Techniques d'enquête, décembre 1991
Vol. 17, n° 2, pp. 135–142
Statistique Canada

Estimation de la variance dans un système de double enregistrement pour des populations hétérogènes

JUHA M. ALHO[1]

RÉSUMÉ

L'estimateur de système dual de la taille d'une population peut être fortement biaisé s'il y a hétérogénéité des probabilités de saisie. Dans cet article, nous étudions le biais de l'estimateur de variance correspondant dans des conditions d'hétérogénéité. Nous montrons que l'estimateur habituel est conservatif, c'est-à-dire qu'il produit des estimations excessives, lorsque les deux systèmes d'enregistrement sont corrélés négativement ou pas du tout ou encore lorsqu'il existe une corrélation positive mais faible. S'il existe une corrélation positive forte entre les deux systèmes, l'estimateur peut produire des sous-estimations. Nous proposons donc deux estimateurs, que nous comparons au premier. L'un est conservatif dans des conditions d'hétérogénéité quelconque, l'autre l'est dans des conditions d'hétérogénéité gaussienne. Nous appliquons ensuite ces estimateurs à des données sur les maladies professionnelles en Finlande.

MOTS CLÉS: Saisie-resaisie; système dual; hétérogénéité; maladies professionnelles.

1. INTRODUCTION

Supposons N individus dans une population fermée. Le problème consiste à estimer N au moyen d'un système de double enregistrement. Nous procédons à un échantillonnage double, selon lequel n_j individus sont prélevés (saisis) au temps $j, j = 1, 2$. Soit m le nombre d'individus prélevés deux fois. Définissons les variables indicatrices u_{ji} et m_i, pour $i = 1, \ldots, N$, telles que $u_{ji} = 1$, si et seulement si l'individu i est prélevé au temps j seulement ($j = 1, 2$), et $m_i = 1$ si et seulement si l'individu i est prélevé aux deux occasions. Autrement, u_{ji} et m_i sont égales à zéro. Définissons $n_{ji} = u_{ji} + m_i$ comme l'indicateur de saisie au temps j, $j = 1, 2$. De plus, posons $M_i = u_{1i} + u_{2i} + m_i$ comme la variable indiquant la saisie d'un individu à au moins une occasion. Enfin, désignons par $p_{ji} = E[n_{ji}]$, $j = 1, 2$; et $p_{12i} = E[m_i]$ les probabilités de saisie individuelles. Posons par hypothèse que les probabilités se situent strictement entre zéro et un. Le fait que les probabilités peuvent varier d'un individu à l'autre indique qu'il peut y avoir hétérogénéité des probabilités de saisie dans la population. Nous terminons la définition du modèle de double enregistrement (ou de saisie-resaisie) en supposant que les saisies sont des événements indépendants pour chaque individu, ou $p_{12i} = p_{1i}p_{2i}$, et que les vecteurs multinomiaux

$$(u_{1i}, u_{2i}, m_i, 1 - M_i) \sim \text{Mult}(1; p_{1i}q_{2i}, p_{2i}q_{1i}, p_{1i}p_{2i}, 1 - \phi_i),$$

où $q_{ji} = 1 - p_{ji}, j = 1, 2$, et $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$, sont indépendants pour $i = 1, \ldots, N$.

On sait très bien que lorsque les probabilités de saisie ne varient pas d'un individu à l'autre, c.-à-d. $p_{ji} = p_j, j = 1, 2$, l'estimateur du maximum de vraisemblance de N est $\hat{N} = n_1n_2/m$ (ou, plus précisément, l'entier le plus près de cette valeur vers le bas; voir Feller 1968, p. 46). Cet estimateur classique peut être fortement biaisé dans des conditions d'hétérogénéité (Seber 1982, p. 565; Burnham et Overton 1979, tableau 4, p. 931-932). Comme on peut le voir plus

[1] Juha M. Alho, Institute for Environmental Studies et Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Dr., Urbana IL, 61801, U.S.A.

bas dans l'exemple 1, lorsque les probabilités de saisie sont homogènes, la variance asymptotique de $\hat{N}$ est $\mathrm{Var}(\hat{N}) = Nq_1q_2/(p_1p_2)$, où $q_j = 1 - p_j$, $j = 1, 2$. On peut alors estimer $\mathrm{Var}(\hat{N})$ par $V_1 = n_1n_2u_1u_2/m^3$ (Sekar et Deming 1949, p. 114-115).

