Plan de sondage pour la sélection de panels d'entreprises

Pierre LAVALLÉE
Eurostat

1. Introduction

Afin de répondre à une demande de plus en plus forte pour des informations d'ordre micro-économique, longitudinale, complexe, variable et ciblé, on propose de plus en plus l'utilisation de panels. Eurostat, par exemple, songe à la création d'un réseau de panels d'entreprises au niveau de la Communauté économique européenne (CEE). On retrouve une description générale du projet dans Defays (1990) et Lavallée (1991). Kish (1965) définit un panel comme étant un échantillon dans lequel les éléments sont consultés à deux occasions ou plus. On parle donc d'un échantillon à caractère longitudinal.

On destine, pour le projet européen, l'usage des panels à la production d'estimations de totaux et de tendances pour des variables d'ordre économique. On prévoit de plus utiliser les données du panel de façon plus intensive pour des études analytiques sur des sujets tels que la compréhension des processus de formation des prix. L'aspect longitudinal du panel permettra de suivre l'évolution des entreprises dans les différents secteurs d'activités économiques.

Après avoir stratifié adéquatement la population et réparti les tailles d'échantillon, l'une des tâches de la mise sur pied d'un panel est la sélection des entreprises constituant le panel lui-même. Un plan de sondage couramment employé pour les enquêtes-entreprises fait usage de l'échantillonnage aléatoire simple (EAS) qui s'avère en général très approprié pour des enquêtes occasionnelles ou répétées. Dans le cas de panels, on note cependant certaines complications avec l'EAS que l'on se doit de considérer.

Une autre méthode utilisée pour l'échantillonnage de panels est l'échantillonnage de Bernoulli (EB), aussi connu sous le nom d'échantillonnage de Poisson. On sélectionne alors chaque entreprise sur la base d'un tirage de Bernoulli où la probabilité d'inclusion n'est autre que la fraction de sondage de la strate. Cette dernière méthode est en pratique de plus en plus utilisée.

* Pierre Lavallée est statisticien à Eurostat, Bâtiment Jean Monnet, L-2920, Luxembourg.

Dans le présent document, on discutera de plans de sondage pour la sélection d'un panel d'entreprises. On débutera par exposer le problème de méthodologie relié à la sélection d'un panel. On touchera ensuite brièvement à la stratification et la répartition de l'échantillon habituellement employées pour les enquêtes auprès des entreprises. Suivra alors une discussion sur l'utilisation de l'EAS suivie d'une autre sur l'EB. Pour chacune des deux méthodes d'échantillonnage, on discutera du traitement des naissances, des morts, des changements de strates et de la modification des fractions de sondage. On terminera par une brève conclusion.

Notons que pour l'ensemble du document, on supposera que les entreprises d'un panel sont consultées annuellement. La discussion est toutefois la même pour des fréquences de consultation trimestrielles ou mensuelles.

2. Définition du problème

L'établissement d'un panel se résume souvent à la sélection d'un ensemble fixe d'entreprises que l'on consulte périodiquement. Chaque entreprise du panel déclarée morte est alors remplacée par une autre entreprise choisie aléatoirement parmi le reste de la population ou sinon parmi les entreprises naissantes.

Dans le cas des enquêtes-entreprises, on entend généralement par une mort une entreprise qui sort de la population cible. Une mort correspond en général à une entreprise qui cesse ses activités. Cependant, toute entreprise qui sort du champ de l'enquête à cause d'un changement d'activité économique ou autre est aussi considérée comme une mort. A l'opposé d'une mort, une naissance correspond à une entreprise qui s'ajoute à la population cible. Bien que la création physique d'une nouvelle entreprise soit le cas le plus typique d'une naissance, on distingue aussi d'autres causes. Par exemple, on considère aussi comme naissance une entreprise qui, par un changement d'activité économique, fait maintenant parti du champ de l'enquête.

Bien qu'adéquat pour les analyses micro-économiques, le mode de sélection décrit plus haut pose des problèmes pour l'estimation de totaux et de tendances macro-économiques. En effet, à partir de cette forme de panel, on ne peut généralement dilater les résultats de l'échantillon au niveau de la population sans engendrer d'importants biais. Le problème vient du fait que cette forme de panel ne tient souvent pas compte des déformations de la population que ce soit au niveau des morts, naissances ou changements de strates.

Pour atteindre les objectifs fixés de production d'estimations de totaux et de tendances, on doit choisir un plan de sondage permettant la sélection d'un ensemble relativement fixe d'entreprises en maximisant le chevauchement entre deux périodes de consultation. De plus, on veut, à partir de cet ensemble d'entreprises, être en mesure de produire des estimations de totaux et de tendances macro-économiques avec un biais négligeable.

Les deux méthodes d'échantillonnage présentées ici, soient l'EAS et l'EB, peuvent être utilisées à l'intérieur d'un plan de sondage pour la création d'un panel de manière à

satisfaire les contraintes exposées ci-dessus. Chacune des méthodes amène des avantages et désavantages qui sont exposés dans le présent article.

Il est à noter qu'après avoir effectué la stratification de la base de sondage, on peut à la rigueur employer une méthode d'échantillonnage différente pour chacune des strates puisque la sélection des entreprises se fait indépendamment d'une strate à l'autre. Il est en général conseillé cependant d'utiliser le plus possible une seule méthode pour ne pas compliquer inutilement la sélection des entreprises.

3. Stratification et répartition de l'échantillon

La stratification est une technique couramment utilisée en échantillonnage. D'après Cochran (1977), elle permet premièrement d'assurer une forme de représentativité de certaines sous-populations de l'échantillon. Deuxièmement, elle aide à se plier à des exigences administratives ou politiques. C'est par exemple le cas de la stratification par pays ou province. Troisièmement, elle permet de répondre indépendamment à des problèmes d'échantillonnage spécifiques à chaque strate. Finalement, elle produit généralement une augmentation de la précision des estimations au niveau global.

