Arbeidsnotater

STATISTISK SENTRALBYRÅ

OSLO: Postboks 8131 Dep, Oslo 1
Tlf. (02) *41 38 20

KONGSVINGER: Postboks 510, Stasjonssida, 2201 Kongsvinger
Tlf. (066) *14 988

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 79/4 7. February 1979

THE USE OF MARKOV CHAIN MODELS IN SAMPLING FROM FINITE POPULATIONS

By

Ib Thomsen

CONTENTS

ABSTRACT

We consider problems of estimating the number of elements having a specific characteristic, when auxiliary information is available. We find the optimal strategy under a simple Markov chain "super-population" model, a model which seems to be relevant in many practical situations, for instance labour force surveys, where the auxiliary information can be taken from the last population census, and election surveys, where the auxiliary information are results from the latest election. We find an optimal estimator if attention is restricted to model-unbiased estimators. This estimator is independent of the design, and we also find the design that minimizes the expected mean square error. The estimator suggested is compared with methods suggested elsewhere to estimate proportions when supplementary information is available. It is shown that when the sampling variance of the estimators is used as the measure of uncertainty, the estimator suggested in the present paper has the smallest variance when a simple random sample is selected. When, however, the conditioned mean square error, given the sample, is used as measure of uncertainty, none of the methods studied are uniformly best.

Key Words and Phrases:

Finite populations. Superpopulation models. Markov chain models. Optimal strategy. Efficient use of supplementary information.

1. INTRODUCTION

Recently a number of papers have been published in which is taken the position that many sampling problems can be fruitfully analyzed by applying appropriate super-population models, i.e. the finite population is assumed to have been generated from a infinite super-population. This approach is not new; in Cochran (1953) is used a super-population model to compare ratio estimation using equal probabilities of selection with unbiased estimation with unequal probabilities. Other similar studies are done in Brewer (1963), Foreman and Brewer (1971), Hanurav (1967), Royall (1970), Scott and Smith (1969, 1975). In these papers various sampling strategies are compared under a linear regression super-population model given by

$$Y_i = \alpha + \beta x_i + U_i, \tag{1.1}$$

where x_i $(i = 1,2,\dots,N)$ are known constants, $E(U_i) = 0$, $E(U_i^2) = \sigma_i^2$, $E(U_i U_j) = 0$, and E denotes expectation over the super-population. The aims are to make a sample design and to estimate $Y = \sum_{i=1}^{N} Y_i$, where Y_i $(i = 1,2,\dots,N)$ is the realization of Y_i in the population.

A sample of size n is defined to be a set of population labels $i_1, i_2, \dots, i_n$, together with the set of their associated observed characteristic values

$$(x,Y) = (x_{i_1}, Y_{i_1}), (x_{i_2}, Y_{i_2}), \dots, (x_{i_n}, Y_{i_n}).$$

A sample design is then defined by some finite set, $\mathscr{S}$, of sets of lables, s, together with a probability measure assigned by choosing a function $p(s) > 0$, $\Sigma\, p(s) = 1$, where $p(s)$ is the probability of choosing the sample s. As estimates of Y is considered the class of estimates

$$\hat{Y} = \sum_{i\varepsilon s} a_i(s)\, Y_i, \tag{1.2}$$

where Y_i $(i\varepsilon s)$ are the observed values of Y_i, and $a_i(s)$ are constants. One strategy, $(\hat{Y}', p')$, is said to be better than another strategy $(\hat{Y}'', p'')$, if the former has the smaller expected mean square error, i.e.

$$E\{\sum_{\mathscr{S}'} p'(s)(\hat{Y}'-Y)^2\} \le E\{\sum_{\mathscr{S}''} p''(s)(\hat{Y}''-Y)^2\}. \tag{1.3}$$

In Royall (1970) is considered the class of model-unbiased estimates, defined in Section 2 below, and it is shown that the best estimate under model (1.1) within this class is

$$\hat{Y} = \sum_{i\varepsilon s} Y_i + (N-n)\hat{\alpha} + \hat{\beta} \sum_{i\notin s} x_i , \qquad (1.4)$$

where $\hat{\alpha}$, $\hat{\beta}$ are the usual weighted least squares estimators of α and β, and $\sum_{i\notin s}$ denotes the sum of all labels not included in the sample. $\hat{Y}$ is independent of the design, and when $\sigma_i^2 = \sigma^2$ its expected mean square error is minimized if the sample is selected such that

$$\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{n}\sum_{j\varepsilon s} x_j\right)^2 \Big/ \sum_{i\varepsilon s}\left(x_i - \frac{1}{n}\sum_{j\varepsilon s} x_j\right)^2$$

is minimum.

In this paper we shall consider situations in which the population under study can be divided into classes on the basis of such things as occupation, social status or voting behaviour. Instead of using models like (1.1), we shall assume that the movements over time between the classes can be adequately described by a Markov chain model. This, we believe, is often the case as such models often are used to study social processes, Bartholomew (1973).

We shall assume that to each element in the population is associated two sets of binary variables $(x_{i1}, \ldots, x_{ik};\ Y_{i1}, \ldots, Y_{ik})$, where the x_{ij} are known and $\sum_{j=1}^{k} x_{ij} = \sum_{j=1}^{k} Y_{ij} = 1$.

We assume that the relationship between the two sets of variables can be expressed in the following simple model

$$p(Y_{im} = 1 \mid x_{i\ell} = 1) = p_{\ell m}, \qquad i = 1,2,\ldots,N; \quad m = 1,2,\ldots,k; \quad \ell = 1,2,\ldots,k \qquad (1.5)$$

$$\sum_{m=1}^{k} p_{\ell m} = 1, \qquad \ell = 1,2,\ldots,k .$$

We now want to select a sample to estimate the number of persons having a specific characteristic $Y_h = \sum_{i=1}^{N} Y_{ih}$. The problems we are facing are how the sample should be selected and how we can utilize the fact that x_{ij} is known for all elements in the population.

