OPTIMIZING SAMPLE ALLOCATION OF THE 2000 NON-RESPONSE FOLLOW-UP

Yves Thibaudeau, Alfredo Navarro
U.S Bureau of the Census, Washington DC 20233

1. Introduction

There is accrued interest in the development of new methods to conduct the population census in the United States. The traditional model of the census is a complete enumeration of the United States population. This is to be achieved through communication by mail or with personal visits from qualified enumerators. This approach appears simplistic but in practice the logistic problems it generates are enormous: Complete mailing list are expensive to produce and a large proportion of the enumeration forms that are mailed-out are not returned. When an enumeration form is sent to an address, but no reply is received, the address is called a "non-mail return". The remaining addresses are called the "mail returns". During the last decennial census, the mail response rate was about 67%.

The size of the population associated with the non-mail returns is obviously very large. Under constitutional law, this population must be counted. During the last census, an attempt was made at enumerating all the non-mail returns: Enumerators were sent to identify the vacant housing units and to enumerate the remaining non-mail returns. The costs involved in are considerable and have contributed significantly to the overall cost increase of the decennial census.

Under these financial pressures one proposition is increasingly attractive: replace the enumeration operation by a sampling operation. Only a fraction of the non-mail returns are designated for follow-up under this proposition. Estimates of the population count can then be produced and the error incurred is measurable. In a 1992 report to Congress, the General Accounting Office specifically argues in favor of sampling for non-response and reports potential savings of the order of 400 million of dollars if the Bureau of the Census were to move in that direction. Two panels of the National Academy of Sciences commissioned by the Bureau of the Census have reiterated this statement.

The paper explores avenues open to us in applying a plan for sampling for non-response. Our research is guided by two principles: The first is the efficiency principle. It is motivated entirely by the goal of providing with a census with maximum accuracy, for a given cost. The second principle is equity. Loosely speaking, equity is the same as "fairness". We believe that equivalent portions of the population should be counted with the same accuracy. In particular, the deployment of the resources should be such that the census count in places with higher rates of response has an accuracy comparable to that of places with lower rates of response.

In section 2 we develop the concept of efficiency and we determine which sampling plan provides maximum accuracy under general conditions. In section 3 we discuss the notion of equity. We present one possible definition of equity and we construct a method to achieve it in the context of sampling for non-response.

In section 4 we apply our theories using data from the 1990 census: For the state of New-Jersey we simulate the census operations in the context of a design featuring sampling for non-response along with traditional enumeration techniques. In section 5 we apply our methods for sampling for non-response when specific levels of truncation of the non-mail returns are mandated.

2. Efficiency

We first reformulate a general statistical principle. The principle is based entirely on an algebraic artifact: It is always more efficient to sample than to enumerate. However, this does not mean that we always want to sample. There might be some practical constraints making it necessary to enumerate part of the non-response units. For instance we could be faced with the requirement that at least 70% of all units of every county be enumerated.

Let N be the size of the non-response universe. Consider the following scenario to follow-up the non-response units: First m non-response units are enumerated. After the enumeration, n non-response units are sampled from the N - m remaining units in the non-response universe. Then the unbiased estimator of the total population is:

$$\hat{P} = \sum_{i=1}^{M} X_i + \sum_{j=1}^{m} Y_j + \frac{(N - m)}{n} \sum_{k=1}^{n} Z_k$$

The implicit expressions in this formula are X_i the person counts of the units in the response (mail-return) universe, Y_j the person counts of the units in the enumeration portion of the non-response universe, and Z_k the person counts of the units in the sample. We refer to the enumerated non-mail returns as the self-represented unit. Let

$$S_m^2 = \sum_{i=1}^{N-m} \frac{(Z_i - \bar{Z})^2}{N - m}$$

Then, conditional on the m self-representing units the variance of $\hat{P}$ is:

$$\mathrm{VAR}(\hat{P}) = (N - m)(N - m - n)S_m^2/n$$

Keeping with the goal to maximize efficiency, we are interested in the relationship between the variance and the cost of follow-up operation. Assume for now that the entire cost of sampling and enumerating is not to exceed C. Suppose that the cost is symmetric in n an m, that is it is not more expensive to enumerate an additional unit than to sample one. It is also reasonable to expect the cost to increases when n or m increases. When trying to minimize the variance for a given cost C, our constraint can be expressed through following equation:

$$G(m + n) = C$$

where G is a strictly increasing cost function.

Under the cost constraint the expression for the variance can be further simplify: Substitute for m in terms of n and C, conditional on the self-representing units, the variance becomes:

$$\mathrm{VAR}(\hat{P}) = \left[\frac{(N - G^{-1}(C))^2}{n} + (N - G^{-1}(C)) \right] S_m^2$$

Maximizing efficiency is equivalent to minimizing this expression in terms of n and m. The difficulty in doing so is the unpredictable behavior of S_m^2 .