Cet article a pour but d'examiner le bien-fondé de l'estimateur de variance V_1 et de comparer le biais de cet estimateur avec celui de $\hat{N}$. Nous avons choisi d'étudier V_1 car il n'avait jamais été donné auparavant de vérifier si cet estimateur est efficace dans des conditions d'hétérogénéité, même lorsque $\hat{N}$ est convergent. Cela a pu être vérifié. De même, il était difficile de savoir dans quelles circonstances V_1 produit des estimations excessives et peut, par conséquent, aboutir à des intervalles de confiance valables malgré le biais de $\hat{N}$. Nous avons vu par la suite que cela était possible dans des circonstances particulières avec des intervalles unilatéraux.

Dans la section 2, nous calculons la variance asymptotique de $\hat{N}$, lorsque $N \to \infty$, et déterminons un estimateur conservatif de cette variance (désigné par V_2) dans des conditions d'hétérogénéité quelconque. Autrement dit, V_2 produit une surestimation de la variance asymptotique réelle. De cette façon, on pourrait espérer compenser le biais généralement négatif de $\hat{N}$ par une surestimation de la variance et obtenir quand même des intervalles de confiance valables. Malheureusement, cela ne semble possible que lorsque le biais $\hat{N}$ de est faible ou que N est petit. Dans la section 3, nous examinons le bien-fondé de V_1 dans des conditions d'hétérogénéité gaussienne et déterminons un estimateur V_3, lequel est conservatif dans ces conditions particulières d'hétérogénéité. La caractéristique gaussienne n'est pas indispensable pour les arguments; ce qui compte, c'est que les moments des paires (p_{1i}, p_{2i}) concordent avec ceux d'une distribution gaussienne à deux variables. Il est alors facile d'étudier l'effet de la corrélation entre p_{1i} et p_{2i} sur l'estimation de la variance car la corrélation ne peut être exprimée qu'en fonction d'un paramètre, soit le coefficient de corrélation de moments ordinaire. Dans la section 4, nous comparons le biais de l'estimateur de variance à celui de $\hat{N}$ en nous servant de données empiriques sur l'enregistrement des cas de maladie professionnelle en Finlande.

2. BIAIS ET VARIANCE DANS DES CONDITIONS D'HÉTÉROGÉNÉITÉ

Définissons $\bar{p}_{jN}$ comme la probabilité moyenne de saisie au temps j, $j = 1, 2$; et posons $\bar{p}_{12N}$ comme la moyenne des produits $p_{1i}p_{2i}$, $i = 1, \ldots, N$. Alors, $C_N = \bar{p}_{12N} - \bar{p}_{1N}\bar{p}_{2N}$ est la covariance des paires (p_{1i},p_{2i}). Supposons que les limites $\bar{p}_{jN} \to \bar{p}_j$, $j = 1, 2$; $\bar{p}_{12N} \to \bar{p}_{12}$, et $C_N \to C$ existent. En conséquence, $\hat{N}/N \to \bar{p}_1\bar{p}_2/\bar{p}_{12}$, de sorte que $\hat{N}/N - 1 \to -C/\bar{p}_{12}$, lorsque $N \to \infty$. C'est ce qu'on appelle le biais asymptotique de l'estimateur classique dans des conditions d'hétérogénéité. Fait digne de mention, ce biais ne dépend que des deux premiers moments de la distribution des paires (p_{1i}, p_{2i}). Il est notoire (Sekar et Deming 1949, p. 105-106; Seber 1982, p. 86) que lorsque la covariance est nulle ($C = 0$), l'estimateur classique est convergent; si $C > 0$, $\hat{N}$ produit une sous-estimation et dans le cas contraire, il produit une surestimation. Comme nous l'avons indiqué plus haut, nous cherchons particulièrement à savoir si V_1 est efficace lorsque p_{ji} varie selon les individus et que C est égal à 0.