In Section 3 below we apply an approach similar to the approach in Royall (1970). Among all model-unbiased estimators we find that the estimator that minimizes the conditioned mean square error can be written as $\hat{Y}_h = \sum_{i \in s} Y_{ih} + \sum_{i \notin s} \sum_{j=1}^{k} x_{ij} \hat{p}_{jh}$, where $\hat{p}_{jh}$ estimates p_{jh}. The best estimator being $\hat{p}_{jh} = (\sum_{i \in s} x_{ij} Y_{ih}) / (\sum_{i \in s} x_{ij})$. We also discuss what design is optimal. We find that an approximately optimal design consists of stratifying the population into k strata according to the values of x_{ij} $(j = 1,2,\dots,k)$, and apply the usual estimate of the population total when a stratified random sample is selected. The approximately optimal allocation for fixed sample-size consists of choosing the number of observations in stratum ℓ proportional with $\sum_{i=1}^{N} x_{i\ell} \sqrt{p_{\ell h}(1-p_{\ell h})}$. This strategy seems promising as it suggests to the sampler that he identifies sub-populations within which good predictions can be made, i.e. with transition probabilities close to 0 or 1, and allocates relatively few observations to these sub-populations.

Formally post-stratification is not a new estimation method, but it seems as if the efficiency of post-stratification in our situation is neglected, and other methods are suggested when the aim is to estimate proportions when auxiliary information is available. In Section 6, $\hat{Y}_h$ is compared with other methods suggested elsewhere for $k = 2$.

In our model we have that

$$E(Y_{i\ell} | x_{i1}, x_{i2}, \dots, x_{ik}) = \sum_{j=1}^{k} x_{ij} p_{j\ell}, \qquad i = 1,2,\dots,N$$

and

$$\text{var}(Y_{i\ell} | x_{i1}, x_{i2}, \dots, x_{ik}) = \sum_{j=1}^{k} x_{ij} p_{j\ell}(1-p_{j\ell}), \qquad i = 1,2,\dots,N.$$

A question of interest is what strategy would have been the result from using the model (1.1), generalized to the case with k dependent variables. In Section 4 we show that the strategy found in Section 3 is identical to the optimal strategy under a generalized version of model (1.1) if one disregards the fact that the variance of Y_{ij} depends on x_{ij}, and uses unweighted least squares estimates instead of weighted least squares estimates.

In Thomsen (1977) it is suggested to use the estimator developed in this paper in connection with the political barometres in Norway.

In this paper we make little real use of Markov chain theory. What we use is a matrix of transition probabilities and the strongest assumption is for the homogeneity of each transition probability for all members (or for any particular subgroup) of the population. In a paper aiming at using the same approach in connection with repeated surveys we intend to make more use of Markov chain theory.

2. DEFINITIONS AND NOTATIONS

We shall use the following notations:

$$\underset{\sim}{x}_i = (x_{i1},\ldots,x_{ik})', \qquad i = 1,2,\ldots,N$$

$$\underset{\sim}{x} = (\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_N),$$

$$\underset{\sim}{p} = \begin{bmatrix} p_{11},p_{12},\ldots,p_{1k} \\ p_{21},p_{22},\ldots,p_{2k} \\ \ldots \\ p_{k1},p_{k2},\ldots,p_{kk} \end{bmatrix}$$

Under model (1.5) we then have that

$$E(Y_{ih}|\underset{\sim}{x}) = \sum_{j=1}^{k} x_{ij}\, p_{jh}\,. \qquad (2.1)$$

Further we assume that

$$\begin{aligned} \operatorname{cov}(Y_{uv}, Y_{w\ell} \mid \underset{\sim}{x}) &= \sum_{j=1}^{k} x_{uj}\, p_{jv}(1-p_{jv}) && \text{if } u=w \text{ and } v=\ell\,, \\ &= -\sum_{j=1}^{k} x_{uj}\, p_{jv}\, p_{j\ell} && \text{if } u=w \text{ and } v\neq\ell\,, \qquad (2.2) \\ &= 0 && \text{if } u\neq w\,. \end{aligned}$$

Following Royall (1970) we shall distinguish between alternative definitions of unbiasednes. An estimate $\hat{Y}_h$ is said to be design-unbiased or p-unbiased for Y_h if

$$\sum_{s\in\mathscr{S}} p(s)\,\hat{Y}_h = Y_h.$$

An estimate is called model-unbiased or ξ-unbiased if for each sample it is unbiased under a given model such as (1.1) or (1.5), i.e. if $\hat{Y}_h$ satisfies $E(\hat{Y}_h - Y_h \mid \underset{\sim}{x}) = 0$ for all $s\in\mathscr{S}$.

3. OPTIMAL STRATEGY WITHIN THE CLASS OF MODEL UNBIASED ESTIMATORS

3.a. Minimalization of the conditioned mean square error

Following Royall (1970, 1971) we shall restrict our attention to the class of model unbiased estimators, i.e. the class of estimators $\hat{Y}_h = \sum_{i \in s} \sum_{\ell=1}^{k} a_{i\ell h}(s) Y_{i\ell}$ for which

$$E\{\hat{Y}_h(s) - Y_h | S = s\} = 0 \quad \text{for all} \quad s \in \mathscr{S}. \tag{3.1}$$

This is equivalent to

$$\sum_{i \in s} \sum_{\ell=1}^{k} a_{i\ell h}(s) \sum_{m=1}^{k} x_{im} p_{m\ell} = \sum_{i=1}^{N} \sum_{m=1}^{k} x_{im} p_{mh}$$

for all s, and all $\underset{\sim}{p}$. This can be written as

$$\sum_{m=1}^{k} \sum_{\ell=1}^{k} p_{m\ell} \sum_{i \in s} a_{i\ell h}(s) x_{im} = \sum_{m=1}^{k} p_{mh} N_m \tag{3.2}$$

for all s, and all $\underset{\sim}{p}$, where

$$N_m = \sum_{i=1}^{N} x_{im}.$$

As (3.2) must be fulfilled for all values of $\underset{\sim}{p}$, (3.2) is equivalent to

$$\sum_{i \in s} a_{i\ell h}(s) x_{im} = N_m \delta_{\ell h} \tag{3.3}$$

for all m, ℓ, and s, where $\delta_{\ell h} = 1$ if $\ell = h$, and $\delta_{\ell h} = 0$ if $\ell \neq h$.

