Techniques d'enquête, décembre 1989
Vol. 15, n° 2, pp. 235–245
Statistique Canada

Échantillonnage pour populations dichotomiques par la méthode des réponses randomisées avec randomisation continue

LeROY A. FRANKLIN[1]

RÉSUMÉ

L'auteur élabore un modèle de randomisation des réponses pour des populations dichotomiques. Ce modèle prévoit l'utilisation de la randomisation continue ainsi que des essais multiples pour chaque répondant. L'auteur se penche sur le cas particulier de la randomisation avec distributions normales et exécute une simulation par ordinateur pour découvrir les effets que peut avoir cette méthode d'échantillonnage sur la quantité d'information dans l'échantillon. Il décrit aussi un appareil électronique portatif qui mettrait en application son modèle. Enfin, il présente les résultats de l'enquête qu'il a réalisée à l'aide de cet appareil. Les résultats illustrent la supériorité de la méthode des réponses randomisées par rapport à l'interview directe, du moins en ce qui a trait à certaines questions délicates.

MOTS CLÉS: Réponses randomisées; randomisation avec distributions continues; simulation par ordinateur.

1. INTRODUCTION

Les enquêtes ont souvent pour but d'estimer la proportion d'individus qui remplissent une condition particulière. Si cette condition a trait à quelque chose de très personnel ou controversé (par ex., recherche d'un nouvel emploi, comportement sexuel) ou à quelque chose d'illégal (par ex., consommation de drogues, activités criminelles), le répondant hésitera peut-être à donner une réponse franche ou refusera peut-être de répondre à une question qui lui serait posée directement sur le sujet. Ainsi, l'estimation de proportions devient problématique lorsqu'on doit se fonder sur une enquête où des questions délicates sont posées directement aux répondants.

Les plans d'échantillonnage fondés sur la méthode des réponses randomisées prévoient un dispositif aléatoire (ou dispositif de randomisation) qui permet à des personnes de répondre à des questions délicates sans se trahir. Le résultat du dispositif est connu du répondant mais non de l'intervieweur. En revanche, l'expérimentateur connaît les propriétés du dispositif et peut, par conséquent, tirer des conclusions sur la proportion à estimer sans savoir quoi que ce soit des personnes qui ont répondu au questionnaire. Le dispositif aléatoire introduit du bruit dans le processus de collecte des données mais il peut être préférable de subir une perte d'information que de devoir composer avec le bruit incontrôlable engendré par la non-réponse ou le mensonge lorsque les questions sont posées directement.

Le modèle des réponses randomisées a été initialement imaginé par Warner (1965); il prévoyait alors une randomisation dichotomique pour une population dichotomique. Winkler et Franklin (1979) ont étudié ce modèle dans une perspective bayesienne. Gould, Shah et Abernathy (1969) ont imaginé le modèle des réponses randomisées avec deux essais ou plus par répondant et Liu et Chow (1976) lui ont apporté des perfectionnements. Dans les deux cas, les auteurs ont prouvé que le modèle à essais multiples produisait des estimations plus efficientes que le modèle à essai unique de Warner. En revanche, ils ont fait remarquer que le

[1] LeRoy A. Franklin, Department of System and Decision Sciences, Indiana State University, School of Business, Terre Haute, Indiana 47809.

modèle à essais multiples était plus susceptible d'éveiller la méfiance du répondant et de l'inciter, par conséquent, à ne pas fournir une réponse honnête. L'article de Horvitz, Greenberg et Abernathy (1976) expose plusieurs autres plans qui prévoient des mécanismes de randomisation discontinue. En outre, l'ouvrage récent de Chaudhuri et Mukerjee (1988) intitulé *Randomized Response: Theory and Techniques* contient un exposé théorique détaillé ainsi qu'une analyse de résultats. Warner (1971) présente un modèle plus général, à randomisation continue ou discontinue, et Pitz (1980), Smouse (1984) et O'Hagen (1987) en font l'analyse dans une perspective bayesienne. Quelques enquêtes ont déjà été réalisées à ce jour; certaines d'entre elles ont montré la supériorité de la méthode des réponses randomisées par rapport aux méthodes d'enquête classiques (par ex.: Gold et coll., 1969 et Liu et Chow 1976) et quelques autres ont produit des résultats incertains (par ex.: Brewer 1981). Cependant, seul Poole (1974) a imaginé une distribution de randomisation continue particulière (uniforme) pour estimer une distribution continue en demandant à des enquêtés d'inscrire une réponse qui avait été préalablement multipliée par un nombre choisi aléatoirement dans une table de nombres aléatoires.

Dans cet article, nous allons considérer un modèle de réponses randomisées pour population dichotomique mais avec distribution de randomisation continue. Selon le modèle original de Warner, la question à laquelle doit répondre l'enquêté est déterminée par le dispositif de randomisation alors que dans le cas de la méthode exposée ici, la question est déterminée par le fait que le répondant appartient ou non au groupe auquel on s'intéresse. Selon cette méthode, des valeurs sont tirées aléatoirement de deux distributions (l'une pour "oui" et l'autre pour "non") et le répondant inscrit la valeur qui correspond à sa situation. En donnant une réponse constituée de plusieurs caractères numériques, le répondant se prête simultanément à plusieurs essais. Cela représente un avantage par rapport aux méthodes habituelles en ce que le répondant croit qu'il a donné une seule réponse alors qu'il s'est soumis, à son insu, à plusieurs essais.

