

COMPARAISONS DE DIFFÉRENTS ESTIMATEURS DE VARIANCE À DEUX PHASES: ÉTUDE MONTE-CARLO BASÉE SUR L'ENQUÊTE DES MARCHANDISES AU DÉTAIL

Wisner Jocelyn, Marie Brodeur and Colin Babyak¹

RÉSUMÉ

Pour cette communication, on compare quatre estimateurs du total et leurs estimateurs de variance: l'estimateur classique à deux phases et l'estimateur repondéré de Kott (1994). Ces estimateurs sont ajustés ou non par le quotient. On rapporte les résultats de l'étude Monte-Carlo servant à comparer ces estimateurs et également le choix de l'estimateur retenu pour l'enquête.

MOTS CLÉS: Estimateurs à deux phases; étude Monte-Carlo; estimateur repondéré.

ABSTRACT

In this paper, we compare four estimators of the total and their respective variances. The first two estimators are the classical double-expansion estimator and the reweighted double expansion estimator as defined by Kott (1994). These estimators can be ratio-adjusted or not. A Monte-Carlo study was performed to compare the results from these estimators so that a choice could be made as to which method would be used for the survey.

KEY WORDS: Double-expansion estimator; Monte-Carlo study; reweighted estimator.

1. INTRODUCTION

L'échantillon de l'enquête des marchandises vendues au détail qui a été mis en place est un sous échantillon de l'enquête mensuelle du commerce de détail qui lui-même est un échantillon stratifié par type d'industrie, province ou territoire et revenu brut d'entreprise. Le plan de sondage de l'enquête des marchandises vendues au détail est donc un plan de sondage stratifié à deux phases. La première phase est constituée de l'échantillon mensuel du commerce de détail tandis qu'à la deuxième phase, on a re-stratifié cet échantillon de sorte que chaque strate de première phase ait été subdivisée selon le type industriel dominant et la province dominante et les ventes de première phase. Cette nouvelle enquête sert à produire des estimations des ventes pour différents types de marchandises.

Pour des raisons opérationnelles, les données seront publiées trimestriellement, cependant, l'échantillonnage et l'estimation sont conduits mensuellement. Au niveau de l'estimation, on a dû faire face au problème du choix d'un estimateur pour le total des ventes par types de marchandises ainsi que

l'estimateur de variance correspondant. Dans ce qui suit, on présente d'ailleurs l'étude monte-carlo qui a été réalisée afin d'évaluer les estimateurs considérés.

Plusieurs estimateurs du total et de la variance peuvent être envisagés pour un plan à deux phases. Comment choisir? Et selon quels critères? Pour répondre à ces deux questions, on envisage une étude Monte-Carlo. On procédera ainsi: Après avoir brièvement décrit les estimateurs considérés, on présente la mise-en-oeuvre de l'étude ainsi que les statistiques considérées. Dans la partie suivante, on discute des résultats de l'étude pour finalement conclure en indiquant l'estimateur retenu pour l'enquête.

2. ESTIMATEURS

On présente ci-dessous les estimateurs de totaux considérés. Les estimateurs de variance basés sur la méthode de linéarisation de Taylor sont donnés dans Binder et al (1997). À noter que la méthode de linéarisation utilisée est donnée dans Binder (1996).

¹Wisner Jocelyn, Marie Brodeur et Colin Babyak, Division des Méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

2.1 Notation

En empruntant la notation de Binder(1997), supposons qu'on désire estimer les ventes totales

$Y = \sum_{i=1}^N y_i$ où y_i est la valeur de la variable d'intérêt pour

l'unité i . Notons les valeurs de la population par y_i , $i = 1, \dots, N$. À la première phase d'échantillonnage, nous tirons un échantillon aléatoire simple de manière indépendante dans chacune des H strates de la population U . Soit a_{ih} une variable indicatrice valant

1 si l'unité i est dans la strate h et 0 autrement. On peut

donc définir $N_h = \sum_{i=1}^N a_{ih}$. Appelons z_i la variable

indicatrice valant 1 si l'unité i est dans l'échantillon de première phase et 0 autrement. La taille d'échantillon

pour la h ème strate est donc $n_h = \sum_{i=1}^N z_i a_{ih}$.

