No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Actes du symposium de Statistique Canada, 2006
Enjeux méthodologiques reliés à la mesure de la santé des populations

Masquage dans le cas de variables discrètes

Myron J. Katzoff et Jay J. Kim[1]

Résumé

L'utilisation de variables discrètes ayant une distribution statistique connue pour le masquage des données de variables discrètes est à l'étude depuis un certain temps. Le présent article fait état de quelques-uns de nos résultats de recherche sur le sujet. Les conséquences du prélèvement, dans des populations finies, d'échantillons avec et sans remise suscitent un intérêt tout particulier. Les estimations des moments de premier ordre et de second ordre qui permettent d'atténuer ou de corriger la variation supplémentaire causée par le masquage d'un type connu sont établies. L'incidence du masquage des données originales sur la structure de corrélation de variables discrètes faisant l'objet de mesures concomitantes est examinée, et la nécessité d'un examen plus poussé des résultats à des fins d'analyse des données multivariées est discutée.

MOTS CLÉS : masquage; données multinomiales; protection contre la divulgation; réponse aléatoire.

1. Introduction

Les données sur des individus obtenues dans le cadre d'enquêtes nationales par sondage, généralement désignées par le terme « microdonnées », sont souvent requises pour les analyses de données secondaires dans lesquelles les personnes échantillonnées constituent les unités conceptuelles d'observation. Même si les chercheurs qui demandent ces données ne tiennent pas à connaître l'identité des répondants aux enquêtes, les enregistrements peuvent renfermer suffisamment d'information pour permettre l'identification des individus. C'est de là que vient le risque de rompre l'engagement de protection de la confidentialité pris envers les répondants aux enquêtes. Les techniques de masquage ont été mises au point précisément pour réduire ce risque.

Parmi les procédures de masquage figurent la permutation des données, la microagrégation et la transformation des variables-réponses. Nous nous intéressons tout particulièrement aux procédures qui permettent de préserver les corrélations entre les variables ou, de façon plus générale, à celles qui ne suppriment pas les relations statistiques et les attributs des données se retrouvant dans les données non masquées. Évidemment, dans les cas d'estimation de paramètres, les techniques qui permettent de supprimer les effets du masquage par de simples ajustements ou transformations des estimateurs se révéleraient fort utiles. Dans le cas de données discrètes, le masquage s'entend d'une méthode de protection contre la divulgation dans laquelle les réponses originales d'un membre de l'échantillon sont modifiées conformément à des transformations et/ou à des mécanismes stochastiques préalablement définis.

Dans les paragraphes qui suivent, nous nous limiterons à la discussion d'une procédure applicable aux variables nominales : la perturbation de la variable nominale par l'ajout d'une autre variable nominale de même dimension, une variable « de bruit ». Le lecteur notera que Warner (1971) a peut-être été le premier à relever une relation entre cette méthode de masquage des variables multinomiales et les réponses aléatoires. Il convient de souligner un point subtil, mais important quant aux différences dans l'application des techniques de réponse aléatoire et de masquage. Dans le cas du masquage, nous connaissons, en tant que fournisseurs des microdonnées, les valeurs non modifiées (c.-à-d. réelles) de l'échantillon; dans le cas de la réponse aléatoire, en revanche, ces valeurs ne sont connues ni des fournisseurs ni des utilisateurs des données. Pour cette raison, l'application d'un mécanisme probabiliste dans les

[1]Les deux auteurs sont affiliés au *National Center for Health Statistics*, *Centers for Disease Control and Prevention*, 3311 Toledo Road, Hyattsville, MD 20782, USA. **Mise en garde :** Le présent article représente les points de vue des auteurs; il ne doit pas être considéré comme l'expression des points de vue, des politiques ou des pratiques des *Centers for Disease Control and Prevention*, *National Center for Health Statistics*.

techniques de réponse aléatoire exige qu'il suive une distribution indépendante de ce que l'on appelle les valeurs réelles, tandis que, pour les procédures de masquage, nous pouvons permettre au mécanisme de dépendre de ces valeurs. Les travaux antérieurs de Kim et Flueck (1977, 1978) présentent également de l'intérêt pour l'étude de la méthode examinée dans le présent article, qui réunit quelques-uns des résultats actuellement disponibles pour les données multinomiales masquées par un bruit multinomial, précise certaines questions relatives à l'inférence statistique quant aux données nominales masquées, fait état de nouveaux résultats et relève les lacunes théoriques qui subsistent encore.

