Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 137–144
Statistique Canada

Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles

ANDREA GIOMMI[1]

RÉSUMÉ

S'inspirant de l'approche de Cassel, Särndal et Wretman (1983), l'auteur aborde le problème de la non-réponse dans l'estimation de la moyenne d'une population finie. L'auteur propose tout d'abord deux méthodes très simples pour estimer les probabilités de réponse individuelles; il applique ensuite ces méthodes à un modèle de superpopulation pour construire des estimateurs de la moyenne de la population. Enfin, au moyen d'une étude de Monte Carlo, il fait une première évaluation des propriétés des méthodes proposées. Les résultats de cette évaluation nous éclairent sur l'efficacité de ces méthodes.

MOTS CLÉS: Non-réponse; probabilité de réponse individuelle; méthodes non paramétriques.

1. INTRODUCTION

S'intéressant à l'estimation de la moyenne (ou total, etc.) d'une population finie en situation de non-réponse, Cassel, Särndal et Wretman (1983) ont imaginé une méthode d'estimation très générale qui repose sur la notion fondamentale de probabilité de réponse individuelle (PRI). Ils ont proposé des estimateurs qui sont déterminés en partie par un modèle de superpopulation et en partie par un modèle de réponse, c'est-à-dire un modèle qui reproduit le mécanisme de réponse et qui permet d'estimer la PRI à partir de données d'échantillon. L'estimation de la PRI est le point central de leur théorie. De fait, si le modèle de superpopulation ne convient pas parfaitement, comme cela est souvent le cas, seul un modèle de réponse approprié évitera que les estimateurs soient entachés d'un biais attribuable au plan de sondage. À l'aide d'une étude de Monte Carlo, Giommi (1985a) a montré qu'un modèle de réponse qui offrait une "bonne approximation" du "vrai" modèle de réponse pouvait éliminer virtuellement toute possibilité d'erreur systématique. Toutefois, on peut difficilement dire ce qu'est une bonne approximation et, de toute façon, le choix d'un modèle de réponse, outre qu'il est arbitraire, peut s'avérer contraignant. Une manière naturelle de contourner le problème est d'estimer la PRI par des méthodes non paramétriques. Dans cet article, nous proposons deux méthodes très simples pour estimer la PRI lorsque l'information supplémentaire dont nous disposons (et qui est censée dépendre des réponses obtenues) est représentée par une variable continue unique. Ces méthodes, qui intègrent certains éléments de la théorie d'estimation de noyaux, peuvent être considérées comme un prolongement de la méthode de compensation de la non-réponse, d'usage courant, qui consiste à repondérer les unités au moyen de cellules de correction.

Dans cet article, nous faisons une évaluation empirique de ces méthodes et analysons le biais et l'efficacité des estimateurs correspondants.

2. ESTIMATION DES PROBABILITÉS DE RÉPONSE INDIVIDUELLES

Considérons une population de N unités identifiées k $(k=1, 2, \ldots, N)$, et une variable Y à l'étude, dont nous voulons estimer la moyenne $\bar{Y}=\Sigma_k\, y_k/N$ à l'aide d'un échantillon s de

[1] Andrea Giommi, Département de statistique, Université de Florence, Via Curtatone, 1, 50123 Florence, Italie.

n unités prélevé suivant un plan d'échantillonage $p(s)$. Pour l'estimation, nous disposons de renseignements supplémentaires représentés par les valeurs connues x_k $(k=1, \ldots, N)$, d'une variable scalaire continue X (en principe, l'application des méthodes proposées pour le cas à plusieurs variables n'est pas censée poser de problèmes).

Dans l'échantillon, Y est observable uniquement pour un sous-ensemble r de n_r répondants; elle ne peut être observée pour les $n - n_r$ non-répondants. Après l'échantillonnage, l'information disponible peut être représentée de la façon suivante:

$$(k, I_k, I_k y_k, x_k) \qquad k \in s; N, n,$$

où I_k est une variable aléatoire indicatrice telle que $E(I_k)=q_k$ et q_k est la PRI.

