A VARIABLE VERSUS A UNIFORM SAMPLING RATE FOR CONDUCTING THE 1980 CENSUS

Paul S. T. Lee, University of Nebraska at Omaha

Background of the Problem

In conducting the 1970 Census of Population and Housing, the Bureau of the Census used three types of questionnaires: the short form, the medium form, and the long form. The short-form questionnaire contained a limited number of questions on population and housing and was distributed to 80 percent of the American households. In addition to having questions identical to those contained in the short-form questionnaire, the medium- and long-form questionnaires contained questions concerning housing facilities, utilities, education, income, employment, etc. The medium form was to a 15 percent sampling unit, and the long form to a 5 percent sampling unit. Since some questions were included on both the 5 percent and 15 percent questionnaires, it appears that a 20 percent sampling fraction was used to generate statistics for some of the subjects.

Responding to the need for income data with greater statistical reliability for smaller places, the Bureau of the Census has been, in conducting the 1980 Census, planning to increase the sampling rate to up to 50 percent of the housing units receiving long-form questionnaires for communities with less than 5,000 population. Elsewhere, the sampling rate is reduced to 16.7 percent. In total, this yields a national sampling rate of about 21.5 percent in the 1980 Census. [7]

The uniform sampling rate used in the 1970 Census and proposed to be used in the 1980 Census represents a misallocation of the nation's resources by ignoring the heterogeneous nature of the "universe" and the statistical theory governing the determination of sample size and, hence, the sampling fraction.

Determinants of Sample Rate

Statistical theories dictate that sample size n (and hence the sampling fraction) is a function of three factors: the variance of the major item of inquiry s^2, the desired level of reliability (in terms of confidence limit) t, and the tolerable difference between a sample estimate and its parameter, d^2. Their functional relationship can be expressed in the following equation:[2] [3]

$$n = \frac{t^2 s^2}{d^2} \qquad \ldots (1)$$

Hence, it can be seen that the magnitude of the sample size is positively related to the level of reliability and the variance of the major item of inquiry but negatively related to the size of the allowable difference between a sample estimate and its parameter.

For a multipurpose and multivariate sample survey such as a national census, information is collected on more than one item. One method of determining the overall sample size is to estimate the sample size needed for a given level of precision separately for each of the items regarded as most vital to the survey. If the largest of the n's falls within the limits of the budget, this n is selected. [3]

Economic costs do enter the decision-making process in the choice of n. In their attempts to provide a general solution of the sample size problems, both Cochran (1968) and Yates (1949) proposed that the n should be chosen so as to minimize

$$C(n) + L(n)$$

where C(n) is the sum of overhead cost (c_0) and unit cost (c_1) in for taking the survey, and L(n) is the expected loss incurred in a decision through an error of amount d in the estimate. For instance, if $\bar{y}$ is the sample estimate of μ, and $d = \bar{y} - \mu$, then the expected loss function is

$$L(n) = \lambda v(y) = \lambda s^2/n$$

if simple random sampling is used and the finite population correction is ignored. By differentiation, the required value of n which minimizes costs plus loss is

$$n = \sqrt{\lambda s^2/c_1} \qquad \ldots (2)$$

where s^2 is the estimated population variance and λ is a constant.

As shown in equation (1), the higher the level of reliability one desires, the larger the sample size that is needed. Inversely, the smaller the allowable difference between a sample estimate and its parameter, the larger the size of sample that is needed. Equation (2) indicates that each additional sample household represents an additional cost for the total survey. Hence, the level of reliability of sample estimates and the allowable differences are largely subject to political decisions. In appropriating the budget for taking the 1980 Census, the U.S. Congress should make these decisions in consultation with sampling specialists to explore all options available.

On the other hand, the magnitude of the sample size is positively related to the variance of the item that is to be estimated as shown in both equations (1) and (2). In other words, the larger the variance, the larger the sample needed to maintain the same level of reliability given the same amount of allowable difference. Even though the variance is largely a natural phenomenon, it can be manipulated for achieving a particular purpose.

Stratification and Economic Efficiency

Stratifying a heterogeneous universe into a series of relatively homogeneous strata tends to reduce the magnitude of variances and thereby the size of sample. Hence, economic efficiency is achieved through stratification because a smaller sample is needed for obtaining sample statistics without reducing the level of precision and/or augmenting the magnitude of the tolerable difference between a sample estimate and its parameter.

