Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Répartition optimale pour une enquête téléphonique à base de sondage double

par Kirk M. Wolter, Xian Tao, Robert Montgomery
et Philip J. Smith

Date de diffusion : le 17 décembre 2015

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2015
Vol. 41, nº 2, pp. 411-425
Statistique Canada, nº 12-001-X au catalogue

Répartition optimale pour une enquête téléphonique à base de sondage double

Kirk M. Wolter, Xian Tao, Robert Montgomery et Philip J. Smith[1]

Résumé

La bonne conception d'une enquête téléphonique par composition aléatoire (CA) à partir d'une base de sondage double requiert de choisir entre de nombreuses options, en faisant la part des différences de coût, de précision, et de couverture, afin d'optimiser la réalisation des objectifs de l'étude. L'un des éléments à prendre en considération est celui de savoir s'il faut présélectionner les ménages équipés de téléphones mobiles et n'interroger que ceux utilisant exclusivement des téléphones mobiles (ménages EXM), donc écarter ceux qui se servent d'un téléphone fixe ainsi que d'un téléphone mobile (ménages F-et-M), ou s'il faut, au contraire, interroger toutes les unités de l'échantillon de ménages équipés de téléphones mobiles. Nous présentons un cadre pour comparer les avantages et les inconvénients de ces deux options, ainsi qu'une méthode pour sélectionner le plan de sondage optimal. Nous établissons la répartition optimale de la taille de l'échantillon entre les deux bases de sondage et en discutons, et nous abordons le choix de la valeur optimale du paramètre de composition p pour le domaine des usagers d'un téléphone fixe ainsi que d'un téléphone mobile (F-et-M). Nous illustrons nos méthodes en les appliquant à la *National Immunization Survey* commanditée par les *Centers for Disease Control and Prevention*.

Mots-clés : Enquête à base de sondage double; répartition optimale; plan de sondage; *National Immunization Survey*.

1 Introduction

Aux États-Unis, les enquêtes téléphoniques par composition aléatoire (CA) modernes font usage de deux échantillons : un échantillon de lignes de téléphone fixes (« échantillon de lignes fixes ») et un échantillon de lignes de téléphone mobiles (« échantillon de lignes mobiles »). Les fondements statistiques de ces enquêtes téléphoniques à base de sondage double ont été posés par Wolter, Smith et Blumberg (2010). Le présent article s'inscrit dans le prolongement de ces travaux et rend compte des méthodes statistiques et des éléments à prendre en considération pour allouer les ressources de l'enquête aux deux bases de sondage.

Parce que son coût à l'unité est plus faible et que son usage est établi de plus longue date, il est fréquent que l'échantillon de lignes fixes soit le plus grand des deux échantillons et que l'on tente de réaliser l'interview de l'enquête auprès de tous ses répondants. Pour l'échantillon de lignes mobiles, plus petit, le protocole d'interview prévoit deux modes d'exécution : 1) entreprendre l'interview de l'enquête auprès de tous les répondants, ou 2) faire une brève interview de présélection pour déterminer la situation d'usage du téléphone du répondant, et n'entreprendre ensuite l'interview que si le répondant rentre dans la catégorie utilisant exclusivement un téléphone mobile (EXM) (c'est-à-dire les répondants qui déclarent à l'interview de présélection qu'il n'y a pas de téléphone fixe dans leur ménage). (La démarche de présélection comporte des variantes, telles qu'interviewer à la fois les répondants EXM et ceux qui déclarent que leur ménage est doté d'une ligne de téléphone fixe, mais qu'ils ne sont pas joignables par ce moyen). Dès lors que la taille de la population utilisant exclusivement un téléphone fixe (EXF) (c'est-à-dire les personnes dont le ménage est équipé d'un téléphone fixe, mais qui n'ont pas accès à un téléphone mobile) va décroissant (Blumberg et Luke 2010), les statisticiens d'enquête pourraient envisager de

1. Kirk M. Wolter, Xian Tao et Robert Montgomery, NORC, University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603. Courriel : wolter-kirk@norc.org; Philip J. Smith, Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Disease, Immunization Services Division, MS A-19, 1600 Clifton Road, NE, Atlanta, GA 30333. Courriel : pzs6@cdc.gov.

nouveaux plans de sondage où l'échantillon de lignes mobiles tiendrait lieu du grand échantillon dans lequel on interroge tous les répondants, tandis que le protocole d'interview pour le petit échantillon de lignes fixes comporterait des modalités de présélection ou d'interview exhaustive de tous les répondants. Ici, cependant, nous nous intéressons surtout à la situation qui prévaut depuis plusieurs années, où l'échantillon de lignes mobiles est le petit échantillon dans lequel l'interview des répondants se fait soit de façon exhaustive soit après présélection, selon ce que prévoit le protocole.

Les méthodes de répartition optimale que nous allons élaborer s'appuieront sur les hypothèses idéales selon lesquelles la taille d'échantillon est égale au nombre d'interviews achevées (absence de non-réponse), qu'il existe essentiellement une correspondance biunivoque entre les unités d'échantillonnage (numéros de téléphone) et les unités analytiques (c'est-à-dire les ménages) dans la population de lignes fixes, qu'il existe essentiellement une correspondance biunivoque entre les unités d'échantillonnage et les unité analytiques dans la population de lignes mobiles, et que toutes les unités de la population cible sont incluses dans au moins l'une des deux bases de sondage. Ainsi, selon ces hypothèses, toute unité analytique est liée à une ligne de téléphone fixe, à une ligne de téléphone mobile, ou à une ligne fixe ainsi qu'une ligne mobile, sans qu'elle soit liée à plus d'une ligne fixe et à plus d'une ligne mobile.

À l'instar des études de Hartley (1962, 1974), Fuller et Burmeister (1972), Skinner et Rao (1996), et Lohr et Rao (2000, 2006), la littérature sur les enquêtes à base de sondage double porte en grande partie sur les procédures d'estimation, et non sur la question de la répartition de la taille de l'échantillon entre les différentes bases de sondage. Biemer (1984), ainsi que Lepkowski et Groves (1986) considèrent la répartition dans le cas des bases de sondage emboîtées, comme elles peuvent l'être lorsqu'on dispose d'une base aréolaire et d'une liste spéciale supplémentaire.

