Survey Methodology, December 1993
Vol. 19, No. 2, pp. 173–182
Statistics Canada

Optimum Two-Stage Sample Design for Ratio Estimators: Application to Quality Control - 1990 French Census

JEAN-CLAUDE DEVILLE[1]

ABSTRACT

This study is based on the use of superpopulation models to anticipate, before data collection, the variance of a measure by ratio sampling. The method, based on models that are both simple and fairly realistic, produces expressions of varying complexity and then optimizes them, in some cases rigorously, in others approximately. The solution to the final problem discussed points up a rarely considered factor in sample design optimization: the cost related to collecting individual information.

KEY WORDS: Census quality control; Superpopulation model; Two-stage sample design optimization; Multiple objective survey.

1. INTRODUCTION

The survey method used for quality control of French census data pointed up a number of new and interesting problems, three of which are dealt with in this paper. After discussing them in general terms, we describe their specific application to the census.

In all cases, the problem is one of optimizing a two-stage survey in which the primary units are census collection districts. Units are selected using an index k that varies in a population U of districts and is, in concrete terms, a processing unit of the census forms collected.

The first problem is that of estimating the frequency of a characteristic in the population of forms (the fact of containing an error). Keeping in mind the accuracy defined for this estimate, an attempt is made to minimize survey cost with a cost function in the form

$$C_T = mC_o + nC_1, \tag{1.1}$$

where m is the number of primary units (districts) sampled, C_o the cost of processing one PU, n the number of final units (forms) sampled and C_1 the cost of processing one final unit. The problem is fairly common when a mean is to be estimated (see for example W. Cochran (1977)). Our solution is more complete as it takes into account the great variability in primary unit size.

The second, more unique, problem is also more significant. The final population (*i.e.* the forms) is made up of G separate groups ($g = 1$ to G). We are looking for an estimate of the frequency of occurrence of a characteristic in each group, with an accuracy defined for each one. The constraint resides in the fact that, because the primary units are common to all groups, sampling within one PU affects all groups.

The objective is to minimize survey cost, which is expressed as

$$C_T = mC_o + \sum_{g=1}^{G} n_g C_g, \tag{1.2}$$

where n_g is the total number of final units in group g and C_g the cost of processing one final unit in group g. In practice the groups are made up of the different types of census forms.

The third problem is related to coding control. We do have an *a priori* measure of the difficulty of coding each form. Formally, therefore, we have, at the level of each individual i in the population, a quantitative variable X_i, such that the probability (within a meaning to be defined) of the individual having the characteristic to be measured is approximately proportional to X_i. We are seeking to use this information to minimize the cost of control (measurement of the frequency of the "coding error" characteristic) subject to a defined survey accuracy.

In each case, plausible and simple superpopulation models allow us to evaluate the anticipated variance of the survey. In a manner of speaking, this is an almost standard illustration of model assisted survey sampling as described in Särndal, Swensson, Wretman (1992).

2. OPTIMUM ESTIMATE OF THE PROPORTION OF RECORDS CONTAINING ERRORS TWO-STAGE SAMPLE DESIGN

Each primary unit k (district) has a known number N_k of individuals (forms). Of this number, D_k display the characteristic of interest (*i.e.* contain an error). The aim is to estimate:

[1] Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Études Économiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

$$P = \sum_U D_k \Big/ \sum_U N_k .$$

The survey is done by drawing a sample s of primary units (PU), with π_k, the probability of inclusion in the first order and $\pi_{k\ell}$ in the second order, to be determined. Subsequently, if primary unit k is drawn in s, n_k individuals drawn by simple random sampling without replacement are checked; d_k denotes the number of forms containing errors that will be found.

Estimator $\hat{P}_k$ of $P_k = D_k/N_k$ is expressed $\hat{P}_k = d_k/n_k$ and $\hat{D}_k = N_k\hat{P}_k$ gives an unbiased estimate of D_k. The estimator of P is expressed

$$\hat{P} = \frac{\sum_s \frac{\hat{D}_k}{\pi_k}}{\sum_s \frac{\hat{N}_k}{\pi_k}}. \tag{2.1}$$

This is the ratio of the unbiased estimators of D and N, the total number of forms. Although this number is known, estimator (4.1) is obviously more accurate than $1/N \sum_s \hat{D}_k/\pi_k$.

We have

$$\operatorname{Var}(\hat{P}) = E \operatorname{Var}(\hat{P} \mid s) + \operatorname{Var} E(\hat{P} \mid s). \tag{2.2}$$

Now

$$\operatorname{Var}(\hat{P} \mid s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right)$$

where $$\hat{N} = \sum_s \frac{N_k}{\pi_k}.$$

Hence

$$E \operatorname{Var}(\hat{P} \mid s) \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right). \tag{2.3}$$

Furthermore,

$$E(\hat{P} \mid s) = \frac{\sum_s \frac{D_k}{\pi_k}}{\sum_s \frac{N_k}{\pi_k}}.$$

The variance of this value is obtained by linearization following introduction of variable $Z_k = D_k - PN_k = N_k(P_k - P)$.