2. Masquage de données multinomiales et relation avec la réponse aléatoire

Kim et Flueck (1977), KF dans la présente section, ont examiné les techniques de réponse aléatoire pour les cas de trois catégories ou plus dans un échantillonnage multinomial. Nous confirmons leurs résultats pour les cas trinomiaux facilement généralisés en apportant quelques modifications mineures à la notation afin de (1) respecter la présentation de la partie du développement pour les variables dichotomiques dans Kim (1987) qui fait intervenir un échantillonnage avec remise et (2) préciser la relation entre ces travaux et ceux de Gouweleeuw et coll. (1998), GKWW (1998) ci-après. Dans leur modèle additif, KF définissent les variables indépendantes discrètes C_j, représentant l'échantillon original ou les valeurs réelles, et a_j, les variables de bruit, pour $j = 1, 2, \ldots, n$ (soit la taille de l'échantillon), toutes deux prenant les valeurs 1, 2 et 3. Nous effectuons les calculs et enregistrons $\left(C_j + a_j\right)_{\text{mod } 3}$. Pour $k = 0, 1, 2$, posant $\pi_k = \Pr\left\{C_j = k\right\}$ et $p_k = \left\{a_j = k\right\}$ pour $j = 1, 2, \ldots, n$ lorsque les p_k *non nuls* $\neq \frac{1}{3}$, KF fournissent des estimateurs sans biais $\hat{\pi}_k$ pour les π_k et les variances de ces estimateurs en recourant à l'algèbre linéaire simple.

Les valeurs 0, 1 et 2 peuvent être correctement associées aux résultats trinomiaux $\vec{v}_1 = (1,0,0)^T$, $\vec{v}_2 = (0,1,0)^T$ et $\vec{v}_3 = (0,0,1)^T$, respectivement. Conformément à GKWW (1998), nous représentons les vecteurs tridimensionnels originaux de données par $\vec{\xi}_j$ et les données perturbées par $\vec{x}_j$. Pour chacun de ces vecteurs, pour $j = 1, 2, \ldots, n$, les trois résultats $\vec{v}_1$, $\vec{v}_2$ et $\vec{v}_3$ sont possibles; et si $\vec{a}$ est l'un d'entre eux, alors

$$\left[\vec{x}_j = \vec{a}\right] = \left[\vec{x}_j = \vec{a}\, et\, \vec{\xi}_j = \vec{v}_1\right] \cup \left[\vec{x}_j = \vec{a}\, et\, \vec{\xi}_j = \vec{v}_2\right] \cup \left[\vec{x}_j = \vec{a}\, et\, \vec{\xi}_j = \vec{v}_3\right].$$

Soit

$$\vec{T}_x = \sum_{j=1}^{n} \vec{x}_j = \left(\#\left[\vec{x}_j = \vec{v}_1\right], \#\left[\vec{x}_j = \vec{v}_2\right], \#\left[\vec{x}_j = \vec{v}_3\right]\right)^T,$$

où $\#\left[\vec{x}_j = \vec{v}_1\right]$ représente le nombre de vecteurs $\vec{x}_j$ dans la somme pour laquelle $\vec{x}_j = \vec{v}_1$. Nous obtenons alors

$$\vec{T}_x = \sum_{j=1}^{n} \begin{pmatrix} I_{[\vec{x}_j = \vec{v}_1 | \vec{\xi}_j = \vec{v}_1]} & I_{[\vec{x}_j = \vec{v}_1 | \vec{\xi}_j = \vec{v}_2]} & I_{[\vec{x}_j = \vec{v}_1 | \vec{\xi}_j = \vec{v}_3]} \\ I_{[\vec{x}_j = \vec{v}_2 | \vec{\xi}_j = \vec{v}_1]} & I_{[\vec{x}_j = \vec{v}_2 | \vec{\xi}_j = \vec{v}_2]} & I_{[\vec{x}_j = \vec{v}_2 | \vec{\xi}_j = \vec{v}_3]} \\ I_{[\vec{x}_j = \vec{v}_3 | \vec{\xi}_j = \vec{v}_1]} & I_{[\vec{x}_j = \vec{v}_3 | \vec{\xi}_j = \vec{v}_2]} & I_{[\vec{x}_j = \vec{v}_3 | \vec{\xi}_j = \vec{v}_3]} \end{pmatrix} \begin{pmatrix} I_{[\vec{\xi}_j = \vec{v}_1]} \\ I_{[\vec{\xi}_j = \vec{v}_2]} \\ I_{[\vec{\xi}_j = \vec{v}_3]} \end{pmatrix} \qquad (1)$$