Tout d'abord, fixons la notation et posons les hypothèses. Soit U^A la population des lignes de téléphone fixes et U^B celle des lignes de téléphone mobiles. La population d'intérêt totale est $U = U^A \cup U^B$. Certaines unités ont à la fois une ligne de téléphone fixe et un téléphone mobile (la population F-et-M), tandis que d'autres n'ont qu'une ligne de téléphone fixe (la population EXF), et d'autres encore n'ont qu'un téléphone mobile (la population EXM). Les deux populations se recoupent donc entre elles comme il suit : $U^{ab} = U^A \cap U^B, U^a = U^A - U^{ab}$, et $U^b = U^B - U^{ab}$. U^a est le domaine EXF, U^b est le domaine EXM, et U^{ab} est le domaine F-et-M. Les tailles des populations sont : $N_A = \text{card}\left(U^A\right), N_B = \text{card}\left(U^B\right), N_{ab} = \text{card}\left(U^{ab}\right), N_a = \text{card}\left(U^a\right)$ et $N_b = \text{card}\left(U^b\right)$. Nous désignons par $\alpha = N_{ab}/N_A$ (resp. $\beta = N_{ab}/N_B$) la proportion de la sous-population mixte (c'est-à-dire de la population F-et-M) relativement à la population U^A (resp. U^B).

Soit s_A un échantillon aléatoire simple tiré sans remise dans U^A, s_B un échantillon aléatoire simple tiré sans remise dans U^B, et $n_A = \text{card}(s_A)$, $n_B = \text{card}(s_B)$ les tailles des échantillons respectifs (c'est-à-dire, les nombre d'interviews achevées). Nous supposons que l'appartenance au domaine (a, ab, b) n'est pas connue au moment du tirage.

Soit Y_i une variable d'intérêt pour la i^{e} unité de la population totale. Les moyennes et les composantes de variance relatives aux domaines de population sont désignées par $\overline{Y}_A, \overline{Y}_B, \overline{Y}_{ab}, \overline{Y}_a, \overline{Y}_b, S_A^2, S_B^2, S_{ab}^2, S_a^2$ et S_b^2. Nous supposons que l'objet du sondage est d'estimer le total Y sur toute la population.

Dans la suite de l'exposé, nous calculons la répartition optimale sous les protocoles d'interview exhaustive et de présélection présentés à la section 2 et à la section 3, respectivement. À la section 4, nous

comparons les deux protocoles sur le plan de l'efficience et du coût, et nous donnons des indications quant aux conditions dans lesquelles l'un serait meilleur que l'autre. Nous examinons aussi le choix de la valeur optimale d'un paramètre de composition p, qui sert à combiner les estimateurs des deux échantillons $\left(s_A \cap U^{ab}\right.$ et $\left. s_B \cap U^{ab}\right)$ qui représentent la population F-et-M. À la section 5, nous appliquons ces méthodes à la *National Immunization Survey*, une grande enquête téléphonique à base de sondage double commanditée par les *Centers for Disease Control and Prevention* (CDC). Un bref sommaire conclut l'article à la section 6.

2 Le protocole d'interview exhaustive

Selon le protocole d'interview exhaustive, l'interview de l'enquête est réalisée auprès de toutes les unités de l'échantillon s_A ainsi que de l'échantillon s_B. Par conséquent, les coûts variables de la collecte des données sont approximés par la formule :

$$C_{TA} = c_A n_A + c_B n_B, \tag{2.1}$$

où c_A est le coût unitaire d'une interview effectuée dans l'échantillon s_A et c_B est le coût unitaire correspondant dans l'échantillon s_B. Les nombres espérés d'interviews d'enquête dans l'échantillon de lignes mobiles sont de $(1-\beta)n_B$ unités EXM et de βn_B unités F-et-M.

L'estimateur sans biais de la population totale (Hartley 1962) est donné par :

$$\ddot{Y} = \hat{Y}_a + p\hat{Y}_{ab} + q\hat{Y}_{ba} + \hat{Y}_b, \tag{2.2}$$

où p est un paramètre de composition, $q = 1 - p$, $\hat{Y}_a = (N_A / n_A)\, y_a$ est un estimateur du total pour les unités EXF, $\hat{Y}_{ab} = (N_A / n_A)\, y_{ab}$ est un estimateur du total pour les unités F-et-M obtenu à partir de l'échantillon de lignes fixes, $\hat{Y}_{ba} = (N_B / n_B)\, y_{ba}$ est un estimateur du total pour les unités F-et-M obtenu à partir de l'échantillon de lignes mobiles, $\hat{Y}_b = (N_B / n_B)\, y_b$ est un estimateur du total pour les unités EXM, y_a est la somme des valeurs observées pour la variable d'intérêt dans s_A et dans le domaine U^a, y_{ab} est la somme des valeurs observées pour la variable d'intérêt dans s_A et dans le domaine U^{ab}, y_{ba} est la somme valeurs observées pour la variable d'intérêt dans s_B et dans le domaine U^{ab}, et y_b est la somme des valeurs observées pour la variable d'intérêt dans s_B et dans le domaine U^b. Nous examinerons le choix de p à la section 4.

