Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 145–156
Statistique Canada

Estimations fondées sur des données arrondies aléatoirement

C.S. WITHERS[1]

RÉSUMÉ

Dans cet article, nous présentons des méthodes servant à estimer les fonctions des probabilités de cellule associées à un tableau de données multinomiales qui ont été arrondies aléatoirement selon des multiples d'un nombre donné l. Nous montrons que: (i) l'arrondissement aléatoire n'a que des effets de second ordre sur le biais et la variance; (ii) l'utilisation des estimateurs naturels des probabilités de cellule entraîne une très faible perte d'efficacité si la fréquence de la cellule est élevée par rapport à $(l^2 - 1)/(6R)$ où R désigne le nombre de cellules dans le tableau; et (iii) il existe des estimateurs dont le biais est apparemment de taille exponentiellement faible pour les moments de ces estimateurs naturels et les polynomes des probabilités de cellule.

MOTS-CLÉS: Arrondissement aléatoire; réduction du biais; efficacité.

1. INTRODUCTION ET SOMMAIRE

Dans cet article, nous décrivons des méthodes qui permettent d'estimer une fonction des probabilités de cellule associées à un tableau de données multinomiales arrondies aléatoirement. L'arrondissement aléatoire est une méthode très répandue qui sert à préserver le caractère confidentiel des données lorsqu'une valeur contenue dans un tableau (par ex. : 1) risque d'être associée à une personne en particulier, ce qui peut entraîner la violation de confidentialité. Cette méthode consiste à arrondir la fréquence N, d'un tableau au multiple d'un nombre donné l supérieur avec une probabilité (a.p.) α ou au multiple inférieur a.p. $1 - \alpha$, où α est choisie de telle manière que la valeur arrondie M satisfait l'équation suivante

$$E(M \mid N) = N.$$

En d'autres termes, si pour un nombre entier j, $jl \leq N < (j + 1)l$, alors

$$M = \begin{cases} jl \text{ a.p. } 1 - \alpha \\ (j + 1)l \text{ a.p. } \alpha \end{cases} \tag{1.1}$$

où $\alpha = r/l$ et $r = N - jl$.

Le département de Statistique de la Nouvelle-Zélande utilise $l = 3$ Penny comme base d'arrondissement l alors que Statistique Canada utiliserait $l = 5$. Voir Penny et Ryan (1986).

Il ne faut pas confondre l'arrondissement aléatoire avec le groupement ou arrondissement non aléatoire de données d'échantillon au multiple entier de l le plus près (associé avec les corrections de Sheppard pour les moments), ni avec la contamination aléatoire, qui est une autre méthode de confidentialité qui consiste simplement à ajouter à la fréquence N une variable aléatoire indépendante de moyenne 0. (Le principal inconvénient de cette méthode est qu'elle peut produire une fréquence négative). Pour obtenir les titres d'ouvrages traitant

[1] C.S. Withers, Applied Mathematics Division, Department of Scientific and Industrial Research, C.P. 1335, Wellington, Nouvelle-Zélande.

de ces méthodes, voir Gastwirth et coll. (1978) et Kendall et Stuart (1977). On trouvera également une bibliographie sur l'arrondissement aléatoire pour les données multidimensionnelles et les données groupées dans Gastwirth et coll. (1978).

Dans cet article, nous allons nous concentrer sur les problèmes liés à l'estimation d'une fonction de probabilités de celule associées à un tableau de R valeurs arrondies aléatoirement. Pour des raisons de commodité, nous désignons les probabilités de cellule par $p_1, \ldots, p_R$ au lieu de $\{p_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\}$, comme cela se fait normalement pour un tableau $I \times J$.

Ainsi, $1 = \Sigma_1^R p_i$ et $n = \Sigma_1^R N_i$ est la somme des fréquences contenues dans le tableau. Soient $\{M_i\}$ les valeurs arrondies de $\{N_i\}$. Étant donné n, nous supposons que $\{N_i\}$ suit une loi multinomiale avec les paramètres n et $\{p_i\}$. L'hypothèse ci-dessus est vraie pour $p_i = m_i / \Sigma_j m_j$ si $\{N_i\}$ sont non conditionnellement des variables de Poisson indépendantes avec des moyennes $\{m_i\}$.

Deux estimateurs sans biais de p_1 sont définis par:

$$p_1^* = N_1/n \text{ et } \hat{p}_1 = M_1/n. \tag{1.2}$$

Le premier n'est pas un vrai estimateur puisqu'on ne connaît pas N_1. Le second est l'estimateur naturel. (Nous supposons que n est connu. S'il ne l'est pas nous pouvons sans difficulté le remplacer par $\Sigma_1^R M_i$.) Il existe toutefois d'autres estimateurs sans biais, notamment "l'estimateur complémentaire"

$$\tilde{p}_1 = -\sum_{j \neq 1} M_j/n, \tag{1.3}$$

d'où

$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda \tilde{p}_1 \text{ pour n'importe valeur } \lambda \text{ donnée.} \tag{1.4}$$

Il faut alors se demander quelle est la meilleure valeur de λ que l'on peut utiliser et quelle perte d'efficacité doit-on subir si l'on utilise l'estimateur naturel c'est-à-dire $\lambda = 0$. Pour répondre à ces questions, il faut connaître les variances de ces estimateurs. Elles sont définies ci-dessous.