Pour estimer q_k, on suppose habituellement un modèle paramétrique (Cassel et coll. 1983) de la forme:

$$q_k=q(\Theta, x_k),$$

où Θ est un paramètre (ou vecteur de paramètres) inconnu et $q(\cdot,\cdot)$ est une fonction à définir. On obtient des estimations de q_k en remplaçant Θ par son estimation $\hat{\Theta}$ dans le modèle paramétrique ci-dessus.

Dans cet article, nous estimons q_k $(k \in r)$ sans utiliser de définition paramétrique de la fonction $q(\cdot,\cdot)$; nous continuons toutefois à supposer que les PRI dépendent des valeurs de x_k. Deux méthodes (méthodes (1) et (2)) sont proposées.

Selon la première méthode, q_k $(k \in r)$ est estimée comme taux de réponse (c'est-à-dire, la proportion de répondants) d'un groupe d'unités centré sur l'unité k et qui correspond à un intervalle approprié de valeurs de x centré sur x_k. En supposant qu'il s'agit d'un intervalle de longueur $2h_k$, q_k est estimée par le ratio suivant:

$$\hat{q}_k = \sum_{j \in r} D(x_k - x_j) \,/\, \sum_{j \in s} D(x_k - x_j), \tag{1}$$

où

$$D(x_k - x_j) = \begin{cases} 1 \text{ si } |x_k - x_j| \leq h_k \\ 0 \text{ dans le cas contraire.} \end{cases}$$

Il est évident que l'estimation $\hat{q}_k$ dépend de h_k ou de h si nous utilisons, comme c'est le cas ici, un intervalle fixe; dans la pratique, la détermination de la valeur de h est un problème majeur.

Selon la seconde méthode, l'estimation de q_k repose sur toutes les unités de l'échantillon et non seulement sur une partie d'entre elles. Cela élimine les inconvénients que pouvait poser une catégorisation des unités répondantes. Autrement dit, on pouvait très difficilement imaginer, pour l'estimation de q_k, que des unités aient une pondération de 1 tandis que d'autres avaient une pondération nulle. Selon la méthode (2), l'estimation est définie par:

$$\hat{q}_k = \sum_{j \in r} D^*(x_k - x_j) \,/\, \sum_{j \in s} D^*(x_k - x_j), \tag{2}$$

où D^* doit être définie. Dans ce cas, chaque valeur x_j contribue au calcul de l'estimation

$\hat{q}_j$ par l'intermédiaire de D^*, qui est une quantité inversement proportionnelle à l'écart $|x_k - x_j|$.

La méthode (2) comporte deux difficultés; il faut en effet définir i) la fonction D^* et ii) les valeurs de ses paramètres. Dans cet article, nous définissons D^* comme une fonction de distribution d'une loi normale:

$$D^*(z) = (h^2 2\pi)^{-\frac{1}{2}} \exp(-z^2/2h^2); \qquad z = x_k - x_j, \tag{3}$$

où l'écart-type, désigné par h, joue le même rôle que le paramètre h de l'expression (1). Lorsque h augmente dans les expressions (1) et (2), $\hat{q}_k$ tend vers la valeur constante n_r/n. Dans (1), il atteint n_r/n lorsque h est au moins égal à l'intervalle des valeurs de x.

Une étude empirique visant à évaluer les propriétés des méthodes proposées a été conçue pour un très grand nombre de valeurs de h. Dans le présent article, nous ne rapportons que les résultats se rattachant à trois de ces valeurs (constantes), soit $h = 1/10$, $3/10$ et $5/10$ de l'intervalle des valeurs de x de l'échantillon. Enfin, soulignons que les équations (1) et (2) ne sont pas que des facteurs de normalisation; chacune d'elles se présente aussi comme le rapport de deux estimateurs noyaux de fonction de densité (selon l'approche de Rosenblatt (1956) pour diverses séries de valeurs de x. Comme le propose Giommi (1985b), on peut donc choisir la valeur de h en tenant compte des propositions contenues dans cette théorie.

