Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une standardisation des stratégies fondées sur la réponse aléatoire

par Andreas Quatember

Décembre 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2009
Vol. 35, N° 2, pp. 153-163
Statistique Canada, N° 12-001-X au catalogue

Une standardisation des stratégies fondées sur la réponse aléatoire

Andreas Quatember [1]

Résumé

Les stratégies fondées sur la réponse aléatoire, qui ont été élaborées au départ à titre de méthodes statistiques destinées à réduire la non-réponse ainsi que la réponse mensongère, peuvent aussi être appliquées dans le domaine du contrôle de la divulgation statistique dans les fichiers de microdonnées à grande diffusion. Le présent article décrit une standardisation des méthodes de réponse aléatoire en vue d'estimer des proportions pour des attributs identificatoires ou sensibles. Les propriétés statistiques de l'estimateur standardisé sont établies dans le cas de l'échantillonnage probabiliste général. Afin d'analyser l'effet du choix des « paramètres de plan » implicites de la méthode sur la performance de l'estimateur, nous incluons dans l'étude des mesures de la protection de la vie privée. Nous obtenons ainsi des paramètres de plan qui optimisent la variance, sachant le niveau de protection de la vie privée. Pour cela, les variables doivent être classées dans diverses catégories de sensibilité. Un exemple fondé sur des données réelles illustre l'application de la méthode à une enquête sur la tricherie chez les étudiants.

Mots clés : Protection de la vie privée ; contrôle de la divulgation statistique ; non-réponse ; réponse mensongère.

1. Introduction

Les cas de refus de répondre ou de donner la vraie réponse sont naturels dans les enquêtes par sondage. Ils peuvent donner lieu à un estimateur des paramètres de population présentant un biais de grandeur inconnue et une forte variance. Par conséquent, un utilisateur sérieux des données ne peut pas ignorer l'existence de la non-réponse et de la réponse mensongère.

Soit U l'univers de N unités de population et U_A, un sous-ensemble de N_A éléments, qui appartiennent à une classe A d'une variable catégorique étudiée. En outre, soit U_A^c le groupe de N_A^c éléments qui n'appartiennent pas à cette classe $(U = U_A \cup U_A^c, U_A \cap U_A^c = \emptyset, N = N_A + N_A^c)$. Soit

$$x_i = \begin{cases} 1 & \text{si l'unité } i \in U_A, \\ 0 & \text{autrement} \end{cases}$$

$(i = 1, 2, ..., N)$ et le paramètre d'intérêt π_A, qui est la taille relative de la sous-population U_A :

$$\pi_A = \frac{\sum_U x_i}{N} = \frac{N_A}{N} \qquad (1)$$

$(\sum_U x_i$ est la notation abrégée de $\sum_{i \in U} x_i)$. Dans le cas d'un échantillon probabiliste s (voir par exemple, Särndal, Swensson et Wretman 1992, page 8f), un estimateur de π_A peut être calculé à partir de l'estimateur d'Horvitz-Thompson de N_A par

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \sum_s \frac{x_i}{\pi_i} \qquad (2)$$

($\pi_i > 0$ est la probabilité que l'unité i soit incluse dans l'échantillon), si la question « Êtes-vous un membre du groupe U_A ? » (ou une question équivalente) est posée directement (dir). Cet estimateur est sans biais si toutes les observations x_i $(i = 1, 2, ..., n)$ sont des réponses sincères. En présence de non-réponse totale ou partielle en ce qui concerne une variable étudiée, l'échantillon s est divisé en un « ensemble de réponses » $r \subset s$ de taille n_r et un « ensemble de réponses manquantes » $m \subset s$ de taille n_m $(s = r \cup m, r \cap m = \emptyset, n = n_r + n_m)$. Dans le cas de variables d'un caractère hautement personnel, embarrassant (comme la toxicomanie, les maladies, le comportement sexuel, la fraude fiscale, l'alcoolisme, la violence familiale ou la criminalité), r est en outre divisé en un ensemble t de n_t unités échantillonnées qui répondent sincèrement, et un ensemble u de taille n_u, d'unités qui répondent de manière mensongère $(r = t \cup u, t \cap u = \emptyset, n_r = n_t + n_u)$. L'estimateur (2) doit alors être réécrit sous la forme :

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \left(\sum_t \frac{x_i}{\pi_i} + \sum_u \frac{x_i}{\pi_i} + \sum_m \frac{x_i}{\pi_i} \right). \qquad (3)$$

Naturellement, les éléments de l'ensemble u ne peuvent pas être identifiés et les x_i de m ne sont pas observables, ce qui introduit des erreurs de mesure et de non-réponse dans l'estimation. Par conséquent, tout doit être fait en vue de maintenir les taux de réponses mensongères et de non-réponses aussi faibles que possible.

Les caractéristiques du plan de sondage, qui ont manifestement une incidence sur la quantité et sur la qualité de l'information demandée aux enquêtés (voir par exemple Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2004, Section 6.7), sont étroitement liées aux préoccupations

1. Andreas Quatember est professeur adjoint à l'IFAS – Département de statistiques appliquées, à l'Université Johannes Kepler de Linz, Altenberger Str. 69, A-4040 Linz, Autriche, Europe. Adresse Internet : www.ifas.jku.at. Courriel : andreas.quatember@jku.at.

de ces derniers quant à la « confidentialité des données » et à la « protection perçue de la vie privée ». La première expression fait référence au désir qu'ont les enquêtés de voir leurs réponses demeurer hors de portée des personnes non concernées, tandis que la deuxième fait référence à leur souhait d'empêcher absolument tout le monde d'avoir accès à l'information. Singer, Mathiowetz et Couper (1993), ainsi que Singer, van Hoewyk et Neugebauer (2003) signalent, à l'occasion de deux enquêtes successives auprès de la population américaine, que plus ces préoccupations sont vives, plus la probabilité de participer à l'enquête est faible (page 470ff et page 375ff).

Que peuvent apporter les statisticiens à ce domaine de recherche important ? Dans le cas de questions sensibles, l'utilisation de *stratégies fondées sur la réponse aléatoire* à l'étape de la conception de l'enquête peut réduire les taux de non-réponses et de réponses mensongères parce qu'elles donnent l'impression d'un accroissement de la protection des renseignements personnels. Une caractéristique commune de ces méthodes est que les questions directes sur le sujet sensible sont remplacées par un questionnaire conçu de telle manière que l'enquêteur n'est pas capable d'identifier la question (sélectionnée aléatoirement) à laquelle l'enquêté a répondu, tout en permettant encore d'estimer le paramètre étudié. L'idée est de réduire de cette façon chez les enquêtés la crainte d'une « révélation » embarrassante et de s'assurer ainsi qu'ils seront disposés à coopérer. Pour atteindre cet objectif, l'enquêté doit comprendre clairement comment la conception du questionnaire protège sa vie privée (voir Landsheer, van der Heijden et van Gils 1999, page 6ff).