Pour une valeur fixée de p, la variance de $\ddot{Y}$ se calcule par la formule :

$$\text{Var}\left\{\ddot{Y}\right\} = N^2 \left(\frac{Q_A^2}{n_A} + \frac{Q_B^2}{n_B} \right), \tag{2.3}$$

où $W_A = N_A / N$, $W_B = N_B / N$,

$$Q_A^2 = W_A^2 \left\{ (1-\alpha) S_a^2 + \alpha p^2 S_{ab}^2 + \alpha(1-\alpha)\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2 \right\},$$

et

$$Q_B^2 = W_B^2 \left\{(1-\beta) S_b^2 + \beta q^2 S_{ab}^2 + \beta(1-\beta)\left(\overline{Y}_b - q\overline{Y}_{ab}\right)^2\right\}.$$

La répartition optimale classique de l'échantillon total entre les deux bases de sondage (Cochran 1977) est définie par :

$$\begin{aligned} n_{A,opt} &= \frac{KQ_A}{\sqrt{c_A}} \\ n_{B,opt} &= \frac{KQ_B}{\sqrt{c_B}}, \end{aligned} \tag{2.4}$$

où K est une constante qui prend une valeur différente selon qu'on cherche à réduire le plus possible les coûts en respectant des contraintes de variance ou à réduire le plus possible la variance en respectant des contraintes de coût. Si l'on se fixe un coût C_{TA} à ne pas dépasser, la variance minimum est :

$$\min\left[\mathrm{Var}\left\{\ddot{Y}\right\}\right] = \frac{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}{C_{TA}}, \tag{2.5}$$

tandis que le coût minimum si la variance est fixée à V_0 est

$$\min\left[C_{TA}\right] = \frac{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}{V_0}. \tag{2.6}$$

3 Protocole de présélection

Dans le protocole de présélection, on procède à l'interview de toutes les unités de l'échantillon de lignes fixes s_A. On réalise une interview de présélection auprès de toutes les unités de l'échantillon de lignes mobiles s_B (pour déterminer la situation d'usage du téléphone), puis on effectue l'interview de l'enquête uniquement auprès des unités retenues comme étant EXM. Par conséquent, les coûts attendus de la collecte des données suivent le modèle :

$$\begin{aligned} C_{SC} &= c_A n_A + c_B' \beta n_B + c_B'' (1-\beta) n_B \\ &= c_A n_A + c_B''' n_B, \end{aligned} \tag{3.1}$$

où c_B' est le coût de l'interview de présélection d'une unité de l'échantillon s_B (en vue de déterminer sa situation d'usage du téléphone), c_B'' est le coût de l'interview de présélection et de l'interview d'enquête d'une unité de l'échantillon s_B et $c_B''' = c_B'\beta + c_B''(1-\beta)$. Dans cette notation, n_A est le nombre d'interviews d'enquête achevées auprès des répondants de l'échantillon de lignes fixes et n_B est le

nombre d'interviews achevées (interview de présélection seulement pour les répondants non EXM, et interviews de présélection et d'enquête pour les répondants EXM) auprès des répondants de l'échantillon de lignes mobiles. Cela veut dire que le nombre espéré d'interviews d'enquête achevées est $n_A + (1 - \beta) n_B$.

L'estimateur sans biais du total pour l'ensemble de la population est :

$$\hat{Y} = \hat{Y}_A + \hat{Y}_b, \tag{3.2}$$

où $\hat{Y}_A = (N_A / n_A) y_A$, $\hat{Y}_b = (N_B / n_B) y_b$, et $y_A = y_a + y_{ab}$. La variance de cet estimateur est :

$$\text{Var}\left\{\hat{Y}\right\} = N^2 \left(\frac{R_A^2}{n_A} + \frac{R_B^2}{n_B} \right), \tag{3.3}$$

où

$$R_A^2 = W_A^2 S_A^2$$

et

$$R_B^2 = W_B^2 S_b^2 \left\{ 1 - \beta + \beta (1 - \beta) \frac{\overline{Y}_b^2}{S_b^2} \right\}.$$

La répartition optimale de l'échantillon total est :

$$\begin{aligned} n_{A,\,opt} &= L R_A / \sqrt{c_A} \\ n_{B,\,opt} &= L R_B / \sqrt{c_B'''}, \end{aligned}$$

où L est une constante qui dépend de la contrainte retenue : contrainte de coût ou de variance. La variance minimale à coût total fixe est :

$$\min\left[\text{Var}\left\{\hat{Y}\right\}\right] = \frac{\left(\sqrt{c_A} R_A + \sqrt{c_B'''} R_B\right)^2}{C_{SC}}, \tag{3.4}$$

et le coût minimal à variance fixe est :

$$\min\left[C_{SC}\right] = \frac{\left(\sqrt{c_A} R_A + \sqrt{c_B'''} R_B\right)^2}{V_0}. \tag{3.5}$$

4 Comparaison du protocole d'interview exhaustive et du protocole de présélection

Nous allons comparer le protocole d'interview exhaustive et le protocole de présélection pour déterminer lequel est le moins coûteux ou le plus efficient. Ce genre de comparaison peut fournir des indications pratiques aux concepteurs des futures enquêtes téléphoniques à base de sondage double.

4.1 Comparaison des variances minimales et des coûts minimaux

Si l'on fixe le coût ou la variance, le rapport suivant donne une mesure d'appréciation de l'efficience :

$$E = \frac{\min\left[\mathrm{Var}\left\{\hat{Y}\right\}\right]}{\min\left[\mathrm{Var}\left\{\ddot{Y}\right\}\right]} = \frac{\min\left[C_{SC}\right]}{\min\left[C_{TA}\right]} = \frac{\left(\sqrt{c_A}R_A + \sqrt{c_B'''}R_B\right)^2}{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}. \quad (4.1)$$

Les valeurs inférieures à 1,0 favorisent l'approche de présélection, tandis que les valeurs supérieures à 1,0 favorisent l'approche exhaustive.

Nous allons illustrer l'efficience à l'aide de six scénarios concernant une enquête auprès d'une population hypothétique d'adultes. Pour tous les scénarios, la taille de la population est tirée de la *Current Population Survey* de mars 2010 (http://www.census.gov/cps/data/), et les proportions de population par situation d'usage du téléphone proviennent de la *National Health Interview Survey* couvrant la période de janvier à juin 2010 (Blumberg et Luke 2010). Les valeurs sont $N_A = 83\,451\,980$, $N_a = 15\,162\,402$, $N_{ab} = 68\,289\,578$, $N_b = 31\,265\,108$, $N_B = 99\,554\,686$, $\alpha = 0{,}818$, et $\beta = 0{,}686$. Pour tous les scénarios, l'objet de l'enquête est l'estimation du nombre total d'adultes ayant un certain attribut.