3. MODÈLE DE SUPERPOPULATION ET ESTIMATEURS

Pour le choix de l'estimateur de $\bar{Y}$, nous supposons un modèle de superpopulation Φ, où les valeurs de la population y_k, $k = 1, 2, \ldots, N$, sont réputées former un échantillon aléatoire de telle sorte que:

$$\begin{aligned} E_\Phi(Y_k) &= \mu_k = \beta x_k, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_k^2 = \sigma^2 x_k, \end{aligned} \tag{4}$$

où β et Φ sont inconnues et x_k est la valeur (connue) de la variable auxiliaire X. Il est clair que ce modèle de superpopulation s'applique plus à des variables quantitatives qu'à des variables qualitatives; dans ces conditions, il conviendrait d'utiliser d'autres modèles. Par ailleurs, nous nous limitons ici à l'étude des échantillons aléatoires simples. À la condition que la variance de Y puisse être définie par l'équation en (4), Cassel et coll. (1983) ont montré que l'estimateur:

$$T = \bar{X}\left(\sum\nolimits_r / y_k / q_k\right) / \left(\sum\nolimits_r / x_k / q_k\right),$$

où Σ_r indique la sommation par rapport à l'ensemble r et $\bar{X} = \Sigma_k^N x_k / N$, est approximativement non biaisé (grâce au facteur de correction q_k) même si la première équation en (4) ne définit pas réellement la relation entre X et Y. Cette situation peut être observée, par exemple, lorsque le "véritable" modèle comporte une ordonnée à l'origine ou deux coefficients de régression (voir modèle (5) ci-dessous), etc.

Malheureusement, l'estimateur T ne se prête pas à des applications pratiques puisque q_k est inconnue. Nous devons donc évaluer les propriétés de cet estimateur en remplaçant q_k par son estimation calculée à l'aide de la méthode (1) ou (2).

Nous allons analyser ce genre d'estimateur pour les trois valeurs de h choisies. Nous désignons les estimateurs par TD_i et TD_i^* où $i = 1, 3, 5$ (voir le tableau 1).

Tableau 1

Définition des estimateurs

h	Estimateurs Méthode (1)	Méthode (2)
0.1	TD_1	TD_1^*
0.3	TD_3	TD_3^*
0.5	TD_5	TD_5^*

Par ailleurs, nous considérons les estimateurs suivants dans l'étude de Monte Carlo:

$$TC = \bar{X}\left(\sum_s y_k / \sum_s x_k\right) \quad \text{and} \quad TI = \bar{X}\left(\sum_r y_k / \sum_r x_k\right).$$

TC est l'estimateur pour l'échantillon complet, c'est-à-dire l'estimateur par quotient selon l'hypothèse de l'absence de non-réponse, et TI est le même estimateur pour le sous-ensemble des répondants sans compensation de la non-réponse. Soulignons que TI est aussi un estimateur qui découle d'une méthode d'imputation bien connue (par régression) (Cassel et coll. 1983) et qui est égal à TD lorsque h est au moins égal à l'intervalle des valeurs de x. TI n'est approximativement sans biais que si (4) est vraie. Nous verrons que le biais dépend de la différence entre les conditions définies en (4) et celles de la population à l'étude. Comme dans la simulation de la section suivante, le modèle (4) sera un ''faux'' modèle (c'est-à-dire que les populations étudiées sont définies par des modèles différents de (4)); par ailleurs, la simulation permet de mieux comprendre cette méthode d'imputation très simple et largement répandue.

4. ÉTUDE DE MONTE CARLO

Par une simulation de Monte Carlo, deux populations, POP1 et POP2, ont été produites selon la méthode utilisée par Särndal et Hui (1981). POP1 et POP2 comportent chacune deux strates ($S1$ et $S2$) de 500 unités chacune, et satisfont aux équations suivantes:

$$\begin{aligned} E_\Phi(Y_k) &= \beta_1 x_{k1} + \beta_2 x_{k2}, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_1^2 x_{k1} + \sigma_2^2 x_{k2}, \end{aligned} \tag{5}$$

où $x_{k1} = x_k \partial_k$ et $x_{k2} = x_k(1 - \partial_k)$, $\partial_\kappa = 1$ si $k \in S1$ et $\partial_k = 0$ si $k \in S2$. La différence entre les modèles (4) et (5) sert à simuler l'une des nombreuses erreurs que l'on peut commettre en définissant le modèle de superpopulation. Les caractéristiques numériques de POP1 et POP2 figurent dans le tableau 2.

