APPLICATION OF SAMPLING IN THE 1960 CENSUSES OF POPULATION AND AGRICULTURE IN THE PHILIPPINES

ELPIDIO D. MAKANAS *

Sampling is now a standard tool used not only in surveys but also in censuses. Statistical agencies concerned with census taking have come to rely upon modern sampling methods to increase efficiency in their operations and to meet the ever expanding demands of users of census results.

In view of the proposal to take the 1960 census of population and agriculture simultaneously as was done in the past, the task of completing both censuses within a reasonable period and publishing timely results is indeed a huge job. Most other countries take the population census at a different time from that of the agriculture census, and still turn to the use of sampling in order to accomplish the job in a most efficient way. With the two censuses being taken at the same time, as in this case, the need for the use of sampling to help achieve the goal becomes exceedingly great.

Gains to be achieved by the use of sampling

The proposed application of sampling methods in collecting the data for some selected items in the 1960 census questionnaire would contribute to a large extent to the achievement of important gains.

1. The first and the most important gain is the major reduction of the time interval between collection and publication of census results. Census results in order to be of maximum utility should be published and made available to users as near as possible to the time of collection. The use of sampling would mean a big reduction of the volume of work in coding, editing, punching and other processing operations, thus resulting in a valuable gain in time for the completion of the processing work. With this gain in time it is expected that most census data can be made available much sooner than in the last 1948 census.

* Supervising Statistician, Division of Surveys, Bureau of the Census and Statistics.

2. The cut in the cost of processing the data for the selected items resulting from the reduction of the volume of work is a substantial gain. Added to this will be the cost gained in the enumeration. This is the cost of enumerating only a fraction of the population for the sample items as compared with the cost if the same items were taken on a complete count. Conservative estimates indicate that the combined amount of these gains in enumeration and processing work are in the magnitude of well over three million pesos.

3. The savings in cost make possible the introduction of some needed measures aimed toward the improvement of the census without going beyond the limitations of the total budget. Among the major needs of census taking, as gathered from the experience of past censuses, to which some of the savings may be committed are

   (a) Improvement of the census coverage. Although it seems almost impossible to really cover each and every individual unit in the population within the specified time of enumeration because the problems involved are many and varied, much can be done to reduce the undercoverage by instituting special measures and funds into such areas or groups for which the undercoverage is expected to be large

   (b) A more complete tabulation and publication of results.

   (c) Inclusion of more items in the questionnaire to meet the increasing demands of users, as compared with the number of items that could be accommodated if all the items were taken on a complete count basis.

   (d) Improvements in the accuracy of the collection and processing of census data, especially in such items as those in which the response errors are large.

The Proposed Population and Agriculture Samples

The proposed application of sampling methods in the next 1960 Censuses of Population and Agriculture calls for the use

of two separate samples. One is a 20 per cent sample of households to measure some characteristics of the population as well as housing (the population sample) and the other is a 20 per cent sample of farms to obtain some characteristics of farms and agriculture (the agriculture sample).[1]

1. **The Population Sample.** The selection of the 20 per cent sample of households would be made by the use of systematic sampling and would be carried out as part of the enumeration procedure. Every fifth household on the schedule would be premarked as "sample household" and as the households are listed and enumerated by the interviewer, the household falling in a sample household line is automatically in the sample.

   The sample household would be asked, in addition to the information on the 100 per cent items, all the sample questions pertaining to that household. The determination of whether information on an item is to be obtained on a complete count basis or on a sample basis depends on a number of factors and will be discussed in detail in another part of this report.

   It is important to note, however, that the sample selection is based on the theory of probability sampling and bias should be prevented from creeping into the sample. The enumerator should not tamper with the sampling system by changing the normal order of listing the households in order to put certain households in or out of the sample.

2. **The Agriculture Sample.** The agriculture sample consists of two groups: (1) all the large farms and (2) 20 per cent of the remaining farms. Hence, the sample as a whole would really be more than 20 per cent of all farms. The definition of what a large farm is would be set later in accordance with the specifications

1 The proposed 20 per cent sample is expected to provide for the municipality fairly reliable summary tabulations of sample items broken down into their basic categories. Subject to the type of sample items that would finally be selected, however, the need for such information may not be realized to be of such importance as to be necessary for the municipality, but rather only for the larger area subdivisions, such as the province or region, in which case, a smaller, say, a 10 percent sample, is adequate.

to be made by the agriculture census personnel, which specifications would be based either on acreage or on value of products or on both.

The selection of farms in the sample, similar to that of the population sample, would be systematic for the non-specified farms, i.e., farms that do not meet the the specifications of a large farm. All the large farms, however, are automatically included in the sample.

As the enumerator lists the farms (or farm households). he determines whether the farm is a specified large farm or not. If the farm is a specified farm. it is in the sample and the sample questions would be asked for this farm. For the non-specified farms, the enumerator assigns to each farm unit the key letter A, B, C, D, or E, in rotation. All farms marked "B" are then automatically designated sample farms.

