Techniques d'enquête, juin 1995
Vol. 21, n° 1, pp. 15–21
Statistique Canada

Estimation de la corrélation dans les plans à réponses randomisées

D.R. BELLHOUSE[1]

RÉSUMÉ

Nous examinons la contribution de Stanley Warner à l'étude des réponses randomisées, avant d'élaborer un modèle linéaire, fondé sur les modèles de permutation aléatoire, qui permet d'intégrer, à titre de cas spéciaux, plusieurs plans connus à réponses randomisées. On obtient avec l'aide de ce modèle des estimateurs optimums des variances et des covariances de populations finies à l'intérieur d'une classe générale d'estimateurs quadratiques non biaisés. Ces résultats conduisent à la mise au point d'un estimateur de la corrélation de la population finie. Trois plans d'expérience particuliers à réponses randomisées sont examinés: (i) le modèle des questions non liées de Greenberg et coll. (1969); (ii) le modèle des constantes additives de Pollock et Bek (1976); le modèle des constantes multiplicatives de Pollock et Bek (1976). Nous proposons des modèles simples du biais dans les réponses afin d'illustrer l'effet de ce biais sur l'estimation de la corrélation.

MOTS CLÉS: Modèle des constantes additives; modèles linéaires; modèle des constantes multiplicatives; biais dans les réponses; modèle des questions non liées; estimation de la variance.

1. APERÇU DE LA CONTRIBUTION DE WARNER À L'ÉTUDE DES RÉPONSES RANDOMISÉES

On a recours à la technique des réponses randomisées pour obtenir des réponses à des questions délicates. Cette méthode a été mise au point il y a trente ans par Stanley Warner (Warner 1965) afin d'estimer une proportion avec un plan d'échantillonnage aléatoire simple avec remise. Ce progrès théorique remarquable a été rendu possible grâce à une démarche intellectuelle d'une grande originalité. Comment obtenir des réponses sincères à des questions délicates? Warner a proposé comme solution d'obtenir ces réponses sans que l'intervieweur n'ait conscience que les questions posées sont réellement délicates. Il a élaboré la structure probabiliste de l'interrogation de manière à obtenir une estimation de la proportion requise. Dans la formule originale de Warner, la population est divisée en deux groupes exhaustifs mutuellement exclusifs: A et B. Il importe d'estimer la proportion π de la population appartenant au groupe A. Pour y arriver, on utilise une simple aiguille pivotante qui pointera en direction d'une lettre A avec une probabilité p, et en direction d'une lettre B avec une probabilité $1 - p$. La personne questionnée fait tourner l'aiguille et n'a qu'à répondre oui ou non selon que cette aiguille s'arrête ou non sur son groupe d'appartenance. Le plan d'échantillonnage avec remise permet d'estimer π par la méthode du maximum de vraisemblance.

Cette idée d'une grande originalité a beaucoup attiré l'attention au cours des trente dernières années. Depuis les travaux originaux de Warner, plusieurs techniques à réponses randomisées ont été proposées pour l'estimation d'une proportion ou d'un groupe de proportions - par exemple, les données polytomiques- ou pour l'estimation de la moyenne d'une population à partir de données continues. Une variante de l'idée originale de Warner, proposée pour la première fois par Greenberg et coll. (1969), consiste à poser la question délicate ou une question non liée avec des probabilités de p et de $1 - p$ respectivement. D'autres variantes possibles utilisant des données continues comprennent l'ajout d'une variable aléatoire à la réponse donnée à la question délicate, ou la multiplication de la réponse par une variable aléatoire. L'idée qui sous-tend toutes ces techniques est de masquer la réponse originale de manière qu'il devienne impossible d'attribuer l'information obtenue à l'un ou l'autre des répondants, tout en permettant l'extraction de l'information portant sur la question délicate à partir de l'échantillon général. Il existe aujourd'hui une abondante documentation sur ces techniques, y compris une monographie de Chaudhuri et Mukerjee (1988). Nathan (1988) a préparé une liste bibliographique assez complète de tous ces ouvrages. Umesh et Peterson (1991) ont pour leur part fourni plusieurs exemples détaillés de l'application et de l'applicabilité des techniques à réponses randomisées.

Compte tenu de la variété actuelle des techniques à réponses randomisées, on peut s'interroger sur la façon de les comparer. La minimisation de la variance n'est pas le seul critère à prendre en compte. Chaque méthode est conçue pour protéger la vie privée du répondant. Un gain d'efficacité, en termes de variance, obtenu grâce au choix de valeurs différentes de la probabilité dans le plan de randomisation ou au choix d'une méthode à réponses randomisées au détriment d'une autre, peut mettre en péril le caractère confidentiel des réponses fournies. Pour répondre à cette préoccupation, Leysieffer et Warner (1976) et Warner (1976) ont proposé des mesures naturelles du risque de non-respect de la confidentialité des répondants. Ces mesures sont liées à la probabilité, pour un intervieweur, de déterminer l'origine d'une réponse à une question délicate. La théorie du risque de non-respect de la confidentialité a été examinée par Chaudhuri et Mukerjee

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario (London) Ontario, N6A 5B7.