Les premiers travaux dans ce domaine ont été publiés par Warner (1965). Dans son questionnaire, chaque personne interrogée devait répondre aléatoirement avec la probabilité p_1 à la question « Êtes-vous membre du groupe U_A ? » ou avec la probabilité $p_2 = 1 - p_1$, à la question alternative « Êtes-vous membre du groupe U_A^c ? » $(0 < p_1 < 1)$. Depuis, différentes méthodes de réponse aléatoire fondées sur divers procédés de randomisation ont été proposées (pour une revue, consulter Chaudhuri et Mukerjee 1987, Nathan 1988, ou Tracy et Mangat 1996). Toutes ces stratégies s'appuient sur des questions ou des réponses sélectionnées aléatoirement, quoique certaines utilisent des procédés de randomisation différents selon que l'enquêté possède ou non un attribut particulier (voir, par exemple, Kuk 1990 ; Mangat 1994 ; Kim et Warde 2005).

Warner (1971) a été le premier à constater que ces méthodes pouvaient aussi s'appliquer pour masquer des ensembles de microdonnées confidentielles afin de permettre leur grande diffusion (voir, ibidem, page 887). Ces ensembles de microdonnées peuvent contenir des variables donnant lieu à l'identification directe des unités étudiées, comme le nom ou un numéro d'identification, mais aussi des variables fournissant des renseignements délicats sur une personne. Afin de protéger les unités étudiées contre la divulgation, il pourrait ne pas suffire de supprimer les variables auxquelles elles sont directement liées, parce que certaines unités pourraient encore être identifiées d'après le reste de leurs enregistrements. Le contrôle de la divulgation statistique n'est rien d'autre qu'un exercice d'équilibre entre la protection de l'anonymat des sujets participant à l'enquête et la préservation de l'information contenue dans les données (voir Skinner, Marsh, Openshaw et Wymer 1994). Les méthodes de masquage des données peuvent être réparties en trois catégories (voir Domingo-Ferrer et Mateo-Sanz 2002 ou Winkler 2004), à savoir 1) le *recodage global* des variables en des catégories moins détaillées ou de plus grands intervalles (voir par exemple, Willenborg et de Waal 1996, page 5f) ou le *recodage local* en utilisant divers scénarios de groupement au niveau de l'unité (voir Hua et Pei 2008, page 215f), 2) la *suppression locale* de certaines variables pour les unités étudiées présentant un risque élevé de réidentification en fixant simplement leur valeur à « manquante » (voir Willenborg et de Waal 1996, page 77) et 3) la *substitution* d'autres valeurs aux valeurs réelles d'une variables.

L'une des stratégies de la troisième catégorie est la *micro-agrégation* des variables (voir Defays et Anwar 1998). Dans ce cas, les vraies valeurs des variables sont, par exemple, triées par taille, puis réparties en (petits) groupes. Pour chaque groupe, des données agrégées sont diffusées au lieu des observations originales. Une autre méthode de ce type est la *permutation des données*, où celles provenant d'unités présentant un risque élevé de réidentification sont interchangées avec des données provenant d'un autre ensemble d'unités étudiées (voir Dalenius et Reiss 1982). Une autre technique de substitution d'information identificatoire ou sensible est l'*ajout d'un bruit* aux valeurs observées, autrement dit l'ajout du résultat d'une expérience aléatoire à chaque données (voir Dalenius 1977 ou Fuller 1993). Enfin, les méthodes de randomisation des réponses peuvent aussi être utilisées pour masquer des variables identificatoires ou délicates. Dans ce cas, soit le masquage des données fournies par les unités échantillonnées est déjà effectué à l'étape de la conception de l'enquête, soit l'organisme statistique applique le mécanisme probabiliste de la méthode avant la diffusion du fichier de microdonnées (voir Rosenberg 1980, Kim 1987, Gouweleeuw, Kooiman, Willenborg et de Wolf 1998, ou van den Hout et van der Heijden 2002).

Toutes les méthodes de contrôle de la divulgation statistique protègent la vie privée des unités étudiées par une perte d'information qui peut être considérée comme le prix à payer pour cette protection. Afin de pouvoir corriger comme il convient le processus d'estimation, l'utilisateur du fichier

de microdonnées doit être informé des détails de la méthode de masquage.

À la section 2 du présent article, nous présentons une nouvelle standardisation des méthodes de réponse aléatoire. En outre, nous établissons les propriétés statistiques de l'estimateur standardisé sous échantillonnage probabiliste général. À la section 3, nous exposons la perspective essentielle de la protection de la vie privée. À la section 4, nous répondons à la question de savoir lequel des cas particuliers inclus dans la standardisation est le plus efficace. À la section 5, nous donnons un exemple fondé sur les données réelles, qui illustre l'application des recommandations de la section 4 dans le contexte d'une enquête sur le comportement de tricherie des étudiants.

2. Standardisation des stratégies fondées sur la réponse aléatoire

Soit la standardisation suivante des stratégies de randomisation des réponses : chaque enquêté doit répondre aléatoirement avec la probabilité

- p_1 à la question « Êtes-vous membre du groupe U_A ? »,
- p_2 à la question « Êtes-vous membre du groupe U_A^c ? » ou
- p_3 à la question « Êtes-vous membre du groupe U_B »

ou reçoit l'instruction de dire simplement

- « oui » avec la probabilité p_4 ou
- « non » avec la probabilité p_5

($\sum_{i=1}^{5} p_i = 1$, $0 \le p_i \le 1$ pour $i = 1, 2, ..., 5$). Les N_B éléments du groupe U_B sont caractérisés par la possession d'un attribut entièrement inoffensif B (par exemple, la saison B de naissance), qui ne devrait pas être relié à la possession ou à la non-possession de l'attribut A. Cette question non sensible sur l'appartenance au groupe U_B a été introduite comme une alternative à la question sur l'appartenance au groupe U_A par Horvitz, Shah et Simmons (1967) afin de réduire encore davantage la perception du caractère sensible de la procédure. $\pi_B = N_B/N$ (avec $0 < \pi_B < 1$) est la taille relative du groupe U_B. π_B et les probabilités $p_1, p_2, ..., p_5$ sont les *paramètres de plan* de notre méthode standardisée de randomisation des réponses.