Adjustment of the Sample. As the census returns from the different areas will have been received at the central office, it is possibble that the numbers of sample units (households or farms) taken in some areas may not be equal to the product of the total number of units and the specified sampling proportion. In other words, the actual sampling fractions in such areas may turn out to be either slightly smaller or larger than the desired sampling ratio. It is obvious that some difficulties would be met in blowing up the sample results to the whole population if different raising factors were used. It is therefore desirable to keep the sampling fraction uniform in order to simplify the computation of estimates for each area subdivision and for the nation as a whole by the use of a single raising factor throughout the entire process.

Since the sampling plan for both the population and agriculture censuses are intended to provide tabulations of results at least at the municipality level it follows that the adjustment of the number of sample units should be made within individual municipalities. For municipalities in which the sample returns are: (a) less than 20 per cent of all households, in the case of the pupolation sample, or (b) less than 20 per cent of all non-specified farms, in case of the agriculture sample, adjjustment would be carried out by randomly duplicating the data from the necessary number of sample units

within the municipality. For municipalities in which the sample taken is more than the required number, the excess sample units would be randomly rejected or eliminated.

Selection of Items to be Taken on a Sample Basis

The determination of whether an item is to be taken on a sample basis or on a complete enumeration basis depends upon a number of considerations: (a) the amount of detail needed in tabulation, (b) precision of the sample data, (c) cost and (d) the expected magnitude of response or nonsampling error.

(a) **Amount of detail needed in tabulation.** The amount of detail needed in tabulation refers to the number of classifications to be used in tabulating the item in question. Most of the items are broken down into their basic categories every time they are tabulated and some of these basic categories are in turn subdivided into secondary categories. For example, from the labor force, the unemployed are tabulated by sex and each is in turn tabulated by age. Along with this, the area for which the item is tabulated should also be considered. Items tabulated in great detail for small areas are clearly to be taken on a 100 percent basis. As a general rule the secondary categories of the items elected for the sample should be tabulated only for large areas, such as provinces, large cities and regions.

(b) **Precision of sample data.** The use of sampling introduces sampling variability in the results which is not present in complete enumeration. For the proposed sample design, the precision of sample results depends primarily on the absolute size of the sample, i, e., the number of sample persons for the population sample, the number of sample households for the housing sample, and the number of sample farms for the agriculture sample. This means that the relative precision of the sample items is less for small areas than for large areas, since the larger areas usually have larger populations and hence provide more units in the sample. By the same logic, the greater the amount of detail used in tabulating the items, the larger will be the relative sampling error. These factors, viz., the pre-

cision of sample data, the amount of detail in tabulations and the area level for which the data is needed are therefore taken jointly in the selection of items for the sample. Basing from the results of the current sample survey of households and partly from the result of studies made in the U. S. 1950 Population and Agriculture Censuses, the items for the proposed 20 per cent samples should be tabulated not in great detail for the municipalities; cross tabulations should be made only for much larger areas such as the provinces, big cities and regions.

(c) **Cost.** The cost involved in the selection of sample items includes the cost of enumeration and the cost of processing. The cost of enumeration depends on the time spent in obtaining and recording the information and the cost of processing depends on the time spent in editing, coding, punching and verifying the data. The gains in cost obtained by the use of sampling have already been discussed earlier in this report. Nevertheless, it is clear that the greater the costs of enumeration and processing, the greater will be the potential savings from taking a given item on a sample basis.

(d) **Magnitude of response error.** This type of error is present regardless of whether the item is taken on a complete enumeration or on a sample basis. Result of studies made on the measurement of response errors have provided information on the magnitude to be expected of this type of error for the different items covered in censuses. The degree of non-sampling error attached to an item can be used as a basis for decision. A high degree of nonsampling error would obviously suggest taking the item on a sample basis.

The above considerations are the important factors that influence the choice of items for the sample. To a large extent it is really a question of precision versus cost and time. The use of sampling would substantially reduce both the time and cost but at the same time would decrease to some extent the overall precision of the results. In many cases, it is easy to

reach a decision, especially so when one consideration outweighs all the others. For instance, if an item were to be tabulated for very small areas, as city districts or sitios, then there is little choice but to have a complete count of this item. If, on the other hand, a high degree of response error were attached to an item which had a relatively very small sampling error, then it would clearly be wise to take it on a sample basis since the sampling contribution to the overall error would be every small.

Estimation Procedures

1. **Population sample.** The proposed sample design may be considered as such that for each of the area subdivisions, i,e., the municipality, the province, the region or the nation as a whole, we have a simple random sample of 20 per cent of all households and all individuals in these sample households are included in the sample. In effect, therefore, we have two samples: (a) a 20 per cent sample of the population selected in household clusters from which estimates of the population are derived and (b) a 20 per cent sample of all households for estimating household characteristics.