où $I_{\left[\vec{\xi}_j = \vec{a}\right]}$ est la variable-indicateur pour l'événement $\left[\vec{\xi}_j = \vec{a}\right]$ et $I_{\left[\vec{x}_j = \vec{b} | \vec{\xi}_j = \vec{a}\right]}$ est la variable-indicateur pour l'événement $\left[\vec{x}_j = \vec{b}\right]$ compte tenu que l'événement $\left[\vec{\xi}_j = \vec{a}\right]$ s'est produit. Il en découle que

$$E\left[\vec{T}_x \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right] = P^T \sum_{j=1}^{n} \begin{pmatrix} I_{[\vec{\xi}_j = \vec{v}_1]} \\ I_{[\vec{\xi}_j = \vec{v}_2]} \\ I_{[\vec{\xi}_j = \vec{v}_3]} \end{pmatrix}, \text{ où } P = \begin{pmatrix} p_0 & p_1 & p_2 \\ p_2 & p_0 & p_1 \\ p_1 & p_2 & p_0 \end{pmatrix}.$$

Nous obtenons l'estimateur sans biais $\vec{T}_\xi$, $\hat{\vec{T}}_\xi = \left(P^T\right)^{-1} \vec{T}_x$, calculé par GKWW (1998) où $\vec{T}_\xi = \left(\#\left[\vec{\xi}_j = \vec{v}_1\right], \#\left[\vec{\xi}_j = \vec{v}_2\right], \#\left[\vec{\xi}_j = \vec{v}_3\right]\right)^T$. Nous obtenons ensuite des estimations sans biais des π_k en divisant les coordonnées de $\hat{\vec{T}}_\xi$ par n.

Pour obtenir les expressions inconditionnelles des variances et des covariances, notons que

$$Var\left(\hat{\vec{T}}_x \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right) = \sum_{k=1}^{3} T_\xi(k)\, B_k ,$$

où, pour $k = 1,2,3$, $T_\xi(k) = \#\left[\vec{\xi}_j = \vec{v}_k\right]$, $B_k = Diag(\vec{p}_k) - \vec{p}_k \vec{p}_k{}^T$ et $\vec{p}_k{}^T$ est la ligne k de P. Puisque $E\left[T_\xi(k)\right] = n\, p_{k-1}$ pour $k = 1,2,3$, nous constatons que

$$E\left\{Var\left[\hat{\vec{T}}_\xi \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right]\right\} = \left(P^T\right)^{-1}\left(n\sum_{k=1}^{3} p_{k-1} B_k\right) P^{-1} \qquad (2)$$

où

$$P^{-1} = \Delta^{-1} \begin{pmatrix} p_0^2 - p_1 p_2 & p_2^2 - p_0 p_1 & p_1^2 - p_0 p_2 \\ p_1^2 - p_0 p_2 & p_0^2 - p_1 p_2 & p_2^2 - p_0 p_1 \\ p_2^2 - p_0 p_1 & p_1^2 - p_0 p_2 & p_0^2 - p_1 p_2 \end{pmatrix}$$

et $\Delta = p_0^3 + p_1^3 + p_2^3 - 3 p_0 p_1 p_2$. Puisque $E\left[\hat{\vec{T}}_\xi \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right] = \vec{T}_\xi$,

$$Var\left\{E\left[\hat{\vec{T}}_\xi \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right]\right\} = Var\left[\vec{T}_\xi\right] = n V(\vec{\pi}) \qquad (3)$$

où $\vec{\pi} = (\pi_0, \pi_1, \pi_2)^T$ et $V(\vec{\pi}) = Diag(\vec{\pi}) - \vec{\pi}\vec{\pi}^T$. Pour les $\hat{\pi}_k$, la variance inconditionnelle ou totale est n^{-2} fois la somme des matrices des membres droits des équations (2) et (3).