We obtain

$$\operatorname{Var} E(\hat{P} \mid s) \simeq N^{-2} \operatorname{Var}\left(\sum_s \frac{z_k}{\pi_k}\right).$$

Taking into account that $\sum_U Z_k = 0$:

$$\operatorname{Var} E(\hat{P} \mid s) = N^{-2}\left(\sum_k \frac{Z_k^2}{\pi_k} + \sum\sum_{k \neq l} \frac{Z_k Z_\ell}{\pi_k \pi_\ell}\pi_{k\ell}\right). \tag{2.4}$$

The sum of (2.3) and (2.4) gives us the variance of estimator (2.1).

2.1 Introduction of a Model

Not only is the variance of $\hat{P}$ difficult to manipulate, it contains unknown parameters. The problem may be circumvented by formulating the hypotheses required to produce a superpopulation model. It is assumed below that the parameters of this model may be estimated from the results of a preliminary test covering a very small portion of the population. In the model, expectation is denoted by E_ξ (variance by Var_ξ) and all the random variables are assumed independent of the sampling process.

The model has the following specifications:

(a) D_k has a binomial distribution (N_k, p_k). *In the model,* P_k is thus an estimator of p_k.

(b) p_k is itself random; we assume p_k to be independent and have the same distribution, with

$$E_\xi p_k = P,$$

$$\operatorname{Var}_\xi p_k = \sigma^2$$

for any k, in particular whatever the value of N_k.

In the model, after conditioning with p_k, we obviously have

$$E_\xi(D_k \mid p_k) = N_k p_k,$$

$$\operatorname{Var}_\xi(D_k \mid p_k) = N_k p_k(1 - p_k).$$

The *anticipated variance* of $\hat{P}$ is $E_\xi \operatorname{Var}\hat{P}$, to which we now turn our attention. For its evaluation, we denote

(a) $E_\xi(P_k - P)^2 = E_\xi(E_\xi(P_k - p_k + p_k - P)^2 \mid p_k)$

$$= \frac{P(1-P) - \sigma^2}{N_k} + \sigma^2,$$

(b) $E_\xi P_k(1 - P_k) = E_\xi(E_\xi((P_k - P_k^2) \mid p_k))$

$$= E_\xi p_k(1 - p_k)\frac{N_k - 1}{N_k}$$

$$= (P(1 - P) - \sigma^2)\frac{N_k - 1}{N_k},$$

(c) $E_\xi Z_k Z_{\ell'} = 0$, because of the independence of Z_k and $Z_{\ell'}$, clearing one extremely cumbersome term and $\pi_{k\ell'}$.

When we combine all the pieces of (2.3) and (2.4), a minor algebraic miracle occurs, producing the expression

$$E_\xi \mathrm{Var}\,\hat{P} \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \quad . \quad (2.1.1)$$

where $\tau^2 = P\,(1 - P) - \sigma^2$
(by nature a positive quantity)

Comment:

The algebraic miracle is easily explained if we are not seeking the variance in the sole context of sample design. It is in fact the result of a model slightly more general than the one suggested.

Suppose we wish to estimate the total $N\bar{Y} = \sum_U Y_i$ of a variable Y and suppose that, to this end, a two-stage sample is drawn: in the first stage, primary units k are drawn with π_k probability and, in the second, n_k final units are drawn by simple random sampling.

We are assuming a model in which:

$$Y_i = \bar{Y} + \alpha_k + \epsilon_i,$$

with α_k a variable linked to the PU of index k. α_k is independent, subject to the same zero expectation and has a variance σ^2. ϵ_i is also independent, centred and has a variance τ^2. With $\pi_i^* = \pi_k n_k / N_k$ (N_k = size of PU number k), the Horvitz-Thompson estimator of the *total* is $\hat{Y} = \sum Y_i/\pi_i^*$, the sum being extended to the sample. In the model, and conditionally in the sample, we have

$$\mathrm{Var}_\xi(\hat{Y} \mid s) = \sum_s \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right).$$

For this expression, expectation is again expressed in the form of equation (2.1.1).

2.2 Search for an Optimum Sample Design

The maximum variance of $\hat{P}$ is set by the criteria selected for quality control. As the survey is repeated for each processing unit, it is only natural to seek to minimize the expected survey cost given in (2.1.1), *i.e.*

$$E \sum_s (C_o + n_k C_1) = \sum_U \pi_k (C_o + n_k C_1). \quad (2.2.1)$$

The problem of optimization is expressed as:

$$\text{To minimize } \sum_U \pi_k (C_o + n_k C_1)$$

with the constraints

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \leq V_o$$

and for any k, $n_k \leq N_k$.

