Les simulations dans l'enseignement des sondages

Avec le logiciel GENESIS sous SAS et la bibliothèque "Sondages" sous R

Yves Aragon*, David Haziza** & Anne Ruiz-Gazen*

**GREMAQ, UMR CNRS 5604, Université des Sciences Sociales, 21 allée de Brienne, 31000 Toulouse et Laboratoire de Statistique et Probabilités, UMR CNRS 5583.*

*** Statistique Canada, 120, avenue Parkdale Ottawa (Ontario) K1A OT6.*

Email : aragon@cict.fr, dhaziza@rogers.com, ruiz@cict.fr

Résumé : L'objectif de cette présentation est de montrer l'intérêt d'utiliser des simulations dans un cours de théorie des sondages. En simulant des tirages d'échantillons dans une population et en calculant des indicateurs Monte Carlo pour les estimations obtenues à partir des échantillons tirés, on peut facilement comparer des méthodes d'échantillonnage ainsi que des méthodes d'estimation. Pour cela, on présente deux solutions : le logiciel GENESIS sous SAS et la bibliothèque "Sondages" sous R. L'utilisation de chacune de ces solutions sera illustrée sur des bases de données réelles.

Mots-clés : Enseignement, Simulations, Théorie des sondages,.

Abstract : The aim of this presentation is to show the use of simulations in a sampling theory course. By simulating sample drawings in a population and by calculating Monte Carlo indicators for the estimations obtained from the drawn samples, we can easily compare sampling methods and also estimation methods. For this purpose, we present two solutions : the software GENESIS for SAS and the library "Sondages" for R. The use of each of these solutions will be illustrated on real data sets.

Keywords : Teaching, Sampling theory, Simulations.

1 Introduction

D'un point de vue pédagogique, utiliser des exemples simulés dans un cours de théorie des sondages est intéressant à plusieurs égards.

Tout d'abord, dans la théorie des sondages probabilistes et à la différence des méthodes usuelles de statistique inférentielle, la population d'étude est considérée comme finie. Le statisticien s'intéresse au processus de tirage de l'échantillon ce qui déconcerte souvent les étudiants. Sur une population d'étude observable exhaustivement, on peut facilement tirer plusieurs échantillons, calculer des estimations à partir de chacun des échantillons et comparer les estimations aux valeurs exactes. Ceci précise de façon convaincante à quel niveau se situe l'aléatoire en théorie des sondages.

D'autre part, si le cours ne consiste qu'à énumérer des plans de sondage et des méthodes d'estimation de moyennes, totaux et variances, il devient rapidement fastidieux et trop abstrait. En utilisant des applications numériques basées sur des simulations, l'enseignant se dote d'un outil précieux pour illustrer les propriétés théoriques des méthodes d'échantillonnage et d'estimation et pour comparer différentes méthodes.

Depuis la version 8.2, SAS propose une procédure performante de tirage d'échantillons selon différents plans de sondages (proc *surveyselect*) ainsi que des procédures d'estimation (proc *surveymeans* et proc *surveyreg*). L'utilisation de ces procédures dans l'enseignement des sondages a été décrit dans Aragon et Ruiz-Gazen (2003). Toutefois, ces procédures présentent divers inconvénients. Tout d'abord, elles supposent une bonne connaissance du logiciel SAS. Ensuite, en ce qui concerne les procédures d'estimation, elles ont été conçues pour permettre d'améliorer les proc *means* et proc *reg*, en tenant compte de plans de sondages plus sophistiqués que le plan simple avec remise dans les estimations, mais ne sont pas adaptées à un cours de sondages. Ainsi, l'enseignant qui utilise ces procédures devra souvent avoir recours à des "astuces" pour obtenir les estimations qu'il attend (voir le problème de l'estimation correcte de la variance dans un plan à deux degrés dans Aragon et Ruiz-Gazen, 2003).

Dans le présent papier, nous envisageons deux solutions alternatives à l'utilisation des procédures SAS qui nous paraissent mieux adaptées à un cours sur la théorie des sondages. Nous proposons d'utiliser le logiciel GENESIS (Generalized Simulation System) développé à Statistique Canada (Haziza, 2003) sous SAS ou la bibliothèque "Sondages" (Aragon, 2003) sous R.