À nouveau nous tirons un échantillon aléatoire simple de taille m_g parmi les M_g unités sélectionnées à la première phase de manière indépendante dans chacune

des G strates de deuxième phase. Soit $a_{ig}^{(2)}$ respectivement la variable indicatrice valant 1 si l'unité i est dans la strate g de seconde phase et 0 autrement

et $z_i^{(2)}$ valant 1 si l'unité est dans l'échantillon de seconde phase et 0 autrement. Donc, on peut écrire

$$M_g = \sum_{i=1}^N z_i a_{ig}^{(2)} \text{ et } m_g = \sum_{i=1}^N z_i^{(2)} a_{ig}^{(2)}.$$

Dans Hidiroglou(1997), on arrive aux même formulations pour chacun des estimateurs et des estimateurs de variance en évoquant cette fois-ci un modèle de régression particulier.

2.2 L'estimateur doublement dilaté

C'est l'estimateur classique pour un plan à deux phases qu'on trouve par exemple dans Särndal (1992) mais adapté pour un plan stratifié à deux phases. Il a la forme suivante:

$$\hat{Y}_{DE} = \sum_{i=1}^N \sum_{h=1}^H \sum_{g=1}^G \frac{N_h M_g}{n_h m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i$$

Où y_{hgi} est la variable d'intérêt. Dans notre cas les ventes par marchandises vendues.

2.3 Estimateur par le quotient

Étant donné qu'on dispose d'information auxiliaire au niveau de la première phase, on se propose d'incorporer cette information sous la forme d'un

quotient d'où l'utilisation d'un estimateur par le quotient combiné ayant la forme suivante:

$$\hat{Y}_{DERAT} = \hat{R} \hat{X}^{(1)}$$

avec

$$\hat{R} = \frac{\sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} \sum_{g=1}^G \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i}{\sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} \sum_{g=1}^G \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} x_i} = \frac{\hat{Y}_{DE}}{\hat{X}_{DE}}$$

et

$$\hat{X}^{(1)} = \sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} z_i a_{ih} x_i$$

Où x_{hi} représente la variable auxiliaire. Pour l'enquête, il s'agit des ventes totales tandis que pour la simulation, c'est le RBE.

2.4 Estimateur repondéré (sans information auxiliaire)

Cet estimateur a été introduit par Kott et Stukel (1994). Son principal intérêt réside dans le fait que la version jackknife de la variance semble produire de meilleurs résultats que celle de l'estimateur doublement dilaté. Cet estimateur peut-être formellement défini ainsi:

$$\text{Soit } \hat{N}_g^{(1)} = \sum_{i=1}^N \sum_{h=1}^H z_i a_{ih} \frac{N_h}{n_h} a_{ig}^{(2)}$$

$$\text{et } \hat{N}_g^{(2)} = \sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} \frac{M_g}{m_g} z_i z_i^{(2)} a_{ih} a_{ig}^{(2)}$$

$$\text{ainsi que } \hat{Y}_g^{(2)} = \sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} \frac{M_g}{m_g} z_i z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i.$$

L'estimateur repondéré du total est donné par:

$$\hat{Y}_{RW} = \sum_{g=1}^G \hat{N}_g^{(1)} \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}}$$

2.5 Estimateur repondéré par le quotient

Tout comme l'estimateur par le quotient présenté auparavant était un ratio de deux estimateurs doublement dilaté, l'estimateur repondéré par le quotient est également un rapport de deux estimateurs repondéré. Il s'exprime ainsi:

$$\hat{Y}_{RWRAT} = \frac{\hat{Y}_{RW}}{\hat{X}_{RW}} \hat{X}^{(1)} = \hat{R}_{RW} \hat{X}^{(1)}$$

3. PLAN DE L'ÉTUDE MONTE-CARLO

Pour comparer les estimateurs décrits préalablement, on a décidé de les comparer en utilisant des données réelles provenant de l'enquête mensuelle du commerce de détail. On a donc choisi comme population de première phase 408 détaillants en produits domestiques provenant des provinces de l'Alberta, de la Saskatchewan et du Manitoba. Les fractions de sondage de première et de seconde phase sont celles qui sont utilisées respectivement pour l'enquête des marchandises vendues au détail et de l'enquête mensuelle du commerce de détail. La méthode de répartition utilisée est donnée dans Jocelyn et Brodeur (1996). La première phase est constituée de 6 strates. Les bornes de strates de première phase sont basées sur le Revenu Brut d'Entreprise (RBE). On a donc choisi 206 unités au niveau de la première phase. Ces unités ont été re-stratifiées selon les ventes et le type industriel de sorte qu'à la deuxième phase on avait 25 strates. Un échantillon de 137 unités devait être choisi au niveau de la deuxième phase. 10,000 échantillons de première phase ont donc été tirés et pour chacun d'eux on tirait un sous échantillon constituant la deuxième phase.

