A GENERALIZED SAMPLING PROGRAM FOR THE CURRENT SURVEYS IN THE INDUSTRY DIVISION

Heinrich Mantel and Preston J. Waite, Bureau of the Census

A. Introduction

This paper describes the methodology and use of a generalized sampling program. This program has been developed for the selection of samples for current surveys of the Industry Division of the Census Bureau. Sections B, C, and D give a brief discussion of the types of universes about which information is required, along with important characteristics of the methodology used by the program, and some of the additional features of the program. In section E, the paper discusses estimation methods, measure of size determination, and the logic of the generalized sampling program.

B. Nature and Objectives of the Current Industrial Survey Sampling Program

The primary objective of most of the Census Bureau's current industrial survey samples is to provide estimates of change at a detailed product level. The frequency of reporting for these surveys is either monthly, quarterly, or annually. Each survey is multivariate in the sense that several items are estimated simultaneously. A typical survey may provide estimates of materials consumed, sales of finished goods, production of finished goods, stocks on hand, and new orders.

From a conceptual point of view, the universe for one of these surveys can be considered as the union of the universes of each of the items within the scope of the survey. All companies or establishments producing one or more of the items within the scope of the survey comprise the universe of sampling units for the survey. The distributions and reliability requirements for individual item universes may differ greatly within the same survey, so it is necessary that any sampling program be sufficiently flexible to meet the requirements of each item individually, as well as the overall requirements of the survey.

The universes of the individual items for most of these surveys have an extremely skewed distribution. This skewness suggests that a probability-proportionate-to-size (PPS) sampling procedure is an efficient approach, assuming a reasonable size measure is available and "correlates well" with the items of interest. This feature applies only to the univariate case and in the initial analysis where "control items" (the definition will be supplied later) are analyzed independently.

The sample design utilized in the program is an application and a simple extension to the multivariate case of theory developed by Jack L. Ogus and Donald F. Clark in a technical paper prepared at the Census Bureau.[1] An important aspect of the design is that each sampling unit is given an independent chance of selection. This method of selection has been termed Poisson sampling.

Some features of Poisson sampling are:

1. Poisson sampling provides flexibility in drawing nonduplicating samples that can yield unbiased estimates. This feature is particularly useful in light of the Census Bureau's continuing efforts to reduce the respondent burden of small companies.

2. Poisson sampling contributes to simplicity in analytical and operational procedures needed to estimate totals, period-to-period changes, and their standard errors.

3. While there is some increase in the variance caused by the variability of the sample size under Poisson sampling, this is not considered to be a major disadvantage, and it is partially compensated for by the use of the difference estimate (see section E).

C. Methodological Consideration

The complexity of the sample design can be reduced by identification of a subset of items for which reliability requirements are specified. Every item from this subset will be referred to as a control item. Since probabilities of selection are a function of only the control items, every sampling unit must have at least one control item. Given the reliability specification for every control item, the program also computes the predicted reliability for each noncontrol item. The designation of the control items is one of the most important parts of the sample design.

The objective of the program is to satisfy the reliability requirements for all control items and to take advantage of the correlations among the control items to improve the efficiency of the sample design. To achieve this, the following major steps are necessary:

1. Each control item i is analyzed independently and an "optimum" probability of selection (one which minimizes the expected cost per unit of information over the noncertainty stratum subject to satisfying the reliability requirement) can be assigned to every sampling unit h producing item i.

2. Let I_h be the set of control items produced by sampling unit h. Then for every item $i \in I_h$, we have by step 1, an assignment of "item probability," $P_{i,h}$. Now, if I_h contains more than one element, then we must define a rule (called the composition rule) to determine a unique probability of selection P_h which is consistent with the objective of the generalized

sampling selection program.

The following constraints on the rule of composition guarantee that the variance specifications for each control item will not be exceeded:

a. $\max_{i \in I_h} \{P_{i,h}\} \leq P_h \leq 1$. This can be phrased in everyday language in a simple way: The probability of selection for sampling unit h should be at least as large as the probability of selection of sampling unit h when only one of its control items is considered.

This rule implies $\frac{1}{P_h} \leq \frac{1}{P_{i,h}}$ for every $i \in I_h$, and this, in turn, implies that the variance contribution of sampling unit h to each item $i \in I_h$ can only decrease. However, this reduction in variance is associated by an increase in the expected sample size for item i by $(P_h - P_{i,h})$.

