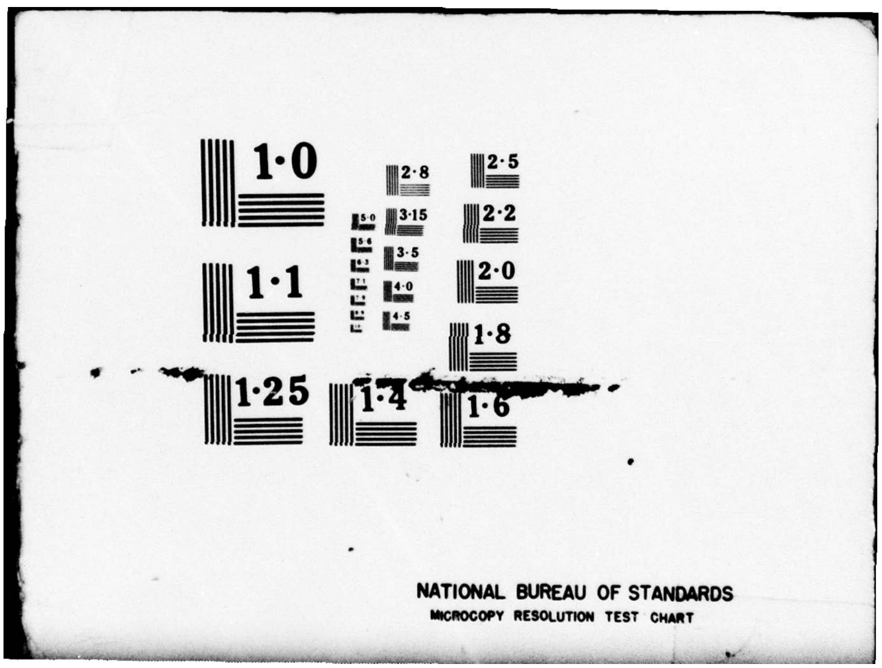
AD-A053 915 TEXAS A AND M UNIV COLLEGE STATION F/G 12/1
THE ESTIMATION OF NON-SAMPLING VARIANCE COMPONENTS IN SAMPLE SU--ETC(U)
MAR 78 H O HARTLEY, J N RAO DAAG29-77-G-0086
UNCLASSIFIED TR-2 ARO-14209.2-M NL

| OF |
ADA
053915

END
DATE
FILMED
6-78
DDC


1·0
2·8
2·5
3·15
2·2
3·5
2·0
1·1
4·0
4·5
1·8
1·25
1·4
1·6
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

ARO 14209.2-M

(12)

ARO-D PROJECT DAAG29-77-G-0086

Technical Report No. 2

THE ESTIMATION OF NON-SAMPLING VARIANCE COMPONENTS IN SAMPLE SURVEYS

by

H. O. Hartley and J. N. K. Rao

March 1978

THE ESTIMATION OF NON-SAMPLING VARIANCE COMPONENTS IN SAMPLE SURVEYS

H. O. Hartley and J. N. K. Rao

1. Introduction

The importance of non-sampling, or measurement errors has long been recognized (for the numerous references see e.g., the comprehensive papers by Hansen, Hurvitz and Bershad (1961) and Bailar and Dalenius (1970)). Briefly the various models suggested for such errors assume that a survey record (recorded content item) differs from its "true value" by a systematic bias, B, and various additive error contributions associated with various sources of errors such as, interviewers, coders, etc. The important feature of these models is that the errors made by a specified error source (say a particular interviewer) are usually 'correlated'. These correlated errors contribute additive components to the total mean square error of a survey estimate which do not decrease inversely proportional to the overall sample size but only inversely proportional to the number of interviewers, coders, etc. Consequently, the application of standard text book formulas for the estimation of the variances of survey estimates may lead to serious underestimates of the real variability which should incorporate the non-sampling errors.

Attempts have, therefore, been made to estimate the components due to non-sampling errors. The early work in this area has concentrated on surveys specifically designed to incorporate features facilitating the estimation of non-sampling components such as reinterviews and/or interpenetrating samples (see e.g. Sukhatme and Seth (1952)). However, the more recent literature (see e.g. Cochran (1968), Fellegi (1969), Nisselson and Bailar (1976), Battese, Fuller and Hickman (1976)) has also treated surveys in which such features are either lacking or limited, but these results are restricted to simple surveys permitting the use of analysis of variance techniques.

In this paper we provide a general methodology applicable to essentially any multistage survey in which the last stage units are drawn with equal probabilities. Specifically our formulas for the estimated variances of target parameter estimates will include all finite population corrections except those in the last stage which are usually negligible. We utilize recent results in the estimation of components of variance in mixed linear models to achieve these results and are able to address the problem of estimability of variance components.

