Arbeidsnotater

STATISTISK SENTRALBYRÅ

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

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 76/3 19. februar 1976

METODEHEFTE NR. 16

INNHOLD

FORORD

Metodehefter i serien Arbeidsnotater

I tilknytning til mange prosjekter i Statistisk Sentralbyrå utarbeides det mindre, uprentesiøse notater for avklaring av spørsmål av metodisk interesse. Det kan dreie seg om utvalgsteknikk, alternative spørsmålsformuleringer, presentasjonsmetoder, begrepsavklaringer, diskusjon av "funn" i data, systemidéer eller andre temaer. Selv om mange slike notater bare har begrenset interesse i ettertid, vil det blant dem være noen som kunne fortjent å bli alminnelig tilgjengelig. Det kan også være nyttig å ha dem registrert sentralt slik at det blir lettere å få oversikt over det stoffet som foreligger, og letter å referere tilbake til det. Byrået publiserer derfor leilighetsvis et passende antall notater av dette slaget samlet i metodehefter i serien Arbeidsnotater.

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

Forsker Per Sevaldson er redaktør av metodeheftene. Fullmektig Liv Hansen er redaksjonssekretær. Medarbeidere i Byrået som lager stoff som kan være aktuelt, bes sende dette til redaksjonen etter hvert som det blir ferdig. Retningslinjer for utformingen av inserater i metodeheftene finnes på side 46 til side 47 i Metodehefte nr. 9 (ANO IO 73/36).

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

by

Ib Thomsen

Contents

1. INTRODUCTION

This note is a continuation of a previous note, [4], in which we studied one way of reducing the bias due to non-response when estimating the population mean in a finite population. The method consists of weighting subclass means in the sample to account for different response rates in the subclasses.

In this note we shall find the variance of the weighted mean, and give an estimate of this variance. Throughout this note we shall assume that we have a simple random sample from the population. From a practical point of view this is a serious limitation, because one often applies more complex sample designs. Further work should therefore be made to study different weighting procedures and different designs simultaneously.

2. THE VARIANCE OF THE WEIGHTED MEAN

Our aim is to estimate the population mean of a variable, say $\overline{Y}$. To do this we select a simple random sample of size n from the population. After the field work is completed the sample size is reduced because of non-response. We shall assume that the population is partitioned into L subclasses before observation of the sample. As suggested by Cochran [1, p. 356] we think of each subclass as divided into two strata, a response stratum, and a non-response stratum. We shall use the same notations as in [4], where N_{i1} denotes the number of units in the response stratum in subclass i, and N_{i2} denotes the number of units in the non-response stratum in subclass i. Furthermore, $W_i = (N_{i1}+N_{i2})/N = N_i/N$, where $N = \Sigma N_i$, and N_i is the number of units in subclass i in the population. Let $h_i = N_{i1}/N_i$, and $\overline{h} = \Sigma W_i h_i$. We shall call h_i the population response rate of subclass i, and $\overline{h}$ the population response rate.

We now have that

$$\overline{Y} = \sum_{i=1}^{L} W_i(h_i\overline{Y}_i' + (1-h_i)\overline{Y}_i''),$$

where $\overline{Y}_i'$ and $\overline{Y}_i''$ are the population means in subclass i of the response stratum and the non-response stratum respectively.

After 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', and the number of units selected from subclass i we shall denote S_i. Throughout this note we shall use the approximations $P(S'>0) = P(S_i>0) = P(S_i>0) = 1$, $(i=1,2,\dots,L)$.

Let $\bar{y}$ denote the sample mean. In [4] it is shown that

(2.1.) $$E(\bar{y} - \bar{Y}) = B + A, \quad \text{where}$$

$$B = (1/\bar{h}) \sum_{i=1}^{L} \bar{Y}_i' W_i (h_i - \bar{h}), \quad \text{and}$$

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

B arises from the fact that different groups in the population have different response rates, while A is due to the biasing effect of non-response within each group.