Applying (2.2) and (3.1) it follows that the conditioned mean square error of $\hat{Y}_h$ given the sample s, is

$$\Delta_h = E\{\sum_{i \in s} \sum_{\ell=1}^{k} a_{i\ell h}(s) Y_{i\ell} - \sum_{i=1}^{N} Y_{ih}\}^2 =$$

$$= \text{var}\{\sum_{i \in s} (a_{ihh}(s) - 1) Y_{ih} + \sum_{i \in s} \sum_{\ell \neq h} a_{i\ell h}(s) Y_{i\ell} -$$

$$- \sum_{i \notin s} Y_{ih}\} =$$

$$= \sum_{i\in s} [a_{ihh}(s) - 1]^2 \sum_{m=1}^{k} x_{im} p_{mh}(1 - p_{mh}) +$$

$$+ \sum_{i\in s} \sum_{\ell \neq h} a_{i\ell h}^2(s) \sum_{m=1}^{k} x_{im} p_{m\ell}(1 - p_{m\ell}) + \tag{3.4}$$

$$+ \sum_{i\in s} \sum_{m=1}^{k} x_{im} p_{mh}(1 - p_{mh}) -$$

$$- \sum_{i\in s} \sum_{\ell \neq h} \sum_{n=\ell} a_{i\ell h}(s)\, a_{inh}(s) \sum_{m=1}^{k} x_{im} p_{m\ell} p_{mn} -$$

$$- 2 \sum_{i\in s} \sum_{\ell \neq h} [a_{ihh}(s) - 1]\, a_{i\ell h}(s) \sum_{m=1}^{k} p_{m\ell} p_{mh} x_{im} .$$

Theorem 1

The estimator $\hat{Y}_h(s)$ that has the smallest conditioned mean square error among all model unbiased estimators is given by

$$\hat{Y}_h(s) = \sum_{i\in s} Y_{ih} + \sum_{i\in s} Y_{ih} \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)} x_{im} ,$$

where

$$N_m(s) = \sum_{i\notin s} x_{im} , \quad \text{and} \quad n_m(s) = \sum_{i\in s} x_{im} .$$

Proof

To minimize (3.4) under conditions (3.1) we apply the Lagrange technique, and find that the following equations must be fulfilled under the side conditions (3.3).

$$2\,[a_{ihh}(s) - 1] \sum_{m=1}^{k} x_{im} p_{mh}(1 - p_{mh}) -$$

$$- 2 \sum_{\ell \neq h} a_{i\ell h}(s) \sum_{m=1}^{k} p_{m\ell} p_{mh} x_{im} - 2 \sum_{k=1}^{k} \lambda_{mhh}(s)\, x_{im} = 0$$

and for $\ell \neq h$

$$2\, a_{i\ell h}(s) \sum_{m=1}^{k} x_{im} p_{m\ell}(1 - p_{m\ell}) - 2 \sum_{n\neq \ell \neq h} a_{inh}(s) \sum_{m=1}^{k} p_{mn} p_{m\ell} x_{im} -$$

$$- 2\,[a_{ihh}(s) - 1] \sum_{m=1}^{k} p_{m\ell} p_{mh} x_{im} - 2 \sum_{m=1}^{k} \lambda_{m\ell h}(s)\, x_{im} = 0 ,$$

where $\lambda_{m\ell h}(s)$ are Lagrangian multipliers. The optimal values of $a_{ij\ell}(s)$ must satisfy (3.2) and

$$[\hat{a}_{ihh}(s)-1]\sum_{m=1}^{k} x_{im}\,p_{mh}(1-p_{mh}) = \sum_{\ell \neq h} \hat{a}_{i\ell h}(s)\sum_{m=1}^{k} p_{m\ell}\,p_{mh}\,x_{im} + \sum_{m=1}^{k}\lambda_{mhh}(s)\,x_{im}, \tag{3.5}$$

and for $\ell \neq h$ we must have

$$\hat{a}_{i\ell h}(s)\sum_{m=1}^{k} x_{im}\,p_{m\ell}(1-p_{m\ell}) = [\hat{a}_{ihh}(s)-1]\sum_{m=1}^{k} p_{m\ell}\,p_{mh}\,x_{im} + \sum_{n\neq\ell\neq h}\hat{a}_{inh}(s)\sum_{m=1}^{k} p_{mn}\,p_{m\ell}\,x_{im} + \sum_{m=1}^{k}\lambda_{m\ell h}(s)\,x_{im}. \tag{3.6}$$

We now multiply with x_{ij}, and use the fact that $x_{im}x_{ij} = \delta_{mj}x_{ij}$. We then must have that

$$\sum_{i\in s}(\hat{a}_{ihh}(s)-1)\,x_{ij}\,p_{jh}(1-p_{jh}) = \sum_{\ell\neq h}\sum_{i\in s}\hat{a}_{i\ell h}(s)\,x_{ij}\,p_{j\ell}\,p_{jh} + \lambda_{jhh}(s)\sum_{i\in s}x_{ij}, \tag{3.7}$$

and for $\ell \neq h$

$$\sum_{i\in s}\hat{a}_{i\ell h}(s)\,x_{ij}\,p_{j\ell}(1-p_{j\ell}) = \sum_{i\in s}[\hat{a}_{ihh}(s)-1]\,x_{ij}\,p_{j\ell}\,p_{jh} + \sum_{n\neq\ell\neq h}\sum_{i\in s}\hat{a}_{inh}(s)\,x_{ij}\,p_{jn}\,p_{j\ell} + \lambda_{j\ell h}\sum_{i\in s}x_{ij} \tag{3.8}$$