Soit

$$y_i = \begin{cases} 1 & \text{si l'unité } i \text{ répond « oui »,} \\ 0 & \text{autrement} \end{cases}$$

($i = 1, 2, ..., n$). Pour un élément i la probabilité d'une réponse « oui » sous le plan d'interrogation à réponse aléatoire R est, sachant x :

$$P_R(y_i = 1) = p_1 \cdot x_i + p_2 \cdot (1 - x_i) + p_3 \cdot \pi_B + p_4 = a \cdot x_i + b \qquad (4)$$

avec $a \equiv p_1 - p_2$ et $b \equiv p_2 + p_3 \cdot \pi_B + p_4$. Alors, le terme

$$\hat{x}_i = \frac{y_i - b}{a}$$

est sans biais pour la valeur réelle $x_i (a \neq 0)$. Si nous utilisons ces « substituts » pour x_i (et en émettant l'hypothèse que la coopération des enquêtés est complète), les théorèmes qui suivent s'appliquent :

Théorème 1 : Sous un plan d'échantillonnage probabiliste avec probabilités d'inclusion π_i, nous avons l'estimateur sans biais du paramètre π_A suivant :

$$\hat{\pi}_A = \frac{1}{N} \cdot \sum_s \frac{\hat{x}_i}{\pi_i}. \qquad (5)$$

Théorème 2 : Sous un plan d'échantillonnage probabiliste P, la variance de l'estimateur standardisé $\hat{\pi}_A$ (5) est donnée par

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left(V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b \cdot (1-b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right). \qquad (6)$$

Les preuves de ces deux théorèmes figurent à l'annexe. Le premier terme de la somme comprise entre les parenthèses externes de (6) fait référence à la variance de l'estimateur d'Horvitz-Thompson pour le total $\sum_U x_i$ sous un plan d'échantillonnage probabiliste P quand la question sur l'appartenance au groupe U_A est posée directement. Le deuxième terme de la somme peut être considéré comme le prix que nous devons payer en perte d'exactitude pour la protection de la vie privée offerte par le plan d'interrogation à réponse aléatoire. Apparemment, cette variance peut être estimée sans biais en insérant l'estimateur sans biais $\hat{V}_P(\sum_s x_i/\pi_i)$ pour $V_P(\sum_s x_i/\pi_i)$ et $\sum_s \hat{x}_i/\pi_i^2$ pour $\sum_U x_i/\pi_i$.

Sous échantillonnage aléatoire simple sans remise, par exemple, l'estimateur (5) est donné par

$$\hat{\pi}_A = \frac{\hat{\pi}_y - b}{a} \qquad (7)$$

avec $\hat{\pi}_y = \sum_s y_i / n$, la proportion de réponses « oui » dans l'échantillon. Dans ce cas, la variance (6) de l'estimateur standardisé $\hat{\pi}_A$ est donnée par

$$V(\hat{\pi}_A) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{1}{n} \cdot \left(\frac{b \cdot (1-b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \pi_A \right). \qquad (8)$$

Cette variance théorique est estimée sans biais par

$$\hat{V}(\hat{\pi}_A) = \frac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n - 1} \cdot \frac{N - n}{N} + \frac{1}{n} \cdot \left(\frac{b \cdot (1 - b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \hat{\pi}_A \right). \quad (9)$$

Afin de pouvoir calculer $\hat{\pi}_A$, la question sur l'appartenance à U_A (ou à U_A^c, mais nous ignorerons cette possibilité subséquemment sans perte de généralité) doit être incluse dans le plan d'interrogation à réponse aléatoire avec la probabilité $p_1 > 0$. Il existe, en tout, 16 combinaisons de cette question avec les quatre autres questions ou réponses (voir le tableau 1). Ces combinaisons peuvent être décrites comme des cas particuliers de notre stratégie standardisée de réponse. Par exemple, choisir $p_1 = 1$ mène à l'interrogation directe sur le sujet. Si nous posons que $0 < p_1 < 1$ et $p_2 = 1 - p_1$, le plan d'interrogation standardisé correspond à la procédure de Warner. Pour $0 < p_1 < 1$ et $p_3 = 1 - p_1$, nous obtenons la méthode d'Horvitz et ses collaborateurs avec la probabilité π_B connue (voir Greenberg, Abul-Ela, Simmons et Horvitz 1969). (Pour d'autres cas particuliers, déjà publiés autant que nous sachions, nous renvoyons le lecteur à la colonne « Références » du tableau 1).

Tableau 1
Tous les cas particuliers de la stratégie standardisée fondée sur la réponse aléatoire

Plan	Questions/réponses					Références
	U_A	U_{A^c}	U_B	Oui	Non	
*ST*1	•					Interrogation directe
*ST*2	•	•				Warner (1965)[1]
*ST*3	•		•			Greenberg et coll. (1969)[2]
*ST*4	•			•		
*ST*5	•				•	
*ST*6	•	•	•			
*ST*7	•	•		•		
*ST*8	•	•			•	Quatember (2007)[3]
*ST*9	•		•	•		
*ST*10	•		•		•	Singh, Horn, Singh et Mangat (2003)[4]
*ST*11	•			•	•	Fidler et Kleinknecht (1977)[5]
*ST*12	•	•	•	•		
*ST*13	•	•	•		•	
*ST*14	•	•		•	•	
*ST*15	•		•	•	•	
*ST*16	•	•	•	•	•	

1. Une version à deux degrés a été présentée par Mangat et Singh (1990).
2. Une version à deux degrés a été présentée par Mangat (1992).
3. Il s'agit d'une version à un degré de Mangat, Singh et Singh (1993).
4. Il s'agit d'une version à un degré de Singh, Singh, Mangat et Tracy (1994).
5. Une version à deux degrés a été présentée par Singh, Singh, Mangat et Tracy (1995).

La question que soulèvent directement ces considérations est celle de savoir comment choisir les paramètres de plan de la méthode standardisée de réponse afin de découvrir les stratégies qui donnent les meilleurs résultats. Nous répondons à cette question à la section 4. Mais pour cela, nous devons prendre en considération le niveau de protection de la vie privée, qui varie selon le choix de ces paramètres.