In [4] we introduced the weighted sample mean,

$$\bar{y}_u^* = \sum_{i=1}^{L} (S_i/n)\bar{y}_i, \quad \text{where } \bar{y}_i \text{ denotes the sample mean in subclass i,}$$

and showed that

(2.2) $$E(\bar{y}_u^* - \bar{Y}) = A.$$

From (2.1) and (2.2) is seen that weighting serves to remove B from the bias. In this section we shall find the variance of $\bar{y}_u^*$. The following two lemmas are usefull:

<u>Lemma 1</u>

Using the approximations $P(S_i'>0) = P(S_i>0) = 1$, $E(\frac{1}{S_i'} / S_i) = 1/S_i h_i$, and ignoring the finite population correction we have that

$$\text{Var} \left(\frac{S_i}{n}\bar{y}_i\right) \approx \frac{1}{n} \{ W_i V_i^2/h_i + \bar{Y}_i'^2 W_i (1-W_i) \},$$

where

$$V_i^2 = \frac{1}{N_{i1}-1} \sum_{j=1}^{N_{i1}} (Y_{ij} - \bar{Y}_i')^2, \quad \text{i.e. the element variance of Y in the response stratum in subclass i.}$$

<u>Proof</u>

In [2; pp 106-107] is shown that when a simple random sample is selected of a finite population then the subsample of any subpopulation is a simple random sample from the subpopulation. From this fact and by using the result that the variance of a stochastic variable is equal to the expectation of the conditioned variance pluss the variance of the conditioned expectation follows that

$$\text{Var}\,(\bar{y}_i|S_i=s) = V_i^2 \sum_{j\geq 1} \frac{1}{j}P(S_i'=j|S_i=s) + \bar{Y}_i'^2 P(S_i'>0|S_i=s)(1-P(S_i'>0|S_i=0))$$

Using the approximations $P(S_i'>0) \simeq P(S_i>0) \simeq 1$ and $E(\frac{1}{S_i'}|S_i) \simeq 1/S_ih_i$ we find that

(2.3.) $\quad \text{Var}\,(\bar{y}_i|S_i=s) \simeq V_i^2/sh_i$

Similarly we find that

$$E(\bar{y}_i|S_i=s) = \bar{Y}_i'P(S_i'>0|S_i=s),$$

again using the approximations $P(S_i'>0) \simeq 1$ this reduces to

(2.4.) $\quad E(\bar{y}_i|S_i=s) \simeq \bar{Y}_i'.$

We now have that

(2.5) $\quad \text{Var}\,(\frac{S_i}{n}\bar{y}_i) = \frac{1}{n^2}\{E\text{Var}(\bar{y}_iS_i|S_i) + \text{Var}\,E(\bar{y}_iS_i|S_i)\}.$

Inserting (2.3) and (2.4) into (2.5) and using $P(S_i>0) = 1$, we find that

$$\text{Var}\,(\frac{S_i}{n}\bar{y}_i) \simeq \frac{1}{n^2}\{E(\frac{V_i^2S_i^2}{S_ih_i}) + \text{Var}\,(S_i\bar{Y}_i')\} \simeq \frac{1}{n}\{V_i^2W_i/h_i + \bar{Y}_i'^2W_i(1-W_i)\} \quad \square$$

Lemma 2

Applying the same approximations as in lemma 1 we have that

$$\text{Cov}\,(\frac{S_i}{n}\bar{y}_i,\frac{S_j}{n}\bar{y}_j) \simeq -\frac{1}{n}\{\bar{Y}_i'\bar{Y}_j'W_iW_j\}$$

Proof

Using the same approach as in lemma 1 we find that

$$E(\bar{y}_i\bar{y}_j|(S_i=s)\cap(S_j=t)) = \bar{Y}_i'\bar{Y}_j'P((S_i'>0)\cap(S_j'>0)|(S_i=s)\cap(S_j=t))$$

Under the approximations $P(S_i'>0) = P(S_j'>0) = 1$ it follows that

(2.6.) $\quad E(\bar{y}_i\bar{y}_j|(S_i=s)\cap(S_j=t)) \simeq \bar{Y}_i'\bar{Y}_j'$