Let us now apply a Lagrange multiplier λ to the first constraint - which will obviously be saturated - and multipliers μ_k to the others. We obtain the solutions

$$C_o + n_k C_1 = \lambda \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \quad (2.2.2)$$

and, for any k:

$$C_1 \pi_k = \lambda \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \quad (2.2.3)$$

with

$$\mu_k = 0 \quad \text{if} \quad n_k < N_k \quad \text{and} \quad \mu_k > 0 \quad \text{if} \quad n_k = N_k.$$

For the use of Lagrange multipliers, see for example Luenberger (1973).

For all primary units in which $\mu_k = 0$ (the largest), we obtain

$$n_k = \frac{\tau}{\sigma}\left(\frac{C_o}{C_1}\right)^{1/2} = n^*. \quad (2.2.4)$$

Each primary unit receives the same allocation, which corresponds to the consistent accuracy principle. Going back to equation (2.2.3), we observe that, again for these primary units, the probability of inclusion π_k must be proportional to size N_k, *i.e.*

$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\tau}{n^*} N_k. \quad (2.2.5)$$

This is the standard proof of a self-weighting one-stage survey in which the first stage is drawn with probabilities proportional to a measure of size. (See for example Cochran 1977).

Since n_k is independent of N_k, it is impossible to have $n_k = N_k$ or $\mu_k > 0$ unless $N_k \leq n^*$. Equation (2.2.2) gives us the probability of inclusion to within one factor:

$$\pi_k = \lambda^{1/2} N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k}\right)^{1/2} = \lambda^{1/2} N_k^{1/2} \left(\frac{N_k\sigma^2 + \tau^2}{N_k C_1 + C_o}\right)^{1/2}. \tag{2.2.6}$$

Relations (2.2.5), valid if $N_k \geq n^*$, and (2.2.6) valid if $N_k \leq n^*$, establish that π_k is proportional to a known variable $T_k = f(N_k)$, for which the graph is given in Figure 1.

To fully define the survey, the number m of primary units to be drawn must still be set. $T = \sum_U T_k$ is also a known quantity.

If we restrict ourselves to fixed size sampling, we have $\pi_k = m\, T_k/T$. m may be determined by importing this value into the variance constraint, *i.e.*

$$N^2 V_o m = T \sum_U \frac{N_k^2}{T_k}(\sigma^2 + \tau^2/n_k).$$

If, as a first approximation, assuming $T_k = N_k$, we obtain the simplified form:

$$m V_o = \sigma^2 + \tau^2/n^*.$$

We now have a full solution to the problem.

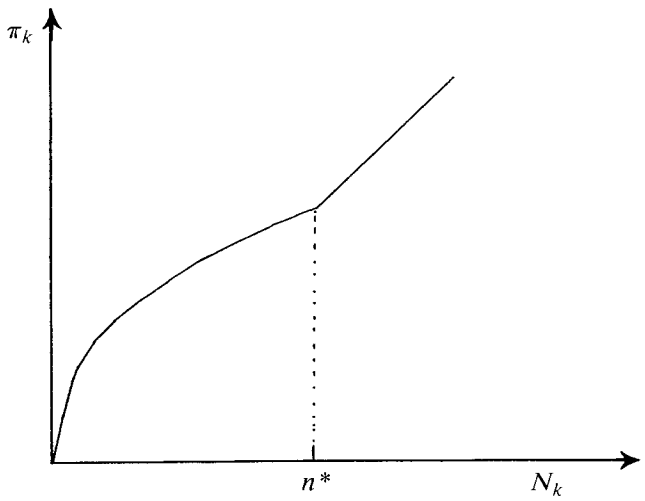

Figure 1. Graph of π_k as a function of N_k

3. OPTIMUM ESTIMATE FOR A TWO-STAGE SURVEY IN WHICH THE PRIMARY UNITS ARE STRATIFIED

The harsh facts of the situation complicate the problem somewhat: because a number of types of forms must be controlled separately, a fairly general problem, described below, arises.

For each primary unit (a district in a processing unit) we know the population N_{kg} of secondary units belonging to G groups. The "population" of PU number k is $N_{k+} = \sum_g N_{kg}$; that of group g is $N_{+g} = \sum_k N_{kg}$. As described above, we are looking for the probability of inclusion π_k with which to sample PU number k, the number of PUs to be drawn and the allocation n_{kg} of the sample among the various groups in PU k, knowing that these n_{kg} units are drawn by simple random sampling from among the N_{kg} units available.

3.1 Search for an Optimum Model Assisted Design

In each group, we postulate a model identical to the one formulated in section (2.1) (or the more general form described in the comment on that section).

For $g = 1$ to G, we have therefore:

$$v_g = E_\xi \operatorname{Var}(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k}(\sigma_g^2 + \tau_g^2/n_{kg}). \tag{3.1.1}$$

The cost function is expressed in the general form (1.2). We are seeking to minimize the expected survey cost

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g\right), \tag{3.1.2}$$

under constraints $V_g \leq \mathcal{V}_g$, where quantities $\mathcal{V}_g$ are externally fixed (*e.g.* quality of data to be obtained, tightness of control.

