TEAMING UP CENSUSES AND SAMPLES

By: Frederick F. Stephan, Princeton University

Census taking and sampling have been recognized as alternative methods of obtaining data about populations and economies for at least two centuries. Throughout this period they have been considered to be wholly distinct and competing methods, notably in the discussions of the International Statistical Institute at the turn of the century. Now we are beginning to understand the great advantage of using them in conjunction.

The practice of sampling to collect data quickly and economically is widespread in industry and science as well as in government statistical work. The practical problems of sampling have stimulated great progress in developing principles of sampling design and techniques for applying them. These principles now lead us to re-examine the assumptions underlying census taking and particularly the detailed planning of census operations.

Such a re-examination is inescapable when census enumeration is combined with sampling. The successive development of sampling from the 1940 Census of Population to the 1960 Census of Population and Housing serves as an excellent example of the benefits and problems inherent in the combination of the two methods.

Coincidental sampling was introduced in the 1940 Census of Population primarily to permit adding to the schedule several questions for which there was considerable pressure from users of census data without requiring the enumerators to these questions for the entire population. The sample design limited the sample questions to 5 per cent of the persons enumerated. Thus there was a 95 per cent saving in interviewing.

A similar saving was made in the subsequent operations of coding and card punching. Further savings in coding and card punching and reductions in tabulating were made by using samples of punch cards for many cross-tabulations of questions which had been completely enumerated. Sampling was also used for quality control of card punching and other processes.

Thus the result of introducing sampling in 1930 was to relieve the enumerators, increase the information obtained for census users, reduce the labor involved in some tabulations, add new tabulations, make possible the publication of some data earlier than previous schedules provided, and control quality better than previous censuses. However, the main body of questions was enumerated completely and the field work procedures were not modified significantly by the introduction of sampling. Hence the potential benefit of employing sampling was not fully realized.

The usefulness of sampling was demonstrated further by the experience of the Census Bureau with the Current Population Survey beginning soon after the 1940 censuses. By the time preparations were made for the 1950 censuses it was clear that it would be advantageous to extend sampling to some of the questions previously enumerated completely. Quite naturally there was considerable difference of opinion in the advisory committees and elsewhere on the choice of questions to be sampled. As a consequence the use of sampling was extended in the 1950 Censuses of Population and Housing somewhat less than it might have been had there been more evidence, more time for study, and less diversity of viewpoints and interests. Nonetheless, great gains were made; the sample was increased to 20 per cent; and a number of major questions were transferred to the sample.

For 1960, as Morris Hansen will tell you, the potentialities of sampling are being exploited to a still greater degree than in the two previous censuses. One of the principal benefits will be the freeing of the complete enumeration from the drag of questions, such as occupation, which are relatively difficult to enumerate and which delay the tabulations until they have gone through a time-consuming process of editing and coding. The transfer of labor force questions to the sample makes great savings of interviewing and processing costs as well as gains in the speed of publication.

Another important step is to take households instead of individual persons as the elementary sampling units. This makes possible sample tabulations which relate data for two or more members of the same household or which form aggregates of individual data, such as income, for each household. The sampling procedure is more difficult and there are other problems in the shift to a household sample but it appears on balance to be preferable to the unit and procedure used in 1950.

This, in brief, is the evolution of the incorporation of sampling into the traditional procedure of census taking. No doubt the experience of the 1960 censuses will lead to further developments and changes. It is worth while to take a broad look at what has been involved in teaming up sampling and census enumeration. After we have done that we will look at some questions and reservations which are of concern to many census users.

Clearly the primary function of the Population Census is to provide an accurate count of the populations of the States for the decennial reapportionment of Congress in fulfillment of the provisions of the Constitution. To the extent that this function is not jeopardized, additional information can be collected for the guidance of government officials and agencies in the performance of their duties and for the enlightenment of the public.

In the past, careful consideration was given to requests for the addition of questions put forth by various groups and the set of questions finally selected for enumeration constituted a compromise. One might almost call it a coalition formed out of the competing interests in obtaining information about the population. The Population Census thus acquired in addition to its Constitutional function, the function of a general-purpose statistical system.

Clearly its capacity to perform this service had some limits. As the demand for additional data increased, the difficulty of choosing the questions to be included increased sharply.

The introduction of sampling alleviated the pressure against the capacity of the system but complicated the problems of planning. Some of the problems are:

(1) Decisions about the questions to be included in the census are complicated by the necessity of deciding which of them are to be in the sample.

(2) Budgeting, scheduling, and preparatory work are complicated by the necessity of allocating and planning for the sample and of seeking an optimal relation between sample and enumeration.