Nous allons maintenant calculer la variance asymptotique de l'estimateur classique selon notre modèle général d'hétérogénéité. Notons qu'il n'existe pas de variance finie car il y a une probabilité réelle que $m = 0$. La "variance asymptotique" désigne donc ici la variance de la distribution limite plutôt que la limite de la variance lorsque $N \to \infty$.

Lemme 1. La variance asymptotique de $\hat{N}$ est

$$\mathrm{Var}(\hat{N}) = N\left\{\frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^3} - \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^2} - \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^2} - \frac{\bar{p}_2^2}{\bar{p}_{12}^2}\bar{S}_1 - \frac{\bar{p}_1^2}{\bar{p}_{12}^2}\bar{S}_2 - \frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^4}\bar{S}_3 + 2\left(\frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3}\bar{S}_4 + \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3}\bar{S}_5\right)\right\},$$

où $\bar{S}_j = S_j/N$ pour $j = 1, \ldots, 5$, et

$$S_1 = \sum_{i=1}^{N} p_{1i}^2, \quad S_2 = \sum_{i=1}^{N} p_{2i}^2, \quad S_3 = \sum_{i=1}^{N} p_{1i}^2 p_{2i}^2,$$

$$S_4 = \sum_{i=1}^{N} p_{1i}^2 p_{2i}, \quad S_5 = \sum_{i=1}^{N} p_{1i} p_{2i}^2.$$

La démonstration de ce lemme est présentée en annexe. Nous constatons que, contrairement au biais de $\hat{N}$, qui dépend uniquement des deux premiers moments des paires (p_{1i}, p_{2i}), $\mathrm{Var}(\hat{N})$ dépend des quatre premiers moments. Dans des cas particuliers comme ceux décrits dans l'exemple 2 et la proposition 2, on peut simplifier la représentation.

Exemple 1. Supposons que les probabilités de saisie ne sont pas hétérogènes, c.-à-d. $p_{ji} = p_j$, $j = 1, 2$. Alors, $\bar{p}_j = p_j, j = 1, 2; \bar{p}_{12} = p_1 p_2; \bar{S}_j = p_j^2, j = 1,2; \bar{S}_3 = p_1^2 p_2^2, \bar{S}_4 = p_1^2 p_2$ et $\bar{S}_5 = p_1 p_2^2$. Par conséquent, la variance asymptotique est $\mathrm{Var}(\hat{N}) = N(1 - p_1 - p_2 + p_1 p_2)/(p_1 p_2) = N q_1 q_2/(p_1 p_2)$. Des estimateurs convergents de Np_1p_2 et de Np_j sont m et n_j, respectivement, $j = 1, 2$. En d'autres termes, $Np_j/n_j \to 1, j = 1, 2$, et $Np_1p_2/m \to 1$, lorsque $N \to \infty$. On obtient ainsi V_1 comme estimateur de $\mathrm{Var}(\hat{N})$.

Exemple 2. Supposons que les paires (p_{1i}, p_{2i}), $i = 1, \ldots, N$, sont indépendantes en ce sens que la distribution des probabilités p_{1i} est la même pour chaque valeur de p_{2i}. Alors, $\bar{p}_{12} = \bar{p}_1 \bar{p}_2, \bar{S}_3 = \bar{S}_1 \bar{S}_2, \bar{S}_4 = \bar{p}_2 \bar{S}_1, \bar{S}_5 = \bar{p}_1 \bar{S}_2$. Par substitution dans l'équation du lemme 1, nous obtenons:

$$\mathrm{Var}(\hat{N}) = N\left(\frac{1}{\bar{p}_1\bar{p}_2} - \frac{1}{\bar{p}_2} - \frac{1}{\bar{p}_1} - \frac{\bar{S}_1\bar{S}_2}{\bar{p}_1^2\bar{p}_2^2} + \frac{\bar{S}_1}{\bar{p}_1^2} + \frac{\bar{S}_2}{\bar{p}_2^2}\right)$$

$$= N\left(\frac{\bar{q}_1\bar{q}_2}{\bar{p}_1\bar{p}_2} - cv(p_{1i})^2\, cv(p_{2i})^2\right),$$