In this form, the problem can prove fairly complex. We write a general form of a Lagrange multiplier:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

The problem sets $\lambda = 1$, λ_g being multipliers to be determined. In a simple variant, values are set for λ_g: we wish to minimize a given linear combination of variances under a cost constraint. In all the hypotheses, by differentiation with respect to n_{kg} (considered a real variable), we obtain

$$\lambda \pi_k^2 C_g = \lambda_g N_{+g}^{-2} N_{kg}^2 \tau_g^2/n_{kg}^2. \tag{3.1.3}$$

π_k being for the moment to be defined to within one factor, we may write

$$\pi_k \, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}}. \qquad (3.1.4)$$

By summing over k, we deduce that

$$E \, n_{+g} = \sum_U \pi_k \, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g. \qquad (3.1.5)$$

The total size of the sample in each group is thus directly linked to multiplier λ_g.

Differentiation of the Lagrange multiplier with respect to π_k gives us new relations which, when combined with (3.1.4), are miraculously simplified to give

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g}\right)^2 n_{kg}^2, \qquad (3.1.6)$$

or, if we introduce the numbers

$$n_g^* = \left(\frac{C_o}{C_g}\right)^{1/2} \frac{\tau_g}{\sigma_g},$$

we write

$$\sum_g \left(\frac{n_{kg}}{n_g^*}\right)^2 = 1. \qquad (3.1.7)$$

As may be seen in equation (2.2.4), n_g^* is the number of secondary units to be drawn per PU if there is a single group; n_{kg} is always less than n_g^*.

From (2.1.4), (3.1.5) and (3.1.7) we obtain the relations:

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}}\right)^2. \qquad (3.1.8)$$

Thus, π_k is proportional to T_k such that $T_k^2 = \sum_g \lambda_g \sigma_g^2 N_{kg}^2/N_{+g}^2$, which appears to be a satisfactory measure of size. The relations (3.1.4) show that, if k is fixed, n_{kg} is proportional to $n_g^* \lambda_g^{1/2} \sigma_g N_{kg}/N_{+g}$; taking into account (3.1.7), we obtain

$$n_{kg} = n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} T_k^{-1}. \qquad (3.1.9)$$

3.2 Explicit Solutions to Two Specific Cases

(a) If λ_g were known, *i.e.* if $\sum_g \lambda_g v_g$ were minimized under a cost constraint, then (3.1.2) and (3.1.9) could be used to calculate T_k. By transfering

$$\pi_k = m \, T_k/T \left(T = \sum_U T_k, \; m \; \begin{matrix}\text{number of primary} \\ \text{units to be drawn}\end{matrix}\right)$$

to budget constraint $C_T \leq C_T^*$, we find that

$$C_T^* = \frac{m}{T} \sum_U \left(C_o T_k + \sum_g C_g n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}}\right)$$

i.e.

$$m = C_T^* \Big/ \left(C_o + \sum_g C_g n_g^* \cdot \frac{\lambda_g^{1/2} \sigma_g}{T}\right).$$

If a single λ_g is not equal to zero, it is fairly easy to check that the result is the one given at the end of section (2.2).

(b) The initial problem (min C_T under $V_g \leq \mathcal{V}_g$) is resolved fairly easily in two specific cases.

b1 - *Maximum dispersion* of the groups. For any PU k, we have $N_{kg} = N_{k+}$ for a given k. The problem is broken down into G separate problems, each being of the type examined in section 2.

b2 - *Minimum dispersion*. The distribution is the same in all the PUs; in other words, for any k and any g, we have

$$N_{kg} = N_{k+} \frac{N_{+g}}{N} \quad \text{with} \quad \left(N = \sum_g N_{+g}\right),$$

T_k is then proportional to N_{k+}, and n_{kg} is quantity $n_g^* u_g$ independent of k.

With $\pi_k = mN_{k+}/N$, we obtain by writing $V_g = \mathcal{V}_g$:

$$m\mathcal{V}_g = \sigma_g^2 + \tau_g^2 / n_g^* u_g$$

i.e.

$$m = \frac{\sigma_g^2}{\mathcal{V}_g} + u_g^{-1} \frac{\tau_g^2}{n_g^* \mathcal{V}_g}.$$

Thus we obtain G-1 linear relations between the u_g^{-1}, in principle permitting full resolution of the problem, knowing that the sum of u_g^2 is equal to 1.

3.3 A Numerical Algorithm for Determining the Optimum Solution to the General Case

An iterative numerical resolution of the problem may be achieved as follows.

Step 1: An approximate sample allocation is set in each group (n_{+g} units in group g). The process may be facilitated by using the approximate solution based on the hypotheses in point (a) or point (b).

Step 2: The value of λ_g is determined from relations (3.1.5):

$$\lambda_g = C_g n_{+g}^2 / \tau_g^2.$$

Step 3: π_k is determined from relations (3.1.8). Specifically, the sum of π_k sets the number of PUs to be drawn.