Les hypothèses propres aux scénarios sont présentées dans le tableau suivant :

Tableau 4.1
Définition des six scénarios pour une population hypothétique d'adultes

Scénarios	$\bar{Y}_A$	$\bar{Y}_a$	$\bar{Y}_{ab}$	$\bar{Y}_b$	$\bar{Y}_B$
1	0,791	0,750	0,800	0,750	0,784
2	0,759	0,800	0,750	0,750	0,750
3	0,500	0,500	0,500	0,500	0,500
4	0,518	0,600	0,500	0,400	0,469
5	0,209	0,250	0,200	0,250	0,216
6	0,241	0,200	0,250	0,250	0,250

Les moyennes correspondent aux proportions d'adultes dotés de l'attribut. Le scénario 1 concerne une population où les moyennes des domaines sont similaires, la moyenne du domaine F-et-M étant légèrement supérieure aux moyennes des populations EXM et EXF. Le scénario 2 concerne une population où la moyenne du domaine EXF est légèrement supérieure aux moyennes des domaines des autres situations d'usage du téléphone. Le scénario 3 correspond à une population où les moyennes de tous les domaines de situation d'usage du téléphone sont égales. Le scénario 4 correspond à une population où la moyenne du domaine EXF est beaucoup plus grande que la moyenne du domaine EXM. Les scénarios 5 et 6 correspondent aux scénarios 1 et 2, respectivement, sauf que les moyennes sont égales à un (1) moins la moyenne correspondante. La moyenne du domaine EXM décroît du scénario 1 au scénario 6.

Nous avons sélectionné les six scénarios pour illustrer les diverses conditions dans lesquelles les moyennes des domaines EXM, EXF et F-et-M diffèrent. Des différences peuvent survenir du fait que les jeunes adultes, les Hispaniques, les adultes en simple cohabitation sans lien de parenté, les adultes locataires, et les adultes qui vivent dans la pauvreté ont tendance à appartenir à la catégorie EXM (Blumberg et Luke 2013). Pour se faire une idée plus précise des efficiences relatives de la démarche de

l'interview exhaustive et de celle de la présélection, les concepteurs des futures enquêtes pourront reprendre nos calculs sur des scénarios de leur cru qu'il leur appartiendra de préciser, quitte à les adapter aux conditions particulières de leurs applications.

Les six scénarios seront évalués à l'aune de trois structures de coûts hypothétiques. Ces structures de coûts servent à éclairer les diverses conditions dans lesquelles le coût unitaire de la présélection est élevé ou peu élevé par rapport au coût unitaire de l'interview de l'enquête, les structures de coûts 1 à 3 prenant en compte, dans cet ordre, des coûts de présélection relatifs croissants. Tous les éléments de coût sont exprimés en heures d'interview :

Structure de coût 1 : $c_B' = 0{,}05$, $c_B'' = 2{,}05$, $c_B = 2{,}00$ et $c_A = 1{,}00$

Structure de coût 2 : $c_B' = 0{,}20$, $c_B'' = 2{,}20$, $c_B = 2{,}00$ et $c_A = 1{,}00$

Structure de coût 3 : $c_B' = 0{,}50$, $c_B'' = 2{,}50$, $c_B = 2{,}00$ et $c_A = 1{,}00$.

Tous ces scénarios tiennent compte du fait qu'une interview par téléphone mobile prend deux fois plus de temps par répondant qu'une interview par téléphone fixe.

Les courbes d'efficience correspondant aux différents scénarios pour la première structure de coût sont tracées dans la figure 4.1. Nous avons dressé des graphiques similaires pour les deuxième et troisième structures de coût, mais nous ne pouvons les présenter ici, faute d'espace.

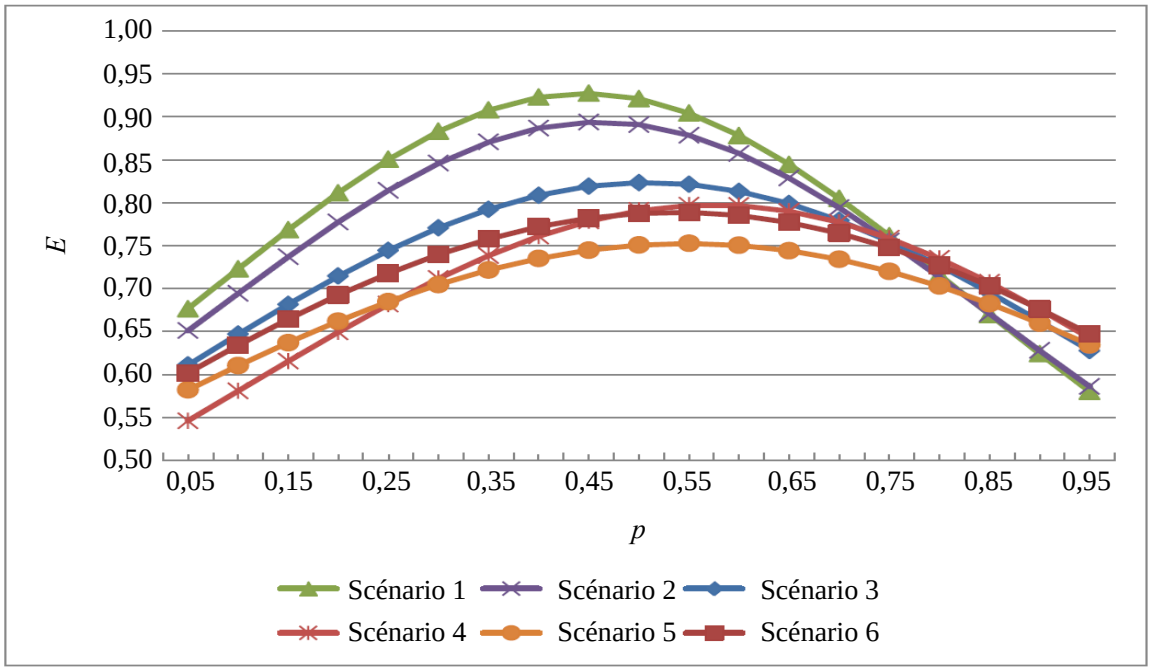

Figure 4.1 Courbe de l'efficience E en fonction du paramètre p pour la structure de coût 1.