La section 2 sert à présenter le modèle général, selon lequel la randomisation peut s'effectuer par n'importe quel type de distribution. Dans la section 3, nous étudions le cas particulier de la randomisation selon une distribution normale et analysons une méthode d'approximation visant à évaluer l'incidence de la randomisation et des essais multiples par répondant. Dans la section 4, nous cherchons à évaluer, au moyen d'une simulation par ordinateur, comment le choix des moyennes et des écarts types influe sur l'efficacité des enquêtes où l'on a recours à la randomisation par distribution normale avec essais multiples. La section suivante montre comment mettre en application une distribution de randomisation normale par l'utilisation d'un appareil électronique informatisé qui génère et affiche des valeurs aléatoires normales. On croit que cet appareil pourrait s'avérer plus efficace que les cartes ou la roulette puisque ces "outils" peuvent être mal utilisés par le répondant ou l'intervieweur. Dans la section 6, nous analysons les résultats d'une enquête où l'appareil en question a servi à répondre à cinq questions délicates. Enfin, la section 7 contient un résumé de l'étude et une brève analyse de questions relative au plan de sondage.

2. PRÉSENTATION DU MODÈLE

Supposons que nous cherchons à connaître θ, la proportion d'individus qui appartiennent au groupe A dans une population particulière. Nous tirons un échantillon aléatoire simple de n personnes dans la population $n \geq 1$, en supposant que n est suffisamment faible par rapport à la population pour que l'on puisse envisager un échantillonnage avec remise. Chaque répondant est soumis à k essais $k \geq 1$. À l'essai j pour le répondant i, des valeurs aléatoires sont tirées des fonctions de distribution G_{ij} et H_{ij}. Le répondant voit ces deux valeurs et doit inscrire celle provenant de G_{ij} s'il appartient au groupe A et celle provenant de H_{ij} dans le cas contraire. L'enquêteur connaît la forme exacte de G_{ij} et H_{ij} mais ne voit que la valeur inscrite par le répondant, qui est désignée par z_{ij}, et ne sait donc pas de quelle distribution elle origine.

Il s'agit d'estimer θ en se fondant sur les kn observations d'échantillon z_{ij}, $i = 1, \ldots, n$ et $j = 1, \ldots, k$. Pour des raisons de commodité, nous supposons dans le reste de cet article que G_{ij} et H_{ij} sont parfaitement continues, avec g_{ij} et h_{ij}; comme fonctions de densité respectives; le même raisonnement s'applique à la distribution discontinue. La fonction de densité conditionnelle de z_{ij} étant donné θ est $\theta\, g_{ij}(z_{ij}) + (1 - \theta)\, h_{ij}(z_{ij})$, et la fonction de vraisemblance pour toute l'expérience est:

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_{ij}(z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_{ij}(z_{ij})\right] \text{ pour } 0 \leq \theta \leq 1, \quad (2.1)$$

où $\underset{\sim}{z} = (\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)$ et $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{ik})$.

À l'aide du théorème du binome, nous pouvons réécrire la fonction de vraisemblance de la façon suivante

$$L(\underset{\sim}{z} \mid \theta) = \sum_{t=0}^{n} \alpha_t\, \theta^t (1 - \theta)^{n-t} \text{ où } 0 \leq \theta \leq 1 \text{ et} \quad (2.2)$$

$$\alpha_t = \sum_{s=1}^{c} \left[\prod_{i \in C_{ts}} \prod_{j=1}^{k} g_{ij}(z_{ij})\right]\left[\prod_{i \notin C_{ts}} \prod_{j=1}^{k} h_{ij}(z_{ij})\right], \quad (2.3)$$

où $C_{t1}, \ldots, C_{tc}$ représentent les $c = \binom{n}{t}$ combinaisons de t éléments parmi n éléments. Dans l'équation (2.2), $\theta^t(1 - \theta)^{n-t}$ est la fonction de vraisemblance de Bernoulli, à condition qu'exactement t répondants appartiennent au groupe A, et α_t est la fonction de vraisemblance de $\underset{\sim}{z}$ étant donné t. La forme mixte de l'équation (2.2) est due à l'impossibilité de connaître la valeur précise de t dans l'échantillon.

Nous obtenons une forme particulière de (2.1) lorsque nous supposons que les mêmes distributions de randomisation sont utilisées pour les n répondants. Dans ce cas, $g_{ij} = g_j$ et $h_{ij} = h_j$ pour $i = 1, \ldots, n$ et l'équation (2.1) se ramène à

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_i(z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_j(z_{ij})\right] \text{ pour } 0 \leq \theta \leq 1. \quad (2.4)$$

Quelle que soit la forme de l'équation, il faut effectuer une recherche par case pour déterminer les estimations les plus vraisemblables. Cela est réalisable, en l'occurrence, puisque θ est une variable unidimensionnelle qui ne peut prendre que les valeurs de 0 à 1. Cette opération peut être exécutée facilement à l'aide de techniques de recherche courantes appliquées au logarithme de la fonction de vraisemblance. (Voir, par exemple, Kennedy et Gentle 1980).