Basic principles and practices of stratification in univariate surveys are well known and have been widely used. Theories and practices of stratification in multipurpose and multivariate situations such as conducting a national survey are unclear. Hagood and Bernert [5] suggested that the criterion for choice of a stratifier is its relation to the item on which observations are to be made in the sample for the purpose of making estimates for the universe. The more closely a single stratifier is correlated with the item or items to be estimated from the sample, the greater will be the improvement in efficiency of estimation from a stratified sample survey than from a simple random sampling with a uniform sample rate. Studies have successfully demonstrated that substantial gains can be achieved using a bivariate normal model. [4, 1, 6] In general, bivariate stratification yields greater gains in reducing variances than univariate stratification, and the gains would even be greater for cases in which two stratifiers are highly correlated. According to Kish and Anderson, [4] the marginal gains in stratification reach a maximum if the universe is stratified into three or four strata. The marginal gains would diminish as the number of strata increase thereafter. For delineating boundaries between strata of continuous variables, both Ekman's rule for creating strata with equal values of $W_h \sigma_h$ and the cumulative $\sqrt{f}$ rule perform well. [3] Problems arise, however, as political or exogenous considerations often enter into the decision-making process of creating boundaries of strata and fail to coincide with optimal choices for the strata.

Hypotheses and Needs for Future Research

Studies have successfully demonstrated that substantial gains in economic efficiency can be achieved through the use of stratification of the universe into a series of strata in conducting sample surveys. Sample sizes would vary among strata in accordance with the magnitudes of the within-stratum variances. Theoretically, this applies to all sample surveys including nationwide surveys conducted by the Bureau of the Census. For conducting multivariate and multipurpose surveys, however, questions as to what is the best choice of the stratification variable or variables, and how the strata should be constructed remain unanswered.

In planning for the 1980 Census, the Bureau of the Census has decided to increase the sampling rate to up to 50 percent of the housing units receiving long-form questionnaires for communities with less than 5,000 population. Elsewhere, the sampling rate is reduced to 16.7 percent. This, in essence, recognizes the heterogenous nature of metropolitan and nonmetropolitan areas. The assumption has been made that geographic differences condition people's life-styles and livelihoods in different regions and thus have an important effect on their social and economic characteristics. For instance, geographic differences along with types of farming were used as the most important single basis for stratification in most of the rural social surveys. [5] Evidence also indicates that vast differences in income, education, value of housing, and other social and economic characteristics exist in nearly all metropolitan areas. The fact that more than 75 percent of the total United States population resides in urban and suburban areas makes it imperative to stratify metropolitan areas into smaller and relatively homogeneous neighborhoods. A variable sampling rate could then be applied to each of these neighborhoods in accordance with the relative magnitude of variances.

Future research should center on the definition of neighborhood and its relationship with population density, types of housing, levels of income, occupations, political structures, and other social and economic characteristics. Should a single variable such as the type of housing or a composite index of a group of variables be used as basis of stratification? And what exactly will the magnitude of economic benefits be from stratification in conducting a national survey such as the 1980 Census?

References

1. Anderson, D. W., L. Kish, and R. G. Cornell
   1976 "Quantifying Gains from Stratification for Optimum and Approximate Strata Using a Bivariate Normal Model." Journal of the American Statistical Association 71: 886-892.
2. Cochran, W. D.
   1961 "Comparison of Methods for Determining Stratum Boundaries." Bulletin of the International Statistical Institute 38: 345-348.
3. ________
   1968 Sampling Techniques (2nd ed.) New York: John Wiley and Sons.
4. Kish, L., and D. W. Anderson
   1978 "Multivariate and Multipurpose Stratification." Journal of the American Statistical Association 73: 24-34.
5. Hagood, M. J., and E. H. Bernert
   1945 "Component Indexes as a Basis for Stratification in Sampling. Journal of the American Statistical Association 40: 330-341.
6. Thomsen, I.
   1977 "On the Effect of Stratification When Two Stratifying Variables Are Used." Journal of the American Statistical Association 72: 149-153.
7. United States Bureau of the Census
   1978 1980 Census Update, No. 7 (July). Washington, D.C.: U.S. Government Printing Office.
8. Yates, F.
   1949 Sampling Methods for Censuses and Surveys. London: Charles Griffin and Company.