Step 4: n_{kg} is determined from relations (3.1.4). Subsequent iteration is possible by returning to step 2, in the expectation that the algorithm will converge toward the optimization solution.

Comment: The probability of drawing a type g unit is

$$\pi_k n_{kg} \Big/ N_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g / N_{+g}.$$

Because it does not depend on primary unit k, it is the same for each unit in a given group g (equal probability survey). Size n_{+g}, or at least its mathematical expectation, may be deduced from the sample in group g. In practice, sample size is sometimes set arbitrarily: this entails determining λ_g or, implicitly, variances $\mathcal{V}_g$. This is another fairly common result.

4. OPTIMUM ESTIMATE ASSISTED BY A MEASURE OF THE DIFFICULTY OF CODING A RECORD

The task is to estimate the proportion of forms containing a coding error in universe U of all forms coded in a given week by one regional branch. The problem is identified by the following characteristic: because all IFs are precoded, it is possible, using information drawn from the trial census, to attribute to each one a positive numerical variable X_i representing its "difficulty". This variable is calibrated in such a way that Y_i (equal to 1 if there is an error and 0 if there is not) has an "expectation" proportional to X_i.

The same cost control considerations suggest a two-stage survey.

- In the first stage of the survey, we draw a sample s_1 of districts k (primary units), with π_k unequal probabilities to be determined. $\pi_{k\ell}$ denotes the probability of inclusion, double in value in this instance.
- In the second stage of the survey, a sample s_k of final units (forms) in primary unit sample k is drawn. $\pi_{i|k}$ denotes the probability of inclusion of the unit in primary unit k, $\pi_{ij|k}$ the probability of inclusion of the pair (i,j) in the primary unit; and $s = U_{k\in s_1} s_k$, the sample of final units.

$X_k = \sum_{i\in k} X_i$ denotes the total of X_i in primary unit k, $X = \sum_{k\in U_o} X_k = \sum_U X_i$ and similar notations are used for all the variables. (U_o denotes the population of primary units - districts, U the population of final units - forms).

The aim is to estimate a quantity in the form $R = \sum_U Y_i / \sum_U W_i$ where W_i is a known variable for each form. This may be $W_i = 1$ or $W_i = X_i$, whichever measure of the error rate seems the more satisfactory.

4.1 Selection of Estimator and Variance

(a) For primary unit k, the total Y_k of the Y_i for $i\in k$ is commonly estimated by the ratio

$$\hat{Y}_k = X_k \left(\sum_{s_k} Y_i/\pi_{i|k} \right) \Big/ \left(\sum_{s_k} X_i/\pi_{i|k} \right) = X_k \hat{a}_k$$

where $\hat{a}_k$ estimates $a_k = Y_k/X_k$ with a slight biais.

(b) To estimate ratio Y/X, we use

$$\hat{a} = \frac{\sum_{s_1} \frac{\hat{Y}_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}.$$

(c) If we wish to estimate R, we note that

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

where X and W are known totals (*e.g.* total difficulty, total number of forms). As variable X_i was selected for its good correlation with Y_i, an *a priori* valuable estimator of R is

$$\hat{R} = \hat{a} \frac{X}{W}$$

and the only real question concerns the estimate of $a = \sum_k a_k X_k / X$.

(d) we have

$$\operatorname{Var}(\hat{a}) = \operatorname{Var} E(\hat{a} \mid s_1) + E \operatorname{Var}(\hat{a} \mid s_1).$$

For the first term, taking into account the fact that $\hat{a}_k$ is an approximate unbiased estimator of a_k, we may write

$$\operatorname{Var} E(\hat{a} \mid s_1) \simeq \frac{1}{X^2} \operatorname{Var} \left(\sum_{s_1} \frac{(a_k - a)X_k}{\pi_k} \right)$$

$$= \frac{1}{X^2} \left(\sum_k \frac{(a_k - a)^2 X_k^2}{\pi_k^2} \right.$$

$$\left. + \sum\sum_{k \neq l} (a_k - a)(a_l - a) \frac{X_k X_l \pi_{kl}}{\pi_k \pi_l} \right). \quad (4.1.1)$$

For the second term, *conditional on* s_1, we have

$$\text{Var}\left(\frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}\right) = \left(\sum_{s_1} \frac{X_k}{\pi_k}\right)^{-2} \cdot \sum_{s_1} \text{Var}(\hat{a}_k) \frac{X_k^2}{\pi_k^2}.$$

For this quantity, the expectation is approximately

$$X^{-2} \sum_k \text{E Var}(\hat{a}_k \mid s_1) \frac{X_k^2}{\pi_k}, \qquad (4.1.2)$$

with

$$\text{Var}(\hat{a}_k \mid s_1) = \text{Var} \frac{\sum_{s_k} \frac{Y_i}{\pi_{i|k}}}{\sum_{s_k} \frac{X_i}{\pi_{i|k}}} \simeq \frac{1}{X_k^2} \text{Var} \sum_{s_k} \frac{Y_i - a_k X_i}{\pi_{i|k}}$$

$$= \frac{1}{X_k^2}\left(\sum_{i \in k} \frac{(Y_i - a_k X_i)^2}{\pi_{i|k}}\right.$$

$$\left. + \sum\sum_{k \neq l} \frac{(Y_i - a_k X_i)(Y_j - a_k X_j)\pi_{ij|k}}{\pi_{i|k}\, \pi_{j|k}}\right).$$

As in the preceding sections, we arrive at formulae that are complex and, in the final analysis, unusable. A model will simplify things somewhat.

