Techniques d’enquête, décembre 1975
Vol. 1, nº 2
Statistique Canada, nº 12-001-X au catalogue

Arrondissement aléatoire contrôlé

I.P. Fellegi[1]

Résumé

L’arrondissement aléatoire est une technique qui vise à assurer la confidentialité des agrégats ou groupes de statistiques. En appliquant cette technique à tous les éléments d’un total, d’une part, et au total lui-même, d’autre part, des divergences importantes peuvent se produire au moment de regrouper les données publiées. La méthode décrite dans ce document permet d’éviter ces divergences tout en assurant la confidentialité des données.

Mots-clés : Confidentialité; arrondissement aléatoire contrôlé.

1 Introduction

L’arrondissement aléatoire est une technique qui vise à prévenir la divulgation statistique directe et résiduelle. Elle consiste à arrondir les agrégats statistiques publiés (ou diffusés de tout autre façon) à un multiple d’un chiffre de base choisi, mais en effectuant l’arrondissement au moyen d’un mécanisme aléatoire qui garantit que chaque agrégat publié arrondi aléatoirement a comme valeur attendue l’agrégat non arrondi correspondant (et bien sûr, non publié). Cela permet de s’assurer que le processus d’arrondissement n’est pas biaisé. Pour obtenir une description plus détaillée de la technique, le lecteur est invité à consulter [2] et [3].

Le problème précis dont il est question dans cette note se résume comme suit. Étant donné que chaque agrégats d’un certain nombre d’agrégats statistiques doit être arrondi aléatoirement, est-il possible de le faire de manière à ce que la somme des chiffres individuels arrondis aléatoirement corresponde à l’arrondissement aléatoire de la somme des chiffres non arrondis? C’est-à-dire, si e_i $(i=1, 2, \ldots, n)$ représente les chiffres non arrondis, et e_i^* représente les chiffres arrondis aléatoirement correspondants, est-il possible de procéder à l’arrondissement aléatoire de manière à ce que

$$\sum_{i=1}^{n} e_i^*$$

1. I.P. Fellegi, Statisticien en chef adjoint, Direction de la méthodologie et des systèmes.

soit équivalent à (c’est-à-dire, ait la même distribution que)

$$\left(\sum_{i=1}^{n} e_i\right)^*.$$

La question ainsi formulée provient d’un problème très concret. Plusieurs pays ont adopté la pratique consistant à diffuser des bandes soi-disant sommaires à la suite de leurs recensements de la population décennaux ou quinquennaux. Ces bandes contiennent des totalisations (agrégats) pour de très petites régions géographiques, qui correspondent habituellement au travail attribué à un seul recenseur. Ces données sur de petites régions géographiques sont utilisées par le personnel de recherche comme des composantes des données agrégées pour leurs domaines d’intérêt respectifs. Au moins deux pays, le Royaume-Uni [4] et le Canada, ont adopté la pratique consistant à introduire une petite perturbation aléatoire dans ces agrégats de petites régions géographiques afin de les protéger contre la divulgation statistique. De plus, le Bureau of the Census est, à tout le moins, en train d’envisager une procédure similaire pour le Recensement de 1980 [1].

Même si la possibilité qu’il y ait de telles erreurs aléatoires est faible, lorsque les chiffres arrondis aléatoirement sont agrégés, leurs variances le sont également. Lorsque les totalisations de plusieurs petites régions géographiques sont agrégées afin d’obtenir la totalisation d’une plus grande région, par exemple une municipalité, la variance peut devenir assez importante (même si la variance relative diminue, bien entendu). Ainsi, lorsque les utilisateurs comparent les totalisations qu’ils ont eux-mêmes préparées à partir des bandes sommaires pour, par exemple, une municipalité, avec les tableaux qui ont réellement été publiés pour les municipalités, ils peuvent remarquer des écarts importants. Cela est dû au fait que les totalisations publiées à l’échelle municipale ont fait l’objet d’un arrondissement aléatoire direct, tandis que les totalisations préparées par les utilisateurs à partir des bandes sommaires ont été arrondies aléatoirement au niveau de la composante de petite région géographique.