Inserting (3.3) into (3.7) and (3.8) it follows that

$$\lambda_{jhh}(s)\;n_j(s) = N_j(s)\,p_{jh}\,(1-p_{jh}),$$

and for $\ell \neq h$

$$\lambda_{j\ell h}(s)\;n_j(s) = -\,N_j(s)\,p_{j\ell}\,p_{jh}.$$

Inserting this into (3.5) and (3.6) gives the following equations:

$$[\hat{a}_{ihh}(s)-1]\sum_{m=1}^{k} p_{mh}(1-p_{mh})\, x_{im} =$$

$$= \sum_{\ell \neq h} \hat{a}_{i\ell h}(s) \sum_{m=1}^{k} p_{m\ell}\, p_{mh}\, x_{im} + \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)}\, p_{mh}(1-p_{mh})\, x_{im} \tag{3.9}$$

and for $\ell \neq h$

$$[\hat{a}_{ihh}(s)-1]\sum_{m=1}^{k} p_{m\ell}\, p_{mh}\, x_{im} =$$

$$= \hat{a}_{i\ell h}(s)\sum_{m=1}^{k} x_{im}\, p_{m\ell}(1-p_{m\ell}) - \sum_{n\neq \ell \neq h} \hat{a}_{inh}(s) \sum_{m=1}^{k} p_{mn}\, p_{m\ell}\, x_{im} + $$

$$+ \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)}\, p_{m\ell}\, p_{mh}\, x_{im} =$$

$$= \hat{a}_{i\ell h}(s)\sum_{m=1}^{k} x_{im}\, p_{m\ell}(1-p_{m\ell}) - \sum_{n\neq h} \hat{a}_{inh}(s) \sum_{m=1}^{k} p_{mn}\, p_{m\ell}\, x_{im} + $$

$$+ \hat{a}_{i\ell h}(s)\sum_{m=1}^{k} p_{m\ell}^{2}\, x_{im} + \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)}\, p_{m\ell}\, p_{mh}\, x_{im} =$$

$$= \hat{a}_{i\ell h}(s)\sum_{m=1}^{k} x_{im}\, p_{m\ell} - \sum_{n+h} \hat{a}_{inh}(s) \sum_{m=1}^{k} p_{mn}\, p_{m\ell}\, x_{im} + $$

$$+ \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)}\, p_{m\ell}\, p_{mh}\, x_{im} . \tag{3.10}$$

Define w such that $x_{iw} = 1$. Then (3.9) and (3.10) can be written as

$$[\hat{a}_{ihh}(s)-1]\, p_{wh}(1-p_{wh}) =$$

$$= \sum_{\ell \neq h} \hat{a}_{i\ell h}(s)\, p_{w\ell}\, p_{wh} + \frac{N_w(s)}{n_w(s)}\, p_{wh}(1-p_{wh}) \tag{3.11}$$

and for $\ell \neq h$

$$[\hat{a}_{ihh}(s)-1]\, p_{w\ell}\, p_{wh} =$$

$$= \hat{a}_{i\ell h}(s)\, p_{w\ell} - \sum_{n\neq h} \hat{a}_{inh}(s)\, p_{wn}\, p_{w\ell} + \frac{N_w(s)}{n_w(s)}\, p_{w\ell}\, p_{wh} . \tag{3.12}$$

From (3.11) and (3.12) follows that

$$\hat{a}_{ihh}(s) - 1 \;=\; \hat{a}_{i\ell h}(s) + \sum_{m=1}^{k} x_{im} \frac{N_m(s)}{n_m(s)} ,$$

from which follows that $\hat{a}_{i\ell h}(s)$ is independent of ℓ for $\ell \neq h$. Inserting this into (3.3), it follows that we must have

$$\sum_{i\in s} \hat{a}_{i\ell h}(s) = 0 \quad \text{for all} \quad \ell \neq h .$$

We now have that

$$\hat{Y}_h = \sum_{i\in s} \sum_{\ell=1}^{k} \hat{a}_{i\ell h}(s) Y_{i\ell} =$$

$$= \sum_{i\in s} \hat{a}_{i\ell h}(s) \sum_{\ell \neq h} Y_{i\ell} + \sum_{i\in s} \hat{a}_{ihh}(s) Y_{ih} =$$

$$= \sum_{i\in s} \hat{a}_{i\ell h}(s)(1 - Y_{ih}) + \sum_{i\in s} (\hat{a}_{i\ell h} + 1 + \sum_{m=1}^{k} x_{im} \frac{N_m(s)}{n_m(s)}) Y_{ih} =$$

$$= \sum_{i\in s} \hat{a}_{i\ell h}(s) - \sum_{i\in s} \hat{a}_{i\ell h}(s) Y_{ih} + \sum_{i\in s} \hat{a}_{i\ell h}(s) Y_{ih} + \sum_{i\in s} Y_{ih} + $$

$$+ \sum_{i\in s} Y_{ih} \sum_{m=1}^{k} \frac{N_m(s)}{n_m(1)} x_{im} =$$

$$= \sum_{i\in s} Y_{ih} + \sum_{i\in s} Y_{ih} \sum_{m=1}^{k} \frac{N_m(s)}{n_m(s)} x_{im} . \tag{3.14}$$

□

The estimate (3.14) has the same structure as (1.4). It includes the observed values of Y_{ih} and predicts the unobserved values.

3.b. Optimal design

The estimator (3.14) minimizes the conditioned mean square error for any sample selected, independent of the design. A natural question to raise is what design is optimal in the sense that it minimizes the expected mean square error, expectation taken over all possible samples.

We find that

$$E\{\sum_{s\in\mathcal{S}} p(s) (\hat{Y}_h^* - Y_h)^2\} = \sum_{s\in\mathcal{S}} p(s) \left[\sum_{m=1}^{k} \frac{p_{mh}(1-p_{mh})}{\sum_{i\in s} x_{im}} \left[\sum_{i\notin s} x_{im}\right]^2 \right] +$$

$$+ \sum_{s\in\mathcal{S}} p(s) \sum_{i\notin s} x_{ij}\, p_{jh}(1 - p_{jh}) .$$

The expected mean square error is minimized if the sample is selected such that

$$\left[\left(\sum_{i=1}^{N} x_{ij} \Big/ \sum_{i\in s} x_{ij}\right)^2 - 1\right] p_{jh}(1-p_{jh})$$

is constant for all values of j. When N is large compared with n, this is approximately equivalent to the conditions that

$$\left(\sum_{i=1}^{N} x_{ij} \Big/ \sum_{i\in s} x_{ij}\right) \sqrt{p_{jh}(1-p_{jh})} \tag{3.15}$$

should be constant for all j.