Bien qu'il existe un nombre infini de possibilités de stratification, certaines variables sont typiques de la stratification d'enquêtes auprès des entreprises. La stratification par région géographique permet de répondre à des contraintes administratives et politiques. Notons de plus que les problèmes d'échantillonnage peuvent être très différents d'une région à l'autre. C'est, par exemple, le cas des pays membres de la CEE. La stratification par région géographique en suivant, par exemple, la Nomenclature des unités territoriales (Code NUTS) au niveau I, II ou III peut être utilisée afin de cerner des zones spécifiques.

Une autre variable de stratification habituellement essentielle est l'activité économique de l'entreprise. Celle-ci est identifiée à partir d'une codification comme la Nomenclature générale des activités économiques dans les Communautés européennes (NACE) disponible au niveau 1, 2, 3 ou 4 chiffres.

Une variable de stratification très utile pour l'augmentation de la précision des estimations est la taille de l'entreprise qui se traduit souvent en termes d'emploi ou de chiffre d'affaires. Cette stratification nécessite la détermination de bornes établissant des catégories. La détermination des bornes peut se faire suivant des contraintes opérationnelles ou bien suivant certains critères d'optimisation. Pour plus de détails sur la stratification dite "optimale", on peut consulter Lavallée (1989). On note finalement que dans la plupart des enquêtes-entreprises, les strates contenant les entreprises de grande taille sont sujettes à un tirage exhaustif, c'est-à-dire à un taux de sondage de 100%.

Il existe finalement une forme de stratification basée sur des concepts de taille englobant, en plus du nombre d'employés et du chiffre d'affaires, des variables décrivant la structure simple ou complexe de l'entreprise. Une telle stratification est, entre autres, employée à Statistique Canada où l'on divise l'univers des entreprises en

deux portions. On retrouve une portion de l'univers dite intégrée contenant les grosses entreprises en termes de revenu brut et celles dont la structure est complexe, et une portion dite non-intégrée contenant les petites entreprises de structure simple (voir Choudhry, Lavallée et Hidiroglou (1989)). Bien que cette forme de stratification permette un meilleur suivi des grosses et/ou complexes entreprises, elle engendre habituellement un mécanisme lourd de gestion. En effet, on doit alors différencier les entreprises de structure simple de celles de structure complexe, établir des règles de passage d'une portion à une autre, etc.

L'emploi des variables de stratification mentionnées plus tôt entraînera la création d'un certain nombre de strates. Comme on désire généralement avoir un minimum d'entreprises sélectionnées dans chaque strate (en général, deux entreprises pour avoir une estimation sans biais de la variance), le nombre total de strates est donc restreint par la taille d'échantillon totale. On doit donc choisir la stratification nécessaire pour la sélection du panel en tenant compte de la taille d'échantillon totale prévue.

Il existe un nombre important de méthodes visant à répartir les tailles d'échantillon entre les différentes strates. Une méthode souvent employée est celle de Neyman qui répartit l'échantillon de manière à minimiser la variance de l'estimation d'un total Y au niveau global. Cette méthode est décrite en détail par Cochran (1977).

La répartition de Neyman tend malheureusement à minimiser la variance au niveau global au détriment des petites strates qui obtiennent alors une portion insuffisante de l'échantillon pour produire des estimations précises. Pour solutionner ce problème, Bankier (1988) propose une répartition dite exponentielle (ou "power allocation") où les inégalités entre les strates sont amoindries par l'utilisation d'une puissance $q \varepsilon [0,1]$. Pour une taille d'échantillon totale n, la taille n_h de chaque strate h est alors déterminée par

$$n_h = n \frac{N_h S_h X_h^q Y_h^{-1}}{\sum_h N_h S_h X_h^q Y_h^{-1}} \qquad (3.1)$$

où N_h, Y_h, S_h et X_h correspondent respectivement, pour la strate h, à la taille de la population, au total pour la variable y, à la variance de la variable y et au total pour une variable auxiliaire x. En pratique, x est souvent choisie égale à y. D'autre part, on note qu'avec $q = 1$ l'expression (3.1) correspond à la répartition de Neyman.

4. Utilisation de l'échantillonnage aléatoire simple

4.1 Description de la méthode

L'EAS (avec remise) constitue l'une des méthodes les plus faciles à concevoir pour la sélection des entreprises d'un panel. Pour une strate h donnée, on peut visualiser cette méthode par un tirage de n_h boules dans un sac contenant N_h boules d'égale grosseur et numérotées de 1 à N_h. Le tirage se fait sans remise de chaque boule choisie. Le sac

de boules correspond en fait à la base de sondage de l'enquête. Pour plus de détails, on peut consulter Kish (1965) et Cochran (1977).

Par définition, chaque échantillon s de la strate h a une même probabilité de sélection $p_h(s)$ donnée par

(4.1) $$p_h(s)=\frac{n_h!(N_h-n_h)!}{N_h!}.$$

A partir de (4.1), on peut obtenir la probabilité π_i que l'unité i soit dans l'échantillon et la probabilité conjointe π_{ij} que les unités i et j soient dans l'échantillon. Celles-ci sont respectivement données par

(4.2) $$\pi_i=\frac{n_h}{N_h} \text{ si } i \varepsilon h$$

(4.3) et $$\pi_{ij}=\frac{n_h(n_h-1)}{N_h(N_h-1)} \text{ si } i,j \varepsilon h$$

$$\frac{n_h}{N_h}\frac{n_{h'}}{N_{h'}} \text{ si } i \varepsilon h \text{ et } j \varepsilon h' \text{ où } h \neq h'.$$

Les probabilités d'inclusion (4.2) et (4.3) sont particulièrement utiles pour le calcul des biais et des variances des estimations.

On envisage habituellement l'utilisation de l'EAS par soucis de simplicité. La littérature entourant cette méthode d'échantillonnage est, de plus, fort abondante.

4.2 Traitement des naissances

Supposons qu'à l'année A-1 on sélectionne par EAS un échantillon d'entreprises devant constituer un panel. En guise de simplification, on suppose que l'on utilise une seule strate de taille N_1 pour laquelle l'échantillon tiré est de taille n_1.

Supposons maintenant qu'à l'année A, N_b naissances soient survenues au sein de la population. On posera pour simplifier qu'il n'y a eu aucun autre changement structurel dans la population; c'est-à-dire qu'aucune mort ni changement de strate n'a eu lieu.