3. RANDOMISATION AVEC DISTRIBUTIONS NORMALES

Bien que n'importe quelle distribution continue (*ex.* Weibull, uniforme, *etc.*) puisse servir de distribution de randomisation dans le modèle exposé ci-dessus, nous allons nous concentrer ici sur la distribution normale. Nous allons, de plus supposer que les mêmes distributions de randomisation sont utilisées pour tous les répondants de sorte que nous aurons la fonction de vraisemblance (2.4). En conséquence, g_j et h_j sont des fonctions de densité normales ayant pour moyennes μ_{gj} et μ_{hj} pour écarts types σ_{gj} et σ_{hj}, respectivement. On peut alors établir un rapport entre ces fonctions de densité normales et la fonction de vraisemblance définie dans la section 2.

La quantité d'information que l'on peut obtenir à propos de θ dépend évidemment du choix des moyennes et des écarts types. D'une part, si $\mu_{gj} = \mu_{hj}$ et $\sigma_{gj} = \sigma_{hj} j = 1, \ldots, k$, θ disparaît de la fonction de vraisemblance et $\underset{\sim}{z}$ (l'échantillon) ne fournit aucune information sur θ. D'autre part, si $|\mu_{gj} - \mu_{hj}| \to \infty$ pour tout j, σ_{gj} et σ_{hj} étant fixes, **ou** si $\sigma_{gj} \to 0$ et $\sigma_{hj} \to 0$ pour tout j, $|\mu_{gj} - \mu_{hj}| \neq 0$, étant fixe et différend de 0, nous sommes effectivement en mesure de savoir à quel groupe appartient chaque répondant et dans ces conditions, l'échantillonnage ressemble à un échantillonnage de Bernoulli dans θ.

En utilisant une formule d'approximation de $L(\underset{\sim}{z} \mid \theta)$ élaborée par Winkler et Franklin (1979), nous pouvons évaluer plus facilement l'effet de la randomisation et des essais multiples en fonction de moyennes et d'écarts types précis. Autrement dit, il est possible, pour chaque échantillon, de faire une approximation de la fonction de vraisemblance définie en (2.4) à l'aide de la fonction de vraisemblance approximative

$$L^*(r^*, n^* \mid \theta) = \theta^{r^*} (1 - \theta)^{n^* - r^*}. \tag{3.1}$$

Si nous calculons les dérivés première et seconde du logarithme de la fonction (3.1) et que nous cherchons à déterminer le maximum ($\hat{\theta}$) et la courbure à ce maximum, nous obtenons:

$$\hat{\theta} = \frac{r^*}{n^*} \tag{3.2}$$

et

$$\left[\frac{\partial^2 \log L^*(r^*, n^* \mid \theta)}{\partial \theta^2}\right]_{\theta = \hat{\theta}} = -\frac{n^*}{\hat{\theta}(1 - \hat{\theta})}. \tag{3.3}$$

Ensuite, si nous calculons la dérivée première du logarithme de la fonction de vraisemblance exacte (2.4) et que nous posons le résultat égal à zéro, nous obtenons l'équation qui permettra de calculer l'estimation la plus vraisemblable de θ:

$$\sum_{i=1}^{n} \frac{\gamma_i - \eta_i}{\theta\gamma_i + (1-\theta)\eta_i} = 0 \text{ où } \gamma_i = \prod_{j=1}^{k} g_j(z_{ij}),\ \eta_i = \prod_{j=1}^{k} h_j(z_{ij}). \tag{3.4}$$

On obtient la solution ($\hat{\theta}_r$) de (3.4) au moyen d'une recherche par case. En calculant la dérivée seconde du logarithme de la fonction de vraisemblance exacte (2.4), nous obtenons:

$$\left[\frac{\partial^2 \log L(\underset{\sim}{z} \mid \theta)}{\partial \theta^2}\right] = -\sum_{i=1}^{n} \frac{[\gamma_i - \eta_i]^2}{[\theta\gamma_i + (1 - \theta)\eta_i]^2}. \tag{3.5}$$

En appliquant $\hat{\theta}_r$ dans l'équation (3.5), nous obtenons la courbure de la fonction de vraisemblance logarithmique à $\hat{\theta}_r$ (le maximum). Les équations (3.2) et (3.3) sont deux équations avec deux inconnues, r^* et n^*. En posant (3.2) $= \hat{\theta}_r$ et (3.3) $=$ (3.5) nous pouvons déterminer les valeurs de r^* et de n^* de manière que le maximum $\hat{\theta} = \hat{\theta}_r$, et la courbure à ce maximum soient les mêmes pour la fonction de vraisemblance logarithmique approximative et la fonction de vraisemblance logarithmique réelle. Par conséquent, on peut apparenter l'échantillon avec réponses randomisées $\underset{\sim}{z}$ à un échantillon avec réponses non randomisées un échantillon de Bernoulli), dont r^* éléments sur n^* appartiennent au groupe auquel s'intéresse l'expérimentateur. Dans ce sens, on peut considérer n^* comme une mesure approximative de la quantité d'information contenue dans l'échantillon avec réponses randomisées de taille n.