3. Protection de la vie privée

Afin de pouvoir comparer l'efficacité des plans d'interrogation caractérisés par des paramètres de plan différents, il paraît inévitable de mesurer la perte de vie privée induite par ces paramètres. Nous pouvons pour cela utiliser les ratios λ_1 et λ_0 des probabilités conditionnelles qui suivent (voir, par exemple, les « mesures de mise en péril » (*measures of jeopardy*) dans Leysieffer et Warner 1976, page 650) :

$$\lambda_j = \frac{\max[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]}{\min[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]} \quad (10)$$

$(1 \le \lambda_j \le \infty;\ j = 1,\ 0)$.

Pour $j = 1$, (10) fait référence à la protection de la vie privée par rapport à une réponse « oui » et pour $j = 0$, par rapport à une réponse « non ». Pour le plan d'interrogation standardisé, ces « mesures λ » de perte de vie privée sont données par

$$\lambda_1 = \frac{\max[a + b;\ b]}{\min[a + b;\ b]} \quad (11)$$

et

$$\lambda_0 = \frac{\max[1 - (a + b);\ 1 - b]}{\min[1 - (a + b);\ 1 - b]}. \quad (12)$$

$\lambda_1 = \lambda_0 = 1$ indique une protection totale de la vie privée. Cela signifie que la réponse donnée par l'unité répondante ne contient absolument aucune information sur le sujet étudié. Cela s'applique pour $a = 0$. Plus les mesures λ diffèrent de l'unité, plus la réponse figurant dans l'enregistrement contient d'information sur la caractéristique étudiée. Parallèlement, l'efficacité de l'estimation augmente (voir plus bas), mais la protection de l'individu contre l'enquêteur diminue. Dans le cas du plan d'interrogation directe avec $p_1 = 1$, où aucun masquage de la variable n'a lieu, ces mesures sont données par $\lambda_1 = \lambda_0 = \infty$.

Soit $\lambda_{1,\,\mathrm{opt}}$ et $\lambda_{0,\,\mathrm{opt}}$ les valeurs λ maximales de (11) et (12) qui, selon l'organisme statistique, permettent d'obtenir une protection suffisante des enregistrements contre la divulgation. En cas d'utilisation d'une stratégie en vue d'éviter la non-réponse et la réponse mensongère dans les enquêtes, nous pouvons aussi modéliser la volonté des enquêtés à collaborer sous forme d'une fonction de la protection perçue de la vie privée. Si la vie privée des

enquêtés est suffisamment protégée par le procédé de randomisation, nous supposons que leur coopération sera totale. Dépasser les limites $\lambda_{1,\text{opt}}$ et/ou $\lambda_{0,\text{opt}}$ introduirait alors automatiquement de la réponse mensongère et de la non-réponse dans l'enquête et, par conséquent, nous ramènerait à la case de départ. Fidler et Kleinknecht (1977) ont montré, dans leur étude du plan $ST11$ (tableau 1) contenant neuf variables dont les niveaux de sensibilité sont très différents, que leur choix des paramètres de plan ($p_1 = 10/16$, $p_4 = p_5 = 3/16$) donnait une réponse quasiment complète et sincère pour chaque variable, y compris le comportement sexuel (ibidem, page 1048). En insérant ces valeurs dans (11) et (12), nous obtenons $\lambda_1 = \lambda_0 = 13/3$. Ce résultat correspond en gros à ceux qui peuvent être tirés de l'expérience de Soeken et Macready (1982) et en suivant les recommandations faites par Greenberg et coll. (1969). Par conséquent, choisir $\lambda_{1,\text{opt}}$ et/ou $\lambda_{0,\text{opt}}$ proche d'une valeur de 4 pourrait être un bon choix pour la plupart des variables, quand la méthode standardisée de réponse aléatoire est utilisée pour éviter les refus et les réponses mensongères des participants à une enquête.

Sans perte de généralité, supposons subséquemment que nous choisirons les deux catégories de la variable étudiée de telle façon que l'appartenance au groupe U_A soit au moins aussi sensible que l'appartenance au groupe U_A^c ($1 \le \lambda_{1,\text{opt}} \le \lambda_{0,\text{opt}} \le \infty$). Partant de (11) et (12), nous pouvons exprimer les termes a et b au moyen des valeurs λ correspondant à λ_1 et λ_0. Leur somme est donnée par

$$a + b = \frac{1 - \frac{1}{\lambda_0}}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}} \qquad (13)$$

avec

$$b = \frac{\frac{1}{\lambda_1} \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}} \qquad (14)$$

et

$$a = \frac{\left(1 - \frac{1}{\lambda_1}\right) \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}}. \qquad (15)$$

Nous gardons les doubles ratios dans le deuxième membre de (14) et de (15) pour trouver facilement les limites de $\lambda_1 \to \infty$ et $\lambda_0 \to \infty$, respectivement.

Cela signifie que, pour un plan d'échantillonnage donné P, la portée du terme $(b \cdot (1-b)/a^2) \cdot \sum_U (1/\pi_i) + (1 - 2 \cdot b - a/a) \cdot \sum_U (x_i/\pi_i)$ dans l'expression (6) de la variance ne dépend pas d'une valeur unique des paramètres de plan, mais de leur effet agrégé sur la perte de vie privée mesurée par λ_1 et λ_0. Les plans d'interrogation possédant les mêmes valeurs λ ont la même efficacité. Les plans pour lesquels la valeur λ_1 et/ou λ_0 est plus grande sont moins efficaces que ceux pour lesquels la valeur λ est plus faible.

4. Plans d'interrogation optimaux

Le cas particulier de stratégie standardisée fondée sur la réponse aléatoire qui, parmi ceux présentés au tableau 1, sera le plus efficace pour des mesures λ données dépendra du type de risque de réidentification ou de caractère sensible du sujet étudié. Les stratégies $ST5$ et $ST8$ ne peuvent jamais donner les meilleurs résultats, parce qu'elles protègent toujours plus une réponse « non » qu'une réponse « oui ».