Théorème 1.1

Soit

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2)\, n^{-1} + \{(l^2 - 1)/6 + \Delta_n(p_1)\}n^{-2} = v_n(p_1), \tag{1.5}$$

où

$$\Delta_n(p_1) = \sum_{i=0}^{l-1} i(l - i)\,\{P(N_1 \bmod l = i) - l^{-1}\}. \tag{1.6}$$

Aussi,

$$\text{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + \{(R - 1)(l^2 - 1)/6 + \sum_{j \neq 1} \Delta_n(p_j)\}n^{-2}, \tag{1.7}$$

et

$$\text{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{\alpha(\lambda)(l^2 - 1)/6 + \nabla_n(p)\}n^{-2}, \tag{1.8}$$

où

$$\alpha(\lambda) = (1 - \lambda)^2 + (R - 1)\lambda^2 \tag{1.9}$$

et

$$\nabla_n(p) = (1 - \lambda)^2 \Delta_n(p_1) + \lambda^2 \Sigma_{i \neq 1} \Delta_n(p_i). \tag{1.10}$$

(Les démonstrations se trouvent à la section 2.)

Dans l'annexe A, nous démontrons que pour $0 < p_1 < 1$, $P(N_1 \bmod l = i) - l^{-1} \to 0$ de façon exponentielle lorsque $n \to \infty$, de sorte que $\Delta_n(p_1) \to 0$ de façon exponentielle lorsque $n \to \infty$, et $\nabla_n(p)$ fait de même à la condition que $p_i \neq 0$ pour tous les i.

Comme $\alpha(\lambda)$ est minimisée par $\lambda_R = R^{-1}$ et $\alpha(\lambda_R) = 1 - R^{-1}$, $\text{var}(p_1(\lambda))$ l'est aussi asymptotiquement. Ainsi, la perte d'efficacité causée par l'utilisation de l'estimateur naturel $\hat{p}_1$ au lieu de l'estimateur asymptotiquement optimal sans biais $p_1(\lambda_R)$ lorsque R est grand) est donnée par

$$\{\text{var}(\hat{p}_1) - \text{var}(p_1(\lambda_R))\}/\text{var}(p_1(\lambda_R)) \approx (l^2 - 1)/\{6Rn(p_1 - p_1^2)\}, \quad (1.11)$$

laquelle valeur est négligeable si $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ est élevée par rapport à $(l^2 - 1)/\{6R\}$.

En règle générale, M_1 est une bonne approximation du membre gauche de (1.11). Nous avons ainsi une façon pratique de vérifier l'efficacité des estimateurs naturels. Si une ou plusieurs valeurs $\{p_i\}$ sont nulles, $p_i = 0$, alors $N_i = M_i = 0$, nous disons que $\Sigma_{i \neq 1}$ ne comprend pas les cellules pour lesquelles $p_i = 0$, et que R désigne le nombre de cellules du tableau pour lesquelles $p_i \neq 0$.)

En nous servant de l'équation (1.5) nous pouvons maintenant faire une brève comparaison de l'arrondissement aléatoire et de la contamination aléatoire. Selon nos informations, les organismes statistiques de l'Australie et du Royaume-Uni arrondissent les fréquences de cellule en leur ajoutant 1 avec une probabilité 1/4, 0 avec une probabilité 1/2 et -1 avec une probabilité 1/4, de sorte que

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

Le coefficient 1/2 passe à 4/3 dans le cas de la Nouvelle-Zélande ($l = 3$) et à 4 pour le Canada ($l = 5$). La méthode de la contamination aléatoire implique une moins grande protection (une variation maximale de 1 par opposition à 2 pour le système de Nouvelle-Zélande et à 4 pour le système canadien) et la possibilité d'une fréquence négative si elle est appliquée à des cellules vides.

Le théorème 1.1 montre que l'arrondissement aléatoire n'a qu'un effet de second ordre sur l'efficacité de l'estimateur de p_1; la variance n'est accrue que par un facteur de l'ordre n^{-2}. La démonstration suivante montre que ce résultat très important est aussi vrai pour l'estimation de n'importe quelle fonction lisse de $\{p_i\}$. Posons $r = R - 1$, $\mathbf{p} = (p_1, \ldots, p_r)$, $\mathbf{N} = (N_1, \ldots, N_r)$, $\mathbf{M} = (M_1, \ldots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ et $\hat{\mathbf{p}} = \mathbf{M}/n$. Alors nous avons $\text{cov}(\mathbf{p}^*) = V/n$ où $V = \text{diag}(\mathbf{p}) - \mathbf{pp}'$. Supposons maintenant que nous voulons estimer $f(\mathbf{p})$, une fonction dont les secondes dérivées sont continues.

Ainsi, $\dot{f}(\mathbf{p}) = \partial f(\mathbf{p})/\partial \mathbf{p}$ est une fonction continue $r \times 1$ et $\ddot{f}(\mathbf{p}) = \partial^2 f(\mathbf{p})/\partial \mathbf{p} \partial \mathbf{p}'$ est une fonction continue $r \times r$.

Théorème 1.2. Lorsque $n \to \infty$, $E(f(\mathbf{p}^*))$ et $E(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } B(\mathbf{p}) = \text{trace}\,(\ddot{f}(\mathbf{p})V/2). \quad (1.12)$$

De plus, $\text{var}(f(\mathbf{p}^*))$ et $\text{var}(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } v(\mathbf{p}) = \dot{f}(\mathbf{p})' V \dot{f}(\mathbf{p}). \quad (1.13)$$

Ce théorème montre que

(a) l'arrondissement aléatoire n'accroît que de $O(n^{-2})$ la variance de l'estimateur naturel de $f(\mathbf{p})$;

(b) l'arrondissement aléatoire n'a qu'un effet du second ordre sur le biais de l'estimateur naturel de $f(\mathbf{p})$.