An important consequence of a, above, is that if the sampling unit is assigned an "item probability" $P_{i,h} = 1$ for at least one control item i, then $P_h = 1$.

b. The composition rule should be a symmetric function of its arguments (the item probabilities). This ensures that the probability of selection of the sampling unit does not depend on the order in which the items are analyzed.

Presently, we have implemented two composition rules:

$$P_h = \max_{i \in I_h} \{P_{i,h}\}, \text{ and } P_h = 1 - \prod_{i \in I_h} (1 - P_{i,h}).$$

Each of these two rules has an additional important property: If $P_h = 1$, then there exists at least one item, $i \in I_h$, such that $P_{i,h} = 1$. This property combined with a, above, implies the equivalence: $P_h = 1$ if, and only if, there exists at least one item i such that $P_{i,h} = 1$.

Following, is a method for deriving new composite rules from an existing one.

Let $P_h^{(1)}$ and $P_h^{(2)}$ be two composite probabilities. Then, for every $0 \leq \alpha \leq 1$, $\alpha P_h^{(1)} + (1 - \alpha) P_h^{(2)}$ is a new composite probability.

3. It is quite common that for a given survey, there is at least one sampling unit h which produces at least two control items ("multiproducer"). For such items, by using the composite rule, we obtain a smaller variance than specified and a larger expected sample size than needed to satisfy the variance requirements.

To overcome this loss of efficiency, an adjustment procedure is required. The adjustment procedure used in the generalized sampling program consists of the following steps:

a. Assign first probabilities of selection for each "multiproducer."

b. For each control item, compute the "residual variance":

Residual variance = Specified variance - Variance contribution of multiproducers.

c. If the residual variance is positive for a given control item, then compute the item probabilities for single producers by setting the variance specification for the single producers of this item equal to the residual variance. If the residual variance is nonpositive, then every single producer is assigned a predetermined probability (minimum probability of selection). This results in oversampling but ensures that each sampling unit has a positive probability of selection.

The procedure outlined in a through c, above, produces a complete adjustment for any control item for which there exists at least one single producer; that is, the variance requirement will be satisfied exactly. The program has a parameter to allow any degree of adjustment starting with no adjustment at all to complete adjustment. A convenient term to describe this adjustment is partial adjustment.

D. Additional Features of the Program

The program has the following additional features:

1. An arbitrary certainty stratum can be designated based on the values of the survey items. Sampling units with values that exceed these arbitrary cutoffs for any item are removed from the noncertainty universes for all items produced by the unit before probabilities of selection are calculated for the remaining sampling units in the universe.

2. Prior to sampling, an arbitrary sampling unit can be specified to be either the company or the establishment.

3. A minimum probability of selection has to be specified.

4. Cost factors reflecting the fixed and variable costs per item of collecting data are specified.

5. Considerable detail on the sample frame and selection process is available (a) to aid in sample verification, and (b) to rerun the program with revised specifications, if that seems necessary or desirable; for example, the sample size, based on the initial sampling specifications, may be too large for the survey budget.

E. Mathematical Developments

1. The univariate case. Let:

$U_1, \ldots, U_N$ be the list of sampling units in the universe at the time of selection,

$X_{i,t}$ be the value associated with unit i at time t (the time of selection is assigned a value 0),

$a_i = 1$ if U_i is selected,

$a_i = 0$ if U_i is not selected,

P_i be the probability of selection of U_i,

$X_{.,t}$ equals $\sum_i X_{i,t}$.

In many CIR (Current Industrial Reports) surveys, the parameter of interest is

$$\Delta X_{.,t} = \sum_1^N (X_{i,t} - X_{i,t-1}).$$

The estimator used is

$$\Delta\hat{X}_{.,t} = \sum_1^N W_i(X_{i,t} - X_{i,t-1})a_i, \text{ where } W_i = \frac{1}{P_i}.$$

(Of course, this is an oversimplification since, undoubtedly, there are changes in the universe: births, deaths, etc.)