Pour une variable non identificatoire (ou non sensible), comme la saison de naissance, où $\lambda_{1,\text{opt}} = \lambda_{0,\text{opt}} = \infty$ s'applique, seul le plan d'interrogation directe ($ST1$ du tableau 1) peut donner les résultats de variance optimaux (voir le tableau 2, qui donne ces valeurs des paramètres de plan qui optimisent la performance de l'estimateur $\hat{\pi}_A$; pour pouvoir utiliser le tableau 2 correctement, la variable catégorique étudiée doit être catégorisée de la manière suivante : C_1 : la variable n'est pas du tout sensible ($\lambda_{1,\text{opt}} = \lambda_{0,\text{opt}} = \infty$) ; C_2 : l'appartenance au groupe U_A est de nature sensible, mais non l'appartenance au groupe U_A^c ($\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}} = \infty$) ; C_3 : l'appartenance aux deux groupes U_A et U_A^c est de nature sensible, mais pas de manière égale ($\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}} < \infty$) ; C_4 : l'appartenance aux groupes U_A et U_A^c est aussi sensible dans un cas que dans l'autre ($\lambda_{1,\text{opt}} = \lambda_{0,\text{opt}} < \infty$), ce qui donne les valeurs des paramètres du plan qui garantissent la meilleure performance de l'estimateur $\hat{\pi}_A$). Bien que les autres plans puissent être utilisés pour ce genre de variables, d'une certaine manière, ils protègent plus que cela n'est nécessaire la vie privée des enquêtés. Le prix à payer est une perte d'exactitude de l'estimation de π_A. Toutefois, pour $p_1 = 1$ ($a = 1$ et $b = 0$), la variance de $\hat{\pi}_A$ (5) devient la formule courante pour l'interrogation directe sous l'hypothèse d'une réponse complète : $V_P(\hat{\pi}_A) = 1/N^2 \cdot V_P(\sum_s x_i/\pi_i)$.

Dans le cas d'une variable pour laquelle seule l'appartenance à U_A, mais non à U_A^c est de nature sensible (par exemple U_A = l'ensemble de toxicomanes l'année précédente, $U_A^c = U - U_A$), nous avons $\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}} = \infty$. Le calcul de (14) et (15) pour $1 < \lambda_1 < \infty$ et $\lambda_0 \to \infty$ donne $a = 1 - b$ et l'introduction de cette expression dans (6) donne l'expression suivante de la variance de l'estimateur :

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left[V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b}{1-b} \cdot \left(\sum_U \frac{1}{\pi_i} - \sum_U \frac{x_i}{\pi_i} \right) \right]. \qquad (16)$$

Si nous recherchons les valeurs des paramètres du plan pour lesquelles la stratégie standardisée de réponse aléatoire peut donner cette variance et pour lesquelles les équations (14) et (15) sont vérifiées, nous constatons que dans ce cas il n'existe qu'une seule solution. Le seul plan d'interrogation capable de donner des résultats optimaux est $ST4$. Les paramètres qui optimisent la variance sont donnés par $p_1 = (\lambda_1 - 1)/\lambda_1$ et $p_4 = 1 - p_1$ (voir le tableau 2). Autrement dit, avec la probabilité $p_1 = (\lambda_1 - 1)/\lambda_1$, on demande à l'enquêté s'il fait partie du groupe U_A et avec la probabilité restante, on lui donne l'instruction de dire « oui ». De cette façon, l'enquêteur ne peut tirer une conclusion que d'une réponse « non » directement à la question non sensible de *non*-possession de la caractéristique A, mais non d'une réponse « oui » à la question de la possession de cet attribut sensible ou identificatoire.

Le plan d'interrogation $ST1$ n'est pas applicable à de tels sujets, parce qu'il ne protège pas du tout la vie privée de l'enquêté dans le cas d'une réponse « oui ». Toutes les autres méthodes protègent une réponse « non » plus qu'il n'est nécessaire. Par conséquent, elles peuvent être utilisées, mais elles ne permettent pas d'atteindre le degré d'efficacité de l'option $ST4$.

Si l'appartenance aux groupes U_A ainsi que U_A^c est de caractère sensible, de sorte que la variable est sensible dans son ensemble (par exemple : U_A = l'ensemble de personnes mariées qui ont eu au moins une relation sexuelle avec leur partenaire la semaine précédente ; la condition $U_A^c = U - U_A$), $\lambda_{1,\text{opt}} \leq \lambda_{0,\text{opt}} < \infty$ s'applique. Dans ce cas, ni l'interrogation directe sur le sujet ni le plan $ST4$ ne peut être utilisé, parce que ces options ne permettent pas de protéger les deux réponses possibles.

Les autres plans sont applicables dans ces conditions, mais le plan de Warner ne permet pas d'atteindre le degré d'efficacité des autres si $\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}}$. Il en est ainsi parce que ce plan protège toujours la vie privée de l'enquêté aussi bien dans le cas d'une réponse « oui » que d'une réponse « non ». Toutefois, si $\lambda_{1,\text{opt}} = \lambda_{0,\text{opt}}$, malgré les allégations faites dans certaines publications dans le passé (voir, par exemple, Greenberg et coll., 1969, page 526f, Mangat et Singh 1990, page 440, Singh et coll., 2002, page 518f), il n'existe *aucune* méthode de réponse aléatoire qui peut donner de *meilleurs* résultats que la méthode $ST2$ de Warner avec les paramètres de plan optimaux p_1 et p_2 conformément au tableau 2. Pour $ST7$, cela n'est valide que si $\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}}$. Par conséquent, $ST7$ est le supplément parfait de $ST2$, pour lequel la situation est tout à fait opposée.

Tous les autres plans du tableau 1, tels que $ST11$ ou $ST14$, peuvent avoir la même efficacité pour $\lambda_{1,\text{opt}} \leq \lambda_{0,\text{opt}} < \infty$ si les paramètres du plan sont choisis conformément aux contraintes (14) et (15). Parmi ces plans, la stratégie de Greenberg et de ses collaborateurs avec π_B connue ($ST3$) a, d'une part, l'avantage par rapport au plan de Warner de donner des résultats optimaux également si $\lambda_{1,\text{opt}} < \lambda_{0,\text{opt}}$. D'autre part, toutefois, il a l'inconvénient (comme $ST6$) que la taille π_B de la sous-population U_B est entièrement prédéterminée (ou du moins bornée par un intervalle) si nous voulons atteindre l'efficacité optimale. En pratique, cela signifie que nous devons trouver une sous-population non reliée à la possession et à la non-possession de l'attribut A *et* de taille relative appropriée pour pouvoir obtenir l'exactitude optimale de l'estimateur. En principe, ces remarques s'appliquent aussi à $ST9$, $ST10$, $ST12$ et $ST13$, mais si nous examinons la prédétermination du paramètre de plan π_B, il s'avère que $ST9$ et $ST10$, ainsi que $ST12$ et $ST13$ se complètent parfaitement l'un l'autre, de sorte que tout sous-ensemble U_B de la population peut en fait être utilisé. Enfin, les cas particuliers les plus complexes, $ST15$ et $ST16$, de notre stratégie standardisée de réponse aléatoire peuvent tous les deux être utilisés avec une sous-population $U_B \subset U$ pour obtenir les meilleurs résultats.