Selon (1.12), l'estimateur naturel de $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, a un biais d'ordre n^{-1}. Nous allons voir maintenant comme réduire ce biais à un ordre n^{-2}.

Corollaire 1.1. Si pour une fonction $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$, alors $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Deux estimateurs possible pour $f_n(\hat{\mathbf{p}})$ dont "l'estimateur delta" pour lequel

$$f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^{r} f_{ii}(\mathbf{p}) p_i - \mathbf{p}' \ddot{f}(\mathbf{p}) \mathbf{p} \right\} / (2n), \tag{1.14}$$

où $f_{ii}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$, et "l'estimateur jack-knife" pour lequel

$$f_n(\mathbf{p}) = nf(\mathbf{p}) - (n-1)\bar{f}, \tag{1.15}$$

où

$$\bar{f} = \Sigma_{i=1}^{r} p_i f([(n\mathbf{p} - e_i)/(n-1)]) + (1 - \Sigma_1^r p_i) f([n\mathbf{p}/(n-1)]),$$

e_i = le i-ième vecteur unitaire dans R^r,

$$\text{et } [x] : R^r \to R^r \text{ est définie par } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \le x_i \le 1. \\ 1, & x_i > 1 \end{cases}$$

Ces estimateurs ont été calculés dans Withers (1987a et 1987b). En particulier, si $f(\mathbf{p})$ est uniquement une fonction de p_1, par exemple $f(\mathbf{p}) = g(p_1)$, alors $f_n(\mathbf{p}) = g(p_1) - \ddot{g}(p_1)(p_1 - p_1^2)/(2n)$ et $\bar{f} = p_1 g([(np_1 - 1)/(n-1)]) + (1 - p_1) g([np_1/(n-1)])$. Si par exemple, $f(\mathbf{p}) = p_1^2$ nous avons, pour l'estimateur delta, l'équation suivante: $f_n(\mathbf{p}) = p_1^2 \{1 - (1 - p_1)/n\}$.

Nous allons maintenant montrer que si $f(\mathbf{p})$ est une fonction polynomiale, il est effectivement possible de trouver un estimateur de $f(\mathbf{p})$ fondé sur l'estimateur naturel avec un biais apparemment de taille exponentiellement faible. Pour cela, nous prenons le cas $f(\mathbf{p}) = p_1^2$.

Théorème 1.3. $\hat{\lambda}_1 = \{\hat{p}_1^2 - n^{-1}\hat{p}_1 - n^{-2}(l^2 - 1)/6\}(1 - n^{-1})^{-1}$ est un estimateur de $\lambda_1 = p_1^2$ avec un biais $\Delta_n(p_1)(n^2 - n)^{-1}$.

De même si $f_n(\mathbf{p})$ est un moment de $\hat{\mathbf{p}}$ il est possible de trouver un estimateur de $f_n(\mathbf{p})$ avec un biais apparemment de taille exponentiellement faible. À des fins d'illustration, nous prenons le cas $f_n(\mathbf{p}) = \text{var}(\hat{p}_1)$.

Théorème 1.4. $\hat{\lambda}_{2n} = n^{-1}(\hat{p}_1 - \hat{\lambda}_1) - n^{-2}(l^2 - 1)/6$ est un estimateur de $\lambda_{2n} = \text{var}(\hat{p}_1)$ avec un biais $-\Delta_n(p_1)(n^2 - n)^{-1}$.

On peut étendre ces résultats à des moments et à des polynômes d'ordre supérieur au moyen de l'expression définie dans l'annexe B pour les moments et les cumulants de $\hat{\mathbf{p}}$. Nous allons maintenant montrer qu'il existe un estimateur sans biais pour le cas particulier où $f(\mathbf{p})$ est collinéaire.

Théorème 1.5. Soit $f_I(\mathbf{p}) = \Pi_{i=1}^{I} p_i$ où $1 \le I \le R$ et

$$a_{nI} = n^{-I} n!/(n-I)! = (1 - n^{-1})(1 - 2n^{-1}) \ldots (1 - \{I - 1\}n^{-1}). \tag{1.16}$$

Alors $$E(f_I(\hat{\mathbf{p}})) = E(f_I(\mathbf{p}^*)) = f_I(\mathbf{p})a_{nI}. \tag{1.17}$$

$f_I(\hat{\mathbf{p}})/a_{nI}$ est donc un estimateur sans biais de $f(\mathbf{p})$.

Corollaire 1.2. $\text{cov}(\hat{p}_1,\hat{p}_2) = -p_1p_2/n$. L'estimateur sans biais correspondant est $-\hat{p}_1\hat{p}_2/(n-1)$. D'une manière plus générale, pour $1 \leq I \leq R$, $\text{E}(\Pi_{i=1}^I(\hat{p}_i - p_i)) = c_{nI}\Pi_{i=1}^I p_i$ avec l'estimateur sans biais $(\Pi_{i=1}^I \hat{p}_i)a_{nI}/c_{nI}$ où $c_{nI} = \Sigma_{j=0}^I(-1)^{I-j}\binom{I}{j}a_{nj}$. (On obtient le même résultat si on remplace $\hat{\mathbf{p}}$ par $\mathbf{p}^*$.)