(3) The effects of sampling on enumerators, respondents, and users add new problems.

(4) Field operations and the processing of data are affected in various ways.

(5) Sampling introduces problems of preparing estimates from the sample and reconciling these estimates with the results of the enumeration.

Offsetting these problems are certain advantages that ease the solution of the problems usually involved in census-taking. For example there is:

(1) Greater freedom in designing the entire data-collecting system.

(2) Opportunity to select a smaller number of better personnel to perform some of the more difficult work.

(3) A greater output of valuable information, or greater economy, or some of both with consequently better command of the allocation of resources.

Similar advantages and problems will arise in other unions of sampling and complete coverage whether they are surveys, inventories, or other canvasses. Some of them arise in the union of two sampling procedures without a complete canvass.

For statisticians, the conjunction of sampling and census-taking brings to the fore a number of technical and procedural questions.

(1) Just how should the sampling and enumeration be coupled? Should sample questions be asked at the same time as as the other questions, by the same enumerators in a separate interview, by a special corps of interviewers, or in some other way?

(2) How can the sample be designated so as to take advantage of the possibility of using the enumeration as the frame but avoiding both the biases that enumerators tend to introduce when they make the selection and the added costs that must be incurred when the sample is selected in the central offices?

(3) How should the sampling proceed in the unusual cases presented by institutions, homeless or mobile persons, and other special groups in the population?

(4) How should the enumeration be used in the preparation of estimates from the sample?

(5) How should biases and sampling errors be estimated and the accuracy of both the enumeration and sample measured?

(6) How should the interests of users in each set of data to be provided by a particular question, and in its accuracy, be given appropriate weight in the planning and processing?

Substantial progress has been made in the solution of these and other problems; interesting questions remain to be answered. The formal analysis of the sixth problem has had perhaps the least attention though the problem has been discussed at length and in great detail by advisory committees and representatives of users. The balance of this paper will sketch a general view of the problem.

We start by assuming that a well-defined purpose is served by information needed to guide certain actions and that we are concerned about the various consequences which might result from these actions - in fact that we can determine in advance the value of each of these possible consequences. Thus we may need information about the economic level of the population in a small area in order to make a decision about an investment in real estate, or the relation of education and fertility to make a decision about the expenditure of funds for further research on fertility differentials.

The consequences of action taken on information depend on the accuracy of the information.

They may be very sensitive to the accuracy of the information or they may not, i.e., moderately inaccurate information may or may not lead to action inappropriate to the objective situation and hence to losses in comparison with the consequences of action based on accurate information.

The mathematical function which expresses the value of consequences in terms of the departure from accuracy of the information may take many forms. Examples are given in Figure 1. In the case of Figure 1a if a census yields completely accurate information, it results in consequences of the highest value. Shifting the question to a sample can only result in a reduction of value.

If the result of a census is not perfectly accurate, the consequences of the action to which it leads will be less valuable than if it were a completely correct answer to the question asked. Shifting the question to a sample will reduce its value further if the sampling errors disperse the estimates under a portion of the value function which is concave downward but if, as in 1c, the census result is at the arrow, the sampling error (if they are not too great) would disperse it under a portion concave upward and would increase the value. In other cases sampling might increase or decrease the value depending on the distribution of sampling errors and the position of the consus result. We conclude that sampling may be expected to decrease the value but not always and not necessarily by a serious amount.

Consider next the lapse of time between the census date and the time census information is put to use. The accuracy of the information from an enumeration will change as the objective situation changes during the time that elapses from the census date to the date of use. Hence we may expect that the value resulting from the use of census information for a particular purpose will decrease with the passage of time somewhat as shown in Figure 2 due to its departure from perfect accuracy. If the question is shifted to a sample, the value will be changed, possibly in a manner that makes it approximately parallel to, or converging toward, the enumeration value function. In those instances in which the sample data become available sooner than they would if they were enumerated completely, the value of sample data may compare favorably to enumeration data both at the time of publication and over the ten year period between censuses. Points A and B in Figure 2 show the value at the time of publication, A' and B' the times at which the information is replaced by new results.

In the foregoing, we have assumed that perfectly accurate information leads to action, producing consequences of the highest value. Actually the utilization of information is not perfect and there is an "error of application" or "error of use" which may be constant or variable but which affects the relation of the value function to the information producing process. Also the relation of the consequences of action to the information on which it is based may be such that the greatest value results from action on information which is not perfectly accurate, i.e., that differs in a certain way and degree from the correct information sought by the questions. In some such cases, the effect of shifting from complete enumeration to sampling may actually be to make no reduction in value or even to increase the expected value of the consequences.