On a auparavant mentionné que le panel doit tenir compte des déformations de la population. Pour assurer à l'année A une certaine représentativité en termes de naissances, on doit ainsi ajouter au panel un échantillon de taille n_b tiré parmi les N_b naissances de l'année A. Cet échantillon doit normalement être sélectionné par EAS.

En rapport avec la théorie de l'échantillonnage, on doit considérer les naissances comme une strate séparée de la strate 1 des unités présentes en A-1. Le problème est que pour l'année $A+1$, en supposant que de nouvelles naissances se produisent, on devra alors considérer une nouvelle strate de naissances. On obtient donc une prolifération du nombre de strates; ce qui engendre rapidement de fortes complications entre autres au niveau de la gestion des strates. Idéalement, on désire ainsi à l'année A pouvoir jumeler la strate des naissances avec l'autre strate de manière à ne traiter

qu'une seule strate de taille $N=N_1+N_b$ lors du calcul des estimations et des variances. Une telle chose peut se faire moyennant certaines conditions.

On désire à l'année A estimer le total $Y=Y_1+Y_b$ d'une variable y où $Y_1=\sum_{i=1}^{N_1} y_{1i}$ et $Y_b=\sum_{i=1}^{N_b} y_{bi}$.

L'estimation du total Y se fait généralement en utilisant l'estimateur de Horwitz-Thompson

(4.4)
$$\hat{Y}_{str} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$
$$= \frac{N_1}{n_1} \sum_{i=1}^{n_1} y_{1i} + \frac{N_b}{n_b} \sum_{i=1}^{n_b} y_{bi}$$

où $n=n_1+n_b$.

En supposant que $\frac{N_1}{(N_1-1)} \doteq 1$ et $\frac{N_b}{(N_b-1)} \doteq 1$, la variance de (4.4) peut s'exprimer par

(4.5)
$$Var(\hat{Y}_{str}) = (\frac{N_1}{n_1}-1) \sum_{i=1}^{N_1} (y_{1i}-\bar{Y}_1)^2 + (\frac{N_b}{n_b}-1) \sum_{i=1}^{N_b} (y_{bi}-\bar{Y}_b)^2$$

où $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$ pour $h=1,b$.

Le problème de la prolifération des strates ne pose pas vraiment de problème au niveau de l'estimation de Y puisqu'elle n'entraîne qu'une hétérogénéité dans les probabilités d'inclusion entrant dans l'expression (4.4). Cependant, le calcul de la variance inclut de plus en plus de termes, d'où le désir de vouloir ignorer la stratification au niveau des naissances.

Afin de réduire les termes de (4.5), on suggère de sélectionner l'échantillon des naissances à un taux proportionnel au reste de la population. En d'autres termes, on choisit n_b tel que

(4.6)
$$\frac{n_b}{N_b} = \frac{n_1}{N_1} = \frac{n}{N} .$$

Avec la répartition (4.6), la variance (4.5) peut alors s'écrire sous la forme

(4.7)
$$Var(\hat{Y}_{str}) = (\frac{N}{n}-1)[\sum_{i=1}^{N} (y_i-\bar{Y})^2 - N_1(\bar{Y}_1-\bar{Y})^2 - N_b(\bar{Y}_b-\bar{Y})^2]$$

où $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$.

Finalement, si l'on peut supposer que

(4.9) $$\bar{Y}_1=\bar{Y}_b=\bar{Y} ,$$

on obtient alors

(4.10) $$Var(\hat{\bar{Y}}_{str})=(\frac{N}{n}-1)\sum_{i=1}^{N}(y_i-\bar{Y})^2 .$$

En échantillonnant les naissances avec la même fraction de sondage que le reste de la population et en supposant que la moyenne $\bar{Y}_b$ des naissances est la même que celle $\bar{Y}_1$ du reste de la population, on peut donc exprimer la variance en ignorant la stratification au niveau des naissances.

Par ailleurs, on peut estimer (4.10) en utilisant

(4.11) $$var(\hat{\bar{Y}}_{str})=\frac{N}{n}(\frac{N}{n}-1)\sum_{i=1}^{n}(y_i-\bar{y})^2$$

où $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

L'hypothèse (4.9) présuppose que les naissances se comportent (en ce qui concerne la moyenne $\bar{Y}_b$) comme le reste de la population. Cependant, on peut aisément concevoir que les entreprises naissantes soient des entreprises possédant de petites valeurs pour la variable y, si celle-ci, par exemple, représente le revenu. Dans ce cas, $\bar{Y}_b$ peut s'avérer de beaucoup inférieur à $\bar{Y}_1$; ce qui introduit un biais dans les formules (4.10) et (4.11). On note que le biais introduit est toutefois toujours positif.Ceci n'est souvent pas considéré comme un problème majeur puisqu'un biais positif dans la variance produit des intervalles de confiance conservateurs.

En utilisant la pondération $w_i = 1/\pi_i$, un estimateur souvent utilisé en pratique pour estimer $Var(\hat{\bar{Y}}_{str})$ est

(4.12) $$\tilde{var}(\hat{\bar{Y}}_{str})=(\frac{N}{n}-1)[\sum_{i=1}^{N}w_i y_i^2-\frac{1}{N}(\sum_{i=1}^{N}w_i y_i)^2]$$

On peut démontrer que cet estimateur est, sous la répartition proportionnelle (4.6), équivalent à l'estimateur (4.11). Cet estimateur est cependant souvent utilisé en pratique même lorsqu'il n'y a pas répartition proportionnelle.

4.3 Traitement des morts

Comme pour le cas des naissances, le panel doit refléter les morts survenant au sein de la population des entreprises. Le nombre de morts et d'entreprises restantes est habituellement connu à travers la base de sondage mise à jour chaque année.

En connaissant le nombre N_r d'unités restantes, l'estimation du total $Y_r=\sum_{i=1}^{N_r}y_{ri}$ peut se faire par post-stratification. A l'année A, on stratifie ainsi à posteriori l'échantillon (ou panel) sélectionné en A-1 en deux post-strates: une contenant les morts et l'autre

les unités restantes. On retrouve une description de la post-stratification dans Cochran (1977) et Holt et Smith (1979).