Voici en bref les étapes de la simulation:

1) un échantillon aléatoire simple s de n ($n = 50$, 100) unités est prélevé dans chaque population;

2) on enregistre ensuite les données de l'échantillon complet puis on crée de la non-réponse à l'aide des deux modèles paramétriques suivants:

$$\text{Modèle A: } q_k = \exp(-\Theta x_k),$$

$$\text{Modèle B: } q_k = \Theta_1^{\partial_k}\,\Theta_2^{1-\partial_k}; \qquad \partial_k = 1\ (0) \text{ si } k \in S1\ (S2),$$

Tableau 2

Propriétés des populations simulées

Population et strates		POP1 Moyenne	E.T.	C.V.	Asym.	POP2 Moyenne	E.T.	C.V.	Asym.
Strate n° 1	x	19.305	12.71	.66	1.30	20.037	14.50	.72	2.25
	y	7.612	5.38	.71	1.62	1.961	2.21	1.13	3.03
Strate n° 2	x	50.325	21.32	.42	.77	49.775	23.28	.47	1.21
	y	30.325	13.38	.44	.72	44.862	21.31	.47	1.04
Total	x	34.815	23.42	.67	.90	34.906	24.44	.70	1.32
	y	18.969	15.26	.80	1.06	23.411	26.25	1.12	1.15

E.T. = Ecart type de la population; Asym. = asymétrie (3-ième moment/(2-ième moment)$^{3/2}$); C.V. = coefficient de variation.

où les paramètres Θ, Θ_1, Θ_2 sont choisis de telle manière que le taux de réponse moyen $\bar{q}$ pour l'ensemble de la population soit de 0.6 ou de 0.7. En pratique, on forme des ensembles de répondants en exécutant un tirage de Bernoulli pour chaque unité $k \in s$ avec une probabilité de résultat favorable q_k (réponse) et une probabilité de résultat défavorable $1-q_k$ (non réponse);

3) on estime la PRI à l'aide des méthodes (1) et (2) et on calcule pour chaque échantillon les valeurs de *TC, TI, TD, TD**;

4) on répète 1000 fois les étapes 1 à 3 et on calcule à la fin le biais, la variance (VAR) et l'erreur quadratique moyenne (EQM) des estimateurs pour chaque combinaison possible des éléments suivants: taille de l'échantillon (50, 100), modèle de réponse (A, B), taux de réponse moyen (0.6, 0.7) et population (POP1, POP2).

Les résultats de la simulation figurent dans les tableaux 3 et 4.

5. RÉSULTATS DE L'ÉTUDE DE MONTE CARLC

L'analyse des tableaux 2 et 3 fait ressortir des points intéressants.

1. Comme prévu, *TC* est approximativement non biaisé dans toutes les simulations.

2. Le biais de *TI* est constamment supérieur à celui de *TD* et de *TD**. Il convient donc, du moins dans les situations qui s'apparentent à celles de la simulation, d'utiliser l'estimateur rajusté de préférence à l'estimateur non-rajusté, qui découle d'une méthode d'imputation par régression.

3. Pour la même valeur de h, le biais de *TD* est toujours moindre que celui de *TD**. Les écarts sont négligeables pour $h=.1$. Lorsque h augmente, *TD** se rapproche de *TI* plus rapidement que *TD*; lorsque $h=.5$, l'écart entre *TD** et *TI* est négligeable sur le plan pratique.

4. La réduction du biais découlant de la substitution de *TD* à *TI* est appréciable dans tous les cas (de 55 à 82% pour le modèle A et de 67 à 92% pour le modèle B). La substitution de *TD** à *TI* entraîne aussi une diminution notable du biais: de 51 à 68% pour le modèle A et de 61 à 84% pour le modèle B.