2. The assumptions made.

In this paper we confine ourselves to what may be regarded as a special case of a more general model which we hope to cover in a subsequent paper. Here we assume that:

(2.1) The survey has a stratified multistage design in which the last stage units are drawn with equal probabilities while any equal - or unequal - probability design may be specified for the remaining stages.

(2.2) Error sources (such as interviewers, or coders, etc.) contribute additive errors to the so called "content items" associated with the last stage units.

(2.3) All "correlations" between the errors contributed by a particular (say the i^{th}) error source are generated through an "additive model". That is the errors have the structure $b_i + \delta b_s$ where b_i is an error contribution from the i^{th} source <u>common to all units affected by the i^{th} source</u> (all units interviewed by the i^{th} interviewer) while δb_s, sometimes referred to as an "elementary non-sampling error", varies randomly from unit to unit (s).

(2.4) The present paper is confined to the case where there is no systematic bias from any of the error sources.

We should state here that the above assumptions (2.2) and (2.3) are quite customary in the literature on non-sampling errors (see e.g., Sukhatme & Seth (1952) and Bailar & Dalenius (1970)).

Although a bias term is usually included in the formulas occurring in the literature it can only be evaluated in special cases. For example, it may be estimated from "special record checks." We do not discuss biasses in this paper.

3. The model formulation.

To fix the ideas expressed in 2 we confine ourselves to two types of error sources without loss of generality described as "interviewers" and "coders". However, generalizations to more than two types of error sources do not afford any difficulties. Moreover, to simplify the notation, we introduce the two index label (p, s) where the index s labels the s^{th} elementary unit (briefly referred to as "secondary") and the index p (briefly called the primary index) is a composite label indexing the last but one stage unit within the next higher stage unit .... within the primary unit within a stratum. Thus, for example, in a three-stage stratified design s will denote the tertiary unit and p will be a composite index for a "secondary within a primary within a stratum."

We may now write the model in the form

$$y_{ps} = \eta_{ps} + b_i + c_c + \delta b_{ps} + \delta c_{ps} \qquad (1)$$

where

y_{ps} = content item recorded for elementary unit labeled (p, s),

η_{ps} = true content item for elementary unit labeled (p, s),

b_i = error variable contributed by i^{th} interviewer common to all (p, s) interviewed by i^{th} interviewer,

c_c = error variable contributed c^{th} coder common to all (p, s) coded by c^{th} coder,

δb_{ps} = elementary interviewer error afflicting the content item of unit (p, s),

δc_{ps} = elementary coder error afflicting the content item of unit (p, s).

We assume that the b_i and c_c are respectively random samples from infinite populations of interviewer and coder errors with

$$E(b_i) = 0 \qquad \text{and } Var\ (b_i) = \sigma_b^2$$
$$E(c_c) = 0 \qquad \text{and } Var\ (c_c) = \sigma_c^2 \ . \tag{2}$$

The assumptions $E(b_i) = E(c_c) = 0$ postulate the absence of systematic interviewer and coder biases.

Likewise we assume that

$$E(\delta b_{ps}) = 0 \qquad Var\ (\delta b_{ps}) = \sigma_{\delta b}^2$$
$$E(\delta c_{ps}) = 0 \qquad Var\ (\delta c_{ps}) = \sigma_{\delta c}^2 \ . \tag{3}$$

The common interviewer errors b_i and common coder errors c_c are assumed to be independent from one another and independent of the true content items η_{ps} and the elementary errors δb_i, δc_c. However η_{ps} and δb_i δc_c are not assumed to be independent.

It should also be noted that $\sigma^2_{\delta b}$ and $\sigma^2_{\delta c}$ apply respectively to the elementary errors of <u>all</u> interviewers and <u>all</u> coders. This means that we do not, in this paper, allow for the possibility of heterogeneity of the interviewers and/or coders elementary error variances.

We may rewrite the model (1) in the form

$$y_{ps} = \bar{\eta}_{p.} + b_i + c_c + e_{ps}$$

where

$$e_{ps} = (\eta_{ps} - \bar{\eta}_{p.}) + \delta b_{ps} + \delta c_{ps} \qquad (4)$$

and where

$$\bar{\eta}_{p.} = \frac{1}{M_p} \sum_{s=1}^{M_p} \eta_{ps} \qquad (5)$$

is the mean of the η_{ps} over the M_p elementary units in the p^{th} primary.