Comme à la section 4.2, on suppose qu'à l'année A-1on a sélectionné un panel de n_1 entreprises choisies parmi N_1 en utilisant l'EAS.

Supposons maintenant qu'à l'année A, N_d des unités (ou entreprises) de la population soient mortes de sorte qu'il reste $N_r=N_1-N_d$ unités. Parallèlement, n_d des unités du panel sont mortes de sorte qu'il reste $n_r=n_1-n_d$ unités échantillonnées où $1 \leq n_r \leq N_r$. On supposera de nouveau qu'aucun autre changement structurel n'est survenu dans la population; c'est-à-dire qu'aucune naissance ni changement de strate n'a eu lieu.

Afin d'estimer Y_r , on peut utiliser l'estimateur suivant obtenu suite à la post-stratification de la population:

(4.13)
$$\hat{Y}_r = \frac{N_r}{n_r} \sum_{i=1}^{n_r} y_{ri} \ .$$

Il est à noter que, contrairement au cas des naissances, le problème de la prolifération des strates ne s'applique pas ici puisque l'on ne s'intéresse qu'à la post-strate des unités restantes et non pas à celle des unités mortes.

On remarque que les tailles d'échantillon n_d et n_r sont des variables aléatoires. En conditionnant sur n_d et n_r, on peut démontrer que $\hat{Y}_r$ est conditionnellement sans biais pour Y_r. Conséquemment,sans conditionner, $\hat{Y}_r$ est aussi sans biais.

La variance conditionnelle de $\hat{Y}_r$ est donnée par

(4.14)
$$Var(\hat{Y}_r \mid n_r,n_d)=\frac{N_r^2}{n_r}(1-\frac{n_r}{N_r})\frac{1}{N_r-1}\sum_{i=1}^{N_r}(y_{ri}-\bar{Y}_r)^2 \ .$$

Par ailleurs, en faisant l'approximation de Stephan (1945)

(4.15)
$$E\,(\frac{1}{n_r}) \doteq \frac{1}{n_1}\frac{N_1}{N_r}+(\frac{1}{n_1}\frac{N_1}{N_r})^2(1-\frac{N_r}{N_1}) \ ,$$

la variance non-conditionnelle de $\hat{Y}_r$ est donnée par

(4.16)
$$Var(\hat{Y}_r) \doteq (\frac{N_1}{n_1}-1)\sum_{i=1}^{N_r}(y_{ri}-\bar{Y}_r)^2+\frac{N_1}{n_1^2}(\frac{N_1}{N_r}-1)\sum_{i=1}^{N_r}(y_{ri}-\bar{Y}_r)^2 \ .$$

Le premier terme de droite de (4.16) reflète le fait que la répartition des morts du panel sera en moyenne proportionnelle au nombre de morts dans la population. Le deuxième terme de droite vient du fait que cette répartition ne sera proportionnelle qu'en espérance.

On retrouve l'estimation de la variance conditionnelle et celle non-conditionnelle dans Rao (1985).

La question d'utiliser l'approche conditionnelle ou non demeure fortement discutée. Holt et Smith (1979) et Rao (1985) suggèrent d'adopter l'approche conditionnelle lors de la production des estimations en tant que telles. L'approche non-conditionnelle n'est appropriée que lorsque l'on compare différents plans de sondage.

4.4 Traitement des changements de strate

L'évolution des entreprises dans le temps rend inévitable les sauts d'une strate à une autre. Ces changements des strates sont directement reliés à la manière dont la stratification est effectuée. En supposant une stratification par région géographique, par exemple, une entreprise effectuera un changement de strate si elle déménage d'une région à une autre. Notons que l'on parle de changement de strate seulement si l'entreprise concernée demeure dans le champ de l'enquête, autrement il s'agit d'une mort.

Supposons qu'à l'année *A-1* on ait stratifié la population en deux strates de tailles N_1 et N_2. On a ensuite sélectionné un panel en tirant respectivement n_1 et n_2 unités en utilisant, dans chaque strate, l'EAS.

Supposons maintenant qu'à l'année *A* , $N_{1,2}$ unités de la population soient passées de la strate 1 à la strate 2 de sorte qu'il reste $N_{1'}=N_{1,1}=N_1-N_{1,2}$ unités dans la strate 1 et que $N_{2'}=N_2+N_{1,2}$ unités sont maintenant dans la strate 2. Parallèlement, $n_{1,2}$ des unités du panel ont changé de strates. Par soucis de simplicité, on suppose que tous les mouvements ont eu lieu de la strate 1 à 2. On suppose de plus qu'aucun autre changement structurel n'est intervenu dans la population; c'est-à-dire qu'aucune naissance ni mort n'a eu lieu.

En rapport avec la théorie de l'échantillonnage, le traitement des changements de strates se fait par post-stratification, comme pour le cas des morts. En effet, en supposant que le panel est sélectionné en *A-1*, les changements de strates survenant à l'année *A* se ramènent à stratifier les unités du panel à posteriori puisque la sélection de l'échantillon a déjà été effectuée. Il est cependant important de noter que la stratification initiale effectuée en *A-1* ne doit pas être ignorée; on doit en fait effectuer la post-stratification à l'intérieur de chaque strate initiale. Dans le cas présent, on considère donc trois post-strates: une contenant les $N_{1,1}$ unités restantes dans la strate 1, une deuxième contenant les $N_{1,2}$ unités passées de la strate 1 à 2, et une troisième contenant les N_2 unités initiales de la strate 2.

On désire à l'année *A* estimer le total $Y=Y_{1'}+Y_{2'}=Y_{1,1}+Y_{1,2}+Y_2$ pour une variable y.

En utilisant les tailles $N_{1,1}$,$N_{1,2}$ et N_2, l'estimateur de Y obtenu par post-stratification est donné par

(4.17)
$$\hat{Y}_{pstr} = \hat{Y}_{1,1} + \hat{Y}_{1,2} + \hat{Y}_2$$
$$= \frac{N_{1,1}}{n_{1,1}} \sum_{i=1}^{n_{1,1}} y_{1,1i} + \frac{N_{1,2}}{n_{1,2}} \sum_{i=1}^{n_{1,2}} y_{1,2i} + \frac{N_2}{n_2} \sum_{i=1}^{n_2} y_{2i} \ .$$

Bien entendu, on suppose dans (4.17) que $1 \le n_{1,1} \le N_{1,1}$, $1 \le n_{1,2} \le N_{1,2}$ et $1 \le n_2 \le N_2$.