It follows that an approximately optimal design consists of stratifying the population into k strata within which the values of $\underset{\sim}{x}_i$ are identical. The optimal number of observations from each stratum is proportional with the size of the stratum and $\sqrt{p_{jh}(1-p_{jh})}$. This allocation is similar to the optimal allocation in stratified sampling, the only difference is that the within stratum variance is substituted with $\sqrt{p_{jh}(1-p_{jh})}$ in (3.15). As in stratified sampling the gains from using optimal allocation instead of proportional allocation are trivial except in cases where p_{jh} assume values less than 0.20 or larger than 0.80. If, however, it is known that certain of the transition probabilities are close to 1 or 0, it is efficient to undersample the corresponding strata.

It should be noted that one does not need to know all $\underset{\sim}{x}_i$ $(i = 1,2,\ldots,N)$ to apply (3.14), as $\hat{Y}^*_h$ can be written in the following way:

$$\hat{Y}^*_h = \sum_{\ell=1}^{k} \left(\frac{\sum_{i\in s} Y_{ih} x_{i\ell}}{\sum_{i\in s} x_{i\ell}} \sum_{i=1}^{N} x_{i\ell} \right).$$

All we need to know are the values of $\underset{\sim}{x}_i$ in the sample and $\sum_{i=1}^{N} x_{i\ell}$, $(\ell = 1,2,\ldots,k)$. This is important if we want to apply $\hat{Y}^*_h$ in election surveys, where x_{ih} is chosen equal to 1 if person i voted for party h at the last election. In such cases $\sum_{i=1}^{N} x_{i\ell}$ $(\ell = 1,2,\ldots,k)$ are known, and knowledge about $\underset{\sim}{x}_i$ for all persons in the sample usually can be found from the questionnaire.

As a conclusion we have found the following results: There are two situations, one where the population can be stratified beforehand into k strata and it is reasonable to expect different values of p_{jh}, and one where these two conditions do not hold. In the former situation we have, in Fisher's terms, "relevant and recognizable subsets". We would then opt for a stratified design and, if reasonable guesses of the p_{jh} were available, differential sample fractions. In the second case, post-stratification is all that is possible. There is no need for randomized selection if the model (1.5) is to be used for inference, but it may be useful to ensure public acceptance of the model's validity. From a practical point of view it is worth noticing that the efficiency of post-stratification when estimating proportions has received very little attention, and that other estimation methods are recommended. In Section 6 below we shall compare these estimators with post-stratification.

4. COMPARING THE OPTIMAL STRATEGY WITH A GENERALIZATION OF THE RESULTS BY ROYALL

Our model can be stated as follows:

$$E(Y_{ih} \mid \underset{\sim}{x}) = \underset{\sim}{x}_i' \begin{pmatrix} p_{1h} \\ p_{2h} \\ \vdots \\ p_{kh} \end{pmatrix} \qquad i = 1,2,\ldots,N,$$

and

$$\mathrm{var}(Y_{ih} \mid \underset{\sim}{x}) = \underset{\sim}{x}_i' \begin{pmatrix} p_{1h}(1-p_{1h}) \\ p_{2h}(1-p_{2h}) \\ \vdots \\ p_{kh}(1-p_{kh}) \end{pmatrix} \qquad i = 1,2,\ldots,N.$$

This model is similar to the model (1.1) extended to k independent variables, $\alpha = 0$, and $\sigma_i^2 = \sum_{j=1}^{k} x_{ij}\, p_{jh}(1-p_{jh})$. In this section we shall generalize model (1.1), and compare the optimal strategy under this generalized model with the strategy in Section 3.

We assume the following linear "super-population" model

$$\underset{\sim}{U} = \underset{\sim}{v}_k\, \underset{\sim}{\beta}_k + \underset{\sim}{\varepsilon}\,; \quad E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}, \quad E(\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}) = \sigma^2 \underset{\sim}{\Sigma}\,, \tag{4.1}$$

where $\underset{\sim}{U}$, the dependent variable, is the object of the survey, and $\underset{\sim}{v}_k$ is a set of k variables. The subscripts in (4.1) and in the rest of this section indicate the dimensions of matrices and vectors. We further assume that a finite population of size N has been drawn

$$\underset{\sim}{U}_N = \underset{\sim}{v}_{Nk}\, \underset{\sim}{\beta}_k + \underset{\sim}{\varepsilon}_N\,.$$

Before the survey is conducted $\underset{\sim}{U}_N$ is unknown, while $\underset{\sim}{v}_{Nk}$ and $\underset{\sim}{\Sigma}_{NN}$ are known. To estimate $U = \sum_{i=1}^{N} U_i$ a sample of size n is selected from the finite population, and the dependent variable is observed for each element. We shall think of the sample procedure as a partitioning of the matrices in (4.1) as follows:

$$\underset{\sim}{U}_N = (\underset{\sim}{U}_n \; \underset{\sim}{U}_{(N-n)})', \qquad \underset{\sim}{\Sigma} = \begin{pmatrix} \underset{\sim}{\Sigma}_{nn} & \underset{\sim}{\Sigma}_{n(N-n)} \\ \underset{\sim}{\Sigma}_{(N-n)n} & \underset{\sim}{\Sigma}_{(N-n)(N-n)} \end{pmatrix},$$

$$\underset{\sim}{v}_{Nk} = \begin{pmatrix} \underset{\sim}{v}_{nk} \\ \underset{\sim}{v}_{(N-n)k} \end{pmatrix}.$$

<u>Theorem 2</u>

Among the model-unbiased estimates $\hat{U} = \sum_{i\varepsilon s} b_i(s)\, U_i$, the estimate that minimizes the expected mean square error (1.3) is given by

$$\hat{\hat{U}} = 1_n' \underset{\sim}{U}_n + 1_{(N-n)}' \underset{\sim}{v}_{(N-n)k} \hat{\underset{\sim}{\beta}}_k, \tag{4.2}$$

where 1_n denotes the n-dimensional vector $(1,1,\ldots,1)'$, $v_{(N-n)k}$ denotes the matrix of v values that are not included in the sample, and $\tilde{\beta}_k$ denotes the weighted least squares estimate of $\underset{\sim}{\beta}_k$, estimates in the selected sample, i.e.