(1988). Umesh et Peterson (1991) abordent également certaines considérations pratiques à ce propos.

Stanley Warner a contribué encore de deux façons à l'étude des réponses randomisées. La première est en rapport direct avec les résultats obtenus ici. Au milieu des multiples nouvelles idées et nouvelles techniques fondées sur le principe des réponses randomisées, Warner (1971) a formulé un modèle linéaire qui a permis d'unifier la théorie et qui permettait, à l'époque de sa formulation, d'intégrer la plupart des techniques à réponses randomisées. La seconde de ses contributions a pris en compte la tendance grandissante à recourir aux entrevues téléphoniques. Stem et Steinhorst (1984) ont décrit des méthodes à réponses randomisées applicables aux entrevues téléphoniques et aux questionnaires postaux. Warner (1986) a suggéré des plans pratiques de randomisation naturelle comme le recours aux numéros de série des billets de banque, aux fins de la réalisation d'entrevues téléphoniques.

Les principaux constituants de la méthodologie de la réponse randomisée sont: l'élaboration des techniques, la comparaison de ces techniques en tenant compte du risque de non-respect de la confidentialité du répondant, l'élaboration de plans de randomisation raisonnablement efficaces, l'élaboration d'une théorie unifiée de la réponse randomisée et la validation des techniques à l'aide d'études en conditions réelles. Stanley Warner s'est penché sur la plupart de ces grandes questions, et sa contribution en ces matières a eu une grande influence. Il est l'inventeur de la technique. Son plan d'expérience original pour une population dichotomique a été rapidement étendu au traitement des populations polytomiques et des populations soumises à des mesures en continu. Le développement des techniques à réponses randomisées s'est poursuivi et Warner a été un chef de file de l'évaluation de ces plans à l'aide de la modélisation du risque de non-respect de la confidentialité du répondant. Ses travaux en vue de l'élaboration d'un modèle linéaire unifié des plans à réponses randomisées sont devenus la pierre d'assise d'une théorie unifiée de la réponse randomisée.

2. INTRODUCTION À L'ESTIMATION DE LA CORRÉLATION

Imaginons une population finie de taille N faisant l'objet de deux mesures x_j et y_j où $j = 1, \ldots, N$. Il est intéressant d'estimer la corrélation de la population finie

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

où $\sigma_{xy} = \sum (x_j - \bar{X})(y_j - \bar{Y})/N$ représente la covariance de la population finie entre les variables x et y, et où σ_x^2 et σ_y^2 sont les variances des variables x et y respectivement. Pour faire l'estimation de ρ, on choisit au sein de la population finie un échantillon de taille définie n assorti d'une probabilité $P(s)$, où s désigne le groupe d'unités de la population finie choisi pour l'échantillon. L'opérateur de l'espérance rattaché à la probabilité $P(s)$ du plan d'échantillonnage est désigné par E_p. Les estimateurs pour ρ sont obtenus en remplaçant σ_x^2, σ_y^2 et σ_{xy} par leurs estimateurs respectifs, biaisés ou non, qui peuvent être optimums dans un sens ou dans l'autre, ou ne pas l'être du tout.

Pour illustrer les résultats généraux obtenus ici dans l'estimation du coefficient de corrélation de la population finie, nous examinerons trois techniques particulières à réponses randomisées:

(i) Le modèle des questions non liées de Greenberg et coll. (1969). La question délicate est posée avec une probabilité p, et une question sans rapport avec la première et qui n'est pas délicate est posée avec une probabilité $1 - p$. Pour l'estimation de la moyenne, on présume que la moyenne $\bar{X}$ de la population finie de la question non liée est connue. Pour les besoins de l'estimation de la variance, on présume également que la valeur de σ_x^2 est connue.

(ii) Le modèle des constantes additives de Pollock et Bek (1976). On ajoute une variable aléatoire issue d'une distribution de probabilité connue à la valeur de la réponse à la question délicate.

(iii) Le modèle des constantes multiplicatives de Pollock et Bek (1976). La valeur de la réponse à la question délicate est multipliée par une variable aléatoire issue d'une distribution de probabilité connue.

Edgell et coll. (1986) ont calculé des estimateurs pour ρ dans le modèle des questions non liées et dans le modèle des constantes additives.

Dans la plupart des plans à réponses randomisées, on a présumé que le plan d'échantillonnage utilisé était l'échantillonnage aléatoire simple, avec ou sans remise. Comme les résultats dont il est question ici proviennent d'un plan d'expérience de taille fixe, on présume que l'échantillonnage aléatoire utilisé est sans remise.

Supposons que x et y sont deux variables délicates. On aura donc recours, pour obtenir des informations sur ces variables, à une technique à réponses randomisées. Désignons par w_j et par z_j pour $j \in s$ les mesures de l'échantillon obtenues. Désignons par u_j et par v_j pour $j = 1, \ldots, N$ les mesures non délicates liées à x_j et y_j respectivement. En vertu du modèle de la question non liée (plan à réponses randomisées (i)), u_j et v_j sont les réponses données aux questions non liées par le j-ième répondant. En vertu du modèle des constantes additives ou du modèle des constantes multiplicatives (plans à réponses randomisées (ii) ou (iii)), u_j et v_j sont les j-ièmes variables aléatoires issues de deux distributions de probabilités connues et peut-être différentes.