4. ANALYSE DE L'INCIDENCE DES MOYENNES ET DES ÉCARTS TYPES À L'AIDE D'UNE SIMULATION PAR ORDINATEUR

Afin d'analyser l'effet de certains moyennes et de certaines écarts types pour les distributions de randomisation normales de même que l'incidence que peut avoir sur r^* et n^* la valeur de θ et de k (nombre d'essais), nous avons simulé par ordinateur un échantillonnage avec réponses randomisées en formant de façon répétée des échantillons par un processus de Bernoulli avec paramètre θ et k séries de réponses à deux chiffres pour chaque échantillon. Pour la simulation, nous avons posé $\mu_{gj} = 50$, $\mu_{hj} = 40$, et $\sigma_{gj} = \sigma_{hj} = \sigma$ pour $j = 1, \ldots, k$. Nous avons considéré deux valeurs de θ (.10 et .25), deux valeurs de σ (6 et 9), trois valeurs de n (50, 200, et 500), et trois valeurs de k (1, 2, et 3). Nous avons choisi ces valeurs car elles permettent de déceler les écarts aléatoires de deux chiffres qui se chevauchent largement dans la distribution; elles peuvent ainsi servir de données-repères lorsque vient le temps de choisir des moyennes et des écarts types pour de vraies enquêtes. Nous avons produit 25 échantillons pour chacune des 36 combinaisons de paramètres. Nous avons déterminé les valeurs de r^* et n^* pour chaque échantillon et avons consigné dans le tableau 1 la moyenne de n^* pour les 25 échantillons pour chaque combinaison de paramètres.

Les valeurs moyennes de n^* varient considérablement. Suivant le pire scénario $\sigma = 9$, $\theta = .10$, et un seul essai par répondant), n^* n'équivaut qu'à 10 ou à 15 % de n. Par contre, lorsque $\sigma = 6$, $\theta = .25$, et qu'il y a trois essais par répondant, n^* équivaut à environ 75 % de n. Comme prévu, la valeur moyenne de n^* (taille effective de l'échantillon) augmente lorsque n (nombre de répondants) ou k (nombre d'essais par répondant) augmente. Par ailleurs, une diminution de σ ou une augmentation de θ entraîne également une augmentation de n^*.

Nous avons aussi déterminé, pour chaque combinaison de paramètres, la moyenne et la variance de $\hat{\theta}$ pour les 25 échantillons répétés. Les valeurs moyennes de $\hat{\theta}$ sont très près des valeurs correspondantes de θ, (écart inférieur à 5 %) et la variance de $\hat{\theta}$ tend à augmenter lorsque la valeur moyenne de n^* diminue, ce qui tend à confirmer la simulation.

Tableau 1

Valeurs moyennes de la taille effective de l'échantillon (n^*) selon diverses tailles d'échantillon (n) et le nombre d'essais par répondant (k)

		$\theta = .10$		$\theta = .25$	
n	k	$\sigma = 6$	$\sigma = 9$	$\sigma = 6$	$\sigma = 9$
	1	16.2	7.0	17.3	9.2
50	2	27.3	13.1	30.6	17.8
	3	32.6	18.1	38.2	23.6
	1	58.3	24.8	79.0	41.2
200	2	103.1	49.6	124.4	72.9
	3	136.6	77.7	151.0	97.7
	1	148.4	59.6	196.9	103.6
500	2	261.1	129.3	309.5	181.2
	3	345.8	193.1	375.6	242.7

5. APPAREIL ÉLECTRONIQUE PORTATIF POUR LA RANDOMISATION

L'échantillonnage fondé sur la randomisation avec distributions normales et essais multiples offre beaucoup de souplesse à l'expérimentateur, celui-ci pouvant choisir à son gré les moyennes et les variances ainsi que le nombre de répondants et le nombre d'essais par répondant. Cet avantage sera toutefois inutile si on ne peut mettre en pratique le plan d'échantillonnage. Le plan d'échantillonnage fondé sur la randomisation de Bernoulli peut être exécuté de diverses façons (par ex. à l'aide de cartes ou de billes de couleur). Or, le modèle que nous élaborons ici prévoit la génération de valeurs aléatoires normales au moyen d'un appareil portatif.

On a construit un appareil électronique autour du microprocesseur Intel 8080 pour générer et afficher des valeurs aléatoires normales. On obtient chaque valeur en additionnant 16 nombres aléatoires distribués uniformément et en transformant cette somme de manière à obtenir un écart aléatoire normal ayant la moyenne et l'écart type voulus. En vertu du théorème limite central, les valeurs obtenues devraient être distribuées approximativement selon une loi normale et des tests sérieux indiquent que les valeurs générées par l'appareil suivent effectivement une distribution comme celle des valeurs aléatoires normales. On a préféré cette méthode à d'autres parce qu'elle est facile à programmer par des instructions machine pour ce qui a trait au microprocesseur Intel 8080. Pour avoir plus de détails sur la génération des valeurs aléatoires normales et l'essai de l'appareil, voir Franklin (1977), Kennedy et Gentle (1980) et Knuth (1969).