La situation est nettement moins compliquée pour les cas bidimensionnels dans GKWW (1998), où $P = \begin{pmatrix} \theta_0 & 1-\theta_0 \\ 1-\theta_1 & \theta_1 \end{pmatrix}$. Nous pouvons constater que

$$Var\left(\hat{\vec{T}}_\xi\right) = \left(n\pi(1-\pi) + \frac{1}{\Delta^2}\left[n\pi\,\theta_0(1-\theta_0) + n(1-\pi)\,\theta_1(1-\theta_1)\right]\right) \bullet \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \text{ où } \Delta = \theta_0 + \theta_1 - 1 .$$

Si $\theta_0 = \theta_1 = p$, par exemple, cette expression concorde parfaitement avec Kim (1987) et Warner (1971).

3. Masquage conditionnel à la réponse

Les travaux de Kim et Flueck (1978) sur la réponse aléatoire indiquent que, pour effectuer le masquage, nous pouvons envisager quatre combinaisons de méthodes d'échantillonnage pour la sélection des cas et des valeurs des variables de masquage correspondant à l'échantillonnage aléatoire simple avec remise (EAS AR) et à l'échantillonnage aléatoire simple sans remise (EAS SR). Nous n'examinerons ici qu'une seule de ces combinaisons produisant de nouveaux résultats pour un échantillon prélevé dans une population finie : la sélection de l'échantillon par EAS SR, mais le masquage par EAS AR, que nous désignons par le terme « masquage de Bernoulli », pour le cas bidimensionnel. Pour cette combinaison, une seule modification des résultats de la variance de l'échantillonnage multinomial est requise : il s'agit de multiplier $Var\left\{E\left[\hat{\bar{T}}_{\xi} \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right]\right\}$ par le facteur $\left(\frac{N-n}{N-1}\right)$. Par conséquent, pour le cas bidimensionnel, le terme $n\pi(1-\pi)$ dans $Var\left(\hat{\bar{T}}_{\xi}\right)$ sera multiplié par ce facteur. Les estimateurs des moments de premier ordre sont élaborés à la dernière section. Un nouvel estimateur sans biais de la variance totale est présenté ci-dessous.

Dans un examen récent de la méthode de la randomisation a posteriori de GKWW (1998), Kim observe que les évaluations des diverses procédures de masquage devraient se fonder sur la variance totale et que les inférences à partir des données d'échantillon masquées exigent des estimateurs de ces variances. Dans le cas bidimensionnel, posons $Z = \sum_{i=1}^{n} z_i$, où z_i est la valeur masquée pour l'élément échantillonné $i = 1, 2, \ldots, n$ et

$$z_i \begin{cases} 1, \text{ si l'élément } i \text{ possède la caractéristique considérée} \\ \quad \text{(c. - à - d. appartient à la catégorie 0)} \\ 0, \text{ dans les autres cas} \end{cases}$$

Alors l'estimation sans biais de π , la proportion des membres de la population possédant la caractéristique est représenté par

$$\hat{\pi} = \frac{Z/n - (1-\theta_1)}{(\theta_0 + \theta_1 - 1)}$$

et

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right) + \frac{\pi\theta_0(1-\theta_0) + (1-\pi)\theta_1(1-\theta_1)}{n(\theta_0 + \theta_1 - 1)^2} \quad .$$

L'estimateur sans biais de cette variance est représenté par

$$\hat{V}(\hat{\pi}) = \frac{1}{nN(N-1)(\theta_0 + \theta_1 - 1)^2}\left[\frac{Z(n-Z)(N-n)}{n(n-1)} + (N^2 - 2N + n)[\hat{\pi}\theta_0(1-\theta_0) + (1-\hat{\pi})\theta_1(1-\theta_1)]\right].$$