La procédure suivante permet de s’assurer que, lorsque les totalisations de petites régions géographiques sont arrondies aléatoirement, les répercussions cumulées de telles erreurs soient contrôlées pour ce qui

est de certaines régions géographiques prédéfinies de plus haut niveau. Bien sûr, à l’exception des régions plus grandes prédéfinies, l’arrondissement aléatoire ne cause probablement aucune répercussion sur la variance.

Une tentative visant à contenir les répercussions cumulatives des erreurs aléatoires est exposée à [4], mais seulement dans une situation où le nombre d’erreurs aléatoire est +1, -1 ou 0.

2 Les contraintes

Si le chiffre de base de l’arrondissement est un nombre entier représenté par b (« B » était équivalent à « 5 » dans le Recensement du Canada de 1971), supposons qu’une entrée dans le tableau est représentée par e. Nous calculons le résidu r de e après une division par b:

$$e = k \times b + r \qquad 0 \leq r < b$$

C’est ce résidu qui est « arrondi » de manière aléatoire à 0 ou à b. Disons que la probabilité que l’arrondissement se fasse à b est représenté par p, et que la probabilité que l’arrondissement soit fait vers le bas à zéro soit représenté par $(1-p)$. Il est alors possible de représenter le e, e^* arrondi aléatoirement comme suit :

$$e^* = k \times b + r^*$$

où $r^* = b$ avec la probabilité p et est égal à zéro avec la probabilité $(1-p)$. La valeur attendue de e^* peut être représentée comme suit :

$$E(e^*) = k \times b + [p \times b + (1-p) \times 0]$$

Si nous voulons que e^* ne soit pas biaisé, nous devons établir

$$E(e^*) = e$$

c’est-à-dire

$$p \times b = r \quad \text{ou} \quad p = \frac{r}{b}$$

Il s'agit de la première contrainte que nous imposons sur une procédure d'arrondissement aléatoire souhaitable. Le raisonnement ci-dessus montre également que si e^* représente une estimation non biaisée de e, la seule façon d'altérer e pour qu'il devienne un multiple de b en le modifiant de manière aléatoire selon une quantité dont la valeur absolue est inférieure à b, est d'utiliser un processus d'arrondissement axé sur les probabilités comme indiqué ci-dessus.

Si nous voulons préserver l'absence de biais de l'arrondissement aléatoire, nous ne devons donc pas enfreindre cette contrainte.

Ensuite, supposons qu'il existe une série de n cellules de totalisations (chacune correspondant à l'agrégat d'une petite région géographique pour une municipalité) qui doivent être arrondies. Celles-ci sont représentées par e_i $(i = 1, 2, \ldots, n)$.

Laissons

$$e_i = k_i x b + r_i \qquad 0 \leq r_i < b$$

et la valeur arrondie aléatoirement correspondante être représentée par

$$e_i^* = k_i x b + r_i^*$$

où $r_i^* = b$ avec la probabilité $p_i = r_i/b$ et est égal à 0 avec la probabilité $1 - p_i$.

Leur somme, e est représentée par

$$e = \sum_{i=1}^{n} e_i$$

qui peut être représenté ainsi

$$e = k \times b + r \qquad 0 \leq r < b \qquad (2.1)$$

et sa valeur arrondie est

$$e^* = k \times b + r^*$$

où r^* est égal à b avec une probabilité de r/b ou sinon est égal à 0. Idéalement, nous souhaiterions avoir

$$e^* = \sum_{i=1}^{n} e_i^*$$

dans le sens que $\Sigma\, e_i^*$ et e^* attendent les mêmes valeurs avec les mêmes probabilités. Il s'agit de la deuxième contrainte que nous imposons pour une procédure d'arrondissement aléatoire souhaitable.