Comme pour le cas des morts, en conditionnant sur les tailles d'échantillon $n_{1,1}$, $n_{1,2}$ et n_2, on peut démontrer que $\hat{Y}_{pstr}$ est conditionnellement sans biais pour Y. La variance conditionnelle de $\hat{Y}_{pstr}$ est donnée par

(4.18)
$$Var(\hat{Y}_{pstr}|n_{1,1},n_{1,2},n_2)=\frac{N_{1,1}^2}{n_{1,1}}\left(1-\frac{n_{1,1}}{N_{1,1}}\right)\frac{1}{N_{1,1}-1}\sum_{i=1}^{N_{1,1}}(y_{1,1i}-\bar{Y}_{1,1})^2$$
$$+\frac{N_{1,2}^2}{n_{1,2}}\left(1-\frac{n_{1,2}}{N_{1,2}}\right)\frac{1}{N_{1,2}-1}\sum_{i=1}^{N_{1,2}}(y_{1,2i}-\bar{Y}_{1,2})^2$$
$$+\frac{N_2^2}{n_2}\left(1-\frac{n_2}{N_2}\right)\frac{1}{N_2-1}\sum_{i=1}^{N_2}(y_{2i}-\bar{Y}_2)^2 .$$

Par ailleurs, en utilisant l'approximation de Stephan (1945), la variance non-conditionnelle de $\hat{Y}_{pstr}$ est donnée par

(4.19)
$$Var(\hat{Y}_{pstr}) \doteq \left(\frac{N_1}{n_1}-1\right)\sum_{i=1}^{N_{1,1}}(y_{1,1i}-\bar{Y}_{1,1})^2+\frac{N_1}{n_1^2}\left(\frac{N_1}{N_{1,1}}-1\right)\sum_{i=1}^{N_{1,1}}(y_{1,1i}-\bar{Y}_{1,1})^2$$
$$+\left(\frac{N_1}{n_1}-1\right)\sum_{i=1}^{N_{1,2}}(y_{1,2i}-\bar{Y}_{1,2})^2+\frac{N_1}{n_1^2}\left(\frac{N_1}{N_{1,2}}-1\right)\sum_{i=1}^{N_{1,2}}(y_{1,2i}-\bar{Y}_{1,2})^2$$
$$+\left(\frac{N_2}{n_2}-1\right)\sum_{i=1}^{N_2}(y_{2i}-\bar{Y}_2)^2 .$$

Afin de simplifier le traitement des changements de strates, on pourrait considérer l'utilisation de la théorie des sous-populations. On considère alors chaque post-strate comme une sous-population (ou domaine) pour laquelle la taille est inconnue. L'estimateur résultant possède la forme suivante:

(4.20)
$$\hat{Y}_{str}=\frac{N_1}{n_1}\sum_{i=1}^{n_1}y_{1i}+\frac{N_2}{n_2}\sum_{i=1}^{n_2}y_{2i}$$
$$=\frac{N_1}{n_1}\sum_{i=1}^{n_{1,1}}y_{1,1i}+\frac{N_1}{n_1}\sum_{i=1}^{n_{1,2}}y_{1,2i}+\frac{N_2}{n_2}\sum_{i=1}^{n_2}y_{2i} .$$

L'estimateur (4.20) suppose que le nombre d'unités changeant de strates au niveau du panel est proportionnel au nombre au niveau de la population.

Bien qu'il soit sans biais, l'estimateur (4.20) ne tient malheureusement pas compte explicitement des changements de strates. A mesure que l'on avance dans le temps, l'évolution des entreprises produit en fait un éloignement de plus en plus prononcé de la stratification initiale et rend ainsi difficile de supposer la proportionnalité des

changements de srates. Ce problème peut malheureusement devenir considérable. Par exemple, une petite entreprise du panel devenue maintenant grosse peut produire d'énormes distorsions au sein des estimations si ce phénomène n'est pas proportionnel au nombre de petites entreprises devenues grosses au sein de la population.

On note finalement que l'on pourrait aussi tenter d'utiliser l'estimateur (4.20) pour le traitement des morts d'entreprises. Dans ce cas, le même raisonnement s'applique à savoir que l'estimateur obtenu par la théorie des sous-populations ne sera efficace que si le nombre de morts au sein du panel est proportionnel à celui de la population.

4.5 Modification des fractions de sondage

Après plusieurs années d'existance d'un panel, on peut concevoir que l'on désire soit augmenter ou diminuer sa taille; ce qui équivaut à une modification des fractions de sondage utilisées pour la sélection du panel. Cette modification peut avoir lieu pour répondre, par exemple, à de nouvelles contraintes budgétaires.

La modification des fractions de sondage doit normalement s'accompagner d'une restratification de la population. On profite alors de l'occasion pour stratifier selon les dernières données disponibles à partir des répertoires. Les changements de strates sont par le fait même corrigés .

Une fois les nouvelles fractions de sondage déterminées, il est alors indispensable de modifier la taille du panel tout en s'assurant un chevauchement maximal avec l'ancien panel. Un chevauchement maximal permet, entre autres, de continuer souvent des séries chronologiques existantes depuis de nombreuses années. Cette contrainte reliée au chevauchement exclut alors la sélection d'un nouveau panel indépendamment du premier. En effet, le sélection indépendante d'un nouvel échantillon n'assurerait pas nécessairement de chevauchement, en particulier si les fractions de sondage sont faibles.