Les ventes par type de marchandises constituent la variable d'intérêt au niveau de l'enquête. Comme on ne disposait pas de cette variable au moment de la simulation, on a utilisé un modèle de régression linéaire pour simuler les ventes. Donc, la variable d'intérêt de la simulation est constituée du modèle ajusté auquel on additionne de manière aléatoire les erreurs. Quant à la variable auxiliaire utilisée c'est le RBE.

3.1 Statistiques de la simulation

Les statistiques suivantes vont permettre de comparer les performances des estimateurs considérés. Soit \hat{Y}_m une estimation des ventes pour la *mième* échantillon. $v_m(\hat{Y})$ est une estimation de la variance $V(\hat{Y}_m)$ pour la *mième* échantillon $m=1, \dots, M$. Le biais relatif $\hat{RB}(\hat{Y})$ est défini comme étant

$$\hat{RB}(\hat{Y}) = \frac{\left(\sum_{m=1}^M \frac{\hat{Y}_m}{M} - Y \right)}{Y}$$

L'erreur quadratique moyenne est donnée par

$$M\hat{SE}(\hat{Y}) = \sum_{m=1}^M \frac{(\hat{Y}_m - Y)^2}{M}$$

La variance Monte-carlo est définie comme étant

$$\hat{V}(\hat{Y}) = \sum_{m=1}^M \frac{\left(\hat{Y}_m - \frac{\sum_{m=1}^M \hat{Y}_m}{M} \right)^2}{M}$$

Le biais relatif de la variance devient donc

$$RB(v(\hat{Y})) = \frac{\left(\sum_{m=1}^M \frac{v(\hat{Y}_m)}{M} - V(\hat{Y}) \right)}{V(\hat{Y})}$$

De plus, afin de vérifier les performances conditionnelles des estimateurs, on a divisé les 10,000 échantillons en 10 groupes de 1000 selon la valeur estimée du RBE à la première phase. Cette façon de procéder décrite dans Royall(1981) revient en quelque sorte à conditionner selon une statistique ancillaire.

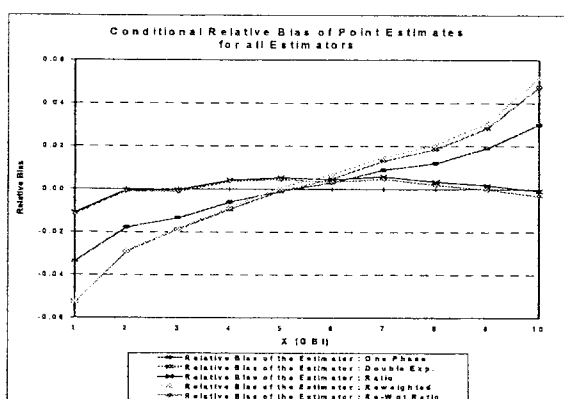
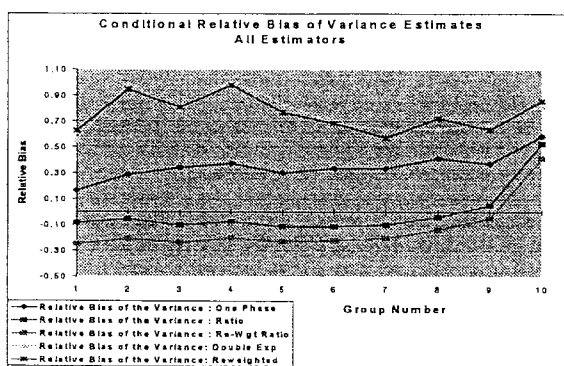
4. ANALYSE DES RÉSULTATS