$$\hat{\underset{\sim}{\beta}}_k = (\underset{\sim}{v}_{nk}' \underset{\sim}{\Sigma}_{nn}^{-1} \underset{\sim}{v}_{nk})^{-1} \underset{\sim}{v}_{nk}' \underset{\sim}{\Sigma}_{nn}^{-1} \underset{\sim}{U}_n . \tag{4.3}$$

<u>Proof</u>

The proof follows the lines of the proof of theorem 1 in Royall (1970). □

If we use the estimate (4.2) with $\underset{\sim}{U}_n = \underset{\sim}{Y}$, $\underset{\sim}{v}_{Nk} = \underset{\sim}{x}$, and $\sigma_i^2 = \sum_{j=1}^{k} x_{ij}\, p_{jh}(1-p_{jh})$, the estimate will depend on p_{ij}, and is therefore of little practical interest. If we, however, disregard the fact that the variance of Y_{ij} depend on x_{ij} and use the unweighted estimate in (4.3) instead of the weighted one, it is easily seen that the two models lead to the same optimal strategies.

5. UNCERTAINTY IN THE ESTIMATE

In Royall (1971) is discussed the appropriateness of the sampling variance as a measure of uncertainty after the sample is selected. Both on theoretical and empirical grounds he ends up with favouring the conditioned mean square error as the most relevant when using ratio estimators.

In what follows we shall give a short discussion concerning the appropriateness of the two measures of uncertainty in our situation with binomial variables.

In the case of $Y_1,\ldots,Y_N$ being independent and $p(Y_i = 1) = p$, the optimal estimator for Y is the simple expansion estimator $\hat{Y} = (N/n) \sum_{i \in s} Y_i$, where n is the sample size. In this case we have that the conditioned mean square error is

$$E(\hat{Y} - Y)^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) p(1-p) . \qquad (5.1)$$

This is correct for any sampling plan. All the $\binom{N}{n}$ samples of size n lead to the same value of (5.1). There seems, however, many good reasons for simple random selection. If such a plan is adopted, the expansion estimate is also the conventional choice, and we have that the sampling mean square error is

$$\sum_{\mathcal{S}} p(s) \, (\hat{Y} - Y)^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left(\frac{1}{N-1}\right) \sum_{i=1}^{N} \left(Y_i - \frac{Y}{N}\right)^2 . \qquad (5.2)$$

Both (5.1) and (5.2) are unknown and must be estimated. A design-unbiased estimate of (5.2) is

$$\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left(\frac{1}{n-1}\right) \sum_{i \in s} \left(Y_i - \left(\sum_{i \in s} Y_i / n\right)\right)^2 ,$$

which is also an ξ - unbiased estimate of (5.1). In this case the two approaches basicly lead to the same estimate of accuracy.

Let the superpopulation model be as given in (1.5), and let $k=2$. In this case we have that $x_{i2} = (1-x_{i1})$, $x_{ij} = 0$; $j>2$, $Y_{i2} = (1-Y_{i1})$, and $Y_{ij} = 0$; $j>2$. Denoting $Y_{i1} = Y_i$ and $x_{i1} = x_i$, the estimate (3.14) can be written as

$$\hat{Y}^* = \frac{\sum_{i\in s} Y_i x_i}{\sum_{i\in s} x_i} \sum_{i=1}^{N} x_i + \frac{\sum_{i\in s} Y_i(1-x_i)}{\sum_{i\in s}(1-x_i)} \sum_{i=1}^{N}(1-x_i). \tag{5.3}$$

It is seen that the conditioned mean square error is

$$E(\hat{Y}^* - Y)^2 = \frac{p_{11}(1-p_{11})}{\sum_{i\in s} x_i} \left(\sum_{i=1}^{N} x_i\right)^2 + \frac{p_{01}(1-p_{01})}{\sum_{i\in s}(1-x_i)} \left(\sum_{i=1}^{N}(1-x_i)\right)^2 \tag{5.4}$$

This measure of accuracy depends on the sample selected, which is consistent with intuition. The smallest value occurs when the sample is allocated according to the rule given by (3.15).

The estimate (5.3) is a post-stratified mean, where the sample is post-stratified according to the values of the x-variable. An approximation to the design variance when a simple random sample is selected is given in Cochran (1963), and can in our notation be written as

$$\sum_{\mathcal{S}} p(s)\,(\hat{Y}^* - Y)^2 \approx N\left\{\sum_{i=1}^{N} x_i \frac{p'_{11}(1-p'_{11})}{n} + \left(\sum_{i=1}^{N}(1-x_i)\right) \frac{p'_{01}(1-p'_{01})}{n}\right\} \tag{5.5}$$

where

$$p'_{11} = \frac{\sum_{i=1}^{N} Y_i x_i}{\sum_{i=1}^{N} x_i} \quad \text{and} \quad p'_{01} = \frac{\sum_{i=1}^{N} Y_i(1-x_i)}{\sum_{i=1}^{N}(1-x_i)}.$$

This quantity is constant for all samples, which contradicts intuition.

In both cases the measure of uncertainty is unknown, and must be estimates from the sample. In (5.4) we need an estimate of p_{11} and p_{01}. Natural estimators are

$$\hat{p}_{11} = \frac{\sum_{i\in s} Y_i x_i}{\sum_{i\in s} x_i}, \quad \text{and} \quad \hat{p}_{01} = \frac{\sum_{i\in s} Y_i(1-x_i)}{\sum_{i\in s}(1-x_i)}.$$

$\hat{p}_{11}$ and $\hat{p}_{01}$ are ξ-unbiased estimates of p_{11} and p_{01} respectively. They are also approximately p-unbiased estimates of p'_{11} and p'_{01} when the sample is a simple random sample. Inserting $\hat{p}_{11}$ and $\hat{p}_{01}$ into (5.4) and (5.5) we find the following estimates of accuracy.

$$V_1 = \frac{\hat{p}_{11}(1-\hat{p}_{11})}{\sum_{i\in s} x_i}\left(\sum_{i=1}^{N} x_i\right)^2 + \frac{\hat{p}_{01}(1-\hat{p}_{01})}{\sum_{i\in s}(1-x_i)}\left(\sum_{i=1}^{N}(1-x_i)\right)^2,$$

and

$$V_2 = N\left\{\frac{\hat{p}_{11}(1-\hat{p}_{11})}{n}\sum_{i=1}^{N} x_i + \frac{\hat{p}_{01}(1-\hat{p}_{01})}{n}\sum_{i=1}^{N}(1-x_i)\right\}.$$

V_1 and V_2 are not identical in general. If we, however, consider the conditioned sampling variance given $\sum_{i\in s} x_i$ instead of the unconditioned sampling variance over all samples, the two approaches lead to the same estimate of uncertainty.