5. Un exemple fondé sur des données réelles

Nous avons exécuté une étude empirique afin d'illustrer l'application de la stratégie à un plan d'interrogation. À cette fin, la population formée des 80 étudiants qui étaient inscrits au cours de « Statistique II » donné par l'auteur à l'Université Johannes Kepler à Linz (Autriche) durant le semestre du printemps 2009 a participé volontairement à une enquête. Le sujet étudié était le comportement de tricherie des étudiants. Pour les besoins de l'étude, la tricherie a été définie comme tout comportement qui n'était pas permis durant les examens écrits (y compris simplement copier les réponses des autres étudiants ou utiliser des documents interdits). Il ne fait aucun doute que le sujet est sensible pour ce genre de population. De surcroît, durant l'enquête, tous les étudiants étaient assis dans une salle de cours. Le paramètre d'intérêt était la proportion de la population d'étudiants qui avaient triché durant au moins l'un des examens du semestre précédent (y compris l'examen du cours Statistique I donné par l'auteur). Par conséquent, nous pouvons supposer d'une manière quasiment certaine que l'interrogation directe sur le sujet aurait donné lieu à une sous-estimation importante de cette proportion. Par exemple, une étude empirique menée par Scheers et Dayton (1987) a révélé de très faibles proportions pour presque tous les comportements de tricherie qu'ils sont examinés quand les questions sur le sujet étaient posées directement. L'utilisation de la stratégie de réponse aléatoire $ST3$ de Greenberg entraînait un accroissement important de ces proportions (ibidem, page 68).

Tableau 2
Paramètres de plan optimaux pour λ_1 et λ_0 donnés et divers types de caractère sensible de la variable étudiée

Plan d'interrogation (catégorie de sujets)	Paramètres de plan qui optimisent la variance
$ST1\ \ (C_1)$	$p_1 = 1$
$ST2\ \ (C_4)$	$p_1 = \frac{\lambda_1}{\lambda_1+1},\ p_2 = 1 - p_1$
$ST3\ \ (C_3, C_4)$	$\pi_B = \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1$
$ST4\ \ (C_2)$	$p_1 = \frac{\lambda_1-1}{\lambda_1},\ p_4 = 1 - p_1$
$ST6\ \ (C_4)$	$\pi_B = 0{,}5,\ p_1 : \frac{\lambda_1-1}{\lambda_1+1} < p_1 < \frac{\lambda_1}{\lambda_1+1},\ p_2 = p_1 - \frac{\lambda_1-1}{\lambda_1+1},$ $p_3 = 1 - p_1 - p_2$
$ST6\ \ (C_3)$	$\pi_B : \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,$ $p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} + \frac{(\lambda_1-1)\cdot\pi_B-(\lambda_0-1)\cdot(1-\pi_B)}{(\lambda_1\cdot\lambda_0-1)\cdot(2\pi_B-1)},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1 - p_2$
$ST7\ \ (C_3)$	$p_1 = \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = \frac{\lambda_1-1}{\lambda_1\cdot\lambda_0-1},\ p_4 = 1 - p_1 - p_2$
$ST9\ \ (C_3, C_4)$	$\pi_B : 0 < \pi_B < \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = 1 - p_1 - p_3$
$ST10\ \ (C_3, C_4)$	$\pi_B : \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_0-1}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},\ p_5 = 1 - p_1 - p_3$
$ST11\ \ (C_3, C_4)$	$p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1},\ p_5 = 1 - p_1 - p_4$
$ST12\ \ (C_3, C_4)$	$p_1 : \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B : 0 < \pi_B < \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)},\ p_3 = \frac{\lambda_1-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},$ $p_4 = 1 - \sum_{i=1}^{3} p_i$
$ST13\ \ (C_3, C_4)$	$p_1 : \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B : \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)} < \pi_B < 1,\ p_3 = \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},$ $p_5 = 1 - \sum_{i=1}^{3} p_i$
$ST14\ \ (C_3, C_4)$	$p_1 : \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2,\ p_5 = 1 - p_1 - p_2 - p_4$
$ST15\ \ (C_3, C_4)$	$\pi_B : 0 < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 : 0 < p_3 < \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_3\cdot\pi_B,$ $p_5 = 1 - p_1 - p_3 - p_4$
$ST16\ \ (C_3, C_4)$	$\pi_B : 0 < \pi_B < 1,\ p_1 : \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 : 0 < p_3 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1} - p_1,$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2 - p_3\cdot\pi_B,\ p_5 = 1 - \sum_{i=1}^{4} p_i$

Apparemment, pour la variable d'intérêt, l'appartenance au groupe U_A formé des « tricheurs » est sensible, mais non l'appartenance à l'ensemble complémentaire U_A^c. Par conséquent, conformément à la recommandation de la section 4, nous avons décidé d'utiliser le plan d'interrogation $ST4$ pour notre enquête et de le comparer à la stratégie $ST2$ de Warner. Nous avons fixé les valeurs λ de perte de vie privée à $\lambda_1 = 4$ et $\lambda_0 = \infty$. D'après le tableau 2, nous avons calculé $p_1 = 0{,}75$ et $p_4 = 0{,}25$ comme étant les paramètres du plan $ST4$ produisant la variance optimale. Pour réaliser ces probabilités, nous avons demandé aux étudiants de jeter deux dés sans montrer le résultat à quelqu'un d'autre et de répondre sur un questionnaire à la question « Avez-vous triché aux examens au moins une fois ? » uniquement si la somme des deux nombres obtenu avec les dés était comprise entre 5 et 10. Sinon, ils devaient simplement répondre « oui ».

Avant le sondage, nous avons essayé d'expliquer les conséquences de cette stratégie de randomisation sur la protection de la vie privée. Après que les étudiants aient donné la réponse sur la première feuille du questionnaire, seules ces feuilles ont été recueillies. En tout, 63 des 80 étudiants ont répondu « oui ». Nous nous attendions à ce que 20 des 80 étudiants le fassent parce qu'ils avaient reçu l'« instruction de dire oui ». Par conséquent, en principe, 43 des 60 autres étudiants ont répondu « oui » à la question sensible. L'estimateur de π_A est donnée par

$$\hat{\pi}_A^{ST4} = \frac{\hat{\pi}_y^{ST4} - p_4}{p_1} = \frac{0,7875 - 0,25}{0,75} = 0,71\dot{6}.$$

Pour ce sondage de population, la variance estimée de $\hat{\pi}_A$ est alors

$$\widehat{V}(\hat{\pi}_A^{ST4}) = \frac{1 - p_1}{n \cdot p_1} \cdot (1 - \hat{\pi}_A^{ST4}) = 1{,}181 \cdot 10^{-3}.$$

Après avoir achevé ce plan d'interrogation, nous avons demandé directement aux étudiants d'indiquer sur la deuxième feuille du questionnaire s'ils avaient répondu sincèrement à la première question ou non. Quatre étudiants seulement ont répondu par la négative. Cela signifie, si cela est vrai, que quatre étudiants de plus ont sans doute effectivement triché. La question suivante à laquelle les étudiants devraient répondre était celle de savoir s'ils continueraient de coopérer si p_1 (de $ST4$) était supérieure à 0,75. En tout, 32 des 80 étudiants ont accepté de poursuivre, mais les autres non. Manifestement (au moins) quatre d'entre eux n'ont pas coopéré quand p_1 était égale à 0,75.