L'équation (1.16) nous permet de calculer des estimateurs sans biais pour d'autres polynômes spéciaux en $\mathbf{p}$ tels que p_1^2, $p_1p_2(p_1 + p_2)$et $\Sigma_{i=1}^R p_i^3$ - mais non pour des polynômes de la forme $p_1^2p_2$ ou p_1^3.

Corollaire 1.3. Pour $1 \leq I < R$ un estimateur sans biais de

$$f_I(\mathbf{p}) \sum_1^I p_i \text{ est } f_I(\hat{\mathbf{p}}^1)\left\{1 - In^{-1} - \sum_{I+1}^R \hat{p}_i\right\}/a_{n,I+1}. \tag{1.18}$$

En particulier, $\hat{p}_1(\tilde{p}_1 - n^{-1})(1 - n^{-1})^{-1}$ est un estimateur sans biais de p_1^2. (1.19)

Nous tenons à souligner que les résultats de cette étude reposent sur l'hypothèse que les fréquences de tableaux sont des variables de Poisson indépendantes ou à tout le moins, des variables multinomiales, étant donné le total. Le modèle de Poisson et le modèle multinomial sont tous deux intéressants parce qu'ils sont d'interprétation simple et parce que la somme de variables de Poisson est elle-même une variable de Poisson. Cependant, la somme de variables multinomiales n'est multinomiale que si les probabilités de cellule $\mathbf{p}$ sont les mêmes pour toutes les variables. Cela donne à penser que les modèles multinomiaux peuvent produire des conclusions moins exactes si les populations étudiées sont constituées de deux groupes non-homogènes ou plus.

2. PREUVES

Preuve du théorème 2.1. Posons $r = N_1 \bmod l$. Alors l'équation (1.1) est vraie pour $N = N_1$, $M = M_1$ avec $jl = N - r$ et

$$E(M_1^2 \mid r) = (N_1 - r)^2(1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

Par conséquent,

$$E(\hat{p}_1^2) = E(p_1^{*2}) + n^{-2}A_n(p_1), \tag{2.1}$$

où

$$A_n(p_1) = E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{i=0}^{l-1}(li - i^2)P(N = i)$$

$$= (l^2 - 1)/6 + \Delta_n(p_1)$$

puisque

$$l^{-1} \sum_{i=0}^{l-1} i(l - i) = (l^2 - 1)/6. \tag{2.2}$$

Or,

$$E(p_1^{*2}) = p_1^2 + (p_1 - p_1^2)n^{-1}, \tag{2.3}$$

ce qui nous amène à l'équation (1.5). Par ailleurs, $\tilde{p}_1 = \hat{p}_1 - \sum (M_j - N_j)/n$,

alors

$$E(\tilde{p}_1^2) = E(\hat{p}_1^2) - 2n^{-2} \sum E(M_1(M_j - N_j)) + n^{-2} \sum E((M_i - N_i)(M_j - N_j))$$

$$= E(\hat{p}_1^2) - 2n^{-2}A_n(p_1) + n^{-2} \sum A_n(p_i)$$

puisque $E(\Pi_i f_i(M_i)|\{N_i\}) = \Pi_i E(f_i(M_i)|N_i)$. (2.4)

Par conséquent $\text{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + n^{-2}\Sigma_{i\neq 1} A_n(p_i)$ ce qui vérifie l'équation (1.7).

Par ailleurs

$$E(\hat{p}_1\tilde{p}_1) = p_1 - n^{-2} \sum_{i\neq 1} E(M_1M_i) = p_1 - \sum_{i\neq 1} E(p_1^*p_i^*)$$

$$= p_1 - \sum_{i\neq 1} p_1p_i(1 - n^{-1}) = p_1 - p_1(1 - p_1)\ (1 - n^{-1}),$$

alors,

$$\text{cov}(\hat{p}_1,\tilde{p}_1) = (p_1 - p_1^2)n^{-1}. \tag{2.5}$$

Par conséquent, $\text{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{(1 - \lambda)^2A_n(p_1) + \lambda^2\Sigma_{i\neq 1} A_n(p_i)\}n^{-2}$ ce qui vérifie (1.8).

Preuve du théorème 1.2. Cette démonstration a été faite pour $\mathbf{p}^*$ dans Withers (1987a). En outre, comme $\dot{f}$ est finie dans un voisinage de $\mathbf{p}$,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)'\dot{f}(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*)|N) = 0,\ E((\hat{p}_1 - p_i^*)^2|N) = 2n^{-2}I(N_1 \bmod l \neq 0),$$

où $I(A) = 1$ ou 0 selon que A est vrai ou faux, autrement dit, $I(\cdot)$ est la fonction indicatrice. Par conséquent $\text{E}(f(\hat{\mathbf{p}})) = \text{E}(f(\mathbf{p}^*)) + O(n^{-2})$ et $\text{var}(f(\hat{\mathbf{p}})) = \text{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Preuve du théorème 1.3. Découle directement de (2.1) et de (2.3).

Preuve du théorème 1.4. Découle de (2.1) et de (1.5).

Preuve du théorème 1.5. La première équation de l'expression (1.16) découle de (2.4) tandis que la seconde découle du théorème multinomial. Le corollaire 1.2 s'ensuit automatiquement.