3. MODÈLES DE PERMUTATION ALÉATOIRE

Plusieurs modèles ont été proposés pour les mesures de populations finies dans la documentation spécialisée sur l'échantillonnage d'enquête. Nous nous attarderons en particulier aux modèles de permutation aléatoire de Rao (1975) et de Rao et Bellhouse (1978). Ces modèles présentent un énorme avantage: leurs paramètres ont une interprétation directe dans la population finie étudiée puisque les

paramètres des modèles de permutation aléatoire sont également des paramètres de populations finies. Dans la situation la plus simple des modèles de permutation aléatoire, on présume que le vecteur N-dimensionnel des mesures dans la population finie est une permutation aléatoire d'un vecteur N-dimensionnel de nombres fixes. Rao (1975) a montré comment cette supposition conduit à un modèle linéaire. Bellhouse (1980) a élargi l'utilisation de ce modèle aux plans à réponses randomisées à échantillonnages avec probabilités inégales.

Le modèle et les plans d'expérience apparentés applicables à l'échantillonnage avec probabilités inégales ne s'appliquent pas facilement à l'estimation des variances et des covariances, avec ou sans réponses randomisées. Nous présentons par conséquent ici un cas spécial du modèle de Bellhouse (1980). Dans ce modèle, deux opérateurs d'espérance différents interviennent et donnent ensemble une espérance composite E_m. Ces opérateurs sont: E_r, l'opérateur d'espérance lié au plan de randomisation, et E_{rp}, l'opérateur d'espérance lié au modèle de permutation aléatoire. L'espérance composite est donc $E_m = E_{rp}E_r$ et $E = E_mE_p$. Pour le modèle de permutation aléatoire, nous présumons que les paires $(x_j, y_j), j = 1, \ldots, N$ correspondent à la permutation aléatoire d'un groupe de N paires de nombres fixes que nous appellerons $(p_j, q_j), j = 1, \ldots, N$. Il s'agit là d'un cas spécial du modèle (4.1) de Rao et Bellhouse (1978); le modèle plus général décrit par Rao et Bellhouse (1978) a servi dans des cas d'échantillonnage double et d'échantillonnages effectués en deux occasions. Avec le modèle des questions non liées (plan à réponses randomisées (i)) il faut supposer également que les quadruplets $(x_j, y_j, u_j, v_j), j = 1, \ldots, N$ constituent une permutation aléatoire d'un groupe de N quadruplets fixes de nombres que nous appellerons (p_j, q_j, r_j, t_j), $j = 1, \ldots, N$.

Supposons que la combinaison du plan de randomisation et du modèle de permutation aléatoire conduit au modèle linéaire suivant:

$$\begin{aligned} w_j &= \alpha_1 + \beta_1 \bar{X} + e_{1j} \\ z_j &= \alpha_2 + \beta_2 \bar{Y} + e_{2j}, \end{aligned} \tag{1}$$

pour $j = 1, \ldots, N$ où $\bar{X}$ et $\bar{Y}$ sont les moyennes des populations finies d'où viennent les mesures x et y respectivement, et où pour $j = 1, \ldots, N$

$$E_m(e_{1j}) = E_m(e_{2j}) = 0,$$

$$E_m(e_{1j}^2) = \phi_1\sigma_x^2 + \psi_{01} + \psi_{11}\bar{X} + \psi_{21}\bar{X}^2,$$

$$E_m(e_{2j}^2) = \phi_2\sigma_y^2 + \psi_{02} + \psi_{12}\bar{Y} + \psi_{22}\bar{Y}^2,$$

$$E_m(e_{1j}\,e_{1k}) = \delta_1\sigma_x^2 + \lambda_1, E_m(e_{2j}\,e_{2k}) = \delta_2\sigma_y^2 + \lambda_2, \quad \text{pour} \quad j \neq k,$$

$$E_m(e_{1j}\,e_{2j}) = \phi_3\sigma_{xy} + \psi_3, \quad \text{et}$$

$$E_m(e_{1j}\,e_{2k}) = \delta_3\sigma_{xy} + \lambda_3, \quad \text{pour} \quad j \neq k. \tag{2}$$

et tous les autres moments plus élevés sont indépendants de j. Dans le modèle illustré en (1) et (2), les paramètres α, λ, ϕ, ψ et δ sont tous des constantes connues. Les variances et les covariances des populations finies des questions délicates, σ_x^2, σ_y^2 et σ_{xy} sont toutes inconnues.