La version finale de l'appareil a à peu près la taille d'une boîte de cigares et tient facilement dans la main. Il peut être alimenté par un bloc-piles ou un cordon rallonge.

Pour l'affichage, les valeurs aléatoires normales ne sont formées que de deux chiffres et l'appareil est conçu de manière à afficher simultanément six nombres de deux chiffres chacun dans des "fenêtres" pouvant contenir chacune six chiffres. Dans une des fenêtres (fenêtre "Oui") apparaissent les valeurs tirées de g_1, g_2, et g_3 sous la forme d'un nombre à six chiffres. L'autre fenêtre (fenêtre "Non") sert à afficher les valeurs tirées de h_1, h_2, et h_3 et qui sont groupées sous la forme d'un nombre à six chiffres. Les six moyennes et les six écarts types sont mémorisés dans l'appareil mais on peut les modifier facilement à l'aide d'un petit clavier mobile.

Voici comment se déroule l'interview. Tout d'abord, l'intervieweur pose au répondant une question délicate au sujet du groupe A. Le répondant appuie alors sur un bouton pour mettre en activité l'appareil et en un quart de seconde, deux nombres de six chiffres apparaissent dans les fenêtres. Si le répondant appartient au groupe A, il inscrit le nombre paraissant dans la première fenêtre (fenêtre "Oui"); dans le cas contraire, il inscrit le nombre paraissant dans la seconde fenêtre (fenêtre "Non"). Pour convaincre le répondant du "caractère aléatoire" des valeurs observées, l'intervieweur l'invite à appuyer plusieurs fois sur le bouton et à observer les valeurs affichées; ensuite, il lui pose la question délicate. Bien que $k = 3$, il faut préciser que le répondant a l'impression de fournir une seule réponse, en l'occurrence un nombre à six chiffres; en réalité, cette réponse est le résultat de trois essais. Ainsi, l'expérimentateur profite des avantages des essais multiples par répondant sans en subir les inconvénients habituels.

6. RÉSULTATS DE L'ENQUÊTE ET CONCLUSIONS

Deux enquêtes indépendantes ont été menées simultanément sur le campus d'une grande université auprès des étudiants de cette université. L'une des enquêtes consistait à poser cinq questions délicates par interview directe. L'autre enquête consistait à poser les mêmes questions délicates à d'autres personnes mais cette fois, en utilisant la méthode des réponses randomisées avec randomisation continue; pour cela, on a utilisé l'appareil électronique décrit dans la section précédente. Pour les besoins de l'étude, nous avons posé $k = 3$, $\mu_{g_1} = \mu_{g_2} = \mu_{g_3} = 40$, $\mu_{h_1} = \mu_{h_2} = \mu_{h_3} = 50$ et $\sigma_{g_j} = \sigma_{h_j} = 5$ pour $j = 1,2,3$. Ces valeurs

ont été choisies en conformité avec les résultats de la simulation par ordinateur exposée dans la section 4. Pour chacune des deux enquêtes, des étudiants ont été choisis systématiquement parmi ceux qui circulaient sur le campus (un à tous les cinq passants) et ont été interviewés individuellement. Lorsqu'on abordait un étudiant, on lui exposait brièvement le but de l'enquête et on lui demandait s'il voulait y participer. Moins de 10 % de toutes les personnes abordées par les deux équipes d'enquêteurs ont refusé de prêter leur concours. Lorsqu'une personne acceptait de participer à l'enquête, on lui demandait de produire son numéro de sécurité sociale afin de vérifier si elle était réellement inscrite à l'université. À cette fin, on a dû consulter un répertoire d'étudiants fourni par l'administration de l'université et on a exclu de l'enquête tous les répondants qui n'étaient pas réellement inscrits comme étudiants (moins de 5 % des personnes interviewées).

Le fait de demander à une personne son numéro de sécurité sociale l'amène à penser que l'enquêteur fera inévitablement le lien entre elle et les réponses qu'elle donnera. Cette attitude se retrouve dans beaucoup d'enquêtes (par ex. : enquêtes téléphoniques, questionnaires envoyés par la poste, interviews sur place, *etc.*) et détermine dans une très large mesure l'honnêteté des réponses de la personne. On croit que c'est justement dans ces circonstances que l'expérimentateur peut tirer le plus grand profit de la méthode des réponses randomisées. La taille des échantillons pour la méthode de l'interview direct et la méthode des réponses randomisées était $n_1 = 473$ et $n_2 = 477$ respectivement. Les cinq questions délicates étaient les suivantes:

Q1 – "Avez-vous déjà triché à un examen que vous avez passé ici à l'université?"
Q2 – "Seriez-vous tenté de falsifier votre déclaration de revenu?"
Q3 – "Voleriez-vous votre employeur?"
Q4 – "Avez-vous fumé de la marijuana dans les 30 derniers jours?"
Q5 – "Vous êtes-vous déjà livré à un acte homosexuel?"