où $cv(p_{ji}) = (\bar{S}_j - \bar{p}_j^2)/\bar{p}_j$, est le coefficient de variation des valeurs $p_{ji}, j = 1, 2$. Évidemment, $\mathrm{Var}(\hat{N}) \leq N\bar{q}_1\bar{q}_2/(\bar{p}_1\bar{p}_2)$. Une comparaison avec l'exemple 1 permet de constater que V_1 est un estimateur conservatif de $\mathrm{Var}(\hat{N})$ (c'est-à-dire que V_1 est asymptotiquement trop élevé) lorsque les p_{1i} sont indépendantes des p_{2i}. Autrement dit, étant donné les moyennes $\bar{p}_j$, $j = 1, 2$, on obtient la variance la plus élevée dans des conditions d'homogénéité. [Cela rappelle la variance du nombre de succès dans les schémas de Bernoulli avec probabilité de succès variable; voir Feller (1968, p. 230-231).] Une comparaison avec l'exemple 1 permet de constater que V_1 est un estimateur conservatif de $\mathrm{Var}(\hat{N})$ (c'est-à-dire que V_1 est asymptotiquement trop élevé) lorsque les paires (p_{1i}, p_{2i}) sont indépendantes. Notons que la condition d'indépendance suppose que $C = 0$.

Si les probabilités ne sont pas indépendantes, il n'est pas sûr que l'estimateur classique soit conservatif. Il existe néanmoins un estimateur conservatif. On l'obtient en augmentant $\mathrm{Var}(\hat{N})$ d'une quantité qui peut être estimée en fonction des variables observables. Nous faisons la démonstration de la proposition générale suivante en annexe.

Proposition 1. Un estimateur conservatif de $\mathrm{Var}(\hat{N})$ est

$$V_2 = (n_1^2 n_2^2 + n_2^2 m u_1 + n_1^2 m u_2)/m^3,$$

où $u_j = n_j - m, j = 1, 2$.

3. HÉTÉROGÉNÉITÉ GAUSSIENNE

Nous allons maintenant analyser le cas particulier où les moments empiriques des paires (p_{1i}, p_{2i}), $i = 1, \ldots, N$, concordent avec ceux d'une distribution normale (ou gaussienne) à deux variables. Cette analyse va nous permettre de définir avec beaucoup plus de précision que nous l'avons fait précédemment ce qu'est un estimateur conservatif de la variance. Supposons que

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} v_1^2\mu_1^2 & \rho v_1 v_2 \mu_1 \mu_2 \\ \rho v_1 v_2 \mu_1 \mu_2 & v_2^2 \mu_1^2 \end{bmatrix}\right),$$

où $|\rho| < 1$, et $0 < \mu_j < 1, j = 1, 2$. Notons que v_j peut être considéré comme le coefficient de variation de la distribution de p_{ji}. Écrivons, comme avant, $\bar{S}_j = S_j/N$, pour $j = 1, \ldots, 5$. Intégrons ensuite les moments de la distribution normale bidimensionnelle à la formule du lemme 1 selon les équivalences suivantes:

$$\bar{p}_j = E[X_j] = \mu_j, j = 1, 2;$$

$$\bar{S}_j = E[X_j^2] = \mu_j^2(1 + v_j^2), j = 1, 2;$$

$$\bar{p}_{12} = E[X_1 X_2] = \mu_1 \mu_2 (1 + \rho v_1 v_2);$$

$$\bar{S}_3 = E[X_1^2 X_2^2] = \mu_1^2 \mu_2^2 (1 + v_1^2 + v_2^2 + 4\rho v_1 v_2 + (2\rho^2 + 1) v_1^2 v_2^2);$$

$$\bar{S}_4 = E[X_1^2 X_2] = \mu_1^2 \mu_2 (1 + 2\rho v_1 v_2 + v_1^2);$$

$$\bar{S}_5 = E[X_1 X_2^2] = \mu_1 \mu_2^2 (1 + 2\rho v_1 v_2 + v_2^2).$$

Alors, par un calcul simple mais quelque peu fastidieux, nous pouvons démontrer la proposition suivante (détails exclus).