Pour le modèle des questions non liées (plan à réponses randomisées (i)), on présume que les processus de randomisation utilisés pour les deux questions délicates sont indépendants et que la question délicate i, $i = 1, 2$, est posée avec une probabilité p_i alors que les questions liées non délicates sont posées avec une probabilité $1 - p_i$. On présume en outre que les questions délicates sont sans lien avec les questions non délicates et que $\sigma_{xu} = \sigma_{yv} = \sigma_{xv} = \sigma_{yu} = 0$. Cette supposition n'est pas nécessaire dans le cas d'un échantillonnage aléatoire simple avec remise. Lorsque, en plus, on suppose un modèle de permutation aléatoire avec le quadruplet (x_j, y_j, u_j, v_j), on obtient dans le modèle illustré en (1) et (2)

$$\alpha_1 = (1 - p_1)\bar{U}, \beta_1 = p_1, \alpha_2 = (1 - p_2)\bar{V}, \beta_2 = p_2,$$

$$\phi_1 = p_1, \psi_{01} = (1 - p_1)\sigma_u^2 + p_1(1 - p_1)\bar{U}^2,$$
$$\psi_{11} = -2p_1(1 - p_1)\bar{U}, \psi_{21} = p_1(1 - p_1),$$

$$\phi_2 = p_2, \psi_{02} = (1 - p_2)\sigma_v^2 + p_2(1 - p_2)\bar{V}^2,$$
$$\psi_{12} = -2p_2(1 - p_2)\bar{V}, \psi_{22} = p_2(1 - p_2),$$

$$\delta_1 = -p_1^2/(N - 1), \lambda_1 = -(1 - p_1)^2\sigma_u^2/(N - 1),$$

$$\delta_2 = -p_2^2/(N - 1), \lambda_2 = -(1 - p_2)^2\sigma_v^2/(N - 1),$$

$$\phi_3 = p_1p_2, \delta_3 = -\phi_3/(N - 1),$$
$$\psi_3 = (1 - p_1)(1 - p_2)\sigma_{uv},$$
$$\text{et} \quad \lambda_3 = -\psi_3/(N - 1). \tag{3}$$

À noter qu'en raison des hypothèses de départ du modèle, il importe que la matrice des variances-covariances de la population finie des questions non délicates soit connue, ainsi d'ailleurs que les moyennes de la population finie.

Pour le modèle des constantes additives (plan à réponses randomisées (ii)), on présume que les variables aléatoires u et v qui sont ajoutées à la valeur des réponses données aux deux questions délicates sont indépendantes et qu'elles sont assorties des moyennes μ_u et μ_v et des variances σ_u^2 et σ_v^2 respectivement. Lorsqu'on suppose un modèle de permutation aléatoire avec le couple (x_j, y_j), on obtient dans le modèle illustré en (1) et (2)

$$\alpha_1 = \mu_u, \beta_1 = 1, \alpha_2 = \mu_v, \beta_2 = 1,$$

$$\phi_1 = \phi_2 = \phi_3 = 1, \psi_{01} = \sigma_u^2, \psi_{02} = \sigma_v^2,$$

$$\delta_1 = \delta_2 = \delta_3 = -1/(N - 1),$$

$$\psi_{11} = \psi_{21} = \psi_{12} = \psi_{22} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \tag{4}$$

Dans le modèle des constantes multiplicatives, deux variables aléatoires indépendantes, u et v, assorties des moyennes μ_u et μ_v et des variances σ_u^2 et σ_v^2 respectivement, sont multipliées respectivement par la valeur de la réponse à la variable x et à la variable y. Lorsqu'on suppose un modèle de permutation aléatoire avec le couple (x_j, y_j), on obtient dans le modèle illustré en (1) et (2)

$$\alpha_1 = \alpha_2 = 0, \beta_1 = \mu_u, \beta_2 = \mu_v,$$

$$\phi_1 = \mu_u^2 + \sigma_u^2, \phi_2 = \mu_v^2 + \sigma_v^2, \phi_3 = \mu_v \mu_v,$$

$$\psi_{21} = \sigma_u^2, \psi_{22} = \sigma_v^2,$$

$$\delta_1 = -\mu_u^2/(N-1), \delta_2 = -\mu_v^2/(N-1),$$
$$\delta_3 = -\mu_u\mu_v/(N-1), \quad \text{et}$$

$$\psi_{01} = \psi_{11} = \psi_{02} = \psi_{12} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \tag{5}$$

4. ESTIMATION DE LA VARIANCE ET DE LA COVARIANCE

Supposons une estimation de σ_y^2 telle que les données appropriées sont z_j pour les unités $j \in s$. La classe générale des estimateurs quadratiques de σ_y^2 prend la forme suivante:

$$e_{bs} = b_{s..} + \sum_{j\in s} b_{sj.}\, z_j + \sum_{j\in s} b_{sjj}\, z_j^2 + \sum\sum_{i\neq j\in s} b_{sij}\, z_i z_j, \tag{6}$$

où les coefficients des valeurs z sont définis pour tous les s, tous les $j \in s$ et toutes les $(i, j) \in s$.