Preuve du corollaire 1.3. Selon (1.16), pour $1 \leq I < i \leq R$ nous avons

$$E(f_I(\hat{\mathbf{p}})\hat{p}_i) = f_I(\mathbf{p})p_i a_{n,I+1}$$

alors

$$E(f_I(\hat{\mathbf{p}}) \sum_{I+1}^{R} \hat{p}_i / a_{n,I+1}) = f_I(\mathbf{p})(1 - \Sigma_1^I p_i)$$

$$= E(f_I(\hat{\mathbf{p}}) / a_{nI}) - f_I(\mathbf{p}) \Sigma_1^I p_i.$$

REMERCIEMENTS

Je tiens à remercier Peter McGavin pour les calculs de l'annexe A.

ANNEXE A

Pour une fonction lisse f on devrait avoir:

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) \to f(\mathbf{p})l^{-s} \tag{A.1}$$

lorsque $n \to \infty$ pourvu que $0 < p_i < 1$ pour $1 \leq i \leq s \leq R$.

Si $E(f(\hat{\mathbf{p}})) = f(\mathbf{p})$, on devrait observer une vitesse de convergence exponentielle, c'est-à-dire $O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Si $f(\hat{\mathbf{p}})$ est biaisé, le biais est alors $O(n^{-1})$, et cette valeur devrait représenter une vitesse de convergence pour (A.1). En règle générale, la convergence cesse lorsque $\mathbf{p}$ approche les limites de $[0,1]^r$, puisque

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l))$$

$$= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \ldots = j_s = 0) & \text{si } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{si } p_1 = 1. \end{cases}$$

Pour vérifier ces hypothèses, nous avons considéré le cas où $s = 1$, $l = 3$, $j = 0$ et les fonctions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, et (c) $f(\mathbf{p}) = \exp(p_1)$. Les calculs ont été effectués à quadruple précision sur un VAX11/780, ce qui a permis de calculer

$$\Delta = E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

à une précision de 112 bits, soit près de 34 décimales. Les figures 1a, 1b et 1c décrivent la relation entre Δ et p_1 pour $n = 6, 18, 54$. Puisque $n \bmod 3 = 0$, Δ est symétrique autour de $p_1 = 1/2$ pour (a).

Insert Figure 1 (a), 1(b)

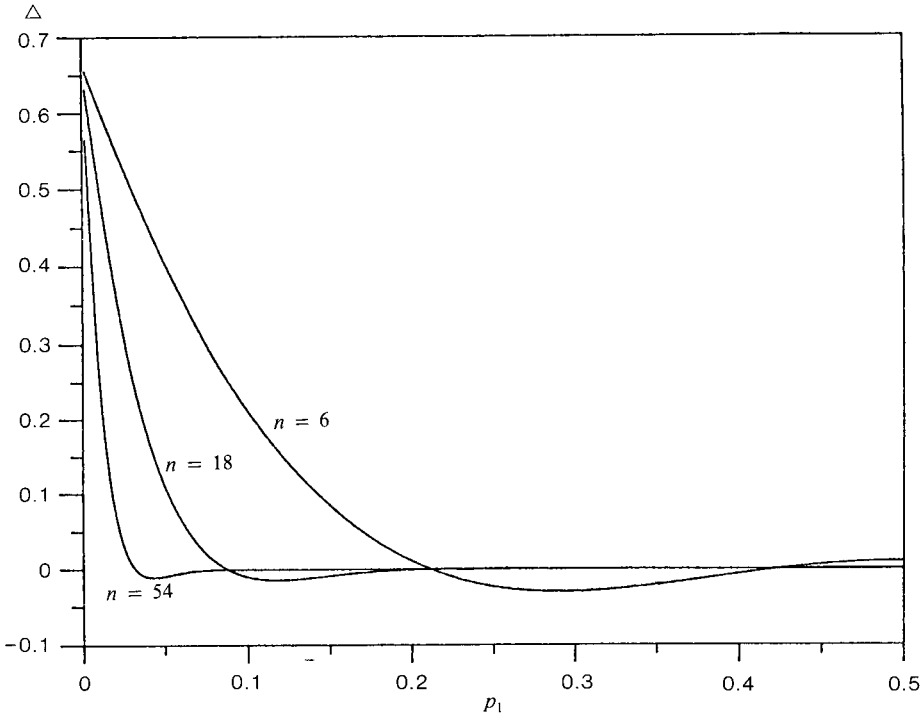

Figure 1a. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = 1$.