On jugeait que les cinq questions étaient suffisamment délicates pour faire ressortir de façon non équivoque les avantages de la méthode des réponses randomisées par rapport à la méthode de l'interview directe. De plus, les personnes qui avaient accepté de répondre aux questions par la méthode des réponses randomisées devaient répondre à la fin à une question libellée comme suit : "Croyez-vous que vos amis seraient plus disposés à donner une réponse honnête s'ils devaient répondre à des questions délicates par cette méthode?" Cette question avait pour but de vérifier si la personne interviewée était vraiment convaincue que cette méthode garantissait l'anonymat et était conforme aux règles de protection des renseignements personnels.

Le tableau 2 donne la proportion estimée de répondants (selon la méthode de l'interview directe ($\hat{\theta}_{id}$) et la méthode des réponses randomisées ($\hat{\theta}_{ir}$) qui appartiennent au groupe A pour chaque question i ainsi que la valeur estimée de n_i^* (taille effective de l'échantillon) établie suivant la méthode analysée dans la section 3. Ce tableau contient également les valeurs z d'un test d'hypothèse unilatéral H_o: $\theta_{id} - \theta_{ir} = 0$ contre H_a: $\theta_{id} - \theta_{ir} < 0$, ainsi que les valeurs p observées. Comme on s'est servi des tailles d'échantillon n_1 et n_i^* pour ces tests, le résultat est beaucoup plus modéré que si on avait utilisé n_1 et n_2 .

Il convient de noter que les valeurs estimées de θ sont plus élevées dans les cinq cas avec la méthode des réponses randomisées qu'avec la méthode de l'interview directe. En outre, dans le cas des questions 1, 2 et 5, l'écart entre les estimations établi à l'aide des deux méthodes était statistiquement significatif (valeur $p < .001$ pour ces trois questions). Il semble donc que l'on puisse affirmer que la méthode des réponses randomisées avec randomisation continue produit de meilleures estimations de proportions, du moins en ce qui concerne certaines questions délicates. Il convient aussi de noter que les valeurs choisies pour μ_{g_j}, μ_{h_j}, σ_{g_j} et σ_{h_j} et $k = 3$ nous ont permis d'obtenir une taille effective n_i^* qui représentait généralement 75 % à 85 % de la taille d'échantillon originale n_2 ce qui nous amène à dire que la méthode des réponses randomisées a permis de "récupérer" la majeure partie de l'information.

Tableau 2

Valeurs estimées de θ et résultats du test d'hypothèse concernant les valeurs de θ's obtenues par la méthode de l'interview directe et la méthode des réponses randomisées avec des échantillons de taille $n_1 = 473$ et $n_2 = 477$ respectivement

Question i	$\hat{\theta}_{id}$	$\hat{\theta}_{ir}$	(taille effective de l'échantillon) n_i^*	valeur z	valeur p
1	.0634	.2013	394.5	6.098	$<.0001$
2	.1797	.2941	408.1	3.997	$<.0001$
3	.1078	.1207	384.8	.583	.2810
4	.1882	.1942	409.5	.234	.4091
5	.0042	.0355	339.0	3.341	.0004

Par ailleurs, il est éclairant d'examiner les résultats non significatifs des questions 3 et 4. Ces résultats pourraient amener un observateur à conclure que la méthode des réponses randomisées n'est pas particulièrement supérieure à la méthode de l'interview directe (si l'on fait abstraction des résultats significatifs observés pour les trois autres questions). Cependant, compte tenu justement des trois écarts significatifs observés, on serait peut-être porté à justifier les résultats non significatifs en disant que la question n'était pas "suffisamment délicate" pour qu'un écart appréciable puisse être observé ou même que la question était "si délicate" que le répondant avait préféré mentir même s'il utilisait l'appareil électronique. En outre, la question 1 (Avez-vous déjà triché à un examen que vous avez passé ici à l'université?) semblait relativement "neutre" aux yeux de l'expérimentateur mais on s'est aperçu après coup que le fait de poser cette question après avoir demandé au répondant son numéro de sécurité sociale indisposait les gens beaucoup plus que nous l'aurions cru au départ. Par conséquent, l'incertitude qui existe à propos de l'efficacité de la méthode des réponses randomisées peut être attribuable à des différences d'opinion entre le répondant d'une part et l'intervieweur ou l'expérimentateur d'autre part en ce qui concerne le "degré de sensibilité" d'une question. Ces aspects méritent un examen plus approfondi.

Enfin, 88.9% des personnes qui ont répondu aux questions au moyen de l'appareil électronique (424 sur 477) croient que leurs amis seraient plus susceptibles de répondre honnêtement à des questions délicates s'ils disposaient de cet appareil. Même si on peut penser que des répondants ont surtout voulu faire plaisir à l'intervieweur en répondant à la dernière question par l'affirmative, le très fort pourcentage observé dans ce sens ainsi que les écarts significatifs mentionnés plus haut tendent à démontrer que la méthode des réponses randomisées est bien reçue et jugée conforme aux règles de protection du secret statistique.

7. ANALYSE

Le modèle élaboré dans cet article permet d'appliquer des distributions de randomisation continues et discontinues à l'échantillonnage dans une population dichotomique. Afin d'exécuter la randomisation avec distribution normale, nous avons mis au point un appareil électronique programmé. Cet appareil est portatif, peut recevoir en mémoire des moyennes et des écarts types pour les six distributions normales et permet de réaliser simultanément trois essais dont les résultats sont groupés sous la forme d'un nombre à six chiffres. Ce système présente des avantages et des inconvénients par rapport à d'autres techniques fondées sur la méthode des réponses randomisées.

Premièrement, comme nous l'avons dit dans l'introduction, un dispositif électronique pourrait être plus efficace que les techniques de randomisation classiques, comme les cartes ou la roulette, puisque ces techniques risquent d'être mal appliquées par le répondant ou l'intervieweur, ce qui aurait pour conséquence d'introduire un élément d'erreur incontrôlable. (Voir Abernathy, Greenberg et Horvitz (1970) pour une analyse des problèmes que soulèvent des "cartes mal battues" ou des "cartes perdues" ainsi qu'une formation déficiente des intervieweurs.) Comme la production des valeurs aléatoires est désormais automatique, on n'a plus les problèmes de distribution que pouvait causer l'utilisation des cartes, des boules ou des roulettes puisque la "sélection aléatoire des valeurs" est une opération qui n'est plus exécutée par l'intervieweur *et* le répondant mais par l'ordinateur. Si l'appareil tombe en panne, ce sera le plus souvent à cause d'une défaillance de la microplaquette, incident facilement détectable et très peu probable.

Le second avantage de l'appareil électronique (et peut-être le plus important) est qu'il offre le choix entre deux nombres formés chacun de six chiffres, lequel choix est déterminé par la réponse qu'a l'intention de fournir le répondant ("oui" ou "non"). Mais cette réponse formée de six chiffres renferme en réalité trois réponses constituées chacune de deux chiffres, ce qui représente les trois essais prévus pour chaque répondant. Ainsi, l'expérimentateur profite des avantages des essais multiples sans devoir en subir les inconvénients habituels (mentionnés par Liu et Chow 1976) étant donné que le répondant ignore qu'il se prête à trois essais.

Par surcroît, l'utilisation d'un appareil électronique offre plus de souplesse que les méthodes de randomisation classiques puisque l'expérimentateur peut déterminer à sa guise les six moyennes et les six écarts types. Par exemple, si l'expérimentateur a l'impression que la différence entre les deux premiers chiffres du premier nombre et les deux premiers du second nombre frappent beaucoup les répondants, il peut rapprocher les moyennes μ_{h_1} et σ_{h_1} et les écarts types μ_{g_1}, et σ_{g_1}, (ou même les faire coïncider). De même, si c'est la différence entre les deux chiffres médians du premier nombre et les deux chiffres médians du second qui attire le moins l'attention, l'expérimentateur peut essayer de tirer le plus d'information possible de ces valeurs en éloignant le plus possible l'une de l'autre les moyennes μ_{h_2} et μ_{g_2} . On peut aussi intervertir les chiffres. Par exemple, la première valeur aléatoire normale pourrait être constituée, dans l'ordre, des cinquième et second chiffres au lieu des deux premiers. Cette caractéristique, alliée au fait de pouvoir choisir librement les paramètres, devrait fournir une méthode d'échantillonnage qui renseignera beaucoup l'enquêteur sans pour autant nuire au répondant.

Précisons aussi que même si la distribution de randomisation normale était indiquée pour le genre de microprocesseur dont était pourvu l'appareil, nous aurions pu utiliser plusieurs autres distributions continues (ex.: uniforme, Weibull) ou même des distributions discontinues plurivalentes (ex.: multinomiale ou Poisson). La recherche devrait être orientée vers la création de nouveaux microprocesseurs et l'analyse d'autres distributions de randomisation.

Cette technique de randomisation présente toutefois certains inconvénients. Par exemple, le coût de production de l'appareil électronique est de l'ordre de 1,500 à 2,000 $ (puisqu'il renferme un microprocesseurs). En revanche, le coût relativement élevé de cet appareil pourrait être compensé par une forte fréquence d'utilisation et une durée de vie utile appréciable, deux caractéristiques attribuables à la polyvalence de l'appareil.

Il est plus difficile d'évaluer l'attitude du répondant devant cet appareil et la confiance ou la méfiance que celui-ci peut lui inspirer. Les répondants craignent-ils que l'appareil garde en mémoire leurs réponses et que celles-ci soient déchiffrées taux ou tard pour les confondre? Les résultats de l'enquête semblent indiquer que l'on a obtenu un plus grande nombre de réponses honnêtes avec l'appareil électronique qu'avec l'interview direct. Néanmoins, il serait bon de pousser plus loin la comparaison entre cette nouvelle technique de randomisation et les techniques plus classiques.