Proposition 2. Étant donné les hypothèses ci-dessus

$$\mathrm{Var}(\hat{N}) = A_1 A_2 + RN,$$

où

$$A_1 = N/(1 + \rho v_1 v_2)^2;$$

$$A_2 = [1 - (\mu_1 + \mu_2)(1 + \rho v_1 v_2) + \mu_1 \mu_2 (1 + \rho v_1 v_2)^2] / [\mu_1 \mu_2 (1 + \rho v_1 v_2)^2];$$

$$R = (2\rho v_1 v_2 + 3\rho^2 v_1^2 v_2^2 - \rho^2 v_1^4 v_2^2 - \rho^2 v_1^2 v_2^4 - v_1^2 v_2^2)/(1 + \rho v_1 v_2)^4.$$

À l'aide des résultats ci-dessus, nous pouvons calculer l'estimateur classique de la variance, $V_1 = n_1 n_2 u_1 u_2 / m^3$. Notons tout d'abord que $\{n_1 n_2 / m\}/A_1 \to 1$, lorsque $N \to \infty$. De même, $\{u_1 u_2 / m^2\}/A_2 \to 1$. Cela démontre le corollaire de la proposition 2, à savoir $(V_1 - \mathrm{Var}(\hat{N}))/N \to -R$, lorsque $N \to \infty$. Par exemple, si $\rho = 0$, alors $-R = v_1^2 v_2^2$, de sorte que V_1 est un estimateur qui produit une surestimation de la variance asymptotique, comme cela a été mentionné dans l'exemple 2.

Quelle valeur devons-nous accorder à l'hypothèse des moments gaussiens? De toute évidence, les probabilités de saisie ne peuvent suivre une distribution strictement gaussienne car avec ce genre de distribution, il y a toujours une partie de la masse de probabilité qui se trouve à l'extérieur de l'intervalle unité. Par ailleurs, supposons que nous obtenions les p_{ji} au moyen de la formule $\text{logit}(p_{ji}) = a_j + b_j Y_{ji}$, où les paires (Y_{1i}, Y_{2i}) forment un échantillon provenant d'une distribution normale bidimensionnelle de moyenne nulle, de variance unité et de corrélation ρ. Si nous avons les relations $a_j = \text{logit}(\mu_j), j = 1, 2$, et $b_j = v_j(1 + \mu_j)^2$, l'hypothèse des moments gaussiens est approximativement juste. De fait, même la distribution des paires (p_{1i}, p_{2i}) est, à peu de choses près, une distribution gaussienne bidimensionnelle.

Examinons maintenant le bien-fondé de l'estimateur V_1 selon l'hypothèse des moments gaussiens. Comme les probabilités doivent prendre une valeur entre 0 et 1, μ_j sera aussi limitée au même intervalle. En outre, pour nous assurer que la plus grande partie de la masse de probabilité se trouve dans le carré-unité, supposons que $0 < v_j \leq \frac{1}{2}, j = 1, 2$. Si les valeurs de μ_j sont proches de un, on peut choisir une borne supérieure beaucoup plus petite. Supposons maintenant que $\rho \leq 0$. On peut alors montrer que

$$-R \geq (\rho^2 v_1^4 v_2^2 + \rho^2 v_1^2 v_2^4)/(1 + \rho v_1 v_2)^4 > 0,$$

de sorte que V_1 produit aussi une surestimation de $\text{Var}(\hat{N})$ lorsque $\rho \leq 0$. Notons que, par le principe de continuité, V_1 doit sûrement produire une surestimation de $\text{Var}(\hat{N})$ pour des valeurs positives de ρ.

On peut montrer que $R = R(\rho)$ est une fonction croissante de ρ à tout le moins pour les valeurs positives de $\rho > 0$. À la limite, nous avons

$$-R(\rho) \to (-2v_1 v_2 - 2v_1^2 v_2^2 + v_1^4 v_2^2 + v_1^2 v_2^4)/(1 + v_1 v_2)^4,$$

lorsque $\rho \to 1$. Lorsque $0 < v_j \leq \frac{1}{2}, j = 1, 2$, la limite ci- dessus atteint sa valeur minimum à $v_1 = v_2 = \frac{1}{2}$. Cette valeur minimum est $-152/625 > -1/4$. Par conséquent, pour $\rho > 0$, V_1 peut produire soit une sous-estimation soit une surestimation de $\text{Var}(\hat{N})$.