Recently Holt and Smith (1978) have given an interesting discussion of the appropriateness of the conditioned mean square error in connection with post-stratification in general.

6. COMPARING $\hat{Y}_h$ WITH METHODS USUALLY APPLIED

Formally the estimate (3.14) is not new, as post-stratification is a wellknown estimation method. However, in the sampling literature and in practical survey work the proposed estimate seems to have received very little attention in connection with estimation of proportions when supplementary information is available. On the contrary, other methods are proposed and used. In this section we shall therefore compare the estimate (3.14) with the commonly used ratio-estimate, the usual mean, and an estimate which is a special case of the modified Horwitz-Thompson estimator discussed by Basu (1971), and suggested by Wynn (1976), and by Rao (1977).

Again we shall restrict our attention to cases with k=2, and denote $x_{i1} = x_i$, $x_{i2} = (1-x_i)$, and $Y_{i1} = Y_i$ $(i = 1,2,\ldots,N)$.

$$R = \frac{\sum_{i\in s} Y_i}{\sum_{i\in s} x_i}\sum_{i=1}^{N} x_i \qquad \text{(the ratio-estimate)}$$

$$\hat{Y} = (N/n)\sum_{i\in s} Y_i \qquad \text{(the inflated sample mean)}$$

$$W = \sum_{i=1}^{N} x_i + \frac{N}{n}\left(\sum_{i\in s} Y_i - \sum_{i\in s} x_i\right)$$

It is worth observing that in cases where one chooses $\sum_{i \in s} x_i = n\,p$, where $p = (1/N) \sum_{i=1}^{N} x_i$, all four estimates are identical.

Comparing the estimates by their sampling variance

First, we shall compare the four estimates by comparing their sample variances. Applying the usual approximations for ratio estimates and post-stratification, it is easily shown that

$$\mathrm{var}\,(R) \approx \frac{1}{nN} \sum_{i=1}^{N} Y_i \left\{ \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} x_i} + 1 - 2\,p_{11}^* \right\},$$

where

$$p_{11}^* = \sum_{i=1}^{N} Y_i\, x_i \Big/ \sum_{i=1}^{N} x_i\,.$$

For large N and n we have that

$$\mathrm{var}\,(\hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} Y_i \,(1 - \frac{1}{N} \sum_{i=1}^{N} Y_i)\,/\,n\,.$$

In Wynn (1976) we find the following result, ignoring the finite population coefficient.

$$\mathrm{var}\,(W) = \left\{ p_y(1 - p_y) + p_x(1 - p_x) - 2\,p_x\,(p_{11}^* - p_y) \right\} / \,n$$

$$\mathrm{var}\,(w) = \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \,(1 - \frac{1}{N} \sum_{i=1}^{N} Y_i) + \frac{1}{N} \sum_{i=1}^{N} x_i (1 - \frac{1}{N} \sum_{i=1}^{N} x_i) - 2\,\frac{1}{N} \sum_{i=1}^{N} x_i (p_{11}^* - \frac{1}{N} \sum_{i=1}^{N} Y_i) \right\} / \,n$$

Applying the usual approximation for the variance of a post-stratified mean, we also find that

$$\mathrm{var}\,(T) = \left\{ \frac{1}{N} \sum_{i=1}^{N} x_i \;p_{11}^*(1 - p_{11}^*) + (1 - \frac{1}{N} \sum_{i=1}^{N} x_i)\,p_{01}^*\,(1 - p_{01}^*) \right\} / \,n\,,$$

where

$$p_{01}^* = \sum_{i=1}^{N} Y_i (1 - x_i) \Big/ \sum_{i=1}^{N} (1 - x_i)\,.$$

Using that

$$\frac{1}{N}\sum_{i=1}^{N} Y_i = p_{11}^{*}\frac{1}{N}\sum_{i=1}^{N} x_i + p_{01}^{*}\frac{1}{N}\sum_{i=1}^{N}(1-x_i),$$

it can be shown that

$$\text{var}(T) = \text{var}(R) - \frac{1}{n}\left\{\frac{p_{01}^{*\,2}}{\frac{1}{N}\sum_{i=1}^{N} x_i} - p_{01}^{*\,2}\right\}, \qquad (6.1)$$

$$\text{var}(T) = \text{var}(\hat{Y}) - \frac{1}{N}\sum_{i=1}^{N} x_i(1-\frac{1}{N}\sum_{i=1}^{N} x_i)(p_{11}^{*}-p_{01}^{*})^2 / n, \qquad (6.2)$$

$$\text{var}(T) = \text{var}(w) - p_x(1-p_x)\{(p_{11}-p_{01}-1)^2+1\}/n \qquad (6.3)$$

From (6.1), (6.2) and (6.3) it follows that T has the smallest sampling variance. - In what follows we shall compare the four estimates in two ways both different from the sampling variance. First, we shall follow Smith (1976) and write the estimates on what has been called their predicting form, and finally we shall compare the estimates by considering their expected mean square error given the sample selected. We shall see from the last comparison that T is not always the estimate with the smallest conditioned mean square error for a given sample.