Enfin, nous avons appliqué la technique de Warner avec la même question sensible que pour le plan $ST4$ appliqué auparavant. Afin de nous approcher d'un niveau λ_1 de 4, indiquant la même perte de vie privée pour une réponse « oui » pour les deux plans d'interrogation, la somme des nombres affichés sur les deux dés devait être comprise entre 3 et 9 pour appliquer un paramètre de plan $p_1 = 0,80\dot{5}$. Les mesures λ de perte de vie privée pour ce choix sont données par $\lambda_1 = \lambda_0 = 4{,}143$, indiquant une perte de vie privée un peu plus élevée que dans le cas du plan $ST4$. Avec une probabilité de $0,80\dot{5}$, les étudiants devaient répondre à la question « Êtes-vous membre du groupe U_A ? » et avec la probabilité restante, à la question alternative « Êtes-vous membre du groupe U_A^c ? ».

Dans ces conditions, 38 seulement des 80 étudiants ont répondu « oui », ce qui donne une proportion estimée de « tricheurs » de

$$\hat{\pi}_A^{ST2} = \frac{\hat{\pi}_y^{ST2} - p_2}{p_1 - p_2} = \frac{0,475 - 0,19\dot{4}}{0,6\dot{1}} = 0,45\dot{9}0.$$

En plus de ce léger accroissement de la perte objective de vie privée, il existe une autre explication raisonnable pour ce résultat nettement plus faible. Quoique λ_1 n'ait pas tellement varié, certaines personnes participant à l'expérience doivent avoir été irritées par l'accroissement de la valeur p_1 jusqu'à $0.80\dot{5}$ après qu'on leur ait demandé pour $ST4$, si elles continueraient de coopérer si p_1 devenait plus élevée que 0,75. En n'étant plus capables de faire la distinction entre la perte de vie privée causée par divers paramètres de plans dans différents plans d'interrogation, certains « tricheurs » ne voulaient plus continuer de répondre sincèrement. Simplement pour illustrer l'effet de divers plans d'interrogation sur l'efficacité du processus d'estimation, nous calculons l'estimateur de la variance de $\hat{\pi}_A^{ST2}$:

$$\widehat{V}(\hat{\pi}_A^{ST2}) = \frac{p_1 \cdot (1 - p_1)}{n \cdot (2p_1 - 1)^2} = 5{,}243 \cdot 10^{-3}.$$

L'accroissement considérable de la variance estimée est dû au fait que la stratégie de Warner ne protège pas toujours une réponse « non » de la même manière qu'une réponse « oui ». Puisque, dans notre cas, une réponse « non » ne devait pas être protégée du tout, cette protection inutile a eu un prix en terme d'exactitude.

6. Sommaire

Les stratégies fondées sur la réponse aléatoire ont été élaborées au départ pour réduire les taux de non-réponses et de réponses mensongères aux questions sur des sujets sensibles dans les enquêtes par sondage, mais elles peuvent aussi être appliquées aux fichiers de microdonnées à grande diffusion comme méthodes de masquage. La standardisation de ces méthodes pour l'estimation de proportions exposée dans le présent article offre une occasion d'établir une

formule générale pour la variance de l'estimateur sous échantillonnage probabiliste. Différents plans d'interrogation, en partie publiés et en partie – autant que nous sachions – non publiés jusqu'à présent, peuvent être considérés comme des cas particuliers de la stratégie standardisée (voir le tableau 1). Afin de comparer l'exactitude de ces plans d'interrogation, il est essentiel de tenir compte du niveau de protection de la vie privée qu'ils offrent. L'utilisation, pour cela, des « mesures λ » de la perte de vie privée décrites à la section 3 brosse un tableau entièrement différent de celui donné par presque toutes les publications antérieures connues de l'auteur. Il s'avère que les sujets identificatoires ou sensibles doivent être répartis en diverses catégories afin de trouver le plan d'interrogation à variance minimale pour un niveau donné de protection de la vie privée (voir le tableau 2). La première catégorie comprend les sujets qui n'ont aucun caractère sensible. La deuxième comprend les sujets pour lesquels la possession, mais non la non-possession, d'un certain attribut est embarrassante pour les enquêtés. La dernière catégorie regroupe les sujets qui, dans leur ensemble, ont un caractère sensible.

Pour les sujets qui rentrent dans la première catégorie, il est tout à fait clair qu'aucune stratégie ne peut être plus efficace que l'interrogation directe ($ST1$ du tableau 1).

En ce qui concerne les sujets de la deuxième catégorie, il n'existe qu'un seul plan d'interrogation permettant d'obtenir la variance minimale de l'estimateur. Il s'agit du plan selon lequel chaque enquêté doit, soit avec la probabilité p_1, répondre à la question sur l'appartenance au groupe ayant l'attribut sensible, soit avec la probabilité $1 - p_1$, répondre « oui » ($ST4$). Tous les autres cas particuliers de la stratégie standardisée protègent la vie privée de la personne interrogée, non seulement dans le cas d'une réponse « oui » comme le fait le scénario $ST4$, mais aussi dans le cas d'une réponse « non ». Par conséquent, leur performance ne peut pas atteindre le niveau de variance minimal réalisable.

Pour les sujets appartenant à la troisième catégorie, nous montrons que, contrairement à ce qu'affirment les auteurs d'autres publications, il n'existe aucune autre stratégie donnant de meilleurs résultats que celle proposée par Warner en 1965, dans les conditions où l'appartenance au sous-groupe étudié a un caractère aussi sensible que l'appartenance au complément de ce sous-groupe. De nombreux autres plans sont *aussi* efficaces que celui de Warner, mais aucun n'est *plus* efficace.