We need to develop definite value functions, measures of bias and error distributions before we can apply these ideas to particular cases. However, even before we obtain this implementation they warrant the following general conclusions:

(1) The ultimate effect of shifting a question to the sample is not always a reduction of value.

(2) When sample data can be published sooner than enumeration data, there may be an advantage in favor of sampling.

(3) Decisions about the choice of sample questions call for better specification of the users' value functions and determination of the departure of both enumeration and sample expected values from perfect accuracy. Judgments based on poorly defined formulations of the users' interests and on other concepts may be inappropriate or even irrelevant to the fundamental issues.

When we turn from a consideration of the effect of sampling on the utility of census statistics for an individual user to the aggregate effect for all users, actual or potential, the problem of finding a relatively precise basis for decisions appears out of reach. There is the question of how the gains and losses of the various users should be weighted in the aggregate, which government functions and which private activities should have priority, and many other considerations. Moreover, the aggregate gain or loss must now be compared with the aggregate gain or loss that would result from an alternative program of allocating questions to the complete enumeration, to the sample, or to the reject pile. These comparisons can only be made by considered judgment at the present time but a clearer understanding of the effects of sampling in individual cases can contribute to the soundness of these judgments.

In summary, the union of census and sample is a fruitful one. Statisticians do well to look for opportunities to use a similar combination of sampling with a complete canvass in other surveys and inventories. The 1960 Censuses will stand as a great demonstration of the value of joining the two methods and the progress made by the statisticians from all the major traditions of statistical work who have joined their efforts in accomplishing it.

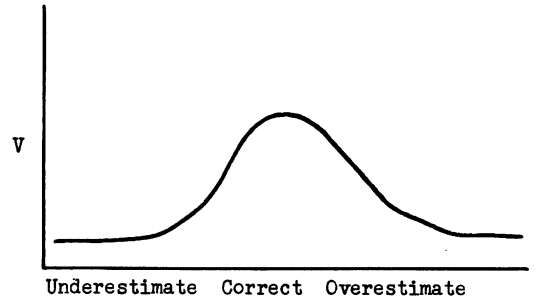

Figure 1a

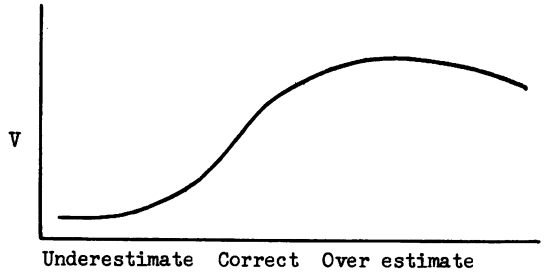

Figure 1b

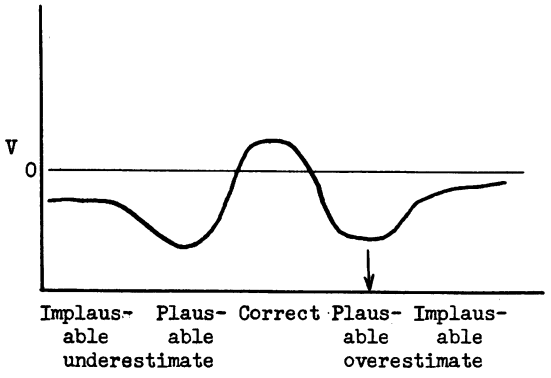

Figure 1c

Three illustrative types of relation between the expected value of the consequences of an action to the actual information on which it is based

Note: The average value of the consequences of the action taken is symmetrically related in Figure 1a, and asymmetrically related in Figure 1b, to the degree of incorrectness of the information on which the action is based. In Figure 1c, the more extreme the degree of under- or overestimate the more implausable it appears and the more it is likely induce further investigation before action is taken.

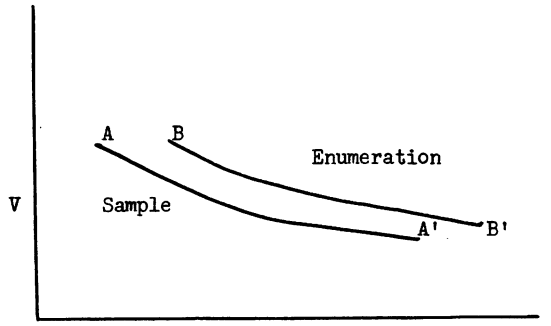

Figure 2

Relation between value of consequences of action and time elapsed since the information on which the action was based was obtained.