Survey Methodology, December 1987
Vol. 13, No. 2, pp. 127–134
Statistics Canada

Nonparametric Methods for Estimating Individual Response Probabilities

ANDREA GIOMMI[1]

ABSTRACT

This paper deals with the nonresponse problem in the estimation of the mean of a finite population, following an approach closely related to that of Cassel, Särndal and Wretman (1983). Two very simple methods are proposed for estimating the individual response probabilities; these are then used, in connection with a superpopulation model, to construct estimators for the population mean. A first evaluation of the properties of the proposed methods is given by a Monte Carlo experiment. The results shed some light on their effectiveness.

KEY WORDS: Nonresponse; Individual response probability; Nonparametric methods.

1. INTRODUCTION

Dealing with the estimation of finite population mean (or total, etc.) in the presence of nonresponse, Cassel, Särndal and Wretman (1983) introduced a very general estimation method based on the fundamental concept of individual response probability (IRP). The authors proposed estimators which are in part determined by a superpopulation model and in part by a response model, i.e., a model formalizing the response mechanism and by which IRP can be estimated from sample data. The estimation of IRP is the crucial point of their theory. In fact, if the superpopulation model is not correctly chosen, as is often the case, only a correct choice of the response model may guard the estimators from design bias. By a Monte Carlo experiment, Giommi (1985a) showed that a response model supplying a "good approximation" of the "true" response model can restore virtual unbiasedness; but little is known about the extent of a good approximation and in any case the choice of a response model may prove cumbersome besides being arbitrary. A natural way of avoiding these difficulties is to estimate the IRP by nonparametric procedures. In the present paper we propose two very simple methods to estimate IRP when available auxiliary information (which is assumed to be related to the response behaviour) is represented by a single continuous variable. The methods which make use of some tools of the kernel estimation theory may be viewed as an extension of the popular correction technique for nonresponse consisting in reweighting units by adjustment cells.

In this paper some empirical evaluations of these methods are described and the results regarding the bias and efficiency of the related estimators are presented.

2. ESTIMATION OF THE INDIVIDUAL RESPONSE PROBABILITIES

Let us consider a population of N units labelled k $(k=1, 2, \ldots, N)$, and let Y be a variable under study, of which we want to estimate the mean $\bar{Y}=\Sigma_k y_k/N$ from a sample s of n units, the selection being based on a given design $p(s)$. For the estimation, auxiliary information is available, represented by known values x_k $(k=1, \ldots, N)$, of a scalar continuous

[1] Andrea Giommi, Department of Statistics, University of Florence, Via Curtatone, 1, 50123 Florence, Italy.

variable X (the extension of the procedures proposed for the multidimensional case is, in principle, straightforward).

In the sample, Y is observable only in a subset r of n_r respondents and not on the $n - n_r$ nonrespondents. After the selection of the sample, the available information can be represented as follows:

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

where I_k is an indicator random variable such that $E(I_k)=q_k$ and q_k is the IRP.

To estimate q_k, a parametric model is generally assumed (Cassel *et al.* 1983) such that:

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

where Θ is an unknown parameter (or vector of parameters) and $q(\cdot,\cdot)$ is a functional form to be specified. Estimated q_k are then obtained replacing in the above parametric model estimated values $\hat{\Theta}$ of Θ.

In this paper the estimates of q_k $(k \in r)$ are obtained by avoiding any parametric specification of the function $q(\cdot,\cdot)$; nevertheless, maintaining the hypothesis that the IRPs depend on the values x_k. Two procedures (methods (1) and (2)) are proposed.

In the first, q_k $(k \in r)$ is estimated as the response rate (i.e. the proportion of respondents) in a group of units centered on the unit k, corresponding to an appropriate interval of x-values centered at x_k. Assuming that $2h_k$ is the length of such an interval, q_k is estimated by the following ratio:

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

where

$$D(x_k - x_j) = \begin{cases} 1 \text{ if } |x_k - x_j| \leq h_k \\ 0 \text{ otherwise.} \end{cases}$$

It is evident that the estimate $\hat{q}_k$ depends on h_k or h if we adopt - as in this paper - a constant interval; the numerical specification of h is a main problem in applications.