4.2 Introduction of a Model

The model has the same structure as those used previously:

(a) a_k is an independent random variable with the same expectation and the same variance:

$$E_\xi a_k = a \qquad \text{Var}_\xi a_k = \sigma^2.$$

The variance takes into account operator influence, which we make no attempt to isolate, and also such factors as day of the week, time of day, day of the month *etc.*...

(b) Conditional on a_k, Y_i in primary unit k is an independent Bernoulli variable with $E_\xi(Y_i \mid k) = a_k X_i$

$$\text{Var}_\xi(Y_i \mid k) = a_k X_i - a_k^2 X_i^2.$$

Comment:

Variable X_i, which has no actual concrete meaning, is defined to within one factor of scale. Conversely aX_i and σX_i, being probabilities, have an invariant physical interpretation. In what follows, one must always keep in mind that the results are invariant if X_i is multiplied by an arbitrary factor, on condition that a and σ are divided by the same factor. $\text{Var}(\hat{a})$ in particular has no concrete meaning; $\text{Var}(\hat{a}X)$ is an exception.

As before, we examine anticipated variance, expectation under the model of the sum of (4.1.1) and (4.1.2).

For the first term, the expectation of the cross products is of course zero. The expectation under the model for this term is thus:

$$X^{-2}\sigma^2 \sum_k \frac{X_k^2}{\pi_k}.$$

For the second term, we find (in light of the definitions given in 4.2.a and 4.2.b)

$$X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i E_\xi \frac{(a_k X_i - a_k^2 X_i^2)}{\pi_{i|k}}$$

$$= X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

Therefore, overall

$$E_\xi \text{Var}(\hat{a}X) = \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k}$$

$$+ \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

No algebraic miracle occurs here. *For simplification*, we assume that $(a^2 + \sigma^2)X_i^2$ is negligible in the face of aX_i. Numerically, we may expect $aX_i = 2$ to 5×10^{-2} and $(a^2 + \sigma^2)X_i^2 = 3$ to 30×10^{-4}: whence the approximation

$$E_\xi \text{Var}(\hat{a}X) \simeq \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k} + a \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{X_i}{\pi_{i|k}}.$$

4.3 Sample Design Optimization

We use the following cost function:

$$C = \sum_{s_1} (C_o + C_1 n_k).$$

Here, $n_k = \sum_{i\in k} \pi_{i|k}$ is the size of the sample drawn in district k (supposedly set at fixed size s_1). Its expectation is

$$C_T = \sum_{k\in U_o} \pi_k (C_o + C_1 n_k).$$

Let

$$\pi_{i|k} = n_k P_i \left(\text{with } \sum_{i\in k} P_i = 1\right) \quad \text{and} \quad Q_k = \pi_k n_k.$$

The problem of optimization is now

$$\text{Min: } C_o \sum_k \pi_k + C_1 \sum_k Q_k$$

$$\text{under: } \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i\in k} \frac{X_i}{P_i} \leq \mathcal{V}_o.$$

In this form, we are pleased to observe that the terms in $\sum_i X_i/P_i$ may be minimized independently of the other terms. In other words, n_k has no impact on this term. Leaving optimization of the second stage of the survey until later, S_k^{*2} denotes the optimized value of $\sum_i X_i/P_i$.

With a Lagrange multiplier λ, by differentiation with respect to π_k and Q_k, we obtain

$$^*C_o = \lambda\sigma^2 \frac{X_k^2}{\pi_k^2} \quad \text{i.e.} \quad \boxed{\pi_k \text{ proportional to } X_k} \quad (4.3.1)$$

$$^*C_1 = \lambda a \frac{S_k^{*2}}{Q_k^2} \quad \text{whence} \quad \boxed{n_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X_k}}. \quad (4.3.2)$$

Specifically, the primary units are drawn with probabilities proportional to total difficulty, a standard resolution (see for example Särndal, Swensson, Wretman, 1992, Chapter 12).

We now move on to sub-district sampling (second stage of survey).

Beginning with a simple, straightforward case, forms are drawn one by one. Minimization produces P_i proportional to $\sqrt{X_i}$. A simple calculation shows that $S_k^* = \sum_{i\in k} \sqrt{X_i}$. We can now calculate n_k using (4.3.2), and our problem is fully resolved.

In practice, things are more complicated. For fairly obvious reasons, only forms for entire households are selected. In other words, the second stage of the survey is a *cluster* survey. The values of P_i are the same (*i.e.* P_m) for all the members of a given cluster (household) m.