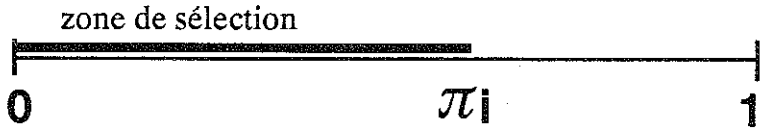
Avec l'EAS, il semble malheureusement qu'il n'existe pas de méthode "parfaite" permettant, pour des nouvelles fractions de sondages données, d'effectuer un chevauchement maximal tout en conservant les probabilités d'inclusion π_i et π_{ij} inhérentes à l'EAS. Kish et Scott (1971) suggèrent de classifier les unités du panel selon la nouvelle stratification tout en conservant des sous-strates où les unités ont été initialement sélectionnées avec la même fraction de sondage. On peut alors modifier à l'intérieur de chaque sous-strate la fraction de sondage initiale en sélectionnant aléatoirement de nouvelles unités dans la sous-strate ou en éliminant, toujours aléatoirement, des unités. On obtient ainsi les nouvelles fractions de sondages désirées tout en maximisant le chevauchement. Cependant, comme le mentionnent Kish et Scott (1971), bien que les probabilités d'inclusion π_i soient respectées par cette dernière méthode, tel n'est pas le cas pour les probabilités conjointes π_{ij}. Les calculs de variance doivent alors tenir compte de ce problème.

5. Utilisation de l'échantillonnage de Bernoulli

5.1 Description de la méthode

L'EB tire son nom du fait que la sélection des unités s'effectue par une succession de N tirages de Bernoulli. A chaque année, on sélectionne les entreprises de l'échantillon (ou du panel) en effectuant un tirage de Bernoulli pour chaque unité contenue dans la base de sondage. Ainsi, chaque unité i est indépendamment sélectionnée en comparant un nombre aléatoire u_i uniformément distribué sur]0,1[à la probabilité d'inclusion π_i. L'unité i est sélectionnée si $u_i \leq \pi_i$. La probabilité d'inclusion correspond en général à la fraction de sondage de la strate où l'unité se retrouve.

Figure 1. Mode de sélection

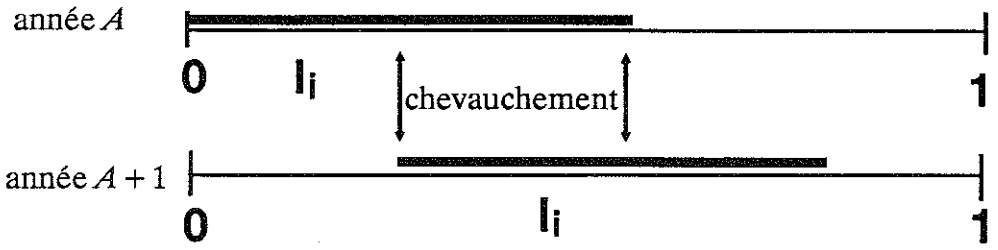


On note que le concept de stratification diffère entre l'EAS et l'EB puisqu'avec ce dernier les unités sont sélectionnées indépendamment autant à l'intérieur des strates qu'entre celles-ci. Pour l'EB, la stratification ne sert en fait qu'à l'assignation des probabilités d'inclusion aux différentes unités de la population.

L'EB fournit un moyen simple pour la création et le maintien d'un panel. En effet, en conservant le nombre aléatoire u_i d'une période à l'autre, une entreprise i sélectionnée à l'année A sera aussi sélectionnée en $A+1$ s'il n'y a pas diminution de la probabilité d'inclusion π_i. On peut ainsi maximiser le chevauchement de l'échantillon entre deux périodes. En pratique, au lieu de conserver le nombre aléatoire u_i, on peut le recalculer à chaque année à partir d'un générateur de nombres pseudo-aléatoires avec comme germe une valeur spécifique à chaque entreprise. Par exemple, on utilise souvent le numéro d'identification de l'entreprise. On note qu'un tel générateur de nombres pseudo-aléatoires est aussi appelé fonction "hash". L'utilisation de cette méthode est décrite, entre autres, par Sunter (1986) et par Choudhry, Lavallée et Hidiroglou (1989).

On peut aussi noter que l'EB fournit un moyen simple d'effectuer de la rotation au sein de l'échantillon. Pour ce faire, on remplace la probabilité d'inclusion π_i par un intervalle d'inclusion I_i défini sur]0,1[et de largeur proportionnelle à π_i. La rotation s'effectue alors en déplaçant l'intervalle I_i d'une année à l'autre. Le nombre d'unités restant dans l'échantillon est en moyenne proportionnel au chevauchement des intervalles.

Figure 2. Rotation de l'échantillon



Si π_i correspond à la fraction de sondage de la strate h, on obtient

(5.1) $$\pi_i = \frac{n_h}{N_h} \text{ si } i \in h$$

(5.2) et $$\pi_{ij} = \pi_i \pi_j \text{ pour } i \neq j \ .$$

Un des problèmes reliés à l'EB est le fait que la taille d'échantillon réalisée (ou finale) est aléatoire. La valeur n_h correspond en fait à la taille d'échantillon espérée et non pas la taille réalisée, que l'on dénote par m_h. Cependant, on peut démontrer que

$$Var\left(\frac{m_h}{N_h}\right) = \frac{n_h}{N_h{}^2}\frac{(N_h - n_h)}{N_h} \ .$$

La quantité $\frac{m_h}{N_h}$ converge donc en probabilité vers $\frac{n_h}{N_h}$.

Par ailleurs, on peut contrôler l'assignation des π_i de manière à éviter, avec un certain niveau de confiance $(1-\alpha)$, des tailles d'échantillon inférieures à un seuil c donné. En notant que, pour la strate h, m_h suit une loi binômiale avec paramètres N_h et $p_h = \frac{n_h}{N_h}$, la fraction de sondage minimale $p_{min,h} = \frac{n_{min,h}}{N_h}$ est obtenue en solutionnant

(5.3) $$\alpha > P(m_h \leq c) = \sum_{d=0}^{c} \frac{N_h!}{d!(N_h - d)!} p_{min,h}^{d} \, (1 - p_{min,h})^{(N_h - d)} \ .$$

Puisque $p_{min,h}$ est compris entre 0 et 1, on peut solutionner (5.3) en utilisant , par exemple, une recherche binaire. Pour plus de détail, on peut consulter Lavallée (1986).

Supposons que la probabilité d'inclusion π_i soit telle que définie par (5.1). En conditionnant sur la taille d'échantillon réalisée m_h, on obtient directement l'expression (4.1) pour la probabilité de sélection de chaque échantillon s de la strate h. Ainsi, en supposant (5.1) et en conditionnant sur m_h, l'EB est équivalent à l'EAS.