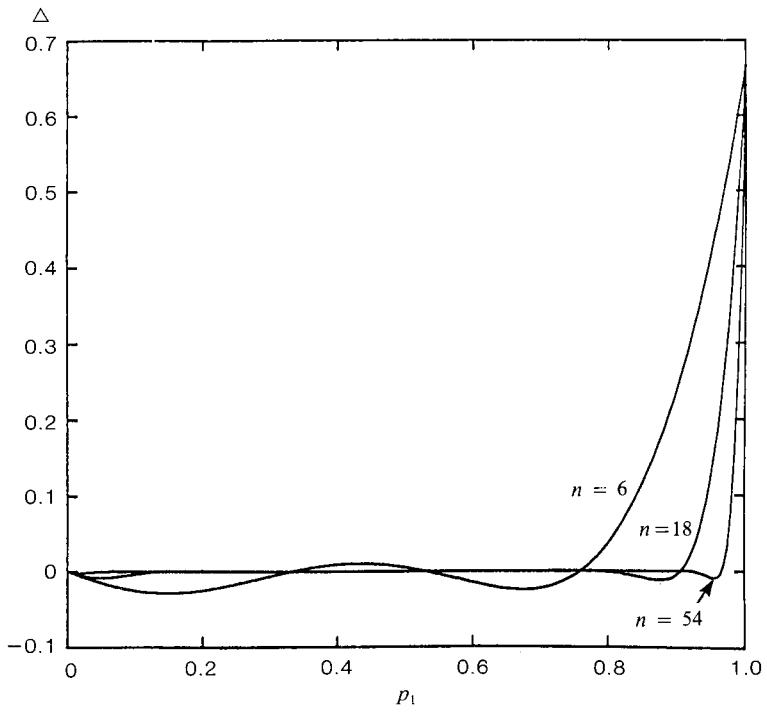

Figure 1b. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = p_1$.

Insert Figure 1(c), 2(a)

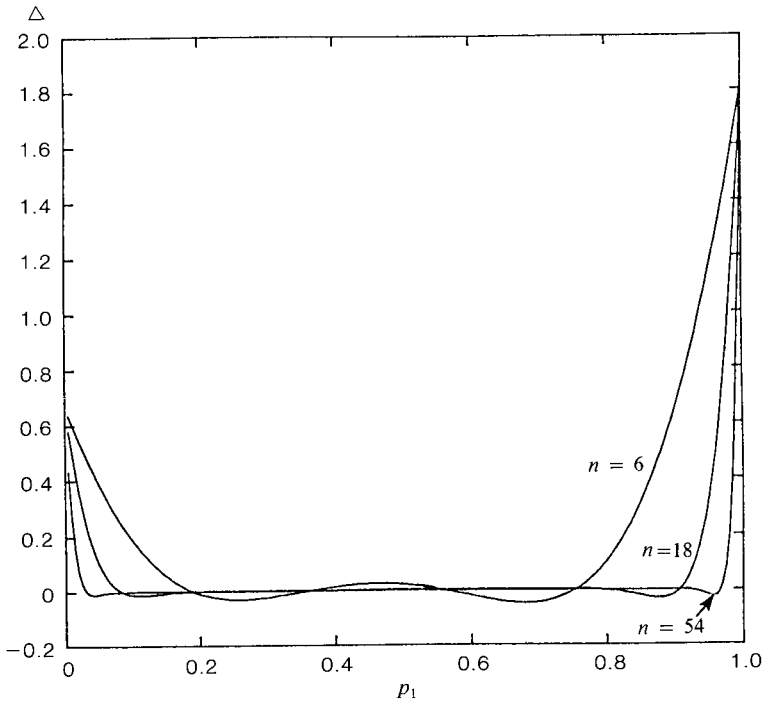

Figure 1c. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = \exp(p_1)$.

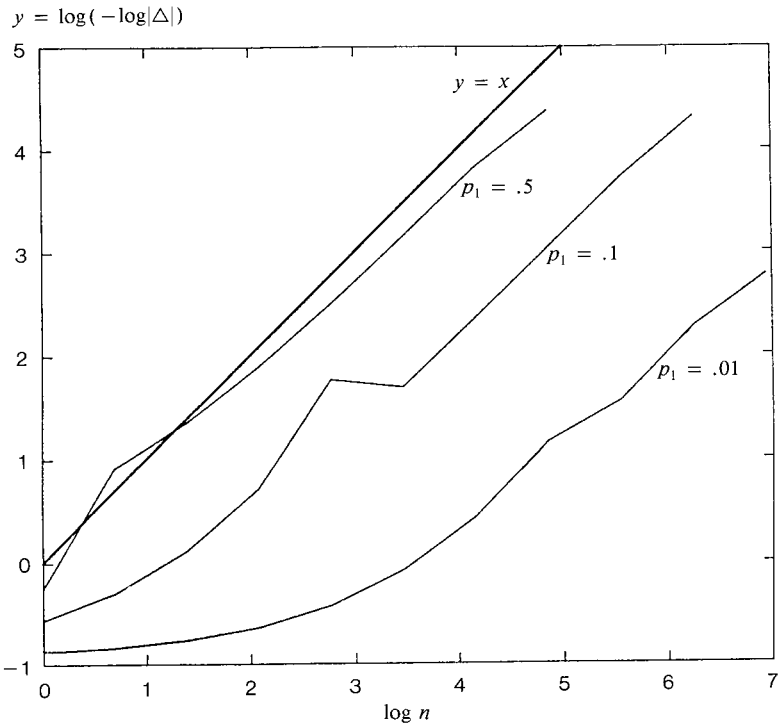

Figure 2a. Confirmation de la convergence exponentielle en (A.1) lorsque $f(\mathbf{p}) = 1$.