The essential concept in our approach is that we shall only estimate the $\sigma^2_{e,p} = Var(e_{ps})$ for each primary, p, but do <u>not</u> obtain separate estimates for the $Var(\eta_{ps} - \bar{\eta}_{p.})$ (the variances of the true sampling errors) or the Var δb_{ps}, Var δc_{ps} (that is, the elementary non-sampling variances). To justify this strategy we shall show that the variance of the estimates of population totals and other target parameters in our finite population likewise only involves the σ^2_{ep} and <u>not</u> its separate components.

4. The complete specification of the survey design.

As stated in (2.1) above we permit any specification of a stratified multistage design in which the last stage units (the 'secondary units' indexed (s)) are drawn with equal probability. This means

(4.1) that the design specifies in advance for any set (p) of sampled primaries the number, m_p, of secondary units to be drawn with equal probability from the M_p units in the population.

Moreover we shall assume for any set of sampled (p)

(4.2) that the design specifies the number of interviewers (I) and number of coders (C) which will be labeled i = 1, ..., I; c = 1, ..., C, and

(4.3) that the design specifies the "work-load assignment" i.e., that it specifies in advance the number of secondary units to be interviewed by interviewer i in each primary p and likewise the numbers to be coded by coder c in each primary p.

Specification (4.1) is quite customary. Specifications (4.2) and (4.3) are only conceptual since in actual practice I and C and the work-load assignment will often not be decided on until after the primary sample (p) has been drawn.

In what follows we shall further assume for the sake of simplifying the argument that the last stage (secondary) sampling fractions m_p/M_p are all negligibly small so that the sampled $n_{ps} - \bar{n}_{p.}$ can be regarded as a random sample of m_p from an ∞ population with mean 0. The inclusion of the finite population corrections will be discussed in the second paper. We do not assume however, that the elementary interviewer and/or coder errors δb_{ps} and/or δc_{ps} are necessarily independent of the sampling errors $n_{ps} - \bar{n}_{p.}$, since we shall, in the next section, estimate the variances of the composite error e_{ps} directly.

5. The conditional estimation of variance components.

Consider a given sample of primaries (p) drawn in accordance with the design. Then under the assumptions made in 4. and conditionally on (p) the model (4) will represent a "mixed analysis of variance model" where the b_i, c_c and e_{ps} are random variables with "variance components" σ_b^2 (for interviewers), σ_c^2 (for coders) and σ_{ep}^2 (for "elementary errors") in primary p. The model also involves "fixed constants" $\bar{n}_p$.

In order to relate the model to the notation customary in variance component estimation methodology we write it in the form

$$y = X\alpha + \sum_{j=1}^{c} U_j b_j \tag{6}$$

where

y is the vector of recorded y_{ps} with number of elements

$$M = \sum_{p=1}^{n} m_{ps},$$

α is the n-vector with elements $\bar{n}_p$, the population means for the sampled primaries, (7)

X is an associated $M \times n$ design matrix with 1's in the column p if y_{ps} is in primary p,

b_1 = I-vector of interviewer variables b_i,

b_2 = C-vector of coder variables c_c,

b_3 to b_{n+3} = m_p-vectors of e_{ps} for $p = 1, \ldots, n$,

U_j = associated design matrices with 1's in those columns that correspond to the interviewer, coder or primary of the unit labeled (p, s).

There is a considerable literature on "component of variance estimation" in the unbalanced mixed ANOVA Model (for a comprehensive bibliography, see e.g., Searle (1971)). For a computationally simple method of computing estimates of the σ_j^2 we refer to the "synthesis based method" by Hartley, Rao and LaMotte (1977) which is a Minque estimate using a particular norm and which enjoys additional optimality properties and provides conditions for estimability as follows: -

Introducing the matrices $V_j = U_j - XX'U_j$, Hartley, Rao and LaMotte show that the σ_j^2 are estimable if the V_jV_j' are not linearly dependent and this condition is usually satisfied by survey designs. In any case the condition can be tested on the computer in advance of the field work and if the V_jV_j' are found to be dependent this can usually be remedied by alteration in the work load assignment to interviewers and/or coders.

Because of the assumptions made in Section 3, the estimates of the variance components σ_j^2 that is, σ_b^2, σ_c^2, and σ_{ep}^2 computed from the sample of y_{ps} conditional on a given set of primaries (p) are universally unbiassed estimates of these variance components and will be available for estimates of variances of target estimators computed directly from the survey data.