Dans une situation de réponses randomisées, un estimateur e_b appartenant à la classe définie par l'équation (6) est non biaisé selon le plan d'échantillonnage pour σ_y^2 si $E_p E_r(e_b) = \sigma_y^2$ et il est non biaisé selon le plan d'échantillonnage et le modèle de permutation (selon p et m) si $E(e_b) = \sigma_y^2$. Les conditions en vertu desquelles un estimateur e_b est non biaisé selon p et m sont réalisées en tirant l'espérance E de (6) des modèles (1) et (2). En établissant un rapport d'égalité entre les coefficients dans $\bar{Y}^0$, $\bar{Y}^1$, $\bar{Y}^2$ et σ_y^2 et on obtient quatre nouvelles équations avec quatre inconnues. La résolution de ces quatre équations conduit aux conditions suivantes en vertu desquelles les estimateurs de la classe définie par (6) sont biaisés selon p et m pour σ_y^2:

$$E_p\left(\sum_{j\in s} b_{sjj}\right) = \frac{\beta_2^2}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = A_2, \tag{7}$$

$$E_p\left(\sum\sum_{i\neq j\in s} b_{sij}\right) = -\frac{\beta_2^2 + \psi_{22}}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = -(A_2 + B_2), \tag{8}$$

$$E_p\left(\sum_{j\in s} b_{sj.}\right) = \frac{(2\alpha_2\psi_{22} - \beta_2\psi_{12})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = C_2, \tag{9}$$

et

$$E_p(b_{s..}) = \frac{\lambda_2(\beta_2^2 + \psi_{22}) - (\alpha_2^2\psi_{22} - \alpha_2\beta_2\psi_{12} + \beta_2^2\psi_{02})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = D_2. \tag{10}$$

Pour obtenir l'estimateur optimal, il faut définir une classe associée d'estimateurs quadratiques de 0. Cette classe est définie par

$$e_{cs} = c_{s..} + \sum_{j\in s} c_{sj.}\, z_j + \sum_{j\in s} c_{sjj}\, z_j^2 + \sum\sum_{i\neq j\in s} c_{sij}\, z_i z_j.$$

Les conditions en vertu desquelles un estimateur e_c de cette classe sera biaisé selon p et m pour 0 sont définies par

$$E_p(c_{s..}) = E_p\left(\sum_{j\in s} c_{sj.}\right) = E_p\left(\sum_{j\in s} c_{sjj}\right) = E_p\left(\sum\sum_{i\neq j\in s} c_{sij}\right) = 0. \tag{11}$$

L'obstruction de l'estimateur quadratique non biaisé selon le plan d'expérience de la variance minimale de σ_y^2 s'inspire du procédé utilisé pour la moyenne d'une population finie par Rao et Bellhouse (1978), pour les cas sans réponse randomisée, et par Bellhouse (1980), pour les cas à réponses randomisées. La covariance $E(e_b e_c)$ pour l'espérance composite est déterminée dans le modèle de manière que seules les espérances de la forme E_p restent à déterminer. À partir de là, les coefficients b sont déterminés de manière que $E(e_b e_c) = 0$ sous les conditions définies par (11). Les valeurs des coefficients b sont alors déterminées en vertu des conditions définies de (7) à (10). Selon un théorème de l'estimation non biaisée de la variance minimale formulé par Rao (1952), l'estimateur résultant est l'estimateur optimal non biaisé selon p et m de σ_y^2. S'il existe un plan d'expérience tel que cet estimateur est également non biaisé selon le plan d'expérience pour σ_y^2, on pourra considérer qu'il s'agit également de l'estimateur optimal non biaisé selon le plan d'expérience de σ_y^2, en invoquant des arguments semblables à ceux du théorème (2.4) de Rao et Bellhouse (1978). Nous présentons d'abord les résultats obtenus pour les estimateurs non biaisés selon p et m (théorèmes 1 et 2), et ensuite les résultats obtenus pour les estimateurs non biaisés selon le plan d'expérience pour les trois processus à réponses randomisées.

Théorème 1. En vertu du modèle défini par (1) et (2) et pour tout plan de taille fixe n, la variance de e_b, selon p et m, $E_{rp}[E_p E_r(e_b - \sigma_y^2)^2] = E(e_b - \sigma_y^2)^2$, est minimisée pour l'estimateur donné par

$$(A_2 + B_2)s_z^2 - B_2\frac{1}{n}\sum_{j\in s} z_j^2 + C_2\,\bar{z} + D_2, \tag{12}$$

où $\bar{z}$ est la moyenne d'échantillon de données et

$$s_z^2 = \frac{1}{n-1} \sum_{j \in s} (z_j - \bar{z})^2$$

est la variance d'échantillon de données obtenue par la technique à réponses randomisées, où A_2, B_2, C_2 et D_2 sont définis par les équations (7) à (10) respectivement.