Insert Figure 2 (b), 3

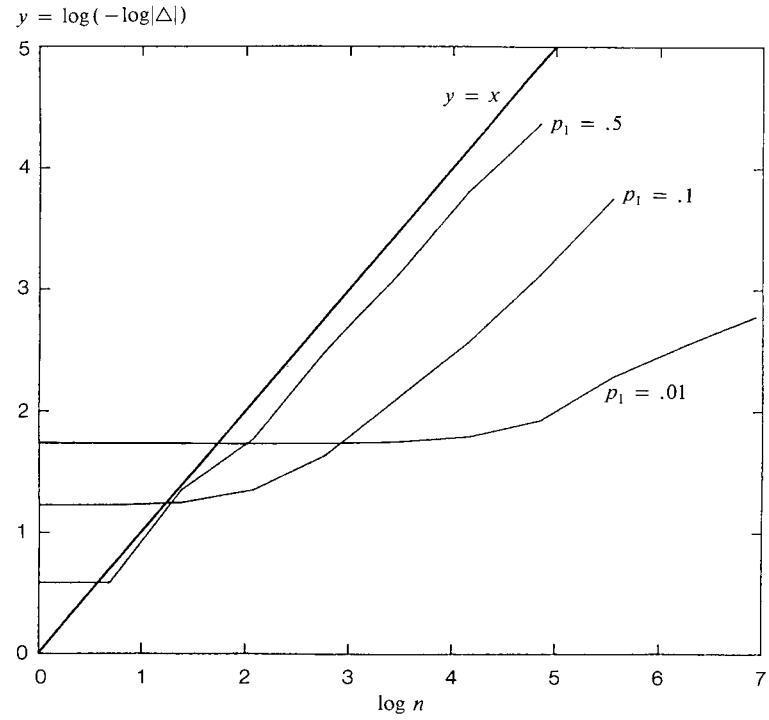

Figure 2b. Confirmation de la convergence exponentielle en (A.1) lorsque $f(\mathbf{p}) = p_1$.

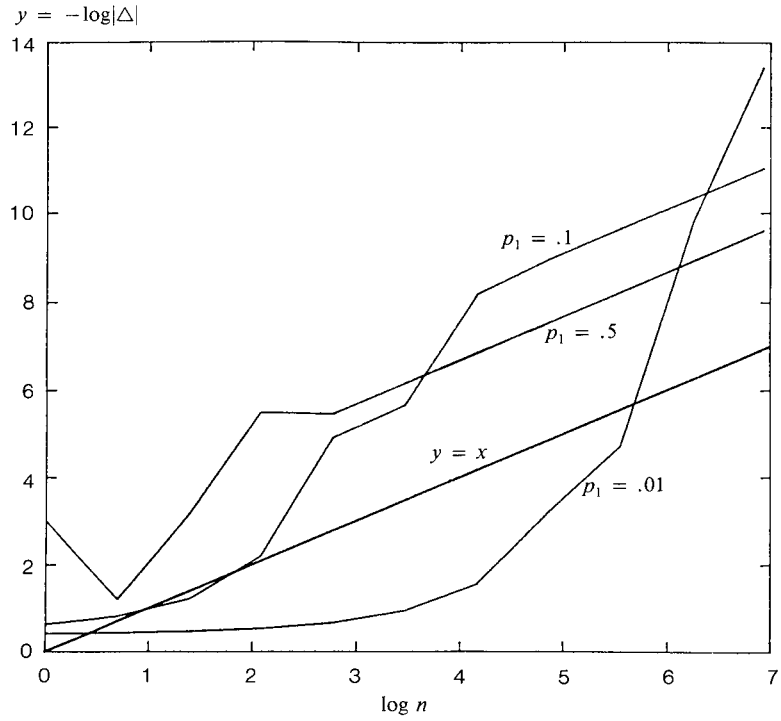

Figure 3. Confirmation de la convergence en (A.1) à un rythme de $\sim n^{-1}$ lorsque $f(\mathbf{p}) = \exp(p_1)$.

Puisque $\Delta = 2/3f(0)$ à $p_1 = 0$, et est égale à 2/3, 0 et 2/3 pour (a), (b) et (c) respectivement. La convergence cesse à $p_1 = 0$ pour (a) et (c), mais non pour (b). Pour $n = 18$, Δ est déjà à peu près nul pour les valeur de p_1 situées dans l'intervalle (.2, .8) pour la fonction (a) et pour les valeurs de p_1 situées dans l'intervalle (.1, .8) pour les (b) et (c). Pour $n = 54$, ces intervalles deviennent (.1, .9) pour (a), (.02, .95) pour (b), et (.07, .95) pour (c).

Les figures 2a et 2b décrivent la relation entre $Y = \log(-\log|\Delta|)$ et $X = \log(n)$ pour (a) $f(\mathbf{p}) = 1$ et (b) $f(\mathbf{p}) = p_1$. Comme prévu abstraction faite des petites valeurs de n, les courbes sont à peu près parallèles à la courbe $Y = X$ (sauf en ce qui concerne la fonction (b) pour $p = .01$), ce qui est conforme à la relation $\Delta = O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Les courbes ne sont pas lisses parce que Δ n'a été calculé qu'en fonction de n puissance de $2 (n = 2^i$ pour $0 \leq i \leq 7)$.

La figure 3 décrit la relation entre $Y = -\log|\Delta|$ et $X = \log(n)$ pour (c) $f(\mathbf{p}) = \exp(p_1)$. Pour les valeurs supérieures de n, les courbes sont parallèles à la courbe $Y = X$ pour $p_1 = .5$ et .1, ce qui est conforme à $\Delta = O(n^{-1})$; pour $p_1 = 0.1$ toutefois, la courbe ne reflète pas du tout une relation linéaire; le rythme d'accroissement est beaucoup plus prononcé. Les graphiques confirment de façon générale nos hypothèses sur la vitesse de convergence de (A.1). Si nous voulions vérifier ces hypothèses par des méthodes analytiques, il nous faudrait recourir à la théorie des nombres.