We have that

$$\text{Cov}\,(\frac{S_i}{n}\bar{y}_i,\frac{S_j}{n}\bar{y}_j) = E[\frac{S_iS_j}{n^2}\bar{y}_i\bar{y}_j] - E[\frac{S_i}{n}\bar{y}_i]\,E[\frac{S_j}{n}\bar{y}_j]$$

$$= E\{\frac{S_iS_j}{n^2}E[\bar{y}_i\bar{y}_j|S_i\ S_j]\} - E[\frac{S_i}{n}\bar{y}_i]\,E[\frac{S_i}{n}\bar{y}_j]$$

Then by (2.6), lemma 1 in [4], and using the approximations from lemma 1 above, we have that

$$\text{Cov}\left(\frac{S_i}{n}\bar{y}_i, \frac{S_j}{n}\bar{y}_j\right) \approx -\frac{1}{n}\{\bar{Y}_i'\bar{Y}_j'W_iW_j\} \square$$

Applying lemma 1 and 2 we find that

$$(2.7) \quad \text{Var}\left(\sum_{i=1}^{L}\frac{S_i}{n}\bar{y}_i\right) \approx \frac{1}{n}\left\{\sum_{i=1}^{L} W_iV_i^2/h_i + \sum_{i=1}^{L} W_i(\bar{Y}_i'-\bar{Y}^*)^2\right\}, \text{ where}$$

$$\bar{Y}^* = \sum_{i=1}^{L} W_i\bar{Y}_i'.$$

Applying the same approximations as above and ignoring the finite population coefficient the variance of the unweighted sample mean is known to be

$$(2.8) \quad \text{Var}(\bar{y}) \approx V^2/n\bar{h}, \text{ where}$$

$$V^2 = \frac{1}{N_1-1}\sum_{j=1}^{N_1}(Y_j-\bar{Y}')^2,\ N_1 = \sum_{i=1}^{L} N_{i1}, \text{ and } \bar{Y}' = \sum_{i=1}^{L} W_ih_i\bar{Y}_i'/\bar{h},$$

i.e., the element variance of Y in the L response strata.

We may decompose (2.8) into the sum of the variances within the response strata and the variance between them. This gives

$$(2.9.) \quad \text{Var}(\bar{y}) = \frac{1}{n}\left\{\sum_{i=1}^{L} W_ih_iV_i^2/\bar{h}^2 + \sum_{i=1}^{L} W_ih_i(\bar{Y}_i'-\bar{Y}')^2/\bar{h}^2\right\}.$$

From (2.7) and (2.9) it is seen that weighting affects both components of the variance, which makes it difficult to compare Var $(\bar{y})$ and Var $(\bar{y}_u^*)$ in general.

Finally in this section we shall consider another estimate of $\bar{Y}$, namely $\bar{y}_u = \Sigma W_i\bar{y}_i$, where W_i is assumed known. Under the same assumptions as in lemma 1, it is known that [1]

$$(2.10) \quad \text{Var}(\bar{y}_u) \approx \frac{1}{n}\sum_{i=1}^{L} W_iV_i^2/h_i .$$

From (2.7) and (2.10) it follows that

$$\text{Var}(y_u^*) - \text{Var}(\bar{y}_u) \approx \frac{1}{n}\sum_{i=1}^{L} W_i(\bar{Y}_i'-\bar{Y}^*)^2.$$

The reduction of the variance due to weighting is substantially larger when W_i is known than when W_i is unknown. In [4] is shown that $\bar{y}_u$ and $\bar{y}_u^*$ have the same bias.

3. ESTIMATION OF THE VARIANCE OF THE WEIGHTED MEAN

Before the researcher chooses to apply a weighted mean to reduce the effects of non-response, it may be of interest to appraise the effect of weighting on the bias and on the variance. In [4] the effect on the bias is studied, and an estimate of the maximum reduction of the bias is given. In this section we shall find an approximately unbiased estimate of Var $(\bar{y}_u^*)$, var $(\bar{y}_u^*)$. The estimate is found by replacing the population parameters in (2.7) with their corresponding sample values. We shall first prove two lemmas.