6. Linear estimates of target parameters and their variances.

The majority of estimators of target parameters (including the population total and means) which are computed from the survey sample data are linear functions of the y_{ps}. Since sampling within primaries is with equal probabilities we confine ourselves to estimators of the form

$$\hat{Y} = c'(p)\bar{y} \tag{8}$$

where $\bar{y}$ is the n-vector of primary-sample means $\bar{y}_p$ $p = 1, \ldots, n$ and the n elements of the coefficient vector c(p) depend on the set of selected primaries (p). We illustrate this estimator by an example. Suppose we have a two stage design with equal probability sampling without replacement at both stages and with the target parameter specified as the population total, then

$$c'(p)\bar{y} = \frac{N}{n} \sum_p M_p \bar{y}_p \quad \text{so that}$$

$$c(p) = \frac{N}{n} M_p \tag{9}$$

where

$$\left.\begin{matrix} N \\ n \end{matrix}\right\} = \text{number of primaries in} \left\{\begin{matrix} \text{population} \\ \text{sample} \end{matrix}\right.$$

and

$$\left.\begin{matrix} M_p \\ m_p \end{matrix}\right\} = \text{number of secondaries in} \left\{\begin{matrix} \text{population} \\ \text{sample} \end{matrix}\right. \tag{10}$$

Clearly

$$Ec'(p)\bar{y} = \underset{p}{E} \; E\Big|_p \; c'(p)\bar{y} = \underset{p}{E} \; c'(p)\bar{n} \tag{11}$$

where $\bar{\eta}$ is the n-vector of true primary means, and $E|_p$ is the conditional expectation given a set (p) of sampled primaries and $\underset{p}{E}$ the expectation over the finite population survey design of primaries. If so called "unbiased extimators" $c'(p)\bar{\eta}$ of target parameters have been chosen then clearly from (11) $c'(p)\bar{y}$ will be unbiassed.

We now turn to the variance formulas. We have

$$\mathrm{Var}\ c'(p)\bar{y} = \underset{p}{\mathrm{Var}}\ E|_p\ c'(p)\bar{y} + \underset{p}{E}\ \mathrm{Var}|_p\ c'(p)\bar{y} \qquad (12)$$

where $\mathrm{Var}|_p$ is a conditional variance given a set of primaries (p) while $\underset{p}{\mathrm{Var}}$ is the variance for the finite population survey design of primary units.

Turning first to the second term in (12) (the "within primary component") we have

$$\mathrm{Var}|_p = c'(p)Sc(p) \qquad (13)$$

where the n × n matrix S is the conditional covariance matrix of the $\bar{y}_p$ whose p, π element is given by

$$S_{p,\pi} = \sum_j \sigma_j^2 \sum_t \nu(p, t; j)\ \nu(\pi, t; j)\ (m_p m_\pi)^{-1} \qquad (14)$$

Here $\nu(p, t; j)$ is the number of 1 elements in the t^{th} column of U_j which are in rows corresponding to units (p, s) for the argument primary p of $\nu(p, t; j)$. The $\nu(p, t; j)$ are parameters which are predetermined through the design and work allocation for any primary sample (p) because of (4.1) to (4.3). An unbiassed estimate of $\underset{p}{E}\ \mathrm{Var}\big|_p\ c'(p)\bar{y}$ is therefore given by

$$\mathrm{var}_W = \sum_j \hat{\sigma}_j^2 \sum_t \{ \sum_p \nu(p,\ t;\ j)\ \frac{c(p)}{m_p} \}^2 \tag{15}$$

where the $\hat{\sigma}_j^2$ are the component of variance estimates whose computation is described in Section 5.

Turning next to the "between primary variance component" in (12) we have

$$\underset{p}{\mathrm{Var}}\ E\big|_p\ c'(p)\bar{y} = \underset{p}{\mathrm{Var}}\ c'(p)\bar{\eta} \tag{16}$$

Now finite population sampling theory for the primary units (p) regarded as units will provide a "variance formula" for the "estimator" $c'(p)\bar{\eta}$ in the form

$$\underset{p}{\mathrm{Var}}\ c'(p)\bar{\eta} = V(\dot{\eta}) \tag{17}$$

(where $\dot{\eta}$ is the N-vector of primary means in the finite population of N primary means) and also provide an unbiassed estimate of $V(\dot{\eta})$ in the form

$$v(c'(p)\bar{\eta}) = \bar{\eta}'A\bar{\eta}$$

with (18)

$$\underset{p}{E}\ \bar{\eta}'A\bar{\eta} = V(\dot{\eta})$$

In the above example of two stage equal probability sampling without replacement we have

$$V(\dot{n}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \sum_{p=1}^{N} \left(M_p\bar{n}_p - \overline{\overline{M_p\bar{n}_p}}\right)^2/(N - 1) \tag{19}$$

and

$$v(c'(p)\bar{n}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \sum_{p=1}^{n} \left(M_p\bar{n}_p - \overline{M_p\bar{n}_p}\right)^2/(n - 1) \tag{20}$$

where $\overline{M_p\bar{n}_p}$ and $\overline{\overline{M_p\bar{n}_p}}$ are respectively the sample and population means of the $M_p\bar{n}_p$. Equations (19) and (20) are the well-known formulas for the variance and variance estimate of the $N(\overline{M_p\bar{n}_p})$ in a simple random sample of n units with characteristics $M_p\bar{n}_p$ drawn from a finite population of N units.