En pratique, l'étude de questions relatives à la conception de l'appareil (c.-à-d., la détermination des moyennes et des écarts types) renvoie à plusieurs autres considérations. Si nous voulons obtenir plus d'information pour une taille d'échantillon donnée, nous devons accroître

$|\mu_{g_j} - \mu_{h_j}|$ et réduire σ_{g_j} et σ_{h_j} pour $j = 1, 2, 3$. Toutefois, si nous faisons cela, le répondant s'apercevra que, malgré la randomisation, la réponse qu'il donne est susceptible de le trahir et donnera donc peut-être une fausse réponse ou n'en donnera pas du tout. La détermination de valeurs optimales pour les moyennes et les écarts types exige d'autres recherches. Les résultats du tableau 1 donnent un aperçu de l'effet d'une modification de l'écart type. Du point de vue pratique toutefois, les résultats de l'enquête semblaient indiquer que le fait d'avoir choisi des moyennes séparées l'une de l'autre par deux écarts types avait permis de gagner la confiance du répondant et de récupérer (grâce aux essais multiples) 75 % à 85 % de la taille d'échantillon originale sans susciter la "méfiance" que l'on observe si souvent chez les personnes qui participent à des enquêtes à essais multiples.

Plus particulièrement, l'enquête sur le terrain nous a permis de comparer la méthode de l'interview directe avec la méthode des réponses randomisées axées sur l'utilisation d'un appareil électronique; à cette fin, nous avons posé pour les distributions de randomisation normales $\mu_{h_j} = 40$ et $\mu_{g_j} = 50$ de même que $\sigma_{h_j} = \sigma_{g_j} = 5$ pour $j = 1, 2, 3$. Cinq questions délicates ont été posées à deux groupes d'étudiants différents; nous avons observé que la méthode des réponses randomisées produisait des estimations beaucoup plus élevées ($p < .001$) que la méthode de l'interview directe dans trois cas sur cinq. De plus, 88.9 % des personnes qui ont répondu aux questions à l'aide de l'appareil électronique étaient d'avis que leurs amis seraient plus susceptibles de donner des réponses honnêtes s'ils devaient répondre à des questions délicates à l'aide de cet appareil. Il semble donc que cette technique de randomisation a fourni des réponses plus justes que la méthode de l'interview directe (du moins en ce qui concerne certaines questions).

La question de la protection de la vie privée du répondant mérite d'être analysée. Il est contraire à la morale de dire au répondant que la randomisation garantit l'anonymat de ses réponses alors que l'enquête dispose d'un moyen discret pour associer la réponse à la personne (par ex. en utilisant uniquement des nombres pairs pour "OUI" et des nombres impairs pour "NON"). Grâce à l'appareil électronique dont nous avons parlé plus haut, il semble désormais possible de respecter la vie privée des répondants sans perdre trop d'information. Si les moyennes et les écarts types sont mémorisés dans l'appareil et inconnus de l'intervieweur, celui-ci pourra très difficilement identifier les personnes du groupe A et celles du groupe B au cours de l'interview, surtout si l'ordre normal des chiffres qui forment la réponse est modifié. Ainsi, le système qui nous permet d'obtenir plus d'information sans nuire au répondant empêche du même coup l'intervieweur de savoir à quel groupe appartient un répondant.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance aux arbitres et à un rédacteur associé pour leurs remarques constructives.

BIBLIOGRAPHIE

ABERNATHY, J.R., GREENBERG, B.G., et HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.

BARNARD, G.A. (1976). Discussion on the invited and contributed papers. *International Statistical Review*, 44, 226.

BREWER, K.R.W. (1981). Estimating marijuana usage using randomized response some parodoxical findings. *Australian Journal of Statistics*, 23, 139-148.

CAMPBELL, C., et JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27, 229- 231.

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

CHOW, L.P., LIU, P.T., et MOSELY, W.H. (1973). A new randomized response technique for study of contemporary social problems. Présenté à la 101ième conférence annuelle de l'American Public Health Association, Statistics Section.

FRANKLIN, L.A. (1977). A Bayesian approach to randomized response sampling. Thèse de doctorat non-publié, Indiana University, Bloomington, IN.

GOULD, A.L., SHAH, B.U., et ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Section on Social Statistics American Statistical Association*, 351-359.

HORVITZ, D.G., GREENBERG, B.G., et ABERNATHY, J.R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistics Review*, 44, 181-196.

KENNEDY, W.J., et GENTLE, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker, Inc..

KNUTH, D.E. (1969). *Semi Numerical Algorithms*, (Volume 2). New York: Addison Wesley.

LIU, P.T., et CHOW, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.

O'HAGAN, A. (1987). Bayes linear estimates for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

PITZ, G.F. (1980). Bayesian analysis of randomized response models. *Psychological Bulletin*, 87, 209- 212.

POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.

SMOUSE, E.P. (1984). A note on Bayesian lease squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390- 392.

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.

WINKLER, R.L., et FRANKLIN, L.A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.