In a manuscript to be submitted for publication, Thibaudeau and Navarro minimizes this variance conditional on the entire non-response universe, assuming that the self-representing units are selected at random. Under this assumption, the variance is minimized when $n = G^{-1}(C)$ and m = 0. The authors also study the behavior of the variance for some particular sequences $\{Z_1, \ldots, Z_m\}$.

3. Equity

Given that a most efficient sample size has been determined, our other priority when designing the follow-up of the non-respondents is equity. In order to evaluate the degree of equity of a particular design we must introduce a device that measure what we consider to be equity and establish consistent rules for the use of that device.

The more obvious comparison in order to measure equity are done along geographical delineations. For instance, under an equitable plan, counties of approximately equal population, should have similar CV's. More generally, we are concerned about equity between regions of similar population size but with different response rate. We show that a constant sampling rate does not provide with an equitable accuracy and areas with low response need a higher rate of sampling to achieve a comparable degree of accuracy. A general formula for measuring the accuracy of the population estimate for the i-th geographical unit is:

$$\mathrm{C.V.}_i^2 = \frac{\mathrm{VAR}(\hat{P}_i)}{\mathrm{E}^2(\hat{P}_i)}$$

$$= \frac{(1 - R_i)(1 - f_i)S_i^2}{n_i \bar{Y}_i^2}$$

That is, the square of the coefficient of variation of $\hat{P}_i$, the estimate of the population count of geographical unit i. $R_i = (M_i - N_i)/M_i$ is the mail response rate in unit i; $f_i = n_i/N_i$ is the sampling fraction; $\bar{Y}_i$ and S_i^2 are respectively the mean and variance of the household size in geographical unit i.

If we use the coefficient of variation as a way to compare accuracy between areas with roughly the same population we also would like to calibrate the sample sizes corresponding to each area so that the estimates have the same accuracy. We have:

$$\mathrm{CV}_j^2 = \frac{(1 - R_j)^2(1 - f_j)S_j^2}{n_j \bar{Y}_j^2}$$

$$= \mathrm{CV}_i^2$$

$$= \frac{(1 - R_i)^2(1 - f_i)S_i^2}{n_i \bar{Y}_i^2}$$

Let's solve this equation for n_j in terms of n_i. The solution allows us to determine the size of the sample from area j yielding the same count accuracy than the sample from area i. We find:

$$n_j = \frac{\bar{Y}_i^2(1 - R_j)^2 S_j^2 n_i}{\bar{Y}_j^2(1 - R_i)^2 S_i^2 - \bar{Y}_j^2(1 - R_i)S_i^2 \frac{n_i}{N_i} + \bar{Y}_i^2(1 - R_j)S_j^2 \frac{n_i}{N_j}} \qquad (1)$$

This formula simplifies considerably when the first sample is small with respect to the size of the population and n_i/N_i and n_j/N_j are negligible.

This formula can be used to determine the allocation of a sample between different geographical units such that the accuracy of the estimate of the count is the same for each area. The next section demonstrate this.

4. Sample Allocation

We now proceed to show how our technical results can be implemented to produce equitable and efficient sample designs in the non-response follow-up. We use the data available from the 1990 census in New Jersey. The state is naturally divided in counties and then in smaller geographical units called tracts. There are 1873 tracts containing housing units. There is a wide range of response rates at the tract level.

We assume that the budgetary constrains are such that we can afford to follow-up exactly 1/6 of the population of non-response units. In this first tentative we are free to follow the principles of section 2, that is the most efficient follow-up is conducted entirely through a sampling plan. Then our concern is the particular of the sample allocation between tracts to obtain an equitable follow-up. To define a sampling rate for each tract would entail the management of possibly 1873 sampling plans, each with its own sampling rate. This is highly unpractical from an operational point of view.

In order to develop a practical sampling design for the follow-up operations we group the tracts in ten classes. We construct the classes in a way to achieve some homogeneity with respect to the response rate at the tract level within each class. Furthermore, to be consistent with our goal of equity, we constructs the classes so that they each cover populations of approximately equal size. Then we allocate the sample between the ten classes so that their coefficients of variation are equal. At an operational level, each tract is sampled individually. Then the rate of sampling for a particular tract is that which has been assigned to the corresponding class. Table 2. summarizes the construction of the classes.

After the classes are clearly delineated, we proceed to execute the sampling strategy. Formula (1) is used to determine the sampling rates. Our goal is equity between classes. To obtain equal CV allocation between classes we imbed formula (1) in the method of Newton. Our constraint is:

$$\sum_{i=1}^{10} n_i = \frac{\sum_{i=1}^{10} N_i}{6}$$

The corresponding sampling rate n_i/N_i is applied uniformly to every track in class i. Table 3 gives the sampling rates for each class along with the average tract CV. Note that the average tract CV does fluctuate from class to class since the average tract size varies from one class to another.