Let X_m be the sum of X_i individuals i in household m. The problem is to minimize $\sum X_m/P_m$ under $\sum n_m P_m = 1$, with n_m the size of household m. We easily reach solution

$$P_m = \sqrt{\bar{X}_m} \Big/ \sum n_m \sqrt{\bar{X}_m},$$

with $\bar{X}_m = X_m/n_m$, *mean difficulty of forms IF in household m*. From this we determine $S_k^* = \sum n_m \sqrt{\bar{X}_m}$.

This solution enables us to determine the number n_k of *final units* to be drawn using (4.3.2). However, the number of clusters (*households*) has not been determined: this snag was predictable. In fact, the cost function does not imply this constraint. To obtain the number m_k of clusters to be drawn, we arrange matters so that the expectation of the number of final units is equal to n_k. Thus,

$$m_k \left(\sum n_m \sqrt{\bar{X}_m}\right) \Big/ \sum \sqrt{\bar{X}_m}$$

whence

$$m_k = n_k \frac{\sum \sqrt{\bar{X}_m}}{\sum n_m \sqrt{\bar{X}_m}}.$$

Taking into account (4.3.2), we also have

$$m_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{\sum \sqrt{\bar{X}_m}}{X_k}$$

and the probability a given household being drawn is thus

$$\frac{m_k \sqrt{\bar{X}_m}}{\sum \sqrt{\bar{X}_m}}.$$

Following a number of algebraic manipulations, the value of the optimum variance is found to be:

$$E_\xi \operatorname{Var} (\hat{a}X)_{\text{OPT}} = \frac{(\sigma X)^2}{m} \left(1 + \frac{a}{\sigma} \frac{a^{-1/2} S^*}{X} \left(\frac{C_1}{C_o}\right)^{1/2}\right).$$

This form respects the homogeneous character of the different factors. In particular, we have $a^{-1/2} S^*/X = a^{1/2} S^*/aX$: the denominator may be interpreted as total number of errors in a lot; the numerator is homogeneous for a given size.

We now have a full solution to the problem.

Comment 1:

In both cases discussed, S_k^* is multiplied by $C^{1/2}$ if X_i is multiplied by C. The formula that gives n_k is thus invariant on the scale of measurement.

Comment 2:

The solution that entails drawing clusters favours small clusters made up of final units with a high index of difficulty.

Comment 3:

As in preceding sections, we determine the probability of single selection, but not the probability of dual selection. Therefore the algorithm for the draw, which sets the latter, has no influence. This is quite common, keeping in mind that the complementary data used to optimize the draw determines π_k and $\pi_{i|k}$ but have no influence on dual probabilities.

5. APPLICATIONS TO CONTROL BY SURVEY OF THE QUALITY OF THE 1990 FRENCH CENSUS

5.1 Problem of Data Capture Control

The sampling techniques described in sections 2 and 3 were designed to control data capture for the 1990 Census. A brief description of the operation would enhance understanding of the nature of the statistical problems involved.

The basic collection unit is the district, which corresponds, in a city, to a block of houses and, in the country, to a village or group of hamlets. It covers a population that ranges from zero inhabitants to approximately 2,000 (the mean values are 150 dwellings and approximately 350 inhabitants).

When collection is completed and the results are audited, the various census forms (specifically individual forms (IF) and dwelling forms (DF)) are meticulously counted for each district. The summary data for a district are computerized; the forms themselves, collated into district files, are forwarded to data capture.

Groups of districts comprising approximately 100,000 dwellings are constructed. The processing units (PU) are processed for INSEE by contractors. INSEE, the "client" in terms of control theory, monitors the quality of each contractor's work by sampling a specific number of forms in each PU.

The aim of the survey described in paragraph 2 is to estimate, to an accuracy (standard deviation) of one point, the proportion of forms containing an error in each PU. The maximum proportion of forms containing an error cannot exceed 4%. A trial census covering approximately 400 districts allows for an estimate of the values of the two model parameters. We find:

$$\sigma^2 \simeq P^2 \simeq 14.10^{-4}$$

$$\tau^2 \simeq P \simeq 4.10^{-2}.$$

Cost function (1.1) is assessed in terms of working time. Based on on-site control measures, 5 minutes is the estimate of the time required to process one district folder (from the time it is taken from the shelf to the time it is returned there) and 30 seconds the estimate of the time required to process one IF. With the numerical data, design optimization based on the hypotheses in section 1 allows for control of 40 districts per processing lot and 16 forms per district.

After discussing the solution with the team responsible for the census, it emerged that two types of documents (individual forms (IF) and dwelling forms (DF)) were to be controlled. The first approximation had taken no account of the latter, which are less likely to contain errors and take only about half as long to code as IFs. However, some districts (*e.g.* a commune with a thriving tourist industry) contain a large majority of secondary dwellings, and so produce many DFs but very few IFs. Because the situation required in-depth study, the theory given in section 3 was developed.