5. Pour $h=.1$, *TD* et *TD** s'équivalent au point de vue de l'EQM quoique le second soit légèrement plus stable (c'est-à-dire, variance moindre). Pour $h=.3$ et $h=.5$, la moins grande stabilité de *TD* par rapport à *TD** est compensée de façon générale par un biais moindre et cette compensation est telle que *TD* devient préférable à *TD** au point de vue de l'EQM.

6. Les estimateurs corrigés par l'estimation de la PRI ne sont pas très stables mais doivent être utilisés de préférence à *TI* si l'on se fonde sur l'EQM.

7. Comme prévu, le biais dépend directement du taux de non-réponse et de la différence entre le vrai modèle de superpopulation et le modèle hypothétique (c'est-à-dire, le "faux" modèle sur lequel reposent les estimateurs). L'utilisation de deux modèles de réponse (A et B) n'a pas d'incidence notable sur les résultats (voir Giommi (1984) pour les effets de divers modèles).

8. L'élargissement de l'échantillon semble se traduire par une légère diminution du biais pour tous les estimateurs étudiés. *TD* et *TD** sont des exceptions: dans ce cas, la réduction du biais n'est pas attribuable à la modification des conditions expérimentales mais à l'amélioration réelle de l'estimation q_k lorsque n augmente.

En conclusion, nous pouvons affirmer que les deux méthodes proposées dans cet article peuvent servir dans des situations comparables à celles que nous venons d'analyser et qu'il convient d'accorder une certaine préférence à la méthode (1) en raison de son application plus simple. Toutefois, nous n'avons pas résolu le problème de la détermination de la meilleure

Tableau 3

Rendement de divers estimateurs selon le modèle de réponse A

Estimateurs		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Taux de réponse moyen $\bar{q}=.60$					
				POP1					
$n=50$	BIAIS	.015	.861	.349	.420	.669	.380	.620	.765
	VAR	.405	.973	1.115	1.036	1.007	1.041	.995	.989
	EQM	.405	1.714	1.237	1.212	1.455	1.185	1.379	1.574
$n=100$	BIAIS	.007	.805	.164	.323	.610	.227	.544	.686
	VAR	.186	.416	.443	.429	.412	.415	.404	.402
	EQM	.186	1.064	.470	.533	.784	.467	.700	.873
				POP2					
$n=50$	BIAIS	.090	3.125	1.433	1.682	2.544	1.544	2.378	2.887
	VAR	3.952	8.744	9.821	9.823	9.743	9.390	9.233	9.118
	EQM	3.960	18.510	11.874	12.652	16.215	11.774	14.888	17.453
$n=100$	BIAIS	.056	2.959	.749	1.387	2.337	1.004	2.104	2.566
	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632	4.518
	EQM	1.713	12.900	5.076	7.046	10.281	5.246	9.059	11.102
				Taux de réponse moyen $\bar{q}=.70$					
				POP1					
$n=50$	BIAIS	.015	.581	.226	.271	.418	.249	.415	.439
	VAR	.405	.765	.794	.750	.738	.754	.752	.753
	EQM	.405	1.103	.845	.823	.913	.816	.924	.946
$n=100$	BIAIS	.007	.531	.099	.205	.396	.143	.357	.457
	VAR	.186	.328	.323	.307	.327	.313	.327	.336
	EQM	.186	.610	.333	.349	.484	.333	.454	.545
				POP2					
$n=50$	BIAIS	.090	2.130	.813	.939	1.542	.887	1.453	1.822
	VAR	3.952	6.996	7.122	6.827	6.991	6.708	6.753	6.871
	EQM	3.960	11.533	7.783	7.709	9.396	7.495	8.864	10.191
$n=100$	BIAIS	.056	1.966	.473	.953	1.541	.658	1.406	1.732
	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008	3.040
	EQM	1.713	6.937	3.229	3.970	5.402	3.359	4.985	6.040