Pour la structure de coût 1, c'est l'approche de présélection qui donne la variance la plus faible à coût fixe, et ce, pour les six scénarios. Pour la structure de coût 3, où le coût de présélection unitaire est beaucoup plus élevé en termes relatifs que dans la structure de coût 1, l'approche exhaustive donne la variance la plus faible dans la moitié des scénarios de population. Pour la structure de coût 2, qui comprend un niveau intermédiaire de coût de présélection, l'approche de présélection l'emporte sur l'approche exhaustive pour tous les scénarios, sauf le scénario 1, où les deux approches sont d'efficience presque égale.

L'expression de l'efficience E donnée en (4.1) permet d'appréhender la comparaison du protocole d'interview exhaustive et du protocole de présélection. Le coût de présélection unitaire n'intervient que dans le terme $\sqrt{c_B'''}R_B$ du numérateur de E. Pour un scénario donné donc, si le coût de présélection augmente, la valeur de E ne peut qu'augmenter. Si les coûts de présélection sont faibles, E peut être inférieur à 1,0, auquel cas on privilégiera le protocole de présélection. De même, des coûts de présélection élevés peuvent faire en sorte que E dépasse 1,0, auquel cas on devra privilégier le protocole d'interview exhaustive.

Il est aussi utile d'étudier le sens de la variation de l'efficience E en fonction des moyennes de domaine (c'est-à-dire des proportions de domaine) pour une structure de coût donnée. L'identité (4.1) et les définitions des composantes de la variance montrent bien que, tant que les écarts entre les moyennes de domaine restent raisonnables, la variation des moyennes de domaine $\overline{Y}_b, \overline{Y}_{ab}$ et $\overline{Y}_a$ n'aura que peu d'incidence, relativement, sur Q_A^2, Q_B^2 et R_A^2, de sorte que E aura tendance à varier directement en fonction de R_B^2, lequel dépend à son tour du rapport $\overline{Y}_b^2 / S_b^2$ du domaine EXM. Plus la moyenne est faible dans le domaine EXM, plus le rapport sera petit, et plus E le sera aussi, par contrecoup. Ainsi, pour chaque structure, E prend des valeurs plus faibles dans les scénarios 5 et 6 que dans les scénarios 1 et 2, et des valeurs moyennes dans les scénarios 3 et 4, comme on peut le constater.

Pour le protocole d'interview exhaustive, on obtient les p optimaux aux points où les courbes d'efficience passent par leur valeur maximale. Le tableau 4.2 donne les tailles d'échantillon optimales et les paramètres p optimaux pour chaque scénario et structure de coût, dans l'hypothèse d'un budget fixe de 1 000 heures d'interview. Pour le protocole de présélection, nous pouvons espérer effectuer en moyenne $(1 - \beta)\, n_B$ interviews par téléphone mobile. Pour tous les scénarios de population et toutes les structures de coût étudiés ici, le nombre d'interviews achevées par téléphone mobile est plus faible dans le cas du protocole de présélection que dans celui du protocole d'interview exhaustive. Ce dernier utilise des ressources pour interroger des cas F-et-M dans les deux échantillons et requiert un plus grand nombre d'interviews par téléphone mobile pour bien représenter les cas EXM. Par ailleurs, si le protocole de présélection est plus efficient pour interviewer les cas EXM, il nécessite, en contrepartie, des ressources pour les interviews de présélection. Les p optimaux se situent dans l'intervalle de valeurs de 0,4 à 0,6, et la variance dans le cas du protocole d'interview exhaustive tombe bien dans cet intervalle. Nous examinerons cette question à la section 4.2.

En résumé, ces illustrations nous permettent de conclure que l'approche de présélection est souvent plus efficiente que l'approche exhaustive. Une augmentation du coût de la présélection par rapport au coût de l'interview peut toutefois faire pencher la balance en faveur de l'approche exhaustive. On privilégiera cette dernière dans les enquêtes où la présélection coûte très cher, en termes relatifs; autrement, la préférence ira à l'approche de présélection. L'approche de présélection aura tendance à être plus efficiente pour les petites valeurs de la moyenne du domaine EXM que pour les grandes valeurs de cette moyenne.

Tableau 4.2
Tailles d'échantillon et valeurs optimales de p pour l'approche exhaustive et l'approche de présélection

Structure de coût	Approche de présélection				Approche exhaustive	
	n_A	n_B	$(1-\beta)n_B$	p_{opt}	n_A	n_B
			Scénario 1			
1	494	747	234	0,45	337	331
2	469	641	201	0,45	337	331
3	431	505	159	0,45	337	331
			Scénario 2			
1	506	728	229	0,45	339	330
2	481	626	197	0,45	339	330
3	443	494	155	0,45	339	330
			Scénario 3			
1	583	615	193	0,50	344	328
2	559	533	167	0,50	344	328
3	520	425	134	0,50	344	328
			Scénario 4			
1	605	582	183	0,55	377	312
2	581	506	159	0,55	377	312
3	543	405	127	0,55	377	312
			Scénario 5			
1	606	581	182	0,55	358	321
2	582	505	159	0,55	358	321
3	544	404	127	0,55	358	321
			Scénario 6			
1	618	563	177	0,55	354	323
2	594	490	154	0,55	354	323
3	557	393	123	0,55	354	323

4.2 Choix du paramètre de composition p pour le protocole d'interview exhaustive

La répartition optimale étant fonction du paramètre de composition, il importe d'aborder le choix de ce paramètre. Nous avons vu que la variance est peu sensible au choix de p dans un voisinage raisonnable du p optimal. Même si la valeur exacte du p optimal n'est jamais connue en pratique, nous allons présenter dans cette partie une méthode pratique que les statisticiens peuvent utiliser pour sélectionner une valeur raisonnable de p qui se rapproche de l'optimum.

L'échantillon de lignes mobiles et l'échantillon de lignes fixes fournissent chacun un estimateur de départ du total du domaine F-et-T. Le paramètre p sert ensuite à composer un estimateur aux performances supérieures pour ce domaine en combinant les estimateurs de départ. Si l'estimateur obtenu à partir de l'échantillon de lignes fixes est le plus précis des deux estimateurs de départ, p devrait être relativement grand, et inversement, si l'estimateur obtenu à partir de l'échantillon de lignes mobiles l'emporte en précision, alors $q = 1 - p$ devrait être relativement grand. Il est donc logique, d'un point de vue statistique, de considérer la valeur de p qui est proportionnelle à l'espérance de la taille de l'échantillon dans le domaine du double usage, c'est-à-dire $p_o = \alpha n_{A,\text{opt}} \big/ \left(\alpha n_{A,\text{opt}} + \beta n_{B,\text{opt}}\right)$, où la répartition optimale est fondée sur ce choix de p. Ainsi, p_o est une solution de l'équation

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{(1-\alpha) S_a^2 + \alpha p^2 S_{ab}^2 + \alpha (1-\alpha) \left(\overline{Y}_a - p \overline{Y}_{ab}\right)^2}{(1-\beta) S_b^2 + \beta (1-p)^2 S_{ab}^2 + \beta (1-\beta) \left\{\overline{Y}_b - (1-p) \overline{Y}_{ab}\right\}^2}, \tag{4.2}$$

et $n_{A,opt}$ et $n_{B,opt}$ sont définis à leur tour en fonction de p_o.

L'équation (4.2) fait apparaître que p_o dépend de la variable d'intérêt y. Or, cette dépendance rend ce choix de p_o inopérant en pratique, puisqu'il en résultera des tailles d'échantillon et des poids de sondage différents selon la variable d'intérêt. Pour avoir une solution pratique, on peut donc envisager de prendre le p_o qui correspond à la variable étudiée $y \equiv 1$ (le total de la population correspondant à cette variable est simplement le nombre total d'unités qui existent dans les deux bases de sondage, chaque unité n'étant comptée qu'une seule fois). Dans cette approche, p_o est une solution de l'équation :

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{\alpha (1-\alpha)(1-p)^2}{\beta (1-\beta) p^2} \ . \tag{4.3}$$

Pour les structures de coût considérées dans cette section, le p_o correspondant est 0,52. Dans la figure 4.1, on voit que cette valeur est très proche du p optimal pour les divers scénarios, sans qu'il y ait de perte sensible d'efficience. Une autre façon de procéder serait d'évaluer (4.2) pour un petit ensemble formé des éléments les plus importants de l'enquête, de choisir une valeur de p qui réalise un bon compromis, et de définir ensuite la répartition optimale en fonction de cette valeur de compromis.

5 Exemple : la *National Immunization Survey*

5.1 Introduction

Les CDC commanditent la *National Immunization Survey* (NIS) depuis 1994 pour surveiller l'état de la vaccination des jeunes enfants âgés de 19 à 35 mois. La NIS comprend deux phases de collecte des données, à savoir une enquête téléphonique de type CA à base double auprès des ménages ayant des enfants dans la tranche d'âge visée, suivie d'une enquête postale auprès des prestataires de vaccination de ces enfants en vue d'obtenir l'historique de la vaccination de ces enfants pour chacun des vaccins recommandés. Pour chaque enfant, on compare le nombre de doses déclarées par le prestataire au nombre de doses recommandées, pour déterminer si l'enfant est sans retard de vaccination (SRV). Pour des renseignements sur la NIS, on peut consulter Smith, Hoaglin, Battaglia, Khare et Barker (2005) et le *Data User's Guide* de 2011 (CDC 2012).

Considérons à présent la NIS, telle qu'elle a été menée en 2011. L'interview principale consistait en six parties. La partie S, celle du début, est un bref questionnaire destiné à déterminer si le ménage a des enfants dans la tranche d'âge visée. Si ce n'est pas le cas, on met fin à l'interview. Pour les répondants retenus qui ont un carnet de vaccination, la partie A recueille l'historique de la vaccination de l'enfant déclaré par le ménage. Pour tous les autres répondants, la partie B recueille des informations plus limitées et moins spécifiques sur les vaccinations des enfants. La partie C recueille les caractéristiques démographiques des enfants, de la mère, et du ménage. La partie D recueille les noms et les coordonnées

des prestataires de vaccination et sollicite le consentement des parents à ce que les prestataires soient contactés, tandis que la partie E recueille l'information sur la couverture actuelle de l'assurance-maladie.

5.2 La répartition optimale pour la NIS

La NIS est destinée à produire des estimations à l'échelle nationale et pour 56 secteurs d'estimation disjoints consistant en 46 États entiers, 6 grands secteurs urbains, et 4 secteurs de type « reste de l'État ». Chacun de ces secteurs constitue une strate dans le plan de sondage de la NIS. Pour chacun de ces secteurs, la NIS est censée réduire le plus possible le coût du sondage sous la contrainte de variance suivante : le coefficient de variation (CV) de l'estimateur du taux de couverture vaccinale (proportion des enfants SRV dans la population des enfants dans la tranche d'âge visée) doit être de 7,5 % à l'échelon du secteur d'estimation, lorsque le vrai taux est de 50 %.

Dans le protocole d'interview exhaustive, on soumet tous les répondants des deux échantillons à l'interview d'enquête en six parties. Dans le protocole de présélection, on soumet, d'une part, tous les répondants de l'échantillon de lignes fixes à l'interview de l'enquête, et d'autre part, on procède en deux étapes pour ce qui est de l'échantillon de lignes mobiles : i) on fait une interview de présélection pour déterminer la situation d'usage du téléphone du répondant et ii) on fait passer l'interview en six parties susmentionnée. Les usagers F-et-M sont éliminés de l'échantillon de lignes mobiles.