In the case of the census, the number of groups G is equal to 2 ($g = 1$ for the IFs and $g = 2$ for the DFs). The numerical data for the two groups are:

$\cdot\ P_1 = 0{,}04 \quad \sigma_1 = P_1 \quad \tau_1^2 = P_1(1 - P_1) - \sigma_1^2 = P_1 - 2P_1^2,$

$\cdot\ P_2 = 0{,}01 \quad \sigma_2 = P_2 \quad \tau_2^2 = P_2 - P_2^2,$

$\cdot\ \nabla_1 = (0{,}0075)^2 \qquad \nabla_2 = (0{,}0150)^2.$

For the cost function, we selected $C_o = 5$ minutes, $C_1 = 0.5$ minute and $C_2 = 0.25$ minute. Optimization of the problem according to the hypotheses in section 3.2.b entailed examining 73 districts per processing unit. In practical terms, it meant processing 15 individual forms (and related DFs) for each district. For the districts that produce fewer than 15 IFs, all IFs were processed. For districts with zero IFs, 4 DFs were processed (if this number was less than the number of DFs in the district).

Comment:

The method described in part 2 seems to have a fairly broad field of application. One example: it was used to sample the 1992 French survey on migration of foreign nationals. For population centres with under 20,000 inhabitants, the sample was drawn in two stages. The first stage of the survey covered the 90 departments in which this type of population centre occurs. The foreign population (based on the census) was divided into 8 nationality groups, for which equally accurate indicators had to be found.

5.2 Problems Related to Coding

The second step in data preparation is known as operation COLIBRI (Codification en Ligne des Bulletins du Recensement des Individus). The operators in the regional branches of INSEE receive forms classified by district and code them for the 25% survey.

In practice, each operator works at a monitor that displays the identifier of the next dwelling to be included in the 25% sample, for which all IFs must be coded.

Coding quality is also controlled by survey. The control unit is all the work done in one week in a regional branch. The entire operation takes a little over one year in the 22 regional branches, and entails more than 1,000 surveys. The household is the unit to be controlled (*i.e.* all the IFs in a household drawn for inclusion in the control sample). The objective is to estimate the proportion of forms containing an error. This is done by automatic detection of forms in which there is a no match situation. The number of errors is determined by reconciliation. The control theory is discussed in section 4 of this paper. The index of difficulty of the forms was developed from the data captured for a study based on the previous census and by test. The procedure and results related to these control measures are described in detail in G. Badeyan (1992).

The practical and numerical application of the theory rests on hypotheses concerning the orders of magnitude of the different parameters (which requires linking them to a simple physical interpretation). In the census preparation phase, without accurate prior measurement, we used the values $\sigma/a = 0.5$ and $C_1/C_o = 0.1$.

Pursuant to a number of hypotheses concerning the other parameters, and after discussing the matter with experts, it was decided that the control would cover 50 districts, with approximately 20 IFs controlled in each one (by region and by week). Since model parameters can be re-estimated at any stage in the process, the initial order of magnitude can obviously be adjusted as the survey proceeds.

Final Comment:

The problem produces somewhat surprising results that are worthy of consideration.

In the first instance, as we assumed it would be possible to separate each form, the forms were drawn with a probability proportional to individual difficulty. We assumed, to some extent, that the cost of using individual information was zero.

In the second instance, the actual control process, it was assumed that cost was infinite and the only information with negligible cost was the information related to an entire household. The solution shows that the probability of drawing an individual (IF) as a function of the mean difficulty of coding the forms for the entire household of which the individual is a member.

The same phenomenon occurs in the district draw. If it is possible to separate the IFs, they are drawn with probabilities proportional to total difficulty; within a district, the difficult IF has a greater probability of selection. Conversely, suppose we are unable to separate IFs within a district. This will be the case, for example, if the designation of IFs to be controlled cannot be implemented in real time because of inadequate processing facilities. Districts would then be selected in proportion to mean difficulty: within a district, it would be necessary to proceed by simple random sampling.

In the first instance, the survey gives precedence to large districts, from which difficult IFs tend to be drawn. In the second instance, precedence is given to small difficult districts, from which forms are selected with equal probability. *In both instances*, we are seeking to increase the probability of surveying difficult IFs. The difference resides simply in the possibility (*i.e.* the cost) of collecting information when we need it.

ACKNOWLEDGEMENTS

The author would like to thank the Editor, the Associate Editor and the Referee for their extremely positive comments. He would also like to thank Claude Thelot, some of whose comments have been incorporated into this paper and Gérard Badeyan, who introduced the methods discussed here at INSEE. Last, but not least, he would like to thank Françoise Hitier, without whom this paper would never have seen the light of day.

REFERENCES

BADEYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, June 17 and 18, 1992, INSEE, Paris.

COCHRAN, W. (1977). *Sampling Techniques*, (3rd Edition). New York: Wiley.

DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.

LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.