In the second procedure, all the sample units, rather than a group, contribute to the estimation of q_k. By this method the possible limitation due to the classification of responding units in groups is removed. In other words, one might consider overly restrictive the fact that in the estimation of q_k some units contribute with weight 1 and some others with weight 0. With method (2), the estimate is given by:

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

where D^* has to be specified. In this case, each value x_j contributes towards the estimate $\hat{q}_k$ through D^*, an amount inversely related to the difference $|x_k - x_j|$.

In (2), the problem is twofold: i) to specify the functional form D^* and ii) to define the values of its parameters. In this paper we adopt a function D^* of the normal type:

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

in which the standard deviation, indicated by h, plays a role analogous to that of the parameter h in the expression (1). In both (1) and (2), when h increases, $\hat{q}_k$ approaches to the constant value n_r/n. In (1), it reaches n_r/n when h covers the whole range of the x-values.

An empirical study was designed to evaluate the properties of the proposed procedures, using a very wide range of h values. In the present paper we have limited ourselves to reporting results for only three (constant) values of h, equal to 1/10, 3/10 and 5/10 of the range of the x-sample values. Finally, we must observe that both expressions (1) and (2), apart from a normalizing factor, show themselves as the ratio of two probability density kernel estimators (in the approach of Rosenblatt (1956)) over different sets of x-values. Therefore, as suggested by Giommi (1985b), the value of h may be selected considering proposals put forward in that theory.

3. SUPERPOPULATION MODEL AND ESTIMATORS

For the choice of the estimator of $\bar{Y}$, we assume a superpopulation model Φ in which the population values y_k, $k=1, 2, \ldots, N$, are considered to be a random sample such that:

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

where β and Φ unknown and x_k is the known value of the auxiliary variable X. It is apparent that the superpopulation model employed here is mainly applicable to quantitative rather than qualitative variables; other models should be employed in such cases. We further limit ourselves to the consideration of simple random samples. Providing the variance of Y may be specified as in (4), Cassel *et al.* (1983) have shown that the following estimator:

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

where Σ_r indicates the sum over the set r and $\bar{X}=\Sigma_k^N x_k/N$, is approximately unbiased, thanks to the q_k correction, even if the first equation in (4) fails to specify the true relationship between X and Y. This may happen, for example, when the "true" model has an intercept or has two regression coefficients (see (5) below), etc.

Unfortunately, in practice the estimator T cannot be used since q_k is unknown. The problem is, therefore, to evaluate its properties when q_k is replaced by its estimate derived either from method (1) or (2).

We shall examine such estimators, for the three chosen values of h. We denote the estimators by TD_i and TD_i^* where $i=1, 3, 5$ as in Table 1.

Table 1

Definition of Estimators

	Estimators	
h	Method (1)	Method (2)
0.1	TD_1	TD_1^*
0.3	TD_3	TD_3^*
0.5	TD_5	TD_5^*

In addition, also the following estimators are considered in the Monte Carlo study:

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

TC is the full sample estimator, that is, the ratio estimator under the hypothesis of complete response and *TI* is the same estimator based on the set of respondents, on which no q_k-correction is made for nonresponse. Note that *TI* is also an estimator derived from a well known procedure of imputation (by regression) of missing values (Cassel *et al.* 1983) and equals *TD* when h covers the whole range of the x-values. *TI* is approximately unbiased only if (4) is true. The bias, as we shall see, depends on the divergence between the conditions in (4) and those of the population under study. As in the experiment of the next section model (4) will be a "false" model (that is, the study populations are specified by models different from (4)), the simulation also contributes to the knowledge of this very simple and widely used imputation method.

4. THE MONTE CARLO EXPERIMENT

In the Monte Carlo experiment two populations, POP1 and POP2, were generated following the same procedure as that of Särndal and Hui (1981). POP1 and POP2 are both composed of two strata, say *S*1 and *S*2, 500 units each and satisfy the following equations:

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

where $x_{k1} = x_k \partial_k$ and $x_{k2} = x_k(1 - \partial_k)$, with $\partial_k = 1$ if $k \in S1$ and $\partial_k = 0$ if $k \in S2$. The difference between (4) and (5) simulates one of the many errors which one can incur in specifying the superpopulation model. The numerical characteristics of POP1 and POP2 are shown in Table 2.

The simulation procedure can briefly be described in the following steps:

1) A simple random sample s of n ($n = 50$, 100) units is selected from each population.

Table 2

Characteristics of Simulated Populations

Population		POP1				POP2			
and strata		Mean	SD	CV	SK	Mean	SD	CV	SK
Stratum 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
Stratum 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

SD = population standard deviation; SK = skewness (3rd moment / (2nd moment)$^{3/2}$); CV = coefficient of variation.

2) The full sample values are recorded and nonresponse is then generated by each of the two following parametric models:

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

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

where the parameters Θ, Θ_1, Θ_2 are chosen in such a way that the average response rate $\bar{q}$ over the whole population is alternatively 0.6 and 0.7. In practice, sets of respondents are obtained by performing a Bernoulli trial for each unit $k \in s$, with probability q_k for "success" (response) and $1-q_k$ for "failure" (nonresponse).

3) The IRP is estimated by method (1) and (2) and, for each sample, the values of *TC, TI, TD, TD** are calculated.

4) Steps 1 to 3 are repeated 1000 times and at the end we calculate: bias, variance (VAR) and mean squared error (MSE) of the estimators for each sample size (50, 100), response model (A, B), average response rate (0.6, 0.7) and population (POP1, POP2).

The experimental results are reported in Tables 3 and 4.

5. RESULTS OF THE MONTE CARLO EXPERIMENT

Some interesting elements emerge from the examination of Tables 3 and 4.

1. As expected, *TC* is approximately unbiased in all of the experimental trials.

2. In this experiment the bias of *TI* is always larger than that of *TD* and *TD**. Therefore, at least in the situations of the experiment, the adjusted estimator is to be preferred over the non-adjusted one, which corresponds to a procedure of imputation by regression.

3. For the same h value, the bias of *TD* is always smaller than that of *TD**. The differences are negligible for $h=.1$. As h increases, *TD** tends toward *TI* faster than *TD*; for $h=.5$ the differences between *TD** and *TI* are irrelevant for practical purposes.

4. The reduction of the bias we are able to obtain using *TD* instead of *TI* is always significant, varying from 55% to 82% for model A, from 67% to 92% for model B. *TD** also experiences a notable reduction of the bias: from 51% to 68% for model A, from 61% to 84% for model B.

5. *TD* and *TD** are equivalent in terms of MSE for $h=.1$, even though TD_1^* is slightly more stable (i.e. has a lower variance). For $h=.3$ and $h=.5$, the lesser stability of *TD* in comparison with *TD** is generally compensated by the smaller bias, more than enough to make *TD* preferable to *TD** in terms of MSE.

6. The estimators adjusted by the estimated IRP are not very stable but, in terms of MSE, must be preferred to *TI*.

7. As expected, the bias is directly related to the increase of the nonresponse rate and to the divergence between the true superpopulation model and the one assumed (i.e. the false model on which the estimators are based). No relevant differences are revealed due to the response models considered in this paper (see Giommi (1984) for the effect of alternative models).

8. The increase of the sample size seems to reduce the bias slightly for all the estimators considered. TD_1 and TD_1^* are exceptions: in this case, the reduction of the bias cannot be attributed to experimental fluctuations but to the actual improvement of the estimate q_k when n increases.