Démonstration. En vertu du modèle donné par (1) et (2), la covariance $E(e_b e_c)$ est algébriquement plutôt longue, mais peut s'exprimer sous la forme

$$\boldsymbol{b}^T \boldsymbol{G} \boldsymbol{c} + H, \tag{13}$$

où $\boldsymbol{b}^T$ est le vecteur

$$\left[E_p(b_{s..}), E_p\left(\sum_{j \in s} b_{sj.}\right), E_p\left(\sum_{j \in s} b_{sjj}\right), E_p\left(\sum\sum_{i \neq j \in s} b_{sij}\right)\right], \tag{14}$$

et $\boldsymbol{c}^T$ est un vecteur identique à (14), où tous les b sont remplacés par des c. La matrice $\boldsymbol{G}$ 4 x 4 en (13) contient les fonctions des moments du premier ordre de z_j et des moments du deuxième ordre de e_{2j} dans (1). Le membre H de (13) est une somme de termes prenant la forme

$$\kappa \sum b_{sij}\, c_{skl}, \tag{15}$$

où le symbole de la somme est limité à un quadruplet, où les indices de b pourraient être remplacés par un point (.), et où κ est une fonction de moments de deuxième à quatrième ordre de e_{2j} dans (1). À noter que ces moments sont tous indépendants de j. En (15), la somme est une somme simple sur $j \in s$ lorsque, par exemple, les indices $i = j = k = l$, ou lorsque $i = k$ et que j et l sont remplacés par des points. La somme est une somme double sur $i \neq k \in s$ lorsque, par exemple, $i \neq k$ et que j et l sont remplacés par des points. Le même procédé est repris jusqu'à la quadruple somme où $i \neq j \neq k \neq l$. Selon (11), $E(e_b e_c)$ se réduit à 0 si $b_{s..} = h_1$, $b_{sj} = h_2$, $b_{sjj} = h_3$, et $b_{sij} = h_4$, où les valeurs h_i sont des constantes. Selon (7) à (10) et compte tenu du fait que le plan d'expérience a une taille fixe, nous obtenons

$$b_{s..} = D_2,\ b_{sj.} = C_2/n,\ b_{sij} = -\frac{A_2 + B_2}{n(n-1)},\ b_{sjj} = A_2/n,$$

de sorte que l'estimateur de (12) minimise la variance de la classe non biaisée selon p et m des estimateurs quadratiques de σ_y^2 Q.E.D.

De la même manière

$$(A_1 + B_1)s_w^2 - B_1 \frac{1}{n} \sum_{j \in s} w_j^2 + C_1 \bar{w} + D_1, \tag{16}$$

est l'estimateur optimal non biaisé selon p et m pour σ_x^2 où

$$A_1 = \frac{\beta_1^2}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}},$$

$$B_1 = \frac{\beta_1^2 + \psi_{21}}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}},$$

$$C_1 = \frac{(2\alpha_1 \psi_{21} - \beta_2 \psi_{11})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}}, \quad \text{et}$$

$$D_1 = \frac{\lambda_1(\beta_1^2 + \psi_{21}) - (\alpha_1^2 \psi_{21} - \alpha_1 \beta_1 \psi_{11} + \beta_1^2 \psi_{01})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}}.$$

La même méthode peut servir à estimer la covariance σ_{xy}. La classe générale des estimateurs quadratiques de σ_{xy} prend la forme de

$$e_{ds} = d_s + \sum_{j \in s} d_{1sj} z_j + \sum_{j \in s} d_{2sj} w_j + \sum\sum_{i \neq j \in s} d_{sij} w_i z_j,$$

où les coefficients des valeurs w et z sont définis pour toutes les valeurs de s, tous les $j \in s$ et toutes les paires $(i, j) \in s$. Le résultat sur la covariance est énoncé sans preuve dans le

Théorème 2. En vertu du modèle défini par (1) et (2) et pour tout plan d'expérience de taille fixe n, la variance de e_d selon p et m, $E_{rp}[E_p E_r(e_d - \sigma_{xy})^2] = E(e_d - \sigma_{xy})^2$, est minimisée pour l'estimateur donné par

$$\frac{s_{wz} - (\psi_3 - \lambda_3)}{\phi_3 - \delta_3}, \tag{17}$$

où

$$s_{wz} = \frac{1}{n-1} \sum_{j \in s} (w_j - \bar{w})(z_j - \bar{z})$$

est la covariance de l'échantillon entre w et z.

On obtient un estimateur pour ρ à partir de (12), (16) et (17). Dans le modèle des constantes additives (plan à réponses randomisées (ii)), l'estimateur de ρ est donné par

$$\hat{\rho}_{ac} = \frac{s_{wz}}{\sqrt{(s_w^2 - \sigma_u^2)(s_z^2 - \sigma_v^2)}}. \tag{18}$$