Concrètement, les résultats ci-dessus nous amènent aux conclusions suivantes. Premièrement, si $\rho \leq 0$, soit que $\hat{N}$ est convergent ou qu'il donne une surestimation de N; de plus, V_1 produit une surestimation de la variance, de sorte que nous pouvons calculer la limite supérieure d'un intervalle de confiance conservatif pour N. Lorsque $\rho > 0$, $\hat{N}$ donne une sous-estimation de N. Si, en outre, ρ est petit, V_1 produit une surestimation et il est alors possible de calculer la limite inférieure d'un intervalle de confiance conservatif pour N. Évidemment, il s'agit là de conditions plutôt particulières qui n'ont pas vraiment d'application pratique.

Selon le présent modèle, le biais asymptotique de V_1 est plus grand que $> -N/4$ pour toutes les valeurs de ρ. Compte tenu de ce que le biais relatif asymptotique de $\hat{N}$ est $-\rho v_1 v_2/(1 + \rho v_1 v_2) \geq -1/5$ dans le cas de la distribution gaussienne, nous pouvons déterminer un estimateur conservatif de la variance. Ainsi, $5\hat{N}/4 \geq N$ asymptotiquement. On a par exemple, comme estimateur conservatif de $\text{Var}(\hat{N})$, $V_3 = V_1 + 5\hat{N}/16$. Cet estimateur peut être beaucoup plus petit que V_2, ce qui donne une très grande valeur à l'hypothèse des moments gaussiens.

4. EXEMPLE D'APPLICATION: DONNÉES SUR L'ENREGISTREMENT DES CAS DE MALADIE PROFESSIONNELLE

Pour avoir une idée de l'ordre de grandeur des biais dans la réalité, nous allons analyser des données sur l'enregistrement des cas de maladie professionnelle en Finlande. Ce pays tient un registre des maladies professionnelles depuis 1964. La tenue de ce registre est confiée à l'Institut de médecine du travail (traduction) à Helsinki. Depuis 1975, le nombre de nouveaux

cas déclarés annuellement à l'institut a varié de 4,000 à plus de 7,000 (soit de 0.2 à 0.4% de la population active occupée). Les principales catégories de maladies enregistrées par l'institut sont la perte d'audition due au bruit, les maladies causées par un travail répétitif ou monotone (épicondylite, bursite, téno-synovite) et les maladies de la peau (voir Vaaranen et coll. 1985). Le registre en question peut être vu comme un système d'enregistrement double car selon les règlements en vigueur, chaque cas de maladie doit être déclaré à l'institut par la compagnie d'assurance d'une part et par le médecin traitant d'autre part.

Il se peut que la probabilité de déclaration d'un cas dépende du diagnostic par exemple. En effet, d'après des données de 1981, nous avons ce qui suit: nombre de cas déclarés par les compagnies d'assurance, $n_1 = 3{,}769$; nombre de cas déclarés par les médecins, $n_2 = 3{,}053$ et nombre de cas déclarés par l'une et l'autre source, $m = 1{,}591$. Ainsi, l'estimation de système dual habituelle est $\hat{N} = 7{,}232$ et $V_1^{1/2} = 97$, $V_2^{1/2} = 222$, et $V_3^{1/2} = 108.0$. La ressemblance entre V_3 et V_1 est frappante. Si on stratifie les données selon quatre catégories de maladies (les trois catégories mentionnées plus haut et la catégorie "autres"), on obtient les estimations suivantes: - perte d'audition due au bruit: $\hat{N} = 2{,}230$, $V_1^{1/2} = 33.4$, $V_2^{1/2} = 47.2$, et $V_3^{1/2} = 42.6$; - maladies causées par un travail répétitif ou monotone: $\hat{N} = 3{,}572$, $V_1^{1/2} = 201.4$, $V_2^{1/2} = 303.8$, et $V_3^{1/2} = 204.2$; - maladies de la peau: $\hat{N} = 1{,}441$, $V_1^{1/2} = 30.9$, $V_2^{1/2} = 86.2$, et $V_3^{1/2} = 37.5$; - autres maladies: $\hat{N} = 1{,}015$, $V_1^{1/2} = 32.7$, $V_2^{1/2} = 79.1$, et $V_3^{1/2} = 37.2$. En regroupant ces résultats, on obtient les estimations globales suivantes: $\hat{N} = 8{,}258$, $V_1^{1/2} = 209.0$, $V_2^{1/2} = 340.3$, et $V_3^{1/2} = 215.2$. Nous remarquons que les maladies causées par un travail répétitif ou monotone sont sous-déclarées dans une assez large mesure.