Le tableau suivant présente les résultats non conditionnels pour les quatre estimateurs considérés. On y a ajouté l'estimateur à une phase pour fins de comparaison. On remarque d'abord que le biais est tout à fait négligeable pour tous les estimateurs. L'efficacité des estimations de l'erreur quadratique moyenne par rapport à la variance de première phase qui représente plus ou moins la perte due au fait qu'un échantillonnage à deux phases a été utilisé, ne montre pas de grandes différences d'un estimateur à l'autre. Le biais de la variance ainsi que la couverture ne permet pas d'avantage de discriminer d'un estimateur à l'autre. On remarque au passage que ces statistiques sont légèrement plus élevées pour l'estimateur par le quotient repondéré, rien cependant que l'erreur Monte-Carlo ne permet de justifier. Les résultats non conditionnels ne permettent donc pas de choisir l'un ou l'autre des estimateurs considérés puisque leurs performances sont trop voisines pour pouvoir discriminer.

Résultats non conditionnels

Estimateur	Biais relatif (%)	Eff de EQM (comparé à 1-phase)	Biais relatif de la variance (%)	Couverture à 95%
1-phase	0.0	1.000	0.3	95.0
Dbl Exp (DE)	0.0	0.680	0.4	95.1
Rat Exp (DERAT)	0.1	0.596	-1.9	95.0
Rwt (RW)	0.2	0.630	0.1	95.0
Rat Rwt (RWRAT)	0.0	0.584	-13.5	95.2

Résultats conditionnels



Les graphiques précédents montrent respectivement le biais relatif conditionnel pour les estimateurs de variance et les estimateurs ponctuels. On remarque que les estimateurs de variance des estimateurs par le quotient quoique proche du point de non biais sous-estiment légèrement la vraie variance. Cependant, ils sont plus stables autour du point de non-biais que les

estimateurs n'utilisant pas l'information auxiliaire qui semblent sur-estimer la vraie variance. On notera également que le biais conditionnel de la variance de l'estimateur par le quotient repondéré semble être systématiquement légèrement supérieure à celui de l'estimateur par le quotient. Si l'on se fie donc au biais relatif conditionnel des estimateurs de variance, on note donc un avantage certain pour les estimateurs utilisant de l'information auxiliaire.

Quant au biais relatif pour les estimateurs ponctuels, il faut surtout retenir le fait que les estimateurs par le quotient n'en exhibent pratiquement pas. Pour le reste, les observations notées pour les estimateurs de variance s'appliquent également ici. On a également considéré les couvertures conditionnelles à gauche et à droite et les mêmes conclusions s'imposent là également.

5. CONCLUSION

Les faibles biais qu'affichent les estimateurs de variance semblent indiquer que la méthode de linéarisation de Taylor fonctionne bien dans ce cas-ci. Les résultats conditionnels indiquent un biais nettement plus faible pour les estimateurs par le quotient. Les résultats conditionnels pour les estimateurs par le quotient ne permettent cependant pas de privilégier un des deux estimateurs. Nous recommandons toutefois l'estimateur par le quotient doublement dilaté à cause de la simplicité de la formule de variance.

RÉFÉRENCES

- Binder, D.A. (1996). Linearization Methods for single phase and two phase samples. A Cookbook Approach. *Survey Methodology*, **22**, 17 - 22.
- Binder, D.A., Brodeur, M., Hidioglou M.A., Jocelyn, W., Babyak, C. (1997). Variance estimation for two-phase stratified sampling. À paraître dans *Proceedings of the Section on Survey Research Methods, Annual American Statistical Association*.
- Hidioglou, M.A. (1997). Notes manuscrites sur les estimateurs du total et de la variance à deux phases.
- Jocelyn, W., Brodeur, M. (1996). Méthodes de répartition multivariées pour l'échantillonnage à deux phases: Application à l'enquête trimestrielle sur les marchandises. *Recueil des communications des XXVIII Journées de Statistiques de l'ASU*, 433-436.

Kott, S., P, Stukel, M., D (1994). Can the Jackknife be used with a two-phase sample? Submitted to *Survey Methodology*.

Royall, R., M., Cumberland, W., G. (1981). An Empirical Study of the Ratio Estimator and Estimators of its Variance. *Journal of the American Statistical Association*, 76, 66-88.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. Springer-Verlag.