La procédure suivante respecte ces deux contraintes.

3 La procédure

Examinez les chiffres r_i et cumulez-les :

$$s_i = \sum_{j=1}^{i} r_j$$

$$S_n = \sum_{j=1}^{n} r_j$$

$$S_o = 0.$$

Sélectionnez un nombre entier entre 1 et b, disons R_1:

$$1 \leq R_1 \leq b.$$

Examinez $S_1, S_2, S_3, \ldots,$ en ordre jusqu'à

$$S_{i_1-1} < R_1 \leq S_{i_1}.$$

Ensuite, laissez

$$R_2 = R_1 + b$$

et sélectionnez i_2 afin que

$$S_{i_2-1} < R_2 \leq S_{i_2}.$$

Ensuite, laissez

$$R_3 = R_2 + b = R_1 + 2b$$

et sélectionnez i_3 afin que

$$S_{i_3-1} < R_3 \leq S_{i_3}$$

etc. Continuez jusqu'à ce que l'étape L soit définie afin que

$$R_L \leq S_n$$

mais que

$$R_{L+1} > S_n .$$

Arrondissez maintenant vers le haut les unités ainsi sélectionnées et vers le bas les autres unités. En d'autres termes,

où

$$e_i^* = k_i x b + r_i^*$$

$$r_i^* = b \qquad \text{si} \qquad i = i_1, i_2, i_3, \ldots$$

$$= 0 \qquad \text{sinon.}$$

La procédure est illustrée au Tableau 1.

4 Preuve que la procédure ainsi définie satisfait aux contraintes

Il est facile de vérifier, en utilisant des arguments qui représentent la norme pour la sélection avec des probabilités proportionnelles à une mesure de taille, si la probabilité

$$P(r_i^* = b) = r_i/b$$

afin que la première contrainte soit satisfaite.

En ce qui concerne la deuxième contrainte, la simple observation suivante montre qu'elle est également satisfaite.

Parce qu'à partir de (2.1)

$$e = \sum_{i=1}^{n} e_i = kb + r$$

et aussi

$$e = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (k_i\, b + r_i) = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i$$

$$= b \sum_{i=1}^{n} k_i + S_n \qquad (4.1)$$

il est logique que le nombre entier restant de S_n, lors d'une division par b, doit également être r. Il faut donc avoir, pour un certain nombre entier m,

$$S_n = mb + r \qquad 0 \leq r < b \qquad (4.2)$$

Ainsi à partir de (4.1), nous obtenons

$$e = \sum_{i=1}^{n} e_i = b \left(\sum_{i=1}^{n} k_i + m \right) + r$$

c'est-à-dire,

$$k = \sum_{i=1}^{n} k_i + m$$

Il découle immédiatement de (4.2) que le nombre d'étapes L, requises pour terminer la procédure est lié à m, r et R_1 comme suit :

$$\text{Prob}\,(L = m + 1) = \text{Prob}\,(1 \leq R_1 \leq r) = \frac{r}{b}$$

$$\text{Prob}\,(L = m) = \text{Prob}\,(r < R_1 \leq b) = 1 - \frac{r}{b}$$

Parce que

$$\sum_{i=1}^{n} e_i^* = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i^* = b \sum_{i=1}^{n} k_i + Lb$$

nous avons

$$\sum_{i=1}^{n} e_i^* = \begin{cases} b \sum_{i=1}^{n} k_i + (m+1)\, b & \text{avec une probabilité } \frac{r}{b} \\ b \sum_{i=1}^{n} k_i + mb & \text{avec une probabilité } 1 - \frac{r}{b} \end{cases}$$

$$= \begin{cases} kb + b & \text{avec une probabilité } \frac{r}{b} \\ kb & \text{avec une probabilité } 1 - \frac{r}{b} \end{cases} \qquad (4.3)$$

aussi,

$$\left(\sum_{i=1}^{n} e_i\right)^* = kb + r^*$$

si

$$P(r^* = b) = \frac{r}{b}$$

$$P(r^* = 0) = 1 - \frac{r}{b}$$

afin que

$$\left(\sum_{i=1}^{n} e_i\right)^* = \begin{cases} kb + b & \text{avec une probabilité } \frac{r}{b} \\ kb & \text{avec une probabilité } 1 - \frac{r}{b}. \end{cases} \qquad (4.4)$$