Pour les variables de cette catégorie où l'appartenance à un groupe a un caractère sensible, mais pas aussi sensible que l'appartenance au groupe complémentaire, la situation change radicalement. Comparée sous les mêmes niveaux de protection de la vie privée, la méthode de Warner ne permet plus d'atteindre le meilleur résultat réalisable du plan randomisé standardisé, tandis que de nombreuses autres stratégies le permettent. Pour certains plans, y compris la question sur l'appartenance à une sous-population non sensible non reliée à l'attribut étudié, il est nécessaire de trouver une sous-population adéquate de taille relative prédéterminée. D'autres plans peuvent être appliqués à des sous-populations de toute taille et sont donc plus pratiques. Par conséquent, une personne qui veut recueillir ou publier des données pourrait choisir, parmi les plans d'interrogation ayant la même efficacité, celui qui semble pouvoir être appliqué plus facilement que les autres.

Remerciements

L'auteur remercie le rédacteur associé et deux examinateurs de leurs précieux commentaires et suggestions.

Annexe

Preuves des théorèmes 1 et 2

Preuve du théorème 1 :

$$E(\hat{\pi}_A) = \frac{1}{N} \cdot E_P\left(E_R\left(\sum_s \frac{\hat{x}_i}{\pi_i} \mid s \right)\right)$$
$$= \frac{1}{N} \cdot E_P\left(\sum_s \frac{x_i}{\pi_i} \right) = \frac{1}{N} \cdot \sum_U x_i = \pi_A.$$

La variance de l'estimateur (5) est donnée par

$$V(\hat{\pi}_A) = V_P(E_R(\hat{\pi}_A \mid s)) + E_P(V_R(\hat{\pi}_A \mid s)).$$

Alors

$$V_P(E_R(\hat{\pi}_A \mid s)) = \frac{1}{N^2} \cdot V_P\left(\sum_s \frac{x_i}{\pi_i} \right).$$

Soit l'indicateur d'inclusion dans l'échantillon

$$I_i = \begin{cases} 1 & \text{si l'unité } i \in s, \\ 0 & \text{autrement.} \end{cases}$$

Comme la covariance $C_R(\hat{x}_i, \hat{x}_j \mid s) = 0 \ \forall \ i \neq j$, pour le deuxième terme de $V(\hat{\pi}_A)$ s'applique

$$E_P(V_R(\hat{\pi}_A \mid s)) = E_P\left(\frac{1}{N^2} \cdot V_R\left(\sum_U I_i \cdot \frac{\hat{x}_i}{\pi_i} \mid s \right)\right)$$
$$= E_P\left(\frac{1}{N^2} \cdot \sum_U \frac{I_i^2}{\pi_i^2} \cdot V_R(\hat{x}_i) \right)$$
$$= \frac{1}{N^2} \cdot \sum_U \frac{V_R(\hat{x}_i)}{\pi_i}.$$

Pour $V_R(\hat{x}_i)$, nous avons

$$V_R(\hat{x}_i) = \frac{1}{a^2} \cdot V_R(y_i)$$

et

$$\begin{aligned} V_R(y_i) &= b + a \cdot x_i - (b + a \cdot x_i)^2 \\ &= (b + a \cdot x_i) \cdot (1 - b - a \cdot x_i) \\ &= b \cdot (1 - b) + a \cdot (1 - 2 \cdot b - a) \cdot x_i. \end{aligned}$$

Alors

$$E_P(V_R(\hat{\pi}_A \mid s)) = \frac{1}{N^2} \cdot \left(\frac{b \cdot (1 - b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right).$$

Cela complète la preuve du théorème 2.

Bibliographie

Chaudhuri, A., et Mukerjee, R. (1987). *Randomized Response*. New York : Marcel Dekker.

Dalenius, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.

Dalenius, T., et Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

Defays, D., et Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14 (4), 449-461.

Domingo-Ferrer, J., et Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.

Fidler, D.S., et Kleinknecht, R.E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin*, 84 (5), 1045-1049.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. et de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14 (4), 463-478.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R. et Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. Hoboken : John Wiley & Sons, Inc.

Horvitz, D.G., Shah, B.V. et Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 65-72.

Hua, M., et Pei, J. (2008). A survey of utility-based privacy-preserving data transformation methods. Dans : *Privacy-preserving Data Mining*: *Models and Algorithms*, (Éds., C.C. Aggarwal et P.S. Yu), New York : Springer, 207-238.

Kim, J. (1987). A further development of the randomized response technique for masking dichotomous variables. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 239-244.

Kim, J.M., et Warde, W.D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211-221.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77 (2), 436-438.

Landsheer, J.A., van der Heijden, P. et van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33, 1-12.

Leysieffer, F.W., et Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

Mangat, N.S. (1992). Two stage randomized response sampling procedure using unrelated question. *Journal of the Indian Society of Agricultural Statistics*, 44, 82-87.

Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, Séries B, 56, 93-95.

Mangat, N.S., et Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

Mangat, N.S., Singh, S. et Singh, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metron*, 51, 211-216.

Nathan, G. (1988). Bibliographie de la méthode des réponses randomisées : 1965-1987. *Techniques d'enquête*, 14, 351-365.

Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series*, 23, www.ifas.jku.at/e2550/e2756/index_ger.html.

Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 311-316.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.

Scheers, N.J., et Dayton, C.M. (1987). Improved estimation of academic cheating behaviour using the randomized response technique. *Research in Higher Education*, 26 (1), 61-69.

Singer, E., Mathiowetz, N.A. et Couper, M.P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *The Public Opinion Quarterly*, 57 (4), 465-482.

Singer, E., van Hoewyk, J. et Neugebauer, R.J. (2003). Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census. *The Public Opinion Quarterly*, 67 (3), 368-384.

Singh, R., Singh, S., Mangat, N.S. et Tracy, D.S. (1995). An improved two stage randomized response strategy. *Statistical Papers*, 36, 265-271.

Singh, S., Horn, S., Singh, R. et Mangat, N.S. (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6 (4), 515-522.

Singh, S., Singh, R., Mangat, N.S. et Tracy, D.S. (1994). An alternative device for randomized responses. *Statistica*, 54, 233-243.

Skinner, C., Marsh, C., Openshaw, S. et Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10 (1), 31-51.

Soeken, K.L., et Macready, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92 (2), 487-489.

Tracy, D.S., et Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4 (2/3), 147-158.

van den Hout, A., et van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *Revue Internationale de Statistique*, 70 (2), 269-288.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

Willenborg, L., et de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York : Springer.

Winkler, W.E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. *Research Report Series of the Statistical Research Division of the U.S. Bureau of the Census*, #2004-06.