Any estimate, $\hat{Y}$, can be written on the form

$$\hat{Y} = \sum_{i\in s} Y_i + \sum_{i\in s} \hat{Y}_i,$$

where $\hat{Y}_i$ is the implied predictor of Y_i under the model when $i\notin s$. If $\hat{R}$ is written on its predictive form, we find that

$$\hat{R} = \sum_{i\in s} y_i + \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\sum_{i\notin s} x_i = \sum_{i\in s} y_i + \sum_{i\notin s}\frac{\sum_{j\in s} y_j}{\sum_{j\in s} x_j} x_i .$$

In this case $$\hat{Y}_i = \frac{\sum_{j\in s} y_j}{\sum_{j\in s} x_j} x_i = \frac{\sum_{i\in s} y_j x_j}{\sum_{j\in s} x_j} x_i + \frac{\sum_{j\in s} y_j(1-x_j)}{\sum_{j\in s} x_j} x_i .$$

While $\sum_{j\in s} y_j x_j / \sum_{j\in s} x_j$ is a reasonable estimate of p_{11}, the second term of $\hat{Y}_i$ is hard to identify as a predictor.

If $\hat{Y}$ is written in a similar way, we find that

$$\hat{Y} = \sum_{i\in s} y_i + \sum_{i\notin s} \left(\frac{1}{n} \sum_{j\in s} y_j\right) .$$

In this case $\hat{Y}_i = \frac{1}{n} \sum_{j\in s} y_j$, $i \notin s$, which is an intuitively sensible predictor, but it makes no use of the x-values.

Compared with (3.14) for $k = 2$, all the three estimates considered seem suspicious when written in their predictive form.

In the section above is given arguments in favour of the conditioned mean square error as the intuitively most appealing measure of uncertainty as sompared with the more usually applied sampling variance. Finally, in this section we shall therefore compare the conditioned mean square error of the estimates R, $\hat{Y}$, and W, and compare with that of (3.14).

We easily find the following:

$$E \{R - Y\}^2 = p_{11}(1-p_{11}) \frac{(\sum_{i\notin s} x_i)^2}{\sum_{i\in s} x_i} + p_{01}(1-p_{01}) \left(\frac{\sum_{i\notin s} x_i}{\sum_{i\in s} x_i}\right)^2 \sum_{i\in s} (1 - x_i) +$$

$$+ \left(\frac{\sum_{i=1}^{N} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i} - N\right)^2 p_{01}^2 + \text{var} \left(\sum_{i\notin s} y_i\right) ,$$

$$E \{T - Y\}^2 = p_{11}(1-p_{11}) \frac{(\sum_{i\notin s} x_i)^2}{\sum_{i\in s} x_i} + p_{01}(1-p_{01}) \frac{(\sum_{i\notin 1} (1 - x_i))^2}{\sum_{i\in s} (1 - x_i)} +$$

$$+ \text{var} \left(\sum_{i\notin s} y_i\right) ,$$

$$E \{\hat{Y} - Y\}^2 = p_{11}(1-p_{11}) \left(\frac{N}{n} - 1\right)^2 \sum_{i\in s} x_i + p_{01}(1-p_{01}) \left(\frac{N}{n} - 1\right)^2 \sum_{i\in s} (1 - x_i) +$$

$$+ (p_{11} - p_{01})^2 \left(\frac{N}{n} \sum_{i\in s} x_i - \sum_{i=1}^{N} x_i\right)^2 + \text{var} \left(\sum_{i\notin s} y_i\right) ,$$

and

$$E \{W - Y\}^2 = p_{11}(1-p_{11}) \left(\frac{N}{n} - 1\right)^2 \sum_{i\in s} x_i + p_{01}(1-p_{01}) \left(\frac{N}{n} - 1\right)^2 \sum_{i\in s} (1 - x_i) +$$

$$+ (1 - p_{11} - p_{01})^2 \left(\sum_{i=1}^{N} x_i - \frac{N}{n} \sum_{i\in s} x_i\right)^2 + \text{var} \left(\sum_{i\notin s} y_i\right) .$$

A thoroughly comparison of the four conditioned mean square errors is very complex, and will not be given here, but it is important to note that, contrary to what was the case when comparing sampling errors, none of the estimates have uniformly smallest conditioned mean square error. This conclusion is in accordance with that of Holt and Smith (1978), who discuss the efficiency of post-stratification in general.

7. REFERENCES

[1] Bartholomew, D.J. (1973): Stochastic Models for Social Processes, J. Wiley & Sons.

[2] Basu, D. (1971): An Essay on the logical foundation of survey sampling. Part 1. In Foundations of Statistical Inference (ed. V.P. Godambe and D.A. Sprott). Holt - Rinehard - Winston.

[3] Brewer, K.R.W. (1963): Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. Aust. J. Statist. **5**, 93-105.

[4] Cochran, W. (1953): Sampling techniques, 1st Edn. 122-24. J.Wiley & Sons.

[5] Foreman, E.K. and Brewer, K.R. (1971): The efficient use of supplementary information in standard sampling procedures. J. R. Statist. Soc. B, **33**, 291-400.

[6] Hanurav, T.V. (1967): Optimum utilization of auxiliary information: π p s sampling of two units from a stratum. J. R. Statist. Soc., B, **29**, 394-91.

[7] Holt, D. and Smith, T.M.F. (1978): Post-stratification. Unpublished manuscript.

[8] Rao, T.J. (1977): Estimation of change in proportions. Proceedings of the 41st session of ISI.

[9] Royall, R.M. (1970): On finite population sampling theory under certain linear regression models. Biometrika, **57**, 377-388.

[10] Royall, R.M. (1971): Linear Regression Models in Finite Population Sampling Theory, in Foundations of Statistical Inference, eds. Godambe, V.P. and Sprott, D.A., Holt, Rinehart and Winston.

[11] Scott, A. and Smith, T.M.F. (1969): A note on estimating secondary characteristics in multivariate surveys. Sankhyā, A, **31**, 497-498.

[12] Scott, A. and Smith, T.M.F. (1975): Comments on "The use of supplementary information in standard sampling procedures". J. R. Statist. Soc., B, **37**, 146-148.

[13] Thomsen, I. (1977): Metoder for politiske meningsmålinger. Sosialøkonomen, 7, 1977.

[14] Wynn, H.P. (1976): An unbiased estimator under SRS of a small change in population. The Statistician, 25, 225-28.