Examinons maintenant le cas d'un échantillonnage avec probabilités arbitraires sans remise et de l'application de la procédure de masquage de Bernoulli. Nous représentons les probabilités d'inclusion par γ_i et les probabilités d'inclusion conjointes par γ_{ij} . Si nous redéfinissons Z ainsi $Z = \sum_{i=1}^{n} z_i / \gamma_i$, alors

$$\hat{\pi}^* = \frac{Z/N - (1-\theta_1)}{\theta_0 + \theta_1 - 1}$$

est l'estimateur sans biais de π . Nous pouvons obtenir l'expression de la variance de $\hat{\pi}^*$ en manipulant chacun de ces deux termes :

$$E\left\{Var\left[Z \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right]\right\} = \left(\sum_{i=1}^{N}\frac{\xi_i}{\gamma_i}\right)\theta_0(1-\theta_0) + \left(\sum_{i=1}^{N}\frac{(1-\xi_i)}{\gamma_i}\right)\theta_1(1-\theta_1)$$

et

$$Var\left\{E\left[Z \mid \left\{\vec{\xi}_j\right\}_{j=1}^{n}\right]\right\} = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\gamma_i\gamma_j - \gamma_{ij}\right)\left(\frac{u_i}{\gamma_i} - \frac{u_j}{\gamma_j}\right)^2$$

où, ici, ξ_i représente la variable-indicateur de l'individu i défini par

$$\xi_i\begin{cases}1, \text{si l'élément } i \text{ possède la caractéristique considérée}\\ \quad \text{(c. - à - d. appartient à la catégorie 0)}\\ 0, \text{dans les autres cas}\end{cases}$$

et $u_i = (\theta_0 + \theta_1 - 1)\xi_i + (1-\theta_1)$, pour $i = 1,2,\ldots,N$. Dans le cadre d'un EAS SR, lorsque $\gamma_i = \frac{n}{N}$ et $\gamma_{ij} = \frac{n(n-1)}{N(N-1)}$, ces expressions se simplifient, comme on peut s'y attendre, de manière à produire des résultats qui sont compatibles avec ceux obtenus plus tôt. Il est plus difficile d'obtenir une expression pour un estimateur sans biais de la variance totale dans ce cas que ce ne l'est dans le cas d'un EAS SR.

4. Moments de second ordre pour les paires de variables nominales soumises à une procédure de masquage

Dans cette section, nous reprenons un résultat important de Kim (1987) en utilisant un cadre de notation plus général que le sien. Le lecteur est invité à consulter Kim (1987) pour les démonstrations.

Examinons un tableau de contingence où $\left\{\pi_{ij} \mid i = 1,2,\ldots,K_1; j = 1,2,\ldots,K_2\right\}$ représente les probabilités de cellule. Si $\pi_{i.} = \sum_{j=1}^{K_2}\pi_{ij}$ et $\pi_{.j} = \sum_{i=1}^{K_1}\pi_{ij}$, l'utilisateur de données pourrait être intéressé par $\text{cov}(\hat{\pi}_{i.}, \hat{\pi}_{.j})$, où les paramètres « chapeautés » sont des estimateurs des probabilités marginales. Dans le cas de l'échantillonnage aléatoire simple avec remise (EAS AR),

$$\text{cov}(\hat{\pi}_{i.}, \hat{\pi}_{.j}) = \frac{\pi_{ij} - \pi_{i.}\pi_{.j}}{n}$$

mais dans celui de l'échantillonnage aléatoire simple sans remise (EAS SR),

$$\text{cov}(\hat{\pi}_{i.}, \hat{\pi}_{.j}) = \frac{\pi_{ij} - \pi_{i.}\pi_{.j}}{n}\left(\frac{N-n}{N-1}\right).$$

Introduisons maintenant des variables de masquage $Y_k, k = 1,2,\ldots,n,$ obtenues par EAS SR à partir d'une population finie de taille $M \geq n$ où pM des membres sont égaux à 1, $p \neq \frac{1}{2}$ et quantifions

$$Z_{ik} = (X_{ik} + Y_k)_{\text{mod}\,2}$$
$$Z_{jk} = (X_{jk} + Y_k)_{\text{mod}\,2}$$

où pour $k = 1,2,\ldots,n$

$$X_{ik}\begin{cases}1, \text{si l'observation appartient à la cellule (i, j)}\\ \quad \text{pour } j = 1,2,\ldots \text{K}_2\\ 0, \text{dans les autres cas}\end{cases} \qquad X_{jk}\begin{cases}1, \text{si l'observation appartient à la cellule (i, j)}\\ \quad \text{pour } i = 1,2,\ldots \text{K}_1\\ 0, \text{dans les autres cas}\end{cases}$$