Nous avons poussé plus loin l'analyse en stratifiant les données selon le genre de maladie (4 classes), l'assureur (11 classes) et le groupe d'activité économique (7 classes). On pourrait croire *a priori* que ces facteurs influent sur les probabilités de déclaration. Or, la stratification n'a pas modifié de façon notable l'estimation ponctuelle. Par contre, elle a fait s'accroître de plus du tiers les écarts types estimés à cause vraisemblablement de l'existence de très petites strates. Cette analyse nous permet de conclure qu'en ce qui concerne cette application, le biais lié au diagnostic représente la principale source d'erreur dans l'estimateur classique.

Nous avons aussi analysé les données à l'aide d'une méthode de régression logistique, laquelle permet de tenir compte de l'hétérogénéité observable d'une population qui résulte de variables explicatives discrètes et continues. Ainsi, nous avons pu observer que pour chaque catégorie de maladies, l'âge avait une incidence sur la probabilité de déclaration pour une source mais non pour l'autre. Les estimations ponctuelles demeuraient donc les mêmes et la conclusion concernant le rôle du diagnostic ne pouvait être réfutée (Alho 1990).

5. ANALYSE

Les résultats théoriques montrent que l'estimateur de la variance habituel, V_1 est conservatif lorsque les deux systèmes d'enregistrement sont corrélés négativement ou qu'ils sont indépendants l'un de l'autre. Par le principe de continuité, V_1, peut aussi être conservatif lorsqu'il y a une corrélation positive mais faible. Lorsqu'il existe une forte corrélation positive, V_1 produit des sous-estimations. Nous avons donc proposé un autre estimateur, V_2, qui est conservatif dans des conditions d'hétérogénéité quelconque, or, cet estimateur semble trop conservatif lorsqu'on le compare à V_3 qui, lui, est assurément conservatif dans des conditions d'hétérogénéité gaussienne. La ressemblance entre V_3 et V_1 donne à penser qu'en pratique, V_1 peut être assez robuste à l'égard de l'hétérogénéité d'une population.

Même en utilisant l'estimateur conservatif V_2 dans notre exemple empirique, nous n'aurions pas réussi à compenser le biais de l'estimateur classique. Cela était peut-être prévisible car le biais de $\hat{N}$ et le degré de surestimation produite par V_2 sont tous deux d'ordre N. L'utilisation de V_2 n'accroîtrait donc la largeur d'un intervalle de confiance que par un facteur d'ordre $N^{1/2}$. Par conséquent, V_2 peut compenser le biais de $\hat{N}$ uniquement si ce biais est

faible ou si N est faible. En conclusion, il semble que la méthode de système dual donnera des résultats satisfaisants uniquement si les systèmes d'enregistrement sont plus ou moins non corrélés ou si l'hétérogénéité est observable. Dans ce dernier cas, nous pouvons recourir à la stratification, comme l'ont déjà proposé Sekar et Deming (1949), ou aux modèles de régression logistique, comme le proposent Huggins (1989) et Alho (1990), pour atténuer l'effet du biais de l'estimateur classique de la taille d'une population.

REMERCIEMENTS

L'auteur tient à remercier Bruce Spencer et un arbitre anonyme, qui ont permis par leurs commentaires d'améliorer la présentation de cet article. Une partie des résultats empiriques ont déjà été présentés au 11$^{\text{ième}}$ congrès des pays nordiques sur la statistique mathématique, tenu à Uppsala, Suède, en juin 1986.

ANNEXE

Démonstration du lemme 1. Appliquons un développement de Taylor à $\hat{N} = n_1 n_2/m$ faisant $E[n_1]\, E[n_2]/E[m] = N\bar{p}_1\bar{p}_2/\bar{p}_{12}$, ou

$$\hat{N} \approx \frac{N\bar{p}_1\bar{p}_2}{\bar{p}_{12}} + \frac{\bar{p}_2}{\bar{p}_{12}}(n_1 - N\bar{p}_1) + \frac{\bar{p}_1}{\bar{p}_{12}}(n_2 - N\bar{p}_2) - \frac{\bar{p}_1\bar{p}_2}{\bar{p}_{12}^2}(m - N\bar{p}_{12}).$$

Nous avons alors

$$E\left[\left(\hat{N} - \frac{N\bar{p}_1\bar{p}_2}{\bar{p}_{12}}\right)^2\right] \approx \left(\frac{\bar{p}_2}{\bar{p}_{12}}\right)^2 \mathrm{Var}(n_1) + \left(\frac{\bar{p}_1}{\bar{p}_{12}}\right)^2 \mathrm{Var}(n_2) + \left(\frac{\bar{p}_1\bar{p}_2}{\bar{p}_{12}^2}\right)^2 \mathrm{Var}(m)$$

$$- \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}} \mathrm{Cov}(n_1, m) - 2\frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3} \mathrm{Cov}(n_2, m).$$

Suivant les hypothèses d'indépendance, $\mathrm{Var}(n_j) = N\bar{p}_j - S_j, j = 1, 2$; $\mathrm{Var}(m) = N\bar{p}_{12} - S_3$, $\mathrm{Cov}(n_1, m) = -S_4 + N\bar{p}_{12}$, et $\mathrm{Cov}(n_2, m) = -S_5 + N\bar{p}_{12}$. En faisant les substitutions nécessaires dans la formule de l'erreur quadratique moyenne, on obtient le résultat recherché.

Démonstration de la proposition 1. Faisons abstraction du terme négatif qui comprend S_3 dans la formule du lemme 1. Comme $0 < p_{ji} < 1$, nous avons $S_4 < N\bar{p}_{12}$, et $S_4 < S_1$. Par conséquent,

$$\frac{2\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3} S_4 < \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3} N\bar{p}_{12} + \frac{\bar{p}_2^2}{\bar{p}_{12}^2} S_1 + \left(\frac{\bar{p}_1 - \bar{p}_{12}}{\bar{p}_{12}}\right)\frac{\bar{p}_2^2}{\bar{p}_{12}^2} N\bar{p}_{12}.$$

De même,

$$\frac{2\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}} S_5 < \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3} N\bar{p}_{12} + \frac{\bar{p}_1^2}{\bar{p}_{12}^2} S_2 + \left(\frac{\bar{p}_2 - \bar{p}_{12}}{\bar{p}_{12}}\right)\frac{\bar{p}_1^2}{\bar{p}_{12}^2} N\bar{p}_{12}.$$

En substituant ces bornes aux termes de l'expression du lemme 1, nous obtenons

$$\mathrm{Var}(\hat{N}) < \frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^3} N + \frac{(\bar{p}_1 - \bar{p}_{12})\bar{p}_2^2}{\bar{p}_{12}^2} N + \frac{(\bar{p}_2 - \bar{p}_{12})\bar{p}_1^2}{\bar{p}_{12}^2} N.$$

En estimant $N\bar{p}_j$ au moyen de $n_j, j = 1, 2$; et $N\bar{p}_{12}$ au moyen de m nous obtenons le résultat recherché.

BIBLIOGRAPHIE

ALHO, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

BURNHAM, P.K., et OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, (Vol. I, 3ième éd.). New York: Wiley.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance*, (2ième éd.). New York: Griffin.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

VAARANEN, V., VASAMA, M., et ALHO, J. (1985). Occupational diseases in Finland in 1984. Reviews 11, Institute of Occupational Health, Helsinki.