Ce résultat s'avère utile car on peut ainsi, sous certaines conditions, utiliser des résultats relatifs à l'EAS.

5.2 Traitement des naissances

Avec l'EB, le traitement des naissances est considérablement simplifié par rapport à l'EAS puisque celles-ci sont en fait considérées automatiquement au cours du processus d'échantillonnage. En effet, pour chaque unité i naissante, on génère alors le nombre aléatoire u_i que l'on compare simplement à la probabilité d'inclusion provenant de la fraction de sondage de la strate où l'unité naissante se retrouve. Il n'y donc pas de création de strates spécifiques aux naissances; ce qui élimine du même coup le problème de prolifération des strates remarqué avec l'EAS. Notons aussi que les unités naissantes seront sélectionnées avec la même fraction de sondage que le reste des unités de la strate et donc avec une répartition proportionnelle comme on le suggérait dans le cas de l'EAS.

On suppose de nouveau qu'à l'année A-1 on a sélectionné un panel de m_1 entreprises choisies parmi N_1 en utilisant l'EB avec $\pi_i = f$ pour $i = 1,...,N_1$. En guise de simplification, on ne considère toujours qu'une seule strate.

Supposons maintenant qu'à l'année A, N_b naissances soient survenues au sein de la population. On pose de nouveau qu'aucun autre changement structurel n'est survenu dans la population.

En utilisant l'EB comme décrit en section 5.1, on sélectionne alors m_b naissances toujours avec $\pi_i = f$ pour $i = 1,...,N_b$.

On peut, pour l'estimation du total $Y=Y_1+Y_b$, utiliser l'estimateur de Horwitz-Thompson

(5.4)
$$\hat{Y}_{str} = \sum_{i=1}^{m} \frac{y_i}{\pi_i}$$

où $m=m_1+m_b$ est la taille d'échantillon réalisée.

La variance de (5.4) est donnée par

(5.5)
$$Var(\hat{Y}_{str}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2$$

$$= \left(\frac{N}{n} - 1\right) \sum_{i=1}^{N} y_i^2$$

où $N=N_1+N_b$ et $n=fN$.

Comme le souligne Sunter (1986), la variance (5.5) ne se compare pas favorablement à la variance de $\hat{Y}_{str}$ obtenue si l'on utilise l'EAS. En utilisant la taille d'échantillon réalisée comme variable auxiliaire, on peut cependant obtenir l'estimateur suivant:

(5.6)
$$\tilde{Y}_{str} = \frac{N}{m} \sum_{i=1}^{m} y_i \ .$$

Ce dernier estimateur correspond en fait à un estimateur par le quotient que l'on décrit, entre autres, dans Cochran (1977). De plus, on peut démontrer que l'estimateur (5.6) est sans biais. Obtenue par linéarisation de Taylor, l'expésssion de la variance de Y_{str} est alors équivalente à l'équation (4.10). On obtient donc directement le résultat désiré à la section 4.2 sans avoir à faire l'hypothèse (4.9).

On note ici que si l'on considère plusieurs strates, on suggère de définir l'estimateur (5.6) en tenant compte de ces strates de manière à garder les probabilités d'inclusion homogènes. On obtient alors l'estimateur suivant:

(5.7)
$$\tilde{Y}_{str} = \sum_{h} \frac{N_h}{m_h} \sum_{i=1}^{m_h} y_{hi} \ .$$

Finalement, on peut estimer la variance de (5.6) en utilisant la formule suivante:

(5.8)
$$var(\tilde{Y}_{str}) = (\frac{N}{m})^2 (1-\frac{n}{N}) \sum_{i=1}^{n} (y_i - \bar{y})^2$$

où $\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$.

5.3 Traitement des morts

Comme pour le cas des naissances, le traitement des morts se fait sans difficulté par rapport à l'EAS. En retirant les unités mortes de la base de sondage, on les soustrait du même coup aux tirages de Bernoulli sous-jacents à l'EB. On élimine donc simplement les morts sans ajouter de traitement spécial au processus de sélection.

Comme à la section 5.2, on suppose qu'à l'année *A-1* on a sélectionné un panel de m_1 entreprises choisies parmi N_1 en utilisant l'EB avec $\pi_i = f$ pour $i = 1,...,N_1$. On suppose de plus qu'à l'année *A*, N_d des unités (ou entreprises) de la population sont mortes de sorte qu'il reste $N_r = N_1 - N_d$ unités où $1 \leq n_r \leq N_r$.

On resélectionne à l'année *A* le panel parmi les N_r unités en utilisant l'EB toujours avec $\pi_i = f$ pour $i = 1,...,N_r$.

Afin d'estimer Y_r, on utilise alors l'estimateur suivant s'inspirant de (5.6):

(5.9)
$$\tilde{Y}_r = \frac{N_r}{m_r} \sum_{i=1}^{m_r} y_{ri}$$

où m_r est la taille d'échantillon réalisée.

La variance et l'estimateur de la variance de $\tilde{Y}_r$ peuvent être obtenus respectivement à partir de (4.10) et (5.8).

5.4 Traitement des changements de strates

Les changements de strates sont de nouveau traités de manière simple avec l'EB. En changeant de strate, une unité i se voit alors assignée une nouvelle probabilité d'inclusion correspondant à la nouvelle strate où elle se retrouve. Si la nouvelle probabilité d'inclusion s'avère supérieure au nombre aléatoire u_i associé auparavant à l'unité, celle-ci est alors sélectionnée pour faire partie de l'échantillon indépendamment du fait qu'elle l'ait ou non été par le passé. Une unité exclue à l'année A de l'échantillon peut donc en faire partie en $A+1$ si l'unité se retrouve maintenant dans une strate possédant une plus grande fraction de sondage.

Supposons qu'à l'année A-1, on ait stratifié la population en deux strates de tailles $N1$ et $N2$. On a ensuite sélectionné un panel de $m1$ et $m2$ entreprises respectivement choisies parmi $N1$ et $N2$ en utilisant l'EB avec $\pi_i = f1$ pour $i \in 1$ et $\pi_i = f2$ pour $i \in 2$.