The variance of $\Delta\hat{X}_{.,t}$ is

$$\sigma^2(\Delta\hat{X}_{.,t}) = \sum_1^N (W_i - 1)(X_{i,t} - X_{i,t-1})^2.$$

Let T be the number of periods for which this panel is selected. Let the average variance be denoted by S^2:

$$S^2 = \sum_1^N (W_i - 1)\frac{1}{T}\sum_{t=1}^T (X_{i,t} - X_{i,t-1})^2.$$

Define

$$D_i^2 = \frac{1}{T}\sum_{t=1}^T (X_{i,t} - X_{i,t-1})^2$$

and

$$X_i^2 = \frac{1}{T}\sum_{t=1}^T X_{i,t}^2.$$

Of course, neither D_i^2 nor X_i^2 are known at the time of selection. Note also that the certainty units do not contribute to the variances (or average variance). Thus, if units $U_1, \ldots, U_n$ are the noncertainty units, then

$$S^2 = \sum_1^n (W_i - 1)D_i^2.$$

Historical evidence indicates that usually there is a simple regression model which relates D_i^2 to X_i^2.

$$D_i^2 = bX_i^2 + \epsilon_i.$$

$$\sigma^2(\epsilon_i) = kX_i^2, \text{ where k is a constant.}$$

$$E(\epsilon_i) = 0.$$

In some cases, the model is valid only if we stratify by size (X_i^2) and for each stratum a separate b has to be computed. Within a given stratum, we obtain

$$b = \frac{\sum D_i^2}{\sum X_i^2}.$$

Of course, at the time of selection, we have only historical information available from the units from the previous panel. At the present time, a provision for estimating b by stratum is not operational.

For each such unit U_i, which is in the current sample already in operation, we know its current weight W_i' and

$$D_i'^2 = \frac{1}{T'}\sum_{t=1}^{T'} (X_{i,-t} - X_{i,-t-1})^2,$$

$$X_i' = \frac{1}{T'}\sum_{t=1}^{T'} X_{i,-t},$$

$$X_i'^2 = \frac{1}{T'}\sum_{t=1}^{T'} X_{i,-t}^2.$$

This allows an estimate of b:

$$\hat{b} = \frac{\sum' W_i' D_i'^2}{\sum' W_i' X_i'^2},$$

which would allow us to estimate D_i^2 by $\hat{D}_i^2$, using

$$\hat{D}_i^2 = \hat{b}X_i^2$$

for all units in the frame.1/ Many CIR surveys have a monthly or quarterly frequency and often their base values X_i are derived by taking an

average over the base year since the frame is often based on the census where annual figures are collected.

The total is estimated by adding the estimated change to the base total $X_{.,0}$:

$$\hat{X}_{.,t} = \sum_{i}^{N} W_i(X_{i,t} - X_{i,0})a_i + X_{.,0}.$$

The variance of this estimator is

$$\sigma^2(\hat{X}_{.,t}) = \sum_{i}^{N} (W_i - 1)(X_{i,t} - X_{i,0})^2.$$

This compares favorably with the simple linear estimator

$$\hat{\hat{X}}_{.,t} = \sum_{i}^{N} W_i X_{i,t} a_i,$$

which has the variance

$$\sigma^2(\hat{\hat{X}}_{.,t}) = \sum_{i}^{N} (W_i - 1)X_{i,t}^2.$$

In reference [1], it is shown that under reasonable assumptions, $\sigma^2(\hat{X}_{.,t}) \leq \sigma^2(\hat{\hat{X}}_{.,t})$ if $\rho(\hat{\hat{X}}_{.,0}, \hat{\hat{X}}_{.,t}) > .5$.

If the primary concern is the estimation of level, then we would consider the average variance:

$$S^2 = \sum_{i}^{N} (W_i - 1)\frac{1}{T} \sum_{t=1}^{T} (X_{i,t} - X_{i,0})^2,$$

and defining

$$D_i^2 = \frac{1}{T} \sum_{t=1}^{T} (X_{i,t} - X_{i,0})^2,$$

$$X_i^2 = \frac{1}{T} \sum_{t=1}^{T} X_{i,t}^2,$$

we can follow the procedure described before.

Note that

$$\Delta\hat{X}_{.,t} = \hat{\hat{X}}_{.,t} - \hat{\hat{X}}_{.,t-1},$$

and so

$$\sigma^2(\Delta\hat{X}_{.,t}) = \sigma^2(\hat{\hat{X}}_{.,t}) + \sigma^2(\hat{\hat{X}}_{.,t-1})$$

$$- 2\rho(\hat{\hat{X}}_{.,t}, \hat{\hat{X}}_{.,t-1})\sigma(\hat{\hat{X}}_{.,t})\sigma(\hat{\hat{X}}_{.,t-1}).$$

Then

$$\text{ave. } \sigma^2(\Delta\hat{X}_{.,t}) \doteq 2 \text{ ave. } \sigma^2(\hat{\hat{X}}_{.,t})(1 - \hat{\rho}),$$

where ρ is the estimated correlation between $\hat{\hat{X}}_{.,t}$ and $\hat{\hat{X}}_{.,t-1}$. This sometimes has an operational advantage since it allows us to determine the average variance of change from the average variance of the simple linear unbiased estimator of level, defining $D_i^2 = X_i^2$. This is very attractive also because it is a necessity for designing new surveys or for redesigning old badly maintained surveys and also for cases where the regression model is not appropriate.