<u>Lemma 3</u>

Applying the same approximations as in lemma 1, we have that

$$E\{\sum_{i=1}^{L} \frac{S_i}{n}(\bar{y}_i-\bar{y}_u^*)^2\} \approx \sum_{i=1}^{L} \frac{(1-W_i)V_i^2}{nh_i} + \sum_{i=1}^{L} W_i(1-\frac{1}{n})(\bar{Y}_i'-\bar{Y}^*)^2.$$

<u>Proof</u>

$$(3.1)\quad E\{\sum_{i=1}^{L} \frac{S_i}{n}(\bar{y}_i-\bar{y}_\mu^*)^2\} = \sum_{i=1}^{L} E(\frac{S_i}{n}\bar{y}_i^2) - E(\sum_{i=1}^{L} \frac{S_i}{n}\bar{y}_i)^2.$$

The first term is found to be

$$\sum_{i=1}^{L} E(\frac{S_i}{n}\bar{y}_i^2) = \sum_{i=1}^{L} E\{\frac{S_i}{n}E(\bar{y}_i^2 | S_i)\} =$$

$$\sum_{i=1}^{L} E\{\frac{S_i}{n}(\mathrm{var}(\bar{y}_i | S_i) + E(\bar{y}_i | S_i)^2)\}.$$

Inserting (2.3) and (2.4) we find that this is equal to

$$(3.2)\quad \sum_{i=1}^{L} E\{\frac{S_i}{n}(\frac{V_i^2}{S_i h_i} + \bar{Y}_i'^2)\} = \sum_{i=1}^{L} \frac{V_i^2}{nh_i} + W_i\bar{Y}_i'^2.$$

From (2.7) and lemma 1 in [4] the second term of (3.1) is found to be

$$E(\sum_{i=1}^{L} \frac{S_i}{n}\bar{y}_i)^2 = \mathrm{var}\,(\bar{y}_u^*) - (E(\sum_{i=1}^{L} \frac{S_i}{n}\bar{y}_i))^2 =$$

$$(3.3)\quad \frac{1}{n}\{\sum_{i=1}^{L} W_iV_i^2/h_i + \sum_{i=1}^{L} W_i(\bar{Y}_i'-\bar{Y}^*)^2\} - (\sum_{i=1}^{L} W_iY_i')^2.$$

Inserting (3.3) and (3.2) into (3.1) we find that

$$E\{\sum_{i=1}^{L} \frac{S_i}{n}(\bar{y}_i-\bar{y}_u^*)^2\} \approx \sum_{i=1}^{L} \frac{(1-W_i)V_i^2}{nh_i} + \sum_{i=1}^{L} W_i(1-\frac{1}{n})(\bar{Y}_i'-\bar{Y}^*)^2 \quad \square$$

Lemma 4

Under the same approximations as in lemma 1, we have that

$$E\{\sum_{i=1}^{L} \frac{S_i}{n}\frac{S_i}{S_i'} v_i^2\} \approx \sum_{i=1}^{L} W_i V_i^2/h_i, \quad \text{where } v_i^2 = \frac{1}{S_i'-1}\sum_{j=1}^{S_i'} (y_{ij}-\bar{y}_i)^2,$$

i.e., the element variance in subclass i in the sample.

Proof

Again applying the fact that the sample from the response stratum in subclass i is a simple random sample of size S_i' we find that

$$E\{\sum_{i=1}^{L} \frac{S_i}{n}\frac{S_i}{S_i'}\{\frac{1}{S_i'}\sum_{j=1}^{S_i'} (y_{ij}-\bar{y}_i)^2\}\} = E\{\sum_{i=1}^{L}\frac{S_i^2}{n} E\{\frac{1}{S_i'} v_i^2 | S_i\}\}$$

$$\approx \sum_{i=1}^{L} E(\frac{S_i^2}{n}\frac{1}{h_i S_i} V_i^2) = \sum_{i=1}^{L} W_i V_i^2/h_i \quad \square$$