En comparant (4.3) et (4.4), nous observons immédiament que les variables aléatoires $\Sigma\, e_i^*$ et $(\Sigma\, e_i)^*$ ont la même distribution.

Ainsi, l'effet net de la procédure sur un agrégat prédéfini de chiffres individuels arrondis aléatoirement équivaut à l'arrondissement aléatoire de l'agrégat lui-même.

Il est également assez facile de montrer que la même observation s'applique à la somme de n'importe quels chiffres consécutifs $e_t, e_{t+1}, \ldots, e_{t+5}$. Ainsi, l'arrondissement aléatoire contrôlé mène à une diminution souhaitable de la variance de l'arrondissement non seul pour un agrégat prédéfini, mais également pour toute région définie par l'utilisateur qui consiste en l'union de régions formant des composantes de base consécutives.

Tableau 1
Exemple d'arrondissement aléatoire contrôlé

											Total
Totalisation non arrondie de l'AE (e_i)	12	23	34	3	49	23	50	17	8	13	232
Base qui ne peut pas être arrondie ($k_i b$)	10	20	30	0	45	20	50	15	5	10	205
Résidu (r_i)	2	3	4	3	4	3	0	2	3	3	27
Résidu cumulatif (s_i)	2	5	9	12	16	19	19	21	24	27	
$R_1 = 1$	*		*	*	*			*		*	
Résidu arrondi (r_i^*)	5	0	5	5	5	0	0	5	0	5	30
Totalisation arrondie de l'AE (e_i^*)	15	20	35	5	50	20	50	20	5	15	235
$R_1 = 2$	*		*	*		*			*	*	
Résidu arrondi (r_i^*)	5	0	5	5	0	5	0	0	5	5	30
Totalisation arrondie de l'AE (e_i^*)	15	20	35	5	45	25	50	15	10	15	235
$R_1 = 3$		*	*		*	*			*		
Résidu arrondi (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Totalisation arrondie de l'AE (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 4$		*	*		*	*			*		
Résidu arrondi (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Totalisation arrondie de l'AE (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 5$		*		*	*			*		*	
Résidu arrondi (r_i^*)	0	5	0	5	5	0	0	5	0	5	25
Totalisation arrondie de l'AE (e_i^*)	10	25	30	5	50	20	50	20	5	15	230
Nombre de fois pour l'arrondissement vers le haut	2	3	4	3	4	3	0	2	3	3	2
Nombre de fois pour l'arrondissement vers le bas	3	2	1	2	1	2	5	3	2	2	3

Bibliographie

[1] Barabba, V.P. et Kaplan, D.C., “U.S. Census Bureau statistical techniques to prevent disclosure - The right to privacy vs. the need to know”. Document présenté à la 2e réunion de l’International Association of Survey Statisticians, Varsovie (1975).

[2] Fellegi, I.P. et Phillips, J.L., “Statisticians Confidentiality: Some Theory and Applications to Data Dissemination”. *Annals of Economic and Social Measurement*, pp. 399-409, 1974.

[3] Nargundkar, M.S. et Saveland, W., “Random Rounding: A Means of Preventing Disclosure of Information about Individual Respondents in Aggregate Data”. *Proceedings of the Social Statistics Section* of ASA, 1972.

[4] Newman, D., “Rounding and Error Injection of Preserving Confidentiality of Census Data”. Document présenté à la 2e réunion de l’International Association of Survey Statisticians, Varsovie (1975).