ANNEXE B

Dans cette annexe, nous comparons les moments et les cumulants de $\mathbf{p}^* = N/n$ et $\hat{\mathbf{p}} = M/n$. Posons $q_1 = 1 - p_1$, $n_i = N_i \bmod l$, et $m_i(j) = E(p_1^{*i} I(n_1 = j)) \to p_1^i / l$ lorsque $n \to \infty$, en supposant que $p_1 \neq 0$ or 1.

Par des calculs élémentaires, nous obtenons

$$\mu(\hat{\mathbf{p}}) = \mu(\mathbf{p}^*) = \mathbf{p},$$

$$\mu_2(\hat{p}_1) = \mu_2(p_1^*) + M_{22}n^{-2} = p_1q_1n^{-1} + O(n^{-2}),$$

où

$$M_{22} = A_n(p_1) = \sum_{i=0}^{l-1} i(l-i)m_0(i) \to (l^2-1)/6$$

lorsque $n \to \infty$,

$$\mu_3(\hat{p}_1) = \mu_3(p_1^*) + 3n^{-2} \sum_{j=0}^{l-1} (lj - j^2)\{m_2(j) - 2p_1m_1(j) + p_1^2m_0(j)\}$$

$$+ n^{-3} \sum_{j=0}^{l-1} a_{jl}m_0(j)$$

$$= \mu_3(p_1^*) + o(n^{-2}) = p_1q_1(1-2p_1)n^{-2} + o(n^{-2}),$$

et

$$a_{jl} = -j^3(1 - j/l) + (l-j)^3 j/l.$$

De même $\mu_4(\hat{p}_1)$ peut s'écrire $\mu_4(p_1^*) + \Sigma_2^4 M_{4i}n^{-i} = O(n^{-2})$ et $\kappa_4(\hat{p}_1)$ peut s'écrire $\Sigma_2^4 k_{4i}n^{-i}$ où $k_{42} = M_{42}$ ne tend pas vers 0 lorsque $n \to \infty$. D'où $\kappa_4(\hat{p}_1) \sim n^{-2}$, et non

n^{-3}. Par conséquent, $\hat{\mathbf{p}}$ ne satisfait pas à l'hypothèse de Cornish-Fisher selon laquelle $\kappa_r(\hat{\mathbf{p}}) = O(n^{1-r})$ pour $r \geq 1$: voir par exemple Kendall et Stuart (1977).

On peut aussi déterminer les moments et les cumulants à l'aide de la f.g.m. (fonction génératrice des moments), que nous définissons ci-sessous.

$$E(\exp(t_1 M_1/n) \mid N_1) = \exp(t_1 N_1/n) S(t_1, n_1),$$

où

$$S(t_1, n_1) = (1 - n_1/l) \exp(-n_1 t_1/n) + (n_1/l) \exp(l - n_1) t_1/n.$$

Par (2.4), la f.g.m. est

$$E(\exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)) S(t) \text{ où } S(t) = \Pi_1^r S(t_i, n_i).$$

De plus, à $t = 0$, $S_1 = 0$ et $S_{ij\ldots} = 0$ si l'indice inférieur ne se repète pas. Par exemple, si nous posons

$$S = S(t),\ \partial_i = \partial/\partial t_i,\ S_i = \partial_i S,\ S_{ij} = \partial_i \partial_j S,\ \ldots$$

nous obtenons

$$E(\hat{p}_1^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2} S + 2p_1^* S_1 + S_{11}\}),$$

$$E(\hat{p}_1^2 \hat{p}_2^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2}(p_2^{*2} S + 2p_2^* S_2 + S_{22}) + 2p_1^*(p_2^{*2} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{*2} S_{11} + 2p_2^* S_{112} + S_{1122})\}).$$

Ainsi $E(\hat{p}_1^2) = E\{p_1^{*2} + S_{11}(0)\}$ et

$$E(\hat{p}_1^2 \hat{p}_2^2) = E\{p_1^{*2} p_2^{*2} + p_1^{*2} S_{22}(0) + p_2^{*2} S_{11}(0) + S_{1122}(0)\}.$$

où $S_{ii}(0) = S_{11}(0, n_i) = n^{-2}(l - n_i) n_i = n^{-2} \Sigma_{k=0}^{l-1} (l - k) k I(n_i = k)$ et $S_{1122}(0) = S_{11}(0)\, S_{22}(0)$. Nous pouvons simplifier davantage en utilisant $N_2 \mid N_1 \sim Bi(\theta, n - N_1)$ où $\theta = p_2/(1 - p_1)$. La f.g.m. multinomiale donne

$$E(p_1^{*2} p_2^{*2}) = n^{-4} p_1 p_2 \{(n)_4 p_1 p_2 + (n)_3 (p_1 + p_2) + (n)_2\}$$

où $(n)_i = n!/(n - i)! = n(n - i) \ldots (n - i + 1)$.

BIBLIOGRAPHIE

GASTWIRTH, J.L., KRIEGE, A.M., et RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 183-188.

KENDALL, M.G., et STUART, A. (1977). *The Advanced Theory of Statistics, Volume* 7. London: Griffin.

NARGUNDAR, M.S., et SAVELAND, W. (1972). Random rounding to prevent statistical disclosures. *Proceedings of the Social Statistics Section, American Statistical Association,* 382-385.

PENNY, R., et RYAN, M. (1986). A problem associated with random-rounding. *New Zealand Statistician,* 21, 43-52.

WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (en voie de rédaction).

WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Document non-publié, Department of Scientific and Industrial Research.