Returning to the general case (12) an unbiassed estimate of Var_B of the between primary component of $Var(c'(p)\bar{y})$ can be computed from the y_{ps} through

$$var_B = \bar{y}'A\bar{y} - \text{tr } AS \tag{21}$$

The above formula (21) cannot, of course, claim any particular properties other than unbiassedness. However numerical experience indicates that the second term will usually be negligible compared with the first.

7. Summary.

To summarize we have provided a method of estimating the overall variance of a linear estimator of the form $c'(p)\bar{y}$ which includes the non-sampling errors for any stratified multistage design in which the last stage is an equal probability selection procedure. The estimate of the variance contains two components, namely a component var_W given by (15) representing variation of the last stage units within the last but one stage units plus elementary measurement errors. The second component var_B given by (21) represents a composite of components due to variation of the higher stage units each within the units of next higher stage. The "within component" var_W involves estimated variance components $\hat{\sigma}_j^2$ computed by simple mixed model ANOVA techniques. The "between component" also involves these $\hat{\sigma}_j^2$ in the correction term $-trAS$ with $S = (S_{p,\pi})$ given by (14). However its leading term $\bar{y}'A\bar{y}$ is a quadratic form in the last but one stage sample means $\bar{y}_p$ directly provided by standard estimation of variance formulas in finite population sampling and including all finite population corrections for the higher stages.

Simple numerical examples will be provided in our next paper.

8. Acknowledgement.

One of us (H.O.H.) wishes to gratefully acknowledge support from the Army Research Office.

References

Bailar, B. A. and Dalenius, T. (1970). "Estimating the response variance components of the U. S. Bureau of the Census survey model". Sankhya Series B, 341-360.

Battese, G. E., Fuller, W. A., Hickman, R.D. (1976). "Estimation of response variance from interview re-interview surveys". Journal Indian Society of Agricultural Statistics, 28, 1-14.

Cochran, W. G. (1968). "Errors of measurements in statistics". Technometrics 10, 637-666.

Fellegi, I. P. (1974). "An improved method of estimating the correlated response variance". Journal of American Statistical Assn., 1969, 496-501

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961). "Measurement errors in censuses and surveys". Bull. International Stat. Inst., 38, 359-374.

Hartley, H. O., Rao, J.N.K., LaMotte, L. (1977). "A simple 'synthesis'-based method of variance component estimation", paper presented at Regional meeting of ENAR, Chapel Hill, April, 1977.

Nisselson, H. and Bailar, B. A. (1976). "Measurement analysis and reporting of non-sampling errors in surveys". Proceedings of the International Biometric Conference (Boston, 1976) 301-321.

Searle, S. R. (1971), Linear Models, John Wiley & Sons, Inc., New York

Sukhatme, P. V. and Seth, G. R. (1952) "Non-sampling errors in surveys". Journal Indian Society of Agricultural Statistics, 4, 5-41.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO 14209.2-M

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): The Estimation of Non-Sampling Variance Components in Sample Surveys.

5. TYPE OF REPORT & PERIOD COVERED: Technical Report.

6. PERFORMING ORG. REPORT NUMBER: TR-2

7. AUTHOR(s):

H. O. Hartley

J. N. K. Rao

8. CONTRACT OR GRANT NUMBER(s): DAAG29-77-G-0086

9. PERFORMING ORGANIZATION NAME AND ADDRESS:

Texas A&M University

College Station, Texas 77843

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:

U. S. Army Research Office

P. O. Box 12211

Research Triangle Park, NC 27709

12. REPORT DATE: Mar[illegible]78

13. NUMBER OF PAGES: 14

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper we provide a general methodology applicable to essentially any multistage survey in which the last stage units are drawn with equal probabilities. Specifically our formulas for the estimated variances of target parameter estimates will include all finite population corrections except those in the last stage which are usually negligible. ~~We utilize~~ recent results are utilized in the estimation of components of variance in mixed linear models to achieve these results and are able to address the problem of estimability of variance components.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified