Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 73/2

18. januar 1973

A NOTE ON THE EFFICIENCY OF WEIGHTING SUBCLASS MEANS TO REDUCE THE EFFECTS OF NON-RESPONSE WHEN ANALYZING SURVEY DATA.

BY

IB THOMSEN

Contents

Ikke for offentliggjøring. Dette notat er et arbeidsdokument og kan siteres eller refereres bare etter spesiell tillatelse i hvert enkelt tilfelle. Synspunkter og konklusjoner kan ikke uten videre tas som uttrykk for Statistisk Sentralbyrås oppfatning.

1. Introduction

In practical work the effect of non-response is an important problem, which has received considerable attention in the sampling literature. The problem is often partitioned into two subproblems: What can be done to reduce non-response? And: What can be done to reduce the effects of non-response after the collection of data has been finished?

In this note we shall study one way of reducing the bias due to non-response. This method is widely recommended and used by researchers, and consists of weighting subclass means in the sample to account for different response rates in the subclasses. [3; pp. 558-559], [4; pp. 233-234], [5; pp. 350-351], [8; p. 4].

It is generally accepted that this method reduces the bias when each subclass is homogeneous and there is some difference between the subclass means in the population. We shall spell out the method in some detail to find the conditions under which the bias is reduced by weighting. We shall also find an upper bound for this reduction of the bias, and shall show that the bound can be estimated from the sample. Two examples are given to illustrate the results.

In this note, non-response is taken as the only cause of missing observations in the sample. Another situation in which weighting is often used occurs when the sampling frame does not include all units in the target population. The effect of weighting subclasses for different coverage rates in this case could be studied by the same methods as those used in the present note.

2. Resolution of the bias due to non-response

Assume that our aim is to estimate the population mean of a variable, say $\bar{Y}$. To do this we select a simple random sample of size n from the population. If measurement is obtained from all selected units it is known that the sample mean is a "good" estimate of $\bar{Y}$. Usually measurement is not obtained from all selected units, and we shall therefore study the bias of the sample mean in this case, and compare it with that of a weighted mean defined below.

We shall decompose the bias of the sample mean into two components. The first component arises from the fact that different groups in the population have different response rates. The second component is due to the biasing effect of non-response within each group.

Assume that the population is partitioned into L subclasses before observation of the sample. We think of each subclass as divided into two "strata" as suggested by Cochran [1, p 356]. The first "stratum", often called the response stratum, consists of all units from which measurement will be obtained if the unit happens to fall in the sample. The second "stratum", often called the non-response stratum, consists of all units from which no measurement will be obtained even in this case. We shall denote by N_{i1} the number of units in the response stratum in subclass i, and by N_{i2} the number of units in the non-response stratum. As N_{i1} and N_{i2} are unknown before observation the response and non-response strata are not strata in the usual sense. We shall however stick to this terminology as it is used throughout the sampling literature. For a futher discussion of Cochran's non-response model the reader is referred to [9].

Denote $W_i = (N_{i1}+N_{i2})/N = N_i/N$, where $N = \Sigma N_i$, and N_i is the number of units in subclass i. Let $h_i = N_{i1}/N_i$, and $\bar{h} = \Sigma W_i h_i$. We shall call h_i the population response rate of subclass i, and $\bar{h}$ the population response rate.

Then

$$\bar{Y} = \sum_{i=1}^{L} W_i\left[h_i\bar{Y}_i'+(1-h_i)\ \bar{Y}_i''\right], \text{ where } \bar{Y}_i' \text{ and } \bar{Y}_i'' \text{ are the}$$

population means in subclass i of the response stratum and the non-response stratum respectively. We select a simple random sample of size n from the whole population. When the field work is completed, we have a simple random sample from the L response strata, but the sample size is a stochastic variable, S'. The sample size of subclass i we shall denote S_i'. Throughout this note we shall use the approximations $P(S'>0) = P(S_i'>0) = 1$ $(i = 1,2,\cdots,L)$.

Let $\bar{y}$ denote the sample mean. Then it is known that using the approximation $P(S'>0) = 1$, we have

$$E(\bar{y}) = \sum_{i=1}^{L} N_{i1}\bar{Y}_i'/\sum_{i=1}^{L} N_{i1} = \sum_{i=1}^{L} h_i W_i \bar{Y}_i'/\bar{h},$$

and it follows that

$$(1)\ E(\bar{y}-\bar{Y}) = \frac{1}{\bar{h}}\sum_{i=1}^{L} \bar{Y}_i' W_i(h_i-\bar{h}) + \sum_{i=1}^{L} W_i(1-h_i)(\bar{Y}_i'-\bar{Y}_i'')$$

Denote

$$(1/\bar{h})\Sigma\bar{Y}_i'W_i(h_i-\bar{h}) = B, \text{ and}$$

$$\Sigma W_i(1-h_i)(\bar{Y}_i'-\bar{Y}_i'') = A.$$

Then (1) simplifies to

$$E(\bar{y}-\bar{Y})\ \cancel{E(\bar{y})} = B+A.$$

In (1) the bias of $\bar{y}$ is partitioned into two components. The first component, B, is large when the h_i vary a lot among large subclasses; added variability in h_i through many strata will not greatly change B, because of their smaller contributions through W_i. We shall call B the bias due to different response rates. As one would except B is zero if $\bar{Y}_i'$ is the same for all i.

We shall now introduce another estimate of $\bar{Y}$,

$$\bar{y}_u = \Sigma W_i\bar{y}_i, \text{ where}$$

$\bar{y}_i$ is the sample mean of subclass i. Using the approximation

$P(S_i'>0) = 1$ it is known that $E(\bar{y}_i) = \bar{Y}_i'$, and therefore

$$(2)\ E(\bar{y}_u-\bar{Y}) = \sum_{i=1}^{L} W_i(1-h_i)(\bar{Y}_i'-\bar{Y}_i'') = A.$$

From (1) and (2) it follows that if W_i is known, weighting serves to remove B, and also acts as post-stratification. If W_i is not known, we use the estimate

$$\bar{y}_u^{*} = \sum_{i=1}^{L} (S_i/n)\ \bar{y}_i, \text{ where}$$

S_i is the number of units selected from subclass i. It follows from lemma 1 in the appendix, that $E(\bar{y}_u^{*}) = E(\bar{y}_u) = A$ when we use the approximation $P(S_i>0) = 1$ $(i = 1,2,\cdot\cdot,L)$. It is known, however, that the variance of $\bar{y}_u^{*}$ may be considerable larger than that of $\bar{y}_u$ unless S_i is reasonable large, [3, p 560]. One should therefore avoid fine divisions of the sample if W_i is unknown.

To determine which of the estimates, $\bar{y}$ or $\bar{y}_u^{*}$, has the smaller bias we compare $|B+A|$ with $|A|$, and find the following rules:

(i) If B and A have the same signs, the bias is never increased by weighting.

(ii) If the two components have different signs, the bias is reduced by weighting if and only if $2|A|<|B|$

(iii) If (i) or (iii) is fullfilled, we have $||B+A|-|A||\leq|B|$.

Remarks (i) and (ii) give the conditions under which the bias is reduced by weighting, and remark (iii) gives an upper bound for the reduction.

In the appendix is shown that

$$\hat{B} = [S'/n]^{-1}\Sigma\bar{y}_i(S_i/n)(S_i'/\overset{\rho}{\cancel{n}}-S'/n)$$ ρs_i

is an unbiased estimate of B.

It should be emphasized that (1) and (2) give the bias for a given partitioning of the sample, and that one can shift part of the bias between B and A by choice of subclasses. A question of interest is whether it is possible to find a partioning that maximizes B before weighting. In section 4 below we shall find a solution to this problem in the cases where $h_i = \alpha\bar{Y}_i'+\beta$.

Two examples will illustrate our results.

3. Examples

Example 1: Norwegian Survey of Expenditures 1967, [6], [8].

The sample has been partioned into subclasses, viz., single member households, and all other households. The reason for choosing this partitioning is that the differences between response rates and group means are both fairly large for this grouping, as is seen from table 1.

Table 1:

	Single Member Household	Households with two or more members
Relative size of subclass	0.174	0.826
Response rate	0.571	0.826
Mean expenditure for food, Nr.Kr.	2.436	6.971

The component due to differential response rates is estimated by $\hat{B}$ to be about Nr. kr. 215. The absolute value of the bias of the weighted mean is then smaller than that of the unweighted mean if $A\geq-108$.

In [6] the standard error of the overall mean expenditure is given to be Nr. kr. 46.7. This is assuming a design effect of 1.

Example 2: Norwegian Election Survey 1969, [7; pp. 215-219]

The sample is partitioned into two subclasses: Persons aged 20-24 years, and persons aged 25-75 years.

Table 2:

	Persons aged 20-24	Persons aged 25-75
Relative size of subclass	0.10	0.90
Response rate	0.86	0.90
Rate of Voting	0.81	0.89

The bias due to differential response rates is estimated to be 0.00032.

In [6; 215-219], it is shown that one should except A to be positive. It follows that weighting reduces the absolute value of the bias. The reduction, however, seems to be negligible.

A continuation of examples 1 and 2 will demonstrate how the size of tne first component, B in (1), varies with the number and sizes of subclasses.

Example 1 (cont.)

When the sample is partioned into three subclasses, the result is as given in Table 3.

Table 3:

	Single Member Households	Two Member Households	Households With Three or More Members
Relative size of subclass	0.174	0.261	0.565
Response rate	0.571	0.742	0.865
Mean expenditure on Food	2.437	5.051	7.908

We find that the estimate, $\hat{B}$, of the component B Nr.kr. 296, which is a substantial increase over the estimate obtained for two subclasses. When the sample is partitioned into four subclasses, the result is as given in table 4.

Table 4:

	One Member Households	Two Member Households	Households with Three cr Four Members	Households with More than Four Mambers
Relative size of subclasses	0.174	0.261	0.390	0.175
Response rate	0.571	0.742	0.854	0.897
Mean Expenditure on Food	2.437	5.051	7.137	9.626

The bias due to differential response rates is estimated by $\hat{B}$ to be Nr. kr. 326, which is a small increase over that given by using only three subclasses. If one introduces more subclasses and calculates $\hat{B}$, one will find very little difference from using four subclasses.

When the data from this survey were published about 30 subclasses were used in the weighting procedure. For this partitioning $\hat{B} = 310$. [8; pp. 16]

Example 2 (cont.)

In this example $\hat{B}$ varies little with the number and construction of subclasses.

In table 5 is given a partitioning of the sample, that is different from the one used in example 2 above.

Table 5:

	Persons Aged 20-29	Persons Aged 30-75
Relative size of subclass	0.18	0.82
Response rate	0.82	0.87
Voting Rate	0.91	0.90

In this case we find that the component due to differential response rates is 0.00045 which is slightly larger than with the first partitioning of the sample. Still it seems to be negligible.

4. On the effect of the choice of subclasses when $h_i = \alpha\bar{Y}_i'+\beta$.

We shall study a little futher how B varies with the choice of subclasses before weighting.

In tables 1, 3, and 4 we observe a rather simple relationship between the sample means and the sample response rates. This suggests that it would be useful to study B assuming that $h_i = \alpha\bar{Y}_i'+\beta$ for any partitioning of the sample using "size of household" as the partitioning variable. We shall rewrite B as $B = (1/\bar{h})\Sigma W_i(h_i-\bar{h})(\bar{Y}_i'-\bar{Y}')$, and insert $h_i = \alpha\bar{Y}_i+\beta$. This gives

$$B = \alpha \sum_{i=1}^{L} W_i(\bar{Y}_i'-\bar{Y}')^2/\bar{h}$$

Both α and $\bar{h}$ are independent of the partitioning, and $\Sigma W_i(\bar{Y}_i'-\bar{Y}')^2$ is known from the analysis of variance as the variance between the group means $\bar{Y}_i'$. In sampling literature the variance within the subclasses has received considerable attention. As the sum of the variance within and the variance between the subclasses is equal to the total variance, which is independent of the partitioning, we can state the following results from this literature:

(i) Several good, practical procedures have been developed to choose the subclasses, for given L, such that $\Sigma W_i(\bar{Y}_i'-\bar{Y}')^2$ is maximized. [1; pp. 128-33], [2].

(ii) $\Sigma W_i(\bar{Y}_i'-\bar{Y}')^2$ is an increasing function of L, but a partitioning into more than 4-8 subclasses gives a relatively small increase [1; pp. 133-35], [2].

Also the variance of $\bar{y}_u$ is affected by the choice of subclasses before weighting. A reasonable partitioning is the one that maximizes B and simultanously minimizes the variance of $\bar{y}_u$. When W_i is known $(i = i,2,\cdot\cdot,L)$ the partitioning that minimizes the variance also maximizes B, which is seen in the following way:

When W_i is known $(i = 1,2,\cdot\cdot,L)$ weighting serves as post-stratification, and when $S_i'(i = 1,2,\cdot\cdot,L)$ is of reasonable size, say > 20, the variance of $\bar{y}_u$ is approximately $\frac{1}{S'}\Sigma W_iV_i^2$, where V_i^2 is the population variance in the response stratum in subclass i. We have that the partitioning that minimizes $\Sigma W_iV_i^2$ also maximizes $\Sigma W_i(\bar{Y}_i'-\bar{Y}')^2 = B$, because the sum of the two terms is constant independent of the choice of subclasses.

5. Acknowledgements

Part of this work was done while the author was on leave of absence at Institute for Social Research, Ann Arbor, Michigan. I am grateful to a number of friends and colleages at that institute, and in particular to professor G. Kalton, University of Southhampton, and Mr. S. Wahlstrøm, Swedish

National Central Bureau of Statistics, for helpful suggestions. The note was finished after my return to the Central Bureau of Statistics of Norway, where I received helpful suggestions from Mr. O.J. Skaugen and Dr. J.M. Hoem, who also improved the proof of lemma 1 in the appendix.

6. References

[1] Cochran, W.G. (1963) Sampling Techniques. John Wiley and Sons, Inc.

[2] Gilje, E. and Thomsen, I. (1970) two methods for splitting data into homogeneous groups. Statistical Review, 4. Swedish National Central Bureau of Statistics, Stockholm.

[3] Kish, L. (1965) Survey Sampling. John Wiley and Sons, Inc.

[4] Lansing, B.L. and Morgan, J. (1971) Economic Survey Methods. Institute for Social Reseach, Ann Arbor, Michigan.

[5] Moser, C.A. and Kalton, G. (1971). Survey Methods in Social Investigation. Heinemann Educational Books Limited. London

[6] NOS A 300. Survey of Consumer Expenditure 1967. Central Bureau of Statistics of Norway.

[7] Thomsen, I. (1971) On the Effect of Non Response in the Norwegian Election Survey 1969. Statistical Review, Vol 3, Swedish National Central Bureau of Statistics, Stockholm.

[8] Vestby, Petra. (1970) Metodisk vurdering av forbruksundersøkelsen 1967. Working Paper, Central Bureau of Statistics of Norway, Oslo.

[9] Wahlstrøm, S. (1968) Några synspunkter på svarsbortfallsproblemet vid stickprovsundersökningar av ändliga populasjoner. Statistiska Institutionen, Lunds Universitet, Sweden.

7. Appendix

We shall prove that $E\hat{B}=B$ when we use the approximations $P(S'>0) = P(S_i'>0) = 1$. The following two lemmas are helpful.

Lemma 1

Let $\bar{y}_i$ be defined as the sample mean of subclass i if $S_i'>0$, and zero otherwise. Then we have

$$E\left[\frac{S_i}{n}\bar{y}_i\right] = W_i\bar{Y}_i' - \frac{N_{i2}}{N-N_{i1}}\bar{Y}_i' P(S_i'=0)$$

Proof:

Let $\delta_o(x) = 0$ or 1 as $x = 0$ or $x \neq 0$, and define $S_i'' = S_i - S_i'$. Then

$$E\left[\frac{S_i}{n}\bar{y}_i\right] = \frac{1}{n}E\{E\left[S_i'+S_i''|S_i'\right]E(\bar{y}_i|S_i')\}$$

$$\doteq \frac{1}{n}\bar{Y}_i'E\{\left[S_i'+(n-S_i')\frac{N_{i2}}{N-N_{i1}}\right]\left[1-\delta_o(S_i')\right]\}$$

$$\doteq \bar{Y}_i'\{W_i - \frac{N_{i2}}{N-N_{i1}}P(S_i'=0)\} \quad \square$$

Lemma 2

Define S_i'/S' as the proportion of the sample from subclass i if $S'>0$, and zero otherwise. Then we have

$$E\left[\frac{S_i'}{S'}\bar{y}_i\right] = \bar{Y}_i'h_iW_iP(S'>0)/\bar{h}$$

Proof:

$$E(\frac{S_i'}{S'} v \bar{y}_i) = EE\left[\frac{S_i'}{S'}\bar{y}_i|S'\right]$$

$$= \sum_{j\geq 1}(1/j)E(S_i'\bar{y}_i|S')P(S'=j)$$

$$= \bar{Y}_i'W_ih_iP(S'>0)/\bar{h} \quad \square$$

From lemma 1 and 2, and by the approximations $P(S'>0) = P(S_i'>0) = 1$ we find that

$$E\{\left[\frac{S'}{n}\right]^{-1}\sum_{i=1}^{L}\bar{y}_i\frac{S_i}{n}(\frac{S_i'}{S_i}-\frac{S'}{n})\} = B.$$