In the end, we may conclude that, in situations similar to the ones considered in this paper, the two methods suggested can be used, with a certain preference for method (1) given its simpler application. The problem of determination of the best value for h (or h_k, in the general case) remains to be examined. We found that, within certain limits, small values for h reduce the bias but also reduce the stability of the adjusted estimator. We have found that, for our experimental examination, the optimum value of h is in the neighbourhood of 0.1. Results obtained from the same experiment but not reported in this paper indicate that a further reduction of h tends to increase the bias. This is to be expected since making h get closer to 0 results in a collection of estimates $\hat{q}_k$ $(k=1, \ldots, n)$, equal to 1 and 0 respectively for the respondents and nonrespondents.

6. ACKNOWLEDGEMENT

I am indebted to Prof. Luigi Biggeri for his support throughout the course of the study. I also wish to thank the referees for their helpful comments on the first draft of this paper.

Table 3

Performance of Different Estimators under Response Model A

Estimators		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Average response rate $\bar{q}=.60$					
				POP1					
$n=50$	BIAS	.015	.861	.349	.420	.669	.380	.620	.765
	VAR	.405	.973	1.115	1.036	1.007	1.041	.995	.989
	MSE	.405	1.714	1.237	1.212	1.455	1.185	1.379	1.574
$n=100$	BIAS	.007	.805	.164	.323	.610	.227	.544	.686
	VAR	.186	.416	.443	.429	.412	.415	.404	.402
	MSE	.186	1.064	.470	.533	.784	.467	.700	.873
				POP2					
$n=50$	BIAS	.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
	MSE	3.960	18.510	11.874	12.652	16.215	11.774	14.888	17.453
$n=100$	BIAS	.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
	MSE	1.713	12.900	5.076	7.046	10.281	5.246	9.059	11.102
				Average response rate $\bar{q}=.70$					
				POP1					
$n=50$	BIAS	.015	.581	.226	.271	.418	.249	.415	.439
	VAR	.405	.765	.794	.750	.738	.754	.752	.753
	MSE	.405	1.103	.845	.823	.913	.816	.924	.946
$n=100$	BIAS	.007	.531	.099	.205	.396	.143	.357	.457
	VAR	.186	.328	.323	.307	.327	.313	.327	.336
	MSE	.186	.610	.333	.349	.484	.333	.454	.545
				POP2					
$n=50$	BIAS	.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
	MSE	3.960	11.533	7.783	7.709	9.396	7.495	8.864	10.191
$n=100$	BIAS	.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
	MSE	1.713	6.937	3.229	3.970	5.402	3.359	4.985	6.040

Table 4

Performance of Different Estimators under Response Model B

Estimators		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Average response rate $\bar{q}=.60$					
				POP1					
$n=50$	BIAS	.015	1.086	.290	.383	.716	.323	.688	.992
	VAR	.405	.966	1.208	1.011	.937	1.050	.907	.928
	MSE	.405	2.145	1.29	1.158	1.450	1.154	1.380	1.912
$n=100$	BIAS	.007	1.079	.120	.349	.732	.196	.668	.902
	VAR	.186	.422	.513	.429	.420	.447	.401	.403
	MSE	.186	1.586	.527	.551	.956	.485	.847	1.217
				POP2					
$n=50$	BIAS	.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
	MSE	3.960	26.655	14.374	15.176	19.996	14.045	18.603	23.682
$n=100$	BIAS	.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
	MSE	1.713	19.338	5.638	7.465	12.431	5.525	10.896	15.166
				Average response rate $\bar{q}=.70$					
				POP1					
$n=50$	BIAS	.015	.584	.179	.221	.409	.196	.376	.499
	VAR	.405	.751	.826	.425	.716	.769	.723	.743
	MSE	.405	1.092	.858	.474	.883	.807	.864	.992
$n=100$	BIAS	.007	.536	.046	.173	.365	.087	.317	.436
	VAR	.186	.307.	318	.295	.295	.299	.295	.302
	MSE	.186	.594	.320	.325	.428	.307	.395	.492
				POP2					
$n=50$	BIAS	.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
	MSE	3.960	10.430	7.253	6.959	8.414	6.986	8.031	9.590
$n=100$	BIAS	.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
	MSE	1.713	6.506	2.922	3.454	4.586	2.936	4.217	5.363

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* (eds. W.G. Madow and 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.

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