que pour la méthode classique pour les quatre populations. L'erreur-type domine la REQM. Le tableau 6 montre que l'erreur-type est également systématiquement plus élevée pour la méthode de régression. Cependant, le biais relatif est légèrement plus faible pour la méthode de régression pour la troisième population. Il convient de souligner que le biais est habituellement négatif pour la méthode classique, même si nous ramenons le ratio estimé à zéro lorsque l'estimation d'échantillon est une valeur négative.

Tableau 5. REQM pour les méthodes classiques et de régression

Population	1	2	3	4
Classique	0,146	0,160	0,171	0,166
Régression	0,220	0,209	0,219	0,221

[2] Les auteurs prévoient rédiger un article où ils discuteront du biais de la variance au niveau de la grappe ultime en tant qu'estimation de la variance totale qu'ils présenteront aux Joint Statistical Meetings de 2004.

Tableau 6. Biais et erreur-type pour les méthodes classique et de régression

	Bais				Erreur-type			
Population	1	2	3	4	1	2	3	4
Classique	0,093	-0,048	-0,075	-0,058	0,145	0,160	0,169	0,165
Régression	0,376	0,080	-0,037	-0,061	0,194	0,208	0,219	0,220

5. DISCUSSION DES RÉSULTATS ET CONCLUSION

La méthode classique d'estimation de la variance entre les unités d'échantillonnage de premier degré dans un plan de sondage à deux degrés n'est pas très satisfaisante, car elle produit habituellement des estimations instables qui peuvent souvent être négatives ou beaucoup trop grandes. Nous avons élaboré ce que nous pensions être une méthode par régression supérieure, mais avons constaté qu'elle donne lieu à des valeurs de l'erreur-type et de la racine de l'erreur quadratique moyenne nettement plus élevées que la méthode classique. Nous pensions que l'utilisation d'un certain nombre de mini échantillons (55 ont été utilisés pour la simulation) produirait des estimations plus stables des composantes de la variance. Malheureusement, puisque les mini échantillons sont des sous-ensembles de l'échantillon original, ils ne semblent ajouter aucune information et ne donnent pas lieu à de meilleures estimations. Par conséquent, nous recommandons de ne pas utiliser la nouvelle méthode.

Nous n'avons pas évalué les diverses méthodes proposées par le personnel du Bureau of Labor Statistics (Baskin, Shoemaker et others) que nous avons mentionnées brièvement à la section 2 de l'article et, par conséquent, nous ne pouvons offrir aucune opinion quant à leur capacité de produire de meilleures estimations des composantes de la variance que la méthode classique.

RÉFÉRENCES

Baskin, R.M. (1993). Estimation of variance components for the U.S. Consumer Price Index via Gibbs sampling, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 808-813.

Baskin, R.M. (1992). Hierarchical Bayes estimation of variance components for the U.S. Consumer Price Index, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 716-719.

Baskin, RM., et Johnson, W.H. (1995). Estimation of variance components for the U.S. Consumer Price Index, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 126-131.

Choudhry, G.H., Lee, H., et Drew, J.D. (1985). Cost-variance optimization for the Canadian Labor Force Survey, *Survey Methodology*, 11, pp. 33-50.

Cochran, W.G. (1977). *Sampling Techniques (3rd Ed.).* New York: John Wiley and Sons.

Leaver, S.G., et Larson, W.E. (in publication 2003). Estimating Components of Variance of Price Change from a Scanner-Based Sample, *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Shoemaker, O.J. (2002). Estimation and Analysis of Variance Components for the Revised CPS Housing Sample. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Shoemaker, O.J. (2001). Estimation of Variance Components for the U.S. Consumer Price Index: A Comparative Study, *Proceedings of the Survey Research Methods Section, American Statistical Association.*