Le logiciel GENESIS est une interface ergonomique sous SAS, utilisable sans connaissance préalable de SAS, et qui permet de simuler des échantillons selon des plans de sondages divers, calculer des estimateurs de totaux ou moyennes par les valeurs dilatées, par le ratio et par régression ainsi que des indicateurs Monte Carlo et éditer des graphiques qui permettent de comparer les différentes méthodes d'estimation. A l'origine, GENESIS a été développé pour des études de simulations en présence de non-réponse. Il est donc aussi

particulièrement performant pour simuler de la non-réponse dans des données complètes et comparer les méthodes d'estimation dans ce cadre.

La bibliothèque "Sondages" sous R comporte plusieurs procédures. Elles permettent de tirer des échantillons suivant la plupart des plans proposés par la proc *surveyselect* de SAS et, pour chacun de ces plans, d'estimer des totaux et moyennes par les valeurs dilatées, par le ratio et par régression. Les variances de ces estimateurs sont aussi estimées, de façon exacte si cela est possible et par approximation sinon.

2 Le logiciel GENESIS

GENESIS version 1.2 (Generalized Simulation System), conçu à Statistique Canada (Haziza, 2003), est une interface ergonomique qui utilise le logiciel SAS version 8. Le système est constitué de macros SAS reliées à des menus par SAS/AF. GENESIS peut être utilisé dans un cours de théorie des sondages par l'enseignant ainsi que par les étudiants afin d'illustrer des concepts théoriques tels que le choix d'un plan de sondage, le choix d'un estimateur ou encore le choix de la répartition dans le cas d'un plan de sondage stratifié. GENESIS est un système relativement facile à utiliser qui ne requiert pas de connaissances préalable du logiciel SAS et est relativement efficace en terme de temps d'exécution. GENESIS comprend trois modules :

(1) Full Response Module

(2) Imputation Module

(3) Class Module

Nous décrivons maintenant le module Full Response qui est approprié dans le contexte d'un cours de théorie des sondages. L'utilisateur commence par fournir à GENESIS une population sous forme de table SAS qui servira de point de départ pour les études de simulation. Ce module permet l'estimation d'un total ou d'une moyenne de population finie. Plusieurs plans de sondage sont disponibles pour la sélection des échantillons : échantillonnage aléatoire simple sans remise, échantillonnage proportionnel à la taille avec ou sans remise, échantillonnage stratifié aléatoire simple, échantillonnage de Poisson, échantillonnage à un et deux degrés, échantillonnage à deux-phases ainsi que la méthode de Rao-Hartley-Cochran. Dans le cas de l'échantillonnage stratifié aléatoire simple, l'utilisateur peut choisir entre une répartition optimale, une répartition de Neyman, une répartition proportionnelle et une répartition quelconque. Pour estimer le paramètre d'intérêt (moyenne ou total), GENESIS calcule l'estimateur de Horvitz-Thompson, l'estimateur par le

ratio et l'estimateur par régression. Dans le cas de l'échantillonnage stratifié aléatoire simple, GENESIS calcule les estimateurs par le ratio séparé et combiné ainsi que les estimateurs par régression séparé et combiné. Finalement, GENESIS calcule plusieurs mesures Monte Carlo telles que le biais relatif des estimateurs, l'erreur quadratique moyenne, le biais relatif des estimateurs de variance ainsi que la probabilité de couverture des intervalles de confiance. Les résultats des simulations sont stockés dans des tableaux SAS, ce qui donne une plus grande flexibilité à l'utilisateur. Par exemple, un tableau contenant les résultats pour chacune des itérations permettra à l'utilisateur de facilement calculer des mesures Monte Carlo autres que celles proposées par défaut dans GENESIS. De plus, GENESIS propose plusieurs graphiques qui permettent une comparaison aisée de la performance des estimateurs.