Pour illustrer la répartition optimale, nous supposerons que les coûts unitaires sont proportionnels aux valeurs suivantes : $c'_B = 0{,}06$, $c''_B = 2{,}03$, $c_B = 1{,}96$ et $c_A = 1{,}00$. Les interviews par téléphone mobile demandent, en gros, deux fois plus de temps de travail que les interviews par téléphone fixe. Nous supposons les proportions de population suivantes concernant les enfants admissibles, selon la situation d'usage du téléphone : $W_A = 0{,}59$, $W_a = 0{,}08$, $W_{ab} = 0{,}51$, $W_b = 0{,}41$, $W_B = 0{,}92$, $\alpha = 0{,}86$ et $\beta = 0{,}55$. Nous avons calculé ces proportions à partir de la NIS de janvier 2010.

Pour estimer le taux de couverture vaccinale dans le cas de l'approche exhaustive, nous nous servirons de la variable

$$Y_i = \begin{cases} 1, \text{ si le } i^{\text{e}} \text{ sujet est un enfant SRV dans la tranche d'âge visée} \\ 0, \text{ sinon.} \end{cases}$$

Le taux de couverture vaccinale sera estimé alors par $\ddot{Y}/N_e$, où N_e désigne le nombre d'enfants dans la tranche d'âge visée dans la population (nombre censé être connu à partir des statistiques de l'état civil et des registres connexes). Conformément à la contrainte de variance, nous posons $\overline{Y}_{ae} = \overline{Y}_{abe} = \overline{Y}_{be} = 0{,}5$, où l'indice e attaché à une moyenne de domaine signifie que la moyenne est prise sur les sujets du domaine qui sont dans la tranche d'âge visée. Ainsi, $\overline{Y}_d = \overline{Y}_{de} P_{de}$ et $S_d^2 = \overline{Y}_d \left(1 - \overline{Y}_d\right)$, où $d = a, ab, b$ désigne les trois domaines relatifs à la situation d'usage du téléphone et $P_{de} = N_{de}/N_d$ désigne la proportion d'enfants de la tranche d'âge visée dans le domaine d. Les valeurs $P_{ae} = 0{,}015$, $P_{abe} = 0{,}03$ et $P_{be} = 0{,}05$ sont fondées sur l'expérience de la NIS et traduisent des pourcentages croissants d'enfants dans la tranche d'âge visée en fonction de la situation d'usage du téléphone, dans le sens où les familles qui élèvent des enfants en bas âge ont tendance à avoir un téléphone mobile et à être EXM par surcroît. Par définition, la variance est le carré du coefficient de variation, multiplié par le carré de la proportion de la population. La contrainte de variance s'exprime donc par $\text{Var}\left\{\ddot{Y}/N_e\right\} = 0{,}075^2 \times 0{,}5^2$.

Pour estimer le taux de couverture vaccinale dans le cas de l'approche de présélection, nous considérons la variable

$$Y_i = \begin{cases} 1, & \text{si } i \in s_A, \text{ et est un enfant dans la tranche d'âge visée qui est SRV} \\ 0, & \text{si } i \in s_A, \text{ et n'est pas un enfant dans la tranche d'âge visée ou n'est pas SRV} \\ 1, & \text{si } i \in s_B, \text{ et est EXM et est un enfant dans la tranche d'âge visée qui est SRV} \\ 0, & \text{si } i \in s_B, \text{ et n'est pas EXM ou n'est pas un enfant dans la tranche d'âge visée ou n'est pas SRV.} \end{cases}$$

Dans ces hypothèses, les valeurs du rapport d'efficience E sont inférieures à 1,0 pour toutes les valeurs de p, d'où il vient que l'approche de présélection peut coûter relativement moins cher que l'approche exhaustive. La valeur optimale de p est de 0,39 environ. Cependant, la courbe de E a tendance à s'aplatir au voisinage de l'optimum, si bien que les valeurs de p dans ce voisinage produisent des coûts totaux similaires.

Selon nos hypothèses, la répartition optimale pour le protocole d'interview exhaustive au p optimal donne $n_A = 3\,069$ et $n_B = 7\,437$, ce qui représente 86 interviews NIS pour le compte des enfants dans la tranche d'âge visée de l'échantillon de lignes fixes, et 289 interviews pour le compte des enfants de l'échantillon de lignes mobiles dans la tranche d'âge visée. Pour le protocole de présélection, la répartition optimale donne $n_A = 5\,858$ et $n_B = 8\,432$, ce qui permet d'espérer 164 interviews NIS pour le compte des enfants dans la tranche d'âge visée de l'échantillon de lignes fixes et 188 interviews NIS auprès de ménages EXM pour le compte de leurs enfants dans la tranche d'âge visée. Ces répartitions s'appliquent à un secteur d'estimation type. Le tableau 5.1 affiche les tailles espérées d'échantillon par domaine de situation d'usage du téléphone pour les répartitions optimales données. Dans le cas du protocole de présélection, l'échantillon de lignes mobiles donne en moyenne 4 674 usagers F-et-M, ce qui correspond à la moyenne de 140 enfants dans la tranche d'âge visée (qui ne doivent pas être interviewés et ne sont donc pas inclus dans le tableau).

Tableau 5.1
Tailles d'échantillon attendues par domaine de situation d'usage du téléphone pour les répartitions optimales

Échantillon et domaine d'usage du téléphone	Protocole d'interview exhaustive		Protocole de présélection	
	Taille de l'échantillon attendue	Nombre de sujets attendus dans la tranche d'âge	Taille de l'échantillon attendue	Nombre de sujets attendus dans la tranche d'âge
s_A	3 069	86	5 858	164
s_B	7 437	289	8 432	188
$s_A \cap U^a$	416	6	794	12
$s_A \cap U^{ab}$	2 653	80	5 064	152
$s_B \cap U^{ab}$	4 122	124	4 674	0
$s_B \cap U^b$	3 314	166	3 758	188

Les répartitions optimales que nous venons de présenter ont été calculées dans des conditions idéales, en faisant abstraction de la non-réponse. Pour élaborer un échantillon à des fins pratiques pour la NIS (ou pour n'importe quelle enquête dans la réalité), il faut ajuster les répartitions par les inverses des taux

attendus de coopération à l'enquête et par l'effet de plan de sondage dû à la pondération et à la corrélation intragrappe.