From now on we will use the generic notation D_i^2 without considering the specific definition or method of estimation of D_i^2.

Let C_i be the cost associated with processing the information from unit U_i. Then the expected cost over the noncertainty stratum is

$$\sum_{i=1}^{n} P_i C_i.$$

The problem can now be stated as follows: Given

$$S^2 = \sum_{1}^{n} (W_i - 1)D_i^2,$$

we want to minimize $\sum_{i=1}^{n} P_i C_i$.

By using the Cauchy-Schwartz inequality, we obtain

$$P_i = t \frac{D_i}{\sqrt{C_i}},$$

where

$$t = \frac{\sum_{j=1}^{n} D_j \sqrt{C_j}}{S^2 + \sum_{j=1}^{n} D_j^2}.$$

The analytical certainty cutoff point can be determined by having sorted $\left\{\frac{D_i}{\sqrt{C_i}}\right\}$ in ascending order and finding the largest number n, such that

$$\frac{\sum_{1}^{n} D_i\sqrt{C_i}}{S_i^2 + \sum_{1}^{n} D_i^2}\left(\frac{D_n}{\sqrt{C_n}}\right) < 1.$$

2. The multivariate case. Let $P_{i,h}$ be the probability of selection of unit h for item i, as obtained from the univariate case. Then, to guarantee that the variance requirements are satisfied,2/ the probability P_h of selection of unit h should be such that

$$P_h \geq \max_i\{P_{i,h}\}.$$

Present options used are $P_h = \max_i\{P_{i,h}\}$ and

$$P_h = 1 - \prod_i(1 - P_{i,h}).$$

To offset the decrease in the expected variances, or conversely, the increase in the expected sample size, an adjustment for the single producers can be made by considering the residual variance:

$$P_{i,h} = (t_i')\frac{D_{i,h}}{\sqrt{C_i}},$$

where

$$t_i' = \begin{cases} \dfrac{\sum_h' D_{i,h}\sqrt{C_h}}{S_i^2 - \sum_h''(W_h - 1)D_{i,h}^2 + \sum_h' D_{i,h}^2} \\ \qquad \text{if } S_i^2 > \sum_h''(W_h - 1)D_{i,h}^2 \\ 0, \text{ otherwise} \end{cases}$$

where $\sum'$ indicates the sum over single producers, and $\sum''$ indicates the sum over multiproducers. There exists, thus, the possibility of no sampling (or too light a sampling) of single producers. Thus, a final adjustment is made possible by redefining

$$\text{New } t_i' = \alpha t_i + (1 - \alpha)\text{old } t_i',$$

$$0 \leq \alpha \leq 1.$$

F. Remarks

Several other procedures are available for the multivariate case, but they generally require a rather lengthy iteration process with the attendant problems of convergence and the rate of convergence, and/or substantial blocks of computer memory. The simplicity and cost effectiveness of this program warrants that it be given serious consideration by the user community.

Acknowledgments:

Kay Bair, CIR Programming Branch, Industry Division, is mainly responsible for the development in 1975 of the present computer program.

Donald Clark, Assistant Chief for Research and Methodology, Industry Division, provided guidance for the completion of this project.

Barbara Bailar, Chief, Research Center for Measurement Methods, and her staff, helped to improve the clarity of the presentation of this paper by providing us their comments.

Footnotes:

1/ If stratification is used, then no weights are needed. Stratification is also operationally advantageous since the weight of a multiproducer is the same for all items produced regardless of the size of the item. Also, it is possible that for some items, the assumption of nonconstant variance around the regression line may not hold. This is more likely to be the case when we use stratification. It is also important that the units from certainty stratum should not be used.

2/ The variance requirements may be specified by upper bound of variance specifications for control items.

Reference:

[1] Ogus, Jack L., and Clark, Donald F., Technical Paper 24. The Annual Survey of Manufactures: A Report on Methodology. U.S. Department of Commerce, Bureau of the Census: February 1971.