Dans le cas des modules Imputation et Class, GENESIS permet d'étudier la performance des estimateurs en présence de non-réponse. Il permet à l'utilisateur d'étudier l'effet du mécanisme de non-réponse (uniforme, ignorable et non-ignorable) sur le biais et la variance des estimateurs. Pour un mécanisme de non-réponse donné, GENESIS permet de comparer plusieurs méthodes d'imputation ainsi que deux méthodes pour la construction de classes d'imputation. Bien que ces deux modules puissent être utilisés dans un cours avancé de théorie des sondages, ils s'adressent plutôt à des méthodologues ou des chargés d'étude afin de répondre à des questions telles que : quelle méthode d'imputation doit-on utiliser ? Comment doit-on former les classes d'imputation ? Combien de classes doit-on utiliser ?

3 La bibliothèque "Sondages"

Cette bibliothèque (Aragon, 2003) a été écrite pour mettre à la disposition des étudiants un outil répondant à différents objectifs :

1. intégration dans un logiciel libre de calcul statistique,
2. disponibilité de fonctions semblables à celles qu'offrent les proc *surveyselect*, *surveymeans* et *surveyreg* de SAS,
3. adoption d'une approche *design based*.

On a choisi de développer les fonctions dans l'environnement R. Dans son état actuel, la bibliothèque contient des fonctions de sélection d'échantillons uniques ou répliqués, avec ou sans stratification, suivant différents plans de sondage : plan simple sans remise, plan simple avec remise, plan systématique, avec ou sans tri suivant une variable auxiliaire, plan proportionnel à la taille

avec remise, plan proportionnel à la taille sytématique, plan proportionnel à la taille par la méthode de Brewer.

L'estimation de totaux et de moyennes se fait par les valeurs dilatées ou par régression. L'estimation de la variance des estimateurs précédents est, suivant les plans, exacte (plan simple et méthode de Brewer) ou approchée pour les plans systématiques (Brewer, 2002).

Remarquons qu'il existe aussi sous R le package *survey* (Lumley, 2003) qui ne contient pas de fonctions pour le tirage d'échantillons mais qui incorpore de nombreuses fonctions pour l'estimation de paramètres d'intérêt tels que moyennes et totaux mais aussi quantiles et coefficients de régression pour des plans de sondages complexes.

4 Exemple

Lors de la présentation, nous considérerons différents jeux de données et différentes applications permettant d'illustrer l'utilisation du logiciel GENESIS et de la bibliothèque "Sondages" dans un cours de sondages.

Par exemple, nous proposons de comparer la méthode de répartition proportionelle à la méthode de répartition optimale dans le plan stratifié. Pour cela, nous considérons la population des 554 communes de moins de 10000 habitants de la Haute-Garonne pour lesquelles nous disposons de données issues du recensement de 1999. Nous nous intéressons à l'estimation du nombre total de logements vacants. Les communes sont réparties en 4 strates d'après la variable nombre de logements, considérée comme auxiliaire. Les strates des petites communes (en nombre de logements) contiennent plus de communes mais avec une dispersion plus faible que les strates des grandes communes. La répartition optimale donne donc des résultats très différents de la répartition proportionnelle à la taille. Le logiciel GENESIS et la bibliothèque "Sondages" permettent de calculer ces répartitions et de comparer, par exemple, la précision des estimateurs par les valeurs dilatées. Ainsi, si on calcule le biais relatif Monte Carlo des estimateurs de variance, on met clairement en évidence le gain en précision de la répartition optimale par rapport à la répartition proportionnelle pour ce type de données.

Bibliographie

[1] Aragon Y. (2003). *Sondages dans R.* Disponible auprès de l'auteur (aragon@cict.fr).

[2] Aragon, Y., Ruiz-Gazen, A. (2003). Utilisation des procédures SAS dans l'enseignement des sondages, soumis pour publication.

[3] Brewer K. R. W. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. Oxford Univ Press.

[4] Haziza, D. (2003). The Generalized Simulation System (GENESIS) : A pedagogical and methodological tool. To appear in the proceedings of the 2003 Survey Research Section of the American Statistical Association.

[5] Lumley, T. (2003). "Analysing Survey Data in R". Rnews vol. 3/1, june 2003, pp. 17-20.