On suppose de nouveau qu'à l'année A, $N1,2$ des unités de la population sont passées de la strate 1 à la strate 2 de sorte qu'il reste $N1'=N1,1=N1-N1,2$ unités dans la strate 1 et que $N2'=N2+N1,2$ unités sont maintenant dans la strate 2.

On resélectionne à l'année A le panel parmi l'ensemble des entreprises toujours en utilisant l'EB avec les mêmes probabilités d'inclusion qu'en A-1 mais appliquées au nouvelles strates 1' et 2'.

Afin d'estimer $Y=Y1'+Y2'=Y1,1+Y1,2+Y2$, on peut employer l'estimateur suivant:

(5.10)
$$\tilde{Y}_{str} = \frac{N1'}{m1'} \sum_{i=1}^{m1'} y1'i + \frac{N2'}{m2'} \sum_{i=1}^{m2'} y2'i \ .$$

La variance et l'estimateur de la variance de $\tilde{Y}_r$ peuvent être obtenus respectivement à partir de (4.10) et (5.8).

On remarque que l'estimateur (5.10) diffère de (4.17) par le fait qu'on ignore ici la stratification initiale utilisée en A-1. Rappelons qu'avec l'EB, la stratification ne sert qu'à l'assignation des probabilités d'inclusion.

En resélectionnant le panel à l'année A, les $N1,2$ unités passées de la strate 1 à 2 ont maintenant la probabilité d'inclusion $f2$ au lieu de $f1$. Notons que la composition même du panel peut avoir été modifée suite à ce changement de probabilié d'inclusion. On constate donc que $m1+m2$ n'est pas nécessairement égal à $m1'+m2'$. Il faut cependant finalement rappeler que le changement de composition du panel ne reflète qu'un ajustement aux déformations de la population et qu'avec l'EB le panel de A possède un chevauchement maximal avec l'année A-1. On peut donc, d'un certain point de vue, considérer ces changements dans le panel comme souhaitables.

5.5 Modification des fractions de sondage

On a vu qu'en utilisant l'EB la sélection d'un panel d'entreprises s'effectue en resélectionnant chaque année le panel à partir des nombres aléatoires u_i gardés en

mémoire d'une année à l'autre. L'EB permet alors d'obtenir un chevauchement maximal entre deux années; ce qui est essentiel à un panel.

Avec l'EB, on peut donc à loisir modifier les fractions de sondages sans avoir de complications opérationnelles comme c'est le cas avec l'EAS. Les probabilités d'inclusion π_i et π_{ij} inhérentes à l'EB seront alors toujours respectées. On bénéficie donc d'une flexibilité qu'on ne retrouve pas avec l'EAS.

6. Conclusion

Dans cet article, on a décrit deux plans de sondage pouvant être utilisés pour la sélection d'un panel d'entreprises. Après avoir exposé le problème de l'adaptation aux déformations de la population, on a décrit la stratification que l'on retrouve le plus souvent pour des enquêtes auprès des entreprises. On a ensuite discuter de l'utilisation de l'EAS et de l'EB en rapport avec le traitement des naissances, des morts, des changements de strates et de la modification des fractions de sondage.

Bien que l'EAS soit un plan de sondage généralement simple à implanter pour des enquêtes occasionnelles, on a vu que certaines précautions sont à prendre lorsqu'il s'agit de panels. Pour le traitement des naissances, on doit contrôler une prolifération éventuelle des strates. Pour les morts et les changements de strates, l'utilisation de la post-stratification semble inévitable pour l'obtention de bonnes estimations. Finalement, on a constaté qu'il n'existe pas de méthode "parfaite" pour l'EAS afin d'effectuer des changements dans les fractions de sondage tout en assurant un chevauchement maximal d'une année à l'autre.

L'utilisation de l'EB semble offrir un bon nombre d'avantages pour la sélection d'un panel. Le traitement des naissances, des morts, des changements de strates et de la modification des fractions de sondage se trouve considérablement simplifié par rapport à l'EAS; ce qui est souvent un atout du point de vue de l'implantation. En plus de permettre un chevauchement maximal entre les années, l'EB rend possible une certaine rotation de l'échantillon. Finalement, on retrouve avec l'EB une flexibilité souhaitable pour la sélection d'un panel construit souvent pour plusieurs années.

Mots clés: panel d'entreprises, échantillonnage de Bernoulli, morts, naissances, changements de strates.

Bibliographie

Bankier, M. (1988), "Power Allocations: Determining Sample Sizes for Subnational Areas", The American Statistician, Vol. 42, No. 3, août 1988.

Choudhry, A.H., Lavallée P., Hidiroglou, M.A. (1989), "Two-Phase Sample Design For Tax Data", Article présenté au Congrès de l'American Statistical Association à Washington, août 1989.

Cochran, W.G. (1977), "Sampling Technique", 3ème édition, John Wiley and Sons, New York, 1977.

Defays, D. (1990), "Définition d'un réseau européen d'entreprises témoins", Document interne à Eurostat, 22 mars 1990.

Holt, D., Smith, T.M.F. (1979), "Post Stratification", Journal of the Royal Statistical Society A, Vol. 142, Part 1, pp. 33-46.

Kish, L. (1965), "Survey Sampling", John Wiley and Sons, New York, 1965.

Kish, L., Scott, A. (1971), "Retaining Units after Changing Strata and Probabilities", Journal of the American Statistical Association, Vol. 66, No. 335, pp. 461-470.

Lavallée, P (1986), "Allocation for Poisson Sampling", Document interne à Statistique Canada, 10 novembre 1986.

Lavallée, P. (1989), "Some contributions to optimal stratification", Masters Abstracts International, Vol. 27, No. 1, 1989.

Lavallée, P. (1991), "Création d'un réseau européen par la sélection d'entreprises témoins - Projet CREUSET", Document EUROSTAT/D5/CREUSET/1, Luxembourg, janvier 1991.

Rao, J.N.K. (1985), "Conditional Inference in Survey Sampling", Survey Methodology, Vol. 11, No. 1, pp. 15-31, juin 1985.

Stephan, F.F. (1945), "The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate", Annals of Mathematical Statistics, Vol. 16, pp. 50-61.

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