Les données existantes ont beau montrer que, à variance constante, le protocole de sélection est légèrement moins coûteux que le protocole d'interview exhaustive, il n'en demeure pas moins que ce dernier offre à la NIS une plateforme permanente pour mettre à l'essai et comparer les deux protocoles. Les auteurs continuent à suivre la composition d'échantillon obtenue et à mener d'autres études spécialisées sur les erreurs de réponse et de non-réponse.

6 Résumé

Nous avons étudié deux plans de sondage destinés aux enquêtes téléphoniques à base de sondage double, à savoir un protocole d'interview exhaustive où l'on interroge tous les répondants de l'échantillon de lignes mobiles, et un protocole de présélection des répondants de l'échantillon selon leur situation d'usage du téléphone, où seuls les répondants EXM sont ensuite interrogés. Pour chaque protocole, nous avons calculé la répartition optimale des ressources de l'enquête entre les bases de sondage.

Nous avons étudié le problème de l'optimisation de la répartition dans les deux sens conventionnels du terme « optimum » : 1) réduire le plus possible la variance en respectant une contrainte de coût de collecte des données, et 2) réduire le plus possible les coûts de collecte des données en respectant une contrainte de variance. À variance fixe, nous constatons que l'approche de présélection tend à coûter moins cher au total que l'approche exhaustive lorsque le coût unitaire de présélection est faible relativement au coût unitaire de l'interview de l'enquête. L'approche exhaustive peut être moins chère si le coût unitaire de présélection est relativement élevé. De même, à coût total fixe, le protocole de présélection tend à l'emporter en efficience lorsque le coût unitaire de présélection est relativement faible, tandis que le protocole d'interview exhaustive peut l'emporter lorsque le coût unitaire de présélection augmente. L'échantillon de lignes mobiles et l'échantillon de lignes fixes ont tous deux la capacité de produire des estimateurs pour le domaine F-et-M, mais seul l'échantillon de lignes mobiles peut produire des estimateurs pour le domaine EXM. Ainsi, si la présélection ne coûte pas trop cher à l'unité, il faudrait l'utiliser pour avoir le plus grand échantillon possible du domaine EXM. Par contre, si la présélection coûte relativement cher, il vaut mieux éviter l'étape de présélection et investir les ressources de l'enquête dans l'élargissement de l'échantillon à interroger. Ces résultats ont été obtenus dans l'hypothèse d'un sondage aléatoire simple, et ne se transposent pas de façon exacte à d'autres plans de sondage.

L'approche exhaustive donne lieu à deux estimateurs pour le domaine F-et-M, lesquels sont ensuite combinés en appliquant les coefficients p et $1 - p$ aux estimateurs issus des échantillons de lignes fixes et de lignes mobiles, respectivement. Nous avons étudié le choix optimal de p et nous avons donné pour p des expressions de valeurs qui réalisent de bonnes solutions de compromis. Lorsque la variance (ou le coût) est considérée comme une fonction de p, nous avons constaté un aplatissement de la courbe de la fonction au voisinage de l'optimum. La répartition optimale elle-même est fonction de p, et nous avons constaté que la répartition est relativement peu sensible au choix de p dans un grand voisinage du p optimal.

Au moment où nous avons entamé cette étude, avant 2010, la population EXM des États-Unis ne représentait que le cinquième ou le quart de la population totale des ménages. Il était donc légitime, à cette époque, d'envisager un protocole dans lequel le grand échantillon de lignes fixes est interrogé

exhaustivement, tandis que le petit échantillon de lignes mobiles est filtré pour retenir les unités EXM. À l'heure actuelle, cependant, la population EXM représente plus du tiers de la population totale des ménages, et elle continue de croître. Il est raisonnable dans ces conditions d'envisager un nouveau protocole de présélection où l'on filtre l'échantillon de lignes fixes pour ne retenir pour les besoins de l'interview que les répondants EXF. Les répartitions et les résultats obtenus ci-dessus s'appliquent, par symétrie, à ce nouveau protocole.

Nous nous sommes servis de la *National Immunization Survey* de 2011 pour illustrer les répartitions optimales et les deux protocoles d'interviews. L'enquête est conçue de façon à réduire le plus possible le coût, à variance constante. Les résultats de la NIS se limitent à la population d'enfants âgés de 19 à 35 mois. Il n'est pas garanti que l'on obtienne des résultats similaires pour une enquête sur une population générale ou pour une enquête présentant une structure de coût différente.

Remerciements

Les auteurs tiennent à remercier le rédacteur en chef adjoint et les examinateurs pour les suggestions qu'ils ont bien voulu leur formuler pour améliorer la lisibilité du texte. Avertissement : les conclusions et constatations de cet article sont celles des auteurs et ne représentent pas nécessairement le point de vue officiel des *Centers for Disease Control and Prevention*.

Bibliographie

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. *Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07* disponible au www.census.gov/srd/papers/pdf/rr84-07.pdf.

Blumberg, S.J., et Luke, J.V. (2010). Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2010. *National Center for Health Statistics*. Décembre 2010. Disponible au http://www.cdc.gov/nchs/nhis/releases.htm.

Blumberg, S.J., et Luke, J.V. (2013). Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2012. *National Center for Health Statistics*. Juin 2013. Disponible au http://www.cdc.gov/nchs/nhis/releases.htm.

CDC (2012). *National Immunization Survey: A User's Guide for the 2011 Public Use Data File*. Disponible au http://www.cdc.gov/nchs/nis/data_files.htm.

Cochran, W.G. (1977). *Sampling Techniques, 3rd Edition*, New York : John Wiley & Sons, Inc.

Fuller, W.A., et Burmeister, L.F. (1972). Estimators for samples from two overlapping frames. *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Hartley, H.O. (1962). Multiple-frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36, 99-118.

Lepkowski, J.M., et Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association,* 81, 930-937.

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. et Barker, L.E. (2005). Statistical Methodology of the National Immunization Survey: 1994-2002. *National Center for Health Statistics, Vital and Health Statistics,* 2(138).

Wolter, K.M., Smith, P. et Blumberg, S.J. (2010). Fondements statistiques des enquêtes par téléphone mobile. *Techniques d'enquête*, 36, 2, 221-234.