Tableau 4

Rendement de divers estimateurs selon le modèle de réponse B

Estimateurs		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Taux de réponse moyen $\bar{q}=.60$					
				POP1					
$n=50$	BIAIS	.015	1.086	.290	.383	.716	.323	.688	.992
	VAR	.405	.966	1.208	1.011	.937	1.050	.907	.928
	EQM	.405	2.145	1.29	1.158	1.450	1.154	1.380	1.912
$n=100$	BIAIS	.007	1.079	.120	.349	.732	.196	.668	.902
	VAR	.186	.422	.513	.429	.420	.447	.401	.403
	EQM	.186	1.586	.527	.551	.956	.485	.847	1.217
				POP2					
$n=50$	BIAIS	.090	4.046	1.362	1.757	2.826	1.562	2.749	3.562
	VAR	3.952	10.285	12.519	12.089	12.010	11.605	11.046	10.994
	EQM	3.960	26.655	14.374	15.176	19.996	14.045	18.603	23.682
$n=100$	BIAIS	.056	3.897	.454	1.531	2.707	.853	2.521	3.284
	VAR	1.710	4.151	5.432	5.121	5.103	4.798	4.541	4.381
	EQM	1.713	19.338	5.638	7.465	12.431	5.525	10.896	15.166
				Taux de réponse moyen $\bar{q}=.70$					
				POP1					
$n=50$	BIAIS	.015	.584	.179	.221	.409	.196	.376	.499
	VAR	.405	.751	.826	.425	.716	.769	.723	.743
	EQM	.405	1.092	.858	.474	.883	.807	.864	.992
$n=100$	BIAIS	.007	.536	.046	.173	.365	.087	.317	.436
	VAR	.186	.307.	318	.295	.295	.299	.295	.302
	EQM	.186	.594	.320	.325	.428	.307	.395	.492
				POP2					
$n=50$	BIAIS	.090	2.057	.682	.891	1.477	.804	1.392	1.822
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093	6.270
	EQM	3.960	10.430	7.253	6.959	8.414	6.986	8.031	9.590
$n=100$	BIAIS	.056	1.918	.157	.755	1.311	.374	1.175	1.562
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836	2.923
	EQM	1.713	6.506	2.922	3.454	4.586	2.936	4.217	5.363

valeur de h (ou de h_k dans le cas général). Notre analyse nous a permis de constater que, selon certaines limites, de faibles valeurs de h réduisent le biais mais réduisent en même temps la stabilité de l'estimateur corrigé. Nous avons pu établir que la valeur optimum de h se situait autour de 0.1 pour notre analyse. Des résultats qui sont tirés de la même simulation mais qui ne sont pas reproduits ici indiquent que si l'on réduit davantage la valeur de h, le biais tend à s'accroître. Ce résultat est prévisible puisque en faisant tendre h vers 0, on obtient une série d'estimations $\hat{q}_k$ $(k=1, \ldots, n)$ égales à 1 ou à 0 selon qu'il s'agit de répondants ou de non-répondants.

6. REMERCIEMENTS

Je tiens à exprimer ma reconnaissance au professeur Luigi Biggeri pour le soutien qu'il m'a accordé au cours de la préparation de cet article. Je veux aussi remercier les arbitres pour les commentaires utiles qu'ils ont exprimés sur la version préliminaire.

BIBLIOGRAPHIE

CASSEL, C.M., SÄRNDAL, C.E., et WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. Dans *Incomplete Data in Sample Surveys* (éd. W.G. Madow et I. Olkin), Vol. 3, New York: Academic Press, 143-160.

GIOMMI, A. (1984). On a simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, 42, 185-200.

GIOMMI, A. (1985a). On estimation in nonresponse situations. *Statistica*, 1, 57-63.

GIOMMI, A. (1985b). On the estimation of the individual response probabilities. *Proceedings of the 45th Session of the International Statistical Institute,* Vol. 2 (Contributed Papers), 577-578.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates for the density function. *Annals of Mathematical Statistics*, 27, 832-837.

SARNDAL, C.E., et T.K. HUI (1981). Estimation for nonresponse situations: to what extent must we rely on models? Dans *Current Topics in Survey Sampling,* (éd. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 227-246.