Les estimateurs de $\pi_{i.}$ et $\pi_{.j}$ sont

$$\hat{\pi}_{i.} = \frac{1}{n}\sum_{k=1}^{n} X_{ik} \qquad \text{et} \qquad \hat{\pi}_{.j} = \frac{1}{n}\sum_{k=1}^{n} X_{jk}$$

et $\left\{\left(X_{ik}, X_{jk}\right) : k = 1, 2, \ldots, n\right\}$ représente l'ensemble des valeurs de l'échantillon pour la cellule (i,j). Si

$$F = \frac{N-n}{N-1} - 4p(1-p)\frac{N(M-n)-n(M-1)}{(N-1)(M-1)} \quad ,$$

Kim (1987) a montré que, compte tenu des procédures précitées d'échantillonnage et de masquage,

$$\operatorname{cov}\left(\hat{\pi}_{i.}, \hat{\pi}_{.j}\right) = \frac{\pi_{ij} - \pi_{i.}\pi_{.j}}{n(1-2p)^2} \cdot F + \frac{p(1-p)}{n(1-2p)^2}\left(\frac{M-n}{M-1}\right)\left(1 - 2\pi_{i.} - 2\pi_{.j} + 4\pi_{ij}\right) \qquad \textit{et}$$

$$\operatorname{Var}\left(\hat{\pi}_{i.}\right) = \frac{\pi_{i.}(1-\pi_{i.})}{n(1-2p)^2} \cdot F + \frac{p(1-p)}{n(1-2p)^2}\left(\frac{M-n}{M-1}\right)$$

avec une expression similaire pour $\operatorname{Var}(\hat{\pi}_{.j})$. Si $M = n$, ces expressions se simplifient de façon évidente, et nous constatons aisément que ce scénario de masquage préservera la structure de corrélation des variables non masquées. S'il s'agit là d'une propriété importante pour l'utilisateur, le scénario de masquage que nous venons de décrire pourrait se révéler supérieur à la permutation ou à la microagrégation. Une discussion plus détaillée des questions traitées dans cette section est présentée dans Kim (1987). Selon les paramètres à prendre en considération, même en tenant compte du champ restreint auquel nous nous sommes confinés, certaines procédures de masquage semblent meilleures que d'autres.

5. Recherches futures

Les travaux futurs de recherche sur le masquage dans le cas de variables discrètes devraient tenir compte des avantages potentiels que présente l'échantillonnage sans remise des valeurs de masquage à partir d'une population finie, comme nous l'avons décrit à la section 4. Il serait aussi opportun d'examiner le potentiel de l'échantillonnage par grappes à plusieurs degrés ainsi que les moyens permettant de réduire ou d'éliminer la complexité de l'estimation associée à l'échantillonnage avec probabilités inégales sans remise. La fonction de masquage qui permet au fournisseur de données de rendre le mécanisme de randomisation dépendant des valeurs observées pourrait se révéler utile dans ces situations.

Évidemment, des estimateurs de variance adaptés aux diverses procédures d'échantillonnage et stratégies de masquage sont nécessaires aux inférences fondées sur l'information fournie aux utilisateurs des données masquées. Cependant, dans de nombreux cas, on pourrait obtenir des approximations satisfaisantes de ces estimateurs pour certaines formes d'inférence statistique et il semble pertinent d'étudier les propriétés de tels estimateurs dans ces situations. Ces estimateurs approximatifs peuvent être particulièrement importants lorsque certains paramètres utilisés dans le masquage doivent être estimés à partir des données disponibles.

Enfin, comme bien d'autres chercheurs qui ont travaillé sur les techniques de protection contre la divulgation, nous reconnaissons la nécessité de telles mesures de protection de même que de critères permettant de déterminer les éléments de données qu'il convient de masquer ou non.

Références

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*, pp. 239-244.

Kim, J. et Flueck, J. (1977), "An Additive Randomized Response Model", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp.351-355.

Kim, J. et Flueck, J. (1978), "Modification of the Randomized Response Techniques for Sampling without Replacement", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp.346-350.

Warner, S. (1971), "The Linear Randomized Response Model", *Journal of the American Statistical Association*, v.66, pp. 884-888.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. et de Wolf, P.-P. (1998), "Post Randomization for Statistical Disclosure Control: Theory and Implementation", *Journal of Official Statistics*, v.14, pp. 463-478. [Referred to in text by GKWW.]