Il s'agit là de l'estimateur obtenu par Edgell et coll. (1986). Dans le modèle des constantes multiplicatives (plan à réponses randomisées (iii)), l'estimateur se réduit à

$$\hat{\rho}_{mc} = \frac{s_{wz}}{\sqrt{s_w^2 - \frac{\sigma_u^2/\mu_u^2}{1 + \sigma_u^2/\mu_u^2}\frac{1}{n}\sum_{j\in s} w_j^2}\sqrt{s_z^2 - \frac{\sigma_v^2/\mu_v^2}{1 + \sigma_v^2/\mu_v^2}\frac{1}{n}\sum_{j\in s} z_j^2}}, \tag{19}$$

pour $\mu_u \neq 0$ et $\mu_v \neq 0$. Lorsque $\mu_u = 0$, le coefficient de $\sum w_j^2$ est $1/n$ et lorsque $\mu_v \neq 0$, le coefficient de $\sum z_j^2$ est $1/n$. Dans le modèle des questions non liées (plan à réponses randomisées (i)), l'estimateur de ρ est

$$\hat{\rho}_{uq} = \frac{s_{wz} - \frac{(1-p_1)(1-p_2)}{p_1 p_2} S_{uv}}{\sqrt{\hat{S}_x^2\,\hat{S}_y^2}}, \tag{20}$$

où $S_{uv} = N\sigma_{uv}/(N-1)$ et où

$$\hat{S}_x^2 = s_w^2 - (1-p_1)\frac{1}{n}\sum_{j\in s} w_j^2 + 2(1-p_1)\bar{U}\bar{w} - (1-p_1)\bar{U}^2 - (1-p_1)\sigma_u^2\left(p_1 + \frac{1-p_1}{N-1}\right)$$

et

$$\hat{S}_y^2 = s_z^2 - (1-p_2)\frac{1}{n}\sum_{j\in s} z_j^2 + 2(1-p_2)\bar{V}\bar{z} - (1-p_2)\bar{V}^2 - (1-p_2)\sigma_v^2\left(p_2 + \frac{1-p_2}{N-1}\right).$$

Lorsque $p_1 = p_2$, ce résultat peut être comparé à l'estimateur de Edgell et coll. (1986). L'estimateur résultant pour $\hat{\rho}_{uq}$ diffère de l'estimateur de Edgell et coll. (1986), lequel suppose que $\sigma_{uv} = 0$. Ces auteurs utilisent également les estimateurs biaisés de σ_x^2 et σ_y^2. L'estimateur de Edgell et coll. pour σ_y^2 s'obtient en écrivant la variance de $\bar{z}$ selon le plan d'expérience dans le plan d'échantillonnage aléatoire simple avec remise sous la forme suivante:

$$\sigma_z^2/n = \sum_{j=1}^{N} (z_j - \bar{Z})^2/(Nn). \tag{21}$$

La variance de $\bar{z}$ dans le plan à réponses randomisées est

$$[p_2\sigma_y^2 + (1-p_2)\sigma_v^2 + p_2(1-p_2)(\bar{Y} - \bar{V})^2]/n. \tag{22}$$

L'expression (22) est tirée de Greenberg et coll. (1971). L'estimateur de σ_y^2 est calculé en l'expression (22) dans le membre de gauche de l'équation (21), en substituant l'estimateur d'échantillon de σ_z^2 et l'estimateur de la réponse randomisée de $\bar{Y}$ dans l'équation ainsi obtenue, et en résolvant pour σ_y^2.

Chacun des estimateurs des variances et covariances de la population finie, qui sont les composantes de $\hat{\rho}$ dans (18), (19) et (20), sont non biaisés selon le plan d'expérience dans le plan à réponses randomisées approprié pour tout plan d'expérience où la probabilité de sélection composée des unités i et j donnée par $\pi_{ij} = n(n-1)/[N(N-1)]$. En conséquence, chaque estimateur est l'estimateur optimal non biaisé selon le plan d'expérience pour le paramètre de la population finie auquel il correspond. Pour obtenir les estimateurs non biaisés appropriés dans (18), on multiplie le numérateur et le dénominateur par $(N-1)/N$. Le numérateur obtenu est non biaisé selon le plan d'expérience pour σ_{xy} et les expressions placées sous le signe de la racine carrée au dénominateur de (18) sont non biaisées pour σ_x^2 et σ_y^2. Dans (19), il est nécessaire de multiplier le numérateur et le dénominateur par $(N-1)/[N\mu_u\mu_v]$ pour obtenir la forme correcte des estimateurs non biaisés selon le plan d'expérience. Les estimateurs appropriés sont obtenus en (20) lorsque le multiplicateur est $(N-1)/(Np_1p_2)$.

Dans n'importe quel des plans à réponses randomisées, l'estimation la plus simple de la variance de $\hat{\rho}$ s'obtient en calculant l'estimation de la variance selon la méthode jackknife. Les estimations jackknife de la variance de $\hat{\rho}$ s'obtiennent à l'aide des formules (4.2.3) ou (4.2.5) fournies par Wolter (1985).