From lemma 3 and 4 we find that

$$E\{\frac{1}{n}\sum_{i=1}^{L}\frac{S_i}{n}\frac{S_i}{S_i'} v_i^2 + \frac{1}{n}\sum_{i=1}^{L}\frac{S_i}{n}(\bar{y}_i-\bar{y}^*)^2\} \approx$$

$$\frac{1}{n}\{\sum_{i=1}^{L} W_i V_i^2/h_i + \sum_{i=1}^{L}\frac{(1-W_i)}{nh_i} V_i^2 + \sum_{i=1}^{L} W_i(1-\frac{1}{n})(\bar{Y}_i'-\bar{Y}^*)^2\}.$$

For large n this is approximately equal to

$$\Sigma \frac{W_i V_i^2}{h_i} + \Sigma W_i(\bar{Y}_i'-\bar{Y}^*)^2, \text{ which is Var } (y_u^*).$$

4. EXAMPLES

In this section we shall give some examples in which the data are taken from actual surveys. It should be noted that complex designs have been applied in the surveys to which we refer, but that we treat the data as if it was collected as a simple random sample.

Example 1:

The following data are taken from [3, pp. 182-83].

The sample has been partioned into two subclasses, viz., men, and women. The reason for choosing this partitioning is that the difference between the group means is fairly large for this grouping, as is seen from table 1.

Table 1

	Men	Women
Relative size of subclass (W_i) ...	0,47	0,53
Response rate ($\hat{h}_i$)	0.83	0,90
Persentage reading daily tabloid ($\bar{y}_i$)	0,80	0.10

In this case we find that

$$\bar{y} = 0{,}41, \text{ var } (\bar{y}) = 0{,}2790/n,$$
$$\bar{y}_u = 0{,}42, \text{ var } (\bar{y}_u) = 0{,}1436/n,$$
$$\bar{y}_u^* = 0{,}43, \text{ var } (y_u^*) = 0.2657/n.$$

Considering that we are estimating proportions the difference between the subclass means is large in this example. In cases where the subclass means do not vary as much as in this example one will typically find less differences between $\bar{y}$ and $\bar{y}_u$, and their variances (see example 2 in [4]).

E x a m p l e 2:

Norwegian Survey of Expeditures 1967 [4]

The sample is partitioned into two subclasses as shown in table 2.

Table 2

	Single member household	Household with two or more members
Relative size of subclass in the sample	0,174	0,826
Response rate ($\hat{h}_i$)	0,571	0,826
Mean expenditure for food, Nr.kr. ($\bar{y}_i$).	2,436	6,971
Variance within the subclass (V_i)	$12{,}335 \cdot 12^2$	$64{,}138 \cdot 12^2$

We find that

$$\bar{y} = 6{,}182, \text{ var } (\bar{y}) = 93{,}843 \cdot \frac{12^2}{n}, \text{ and}$$
$$\bar{y}_u = 5{,}967, \text{ var } (\bar{y}_u) = 88{,}426 \cdot \frac{12^2}{n}.$$

To demonstrate how var ($\bar{y}_u$) varies with the number and sizes of subclasses, we shall divide the sample into three subclasses as given in table 3.

Table 3

	Single member households	Two member households	Households with two or more members
Relative size of subclass in the sample ...	0,174	0,261	0,565
Response rate $\hat{h}_i$	0,571	0,742	0.865
Mean expenditure on food ($\bar{y}_i$)	2,437	5,051	7,908
Variance within the subclass	12,335 12^2	29,159 12^2	63,494 $\cdot 12^2$

In this case we find that

$\bar{y}\mu$ = 5,886 and var ($\bar{y}_u$) = 86,217 $12^2/n$.

REFERENCES

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

[2] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953) Sampling survey Methods and theory Vol. II, John Wiley and Sons, New York.

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

[4] Thomsen, I (1973) A Note on the Efficiency of Weighting Subclass Means to Reduce the effects of non-reponse when Analyzing Survey Data. Statistical Review, Vol. 4, Swedish National Central Bureau of Statistics, Stockholm.