5. Sampling under Administrative Requirements

The mathematics show that sampling is operationally optimal in conducting the non-response follow-up. Nevertheless, administrative constraints may be mandated. One possible administrative constraint is the demand to have every tract enumerated at least at 70% before doing any sampling. Table 1 shows the minimum number of units that needs to be enumerated to satisfy the constraint. The sampling procedure can be carried out, after the administrative constraints are satisfied. I.e. after at least 70% of units in every tract are enumerated by mail or otherwise.

For the New-Jersey example, considerable sampling must be done to bring down the sampling error to the same levels obtained when sampling 1/6 of the non-response universe under no constraint, in section 4. Table 1 shows how many units should be sampled after satisfying the constraint to recover the same accuracy (in terms of CV's) we had when sampling each class under the allocation scheme developed in section 4.

Note that, if each tract must be enumerated at 70% and if we want to have sampling errors as small as we had under the sampling plan of section 4, the entire follow-up (sampled and non sampled units) is almost twice the size of the sample designed in section 4. The situation is alleviated somewhat if our requirement is that every county be enumerated at 70%, rather than every tract.

6. Conclusion and Current Research

In the absence of administrative constraints, we advocate direct sampling to maximize efficiency. We also show that the sampling follow-up can be conducted equitably in the sense that census count estimates are equally accurate between tracts/counties of different response rate. The integration of administrative constraints is also feasible with this technique.

Current research indicates that our method of allocating the non-response sample equitably between regions of different response rates also implicitly induces an equitable allocation with respect to other criterions, such as race. If this is confirmed, our technique could be very useful to implement impartiality in the sampling process and in the decennial census in general. Under our approach, socio-demographic considerations are specifically excluded from the design of the operations, but the product turns out to be equitable with respect to these considerations as well.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

References

United States General Accounting Office, "DECENNIAL CENSUS: 1990 Results Show Need for Fundamental Reform", Report to Congressional Requesters, GAO/GGD-92-94.

Thibaudeau, Y., Navarro A., "Optimal Non-Response Follow-Up with Sampling for the Year 2000", unpublished manuscript.

	TRACT LEVEL	COUNTY LEVEL
MAIL RETURNS	2088628	2088628
NON-MAIL RETURNS	986682	986682
SAMPLE SIZE - 1/6 OF NON-MAIL RETURN UNIVERSE	164447	164447
NO. H.U.'S TO BE ENUMERATED TO GUARANTEE A 70% RESPONSE RATE	207280	125649
NO. H.U.'S TO BE SAMPLED AFTER ENUMERATION TO RECOVER THE ORIGINAL CV'S / TOTAL FOLLOW-UP	89552 / 296832	115684 / 241333

TABLE 2: CLASS DEFINITION BY RESPONSE RATE AT THE TRACT LEVEL

CLASS	AVGE POP'N SIZE PER TRACT	NO. OF TRACTS IN THE CLASS	AVGE NO. OF HU'S PER TRACT	POPULATION SIZE OF THE CLASS	MINIMUM RESPONSE RATE OF THE TRACTS IN THE CLASS %	MAXIMUM RESPONSE RATE OF THE TRACTS IN THE CLASS %
1	2914	259	1472	754646	0	49.8
2	3817	198	1626	755797	49.9	58.6
3	3890	194	1653	754730	58.7	64.2
4	4152	181	1729	751622	64.3	68.6
5	4223	180	1686	760485	68.7	72.4
6	4280	177	1711	752509	72.5	75.5
7	4655	163	1838	758840	75.6	78.6
8	4592	165	1753	757686	78.7	81.5
9	4558	165	1660	752066	81.6	84.2
10	3981	191	1409	760439	84.3	100

TABLE 3: AVERAGE C.V. OF THE ESTIMATOR OF THE POPULATION COUNT OF A TRACT IN EACH CLASS UNDER TWO ALLOCATION SCHEMES

CLASS	NO. OF NON-MAIL RETURNS IN THE CLASS	SAMPLING RATE UNDER EQUAL C.V. ALLOC.	SAMPLING RATE UNDER PROP. ALLOC.	AVERAGE C.V. UNDER EQUAL C.V. ALLOC.	AVERAGE C.V. UNDER PROP. ALLOC.
1	233879	.295	.167	.0305	.0442
2	146703	.200	.167	.0267	.0299
3	122824	.156	.167	.0264	.0253
4	104256	.129	.167	.0255	.0220
5	89374	.111	.167	.0254	.0202
6	78811	.101	.167	.0252	.0189
7	68923	.087	.167	.0242	.0167
8	57902	.076	.167	.0244	.0156
9	46899	.066	.167	.0244	.0145
10	37111	.050	.167	.0262	.0135