5. EFFET DU BIAIS DANS LES RÉPONSES

Dans le modèle des constantes additives, on demande au répondant d'ajouter une variable aléatoire de u à x et une variable aléatoire indépendante de v à y. Au lieu de cela, le répondant peut décider d'ajouter des variables aléatoires indépendantes différentes, par exemple u' et v'. Les moyennes et les variances de u' et v' peuvent être différentes de celles de u et v. Il est raisonnable de penser, toutefois, que $\sigma_{u'}^2 \geq \sigma_u^2$ et $\sigma_{v'}^2 \geq \sigma_v^2$. Nous présentons ci-après un exemple d'une telle situation. Le répondant ne souhaite pas ajouter une variable aléatoire qui se situe près de la moyenne de la distribution de cette variable aléatoire. Dans ce cas, la distribution du biais dans la réponse pourrait être modélisée par la distribution originale avec un intervalle autour de la moyenne déterminé de façon que tout résultat de la distribution originale tombant dans l'intervalle choisi soit placé à l'une des extrémités de l'intervalle. En prenant séparément les espérances du numérateur et de l'expression placée sous chacun des signes de racine carrée dans le dénominateur de (18), on obtient l'expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 + \sigma_{u'}^2 - \sigma_u^2}\sqrt{\sigma_y^2 + \sigma_{v'}^2 - \sigma_v^2}}, \tag{23}$$

On peut noter dans cette expression que le biais dans les réponses conduit à une estimation de la corrélation inférieure à la valeur véritable.

Le modèle des constantes multiplicatives se comporte comme le modèle des constantes additives, sauf que les réponses aux questions délicates sont multipliées par les

variables aléatoires. Comme dans le cas du modèle des constantes additives, on présume que le répondant utilise u' et v' au lieu de u et v. Ensuite, en prenant séparément les espérances du numérateur et de l'expression placée sous chacun des signes de racine carrée dans le dénominateur de (19), on obtient l'expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2+\frac{\sum_{j=1}^{N}x_j^2}{N\mu_{u'}^2}\,\frac{\sigma_u^2\mu_{u'}^2-\sigma_{u'}^2\mu_u^2}{\sigma_u^2+\mu_u^2}}\sqrt{\sigma_y^2+\frac{\sum_{j=1}^{N}y_j^2}{N\mu_{v'}^2}\,\frac{\sigma_v^2\mu_{v'}^2-\sigma_{v'}^2\mu_v^2}{\sigma_v^2+\mu_v^2}}}, \tag{24}$$

Si $\mu_u = \mu_{u'}$, $\mu_v = \mu_{v'}$, $\sigma_{u'}^2 \geq \sigma_u^2$ et $\sigma_{v'}^2 \geq \sigma_v^2$, comme c'est le cas avec le modèle des constantes additives, il en découle que selon l'expression (24), le biais dans les réponses conduit à une surestimation de la corrélation.

Dans le cas du modèle des questions non liées, on peut raisonnablement présumer, quand on se penche sur la question du biais dans les réponses, que les répondants répondent aux questions délicates avec une probabilité $p_1' < p_1$ et $p_2' < p_2$. En général, l'effet de ce biais dépend des valeurs relatives des diverses probabilités, des moyennes et des variances des questions délicates, et des moyennes et des variances des questions non délicates. En vertu du plan d'échantillonnage aléatoire simple sans remise et du modèle du biais dans les réponses, la valeur de l'espérance du numérateur de (20) est donnée par

$$p_1'p_2'\left[S_{xy}+\frac{(1-p_1')(1-p_2')-(1-p_1)(1-p_2)}{p_1'p_2'}\,S_{uv}\right],$$

qui est plus grand que $p_1'p_2'S_{xy}$. De la même façon, l'espérance de l'expression $\hat{S}_x^2$ dans (20) est donnée par

$$\begin{aligned}&S_x^2[p_1'^2+(N-1)p_1'(p_1-p_1')/N]\\&\quad+(p_1-p_1')S_u^2[p_1'-(p_1'+2p_1-2)/N]\\&\quad+p_1'(p_1-p_1')(\bar{X}-\bar{U})^2,\end{aligned}$$

qui est plus grand que $p_1'^2\,\hat{S}_x^2$ lorsque N est grand. Si $S_{uv} = 0$, le biais dans les réponses a tendance à sous-estimer la corrélation.

REMERCIEMENTS

Cet article est dédié à la mémoire de Stan Warner. Il a été préparé en vue de la séance du *Stanley Warner Memorial* tenue à l'occasion des rencontres de la Société statistique du Canada tenues à Banff (Alberta), en mai 1994. Ces travaux ont bénéficié d'une aide financière du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.

CHAUDHURI, A., et MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

EDGELL, S.E., HIMMELFARB, S., et CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.

GREENBERG, B.G., ABUL-ELA, A.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.

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., et HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.

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.

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

POLLOCK, K.H., et BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.

RAO, C.R. (1952). Some theorems on minimum variance unbiased estimation. *Sankhyā* (A), 12, 27-42.

RAO, J.N.K. (1975). On the foundations of survey sampling. Dans *A Survey of Statistical Design and Linear Models*. (Éd. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., et BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.

STEM, D.E., et STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.

UMESH, U.N., et PETERSON, R.A. (1991). A critical evaluation of the randomized response method. *Sociological Methods and Research*, 20, 104-138.

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.

WARNER, S.L. (1976). Optimal randomized response models. *Revue Internationale de Statistique*, 44, 205-212.

WARNER, S.L. (1986). The omitted digit randomized response model for telephone applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.