   It is to be noted, however, that the proposed sample is actually a systematic sample. The assumption of simple random sampling is used here for the purpose of simplifying the computation of the variances of the estimates. This assumption, however, will not give serious differences in the results since systematic and random sampling are expected in general to give results of about the same precision.

   (a) **Population Estimates.** Estimates of the population are made by computing percentages of ratios from the sample and applying these ratios to the total population counts from the census. Hence, the estimate of the number of persons having a certain characteristic is given by

$$\hat{N}_1 = \frac{n_1}{n} \cdot N = pN \quad \cdots \quad (1)$$

and the sampling variance of this estimate is given by the formula

$$S^2_{\hat{N}_1} \doteq N^2 (1-f) \frac{pq}{n} \left[1 + \delta(\bar{n} - 1) \right] \quad \cdots \quad (2)$$

where

n_1 is the number of persons in the sample having the characteristic;

n is the total number of persons in the sample;

N is the total number of persons in the population from the census;

p is the proportion of total persons having the characteristic;

q is $(1-p)$;

f is the sampling fraction;

$\bar{n}$ is the average size of household;

and s is the measure of homogeneity between persons within households.

The term $1+\delta(\bar{n}-1)$ represents the factor by which the variance of simple random sampling must be multiplied to obtain the variance for cluster sampling, and δ is the intraclass correlation among the persons listed in sample households.

A convenient equation for δ is [1]

$$\delta = \frac{\frac{M-1}{M} S_1^2 - \bar{N} S_2^2}{\frac{M-1}{M} S_1^2 + \bar{N}(\bar{N}-1) S_2^2}$$

1 See Chapter 6, Volume I, 'Sample Survey Methods" by Hansen, Hurwitz and Madow.

which for large M is approximately equal to

$$\delta \doteq \frac{S_1^2 - \bar{N}S_2^2}{S_1^2 + \bar{N}(\bar{N}-1)S_2^2}$$

A simple estimate of s from the sample is then

$$\hat{\delta} = \frac{s_1^2 - \bar{n}s_2^2}{s_1^2 + \bar{n}(\bar{n}-1)s_2^2}$$

where s_1^2 and s_2^2 are the variations between clusters and within clusters, respectively, and are defined as

$$s_1^2 = s_{1x}^2 + p^2 s_{1y}^2 - 2ps_{1xy} \quad \ldots\ldots (4)$$

$$s_{1x}^2 = \frac{\sum^{m} (x_i - \bar{x})^2}{m-1} \quad \ldots\ldots (4.1)$$

$$s_{1xy} = \frac{\sum^{m} (x_i - \bar{x})(y_i - \bar{y})}{m-1} \quad \ldots\ldots (4.2)$$

and

$$s_2^2 = s_{2x}^2 + p^2 s_{2y}^2 - 2ps_{2xy} \quad \ldots\ldots (5)$$

$$s_2^2 = \frac{1}{n} \sum_{i}^{m} \frac{n_i}{n_i - 1} \sum_{j}^{n_i} (x_{ij} - \bar{\bar{x}}_i)^2 \quad \ldots (5.1)$$

$$s_{2xy} = \frac{1}{n} \sum^{m} \frac{n_i}{n_i - 1} \sum_{j}^{n_i} (x_{ij} - \bar{\bar{x}}_i)(y_{ij} - \bar{\bar{y}}_i) \ldots (5.2)$$

where the term p in equations (4) and (5) above is the ratio of the two variables x and y and has the same meaning as the p defined in equations (1) and (2). It is an estimate of the proportion of total persons having a certain characteristic x from the sample and is obtained by any of the following relationships:

$$p = \frac{\bar{x}}{\bar{y}} = \frac{x}{y} = \frac{\sum_{i}^{m} \sum_{j}^{n_i} x_{ij}}{\sum_{i}^{m} \sum_{j}^{n_i} y_{ij}}$$

where

x_{ij} is the value of the variate for the j^{th} individual in the i^{th} sample household;

y_{ij} is the value of the ancillary variate for the same individual in the same sample household.

In this case now, x_{ij} only takes the value of 1 or 0 depending on whether the individual has the characteristic under study or not. On the other hand, y_{ij} always takes the value 1 since the variate y refers to the total persons in the sample. This means that the term $(y_{ij} - \bar{y}_i)$ is equal to zero, since $\bar{y}_i = 1$. Hence, equation (5) becomes simply

$$S_z^2 = S_{zx}^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5')$$

(b) **Household Estimates.** The three types of characteristics of the households that are estimated from the 20 per cent sample of households are: (1) the average value $\bar{X}$ of some characteristic, (2) the proportion P of households having a certain characteristic, and (3) the total value X for some characteristic. The estimates of $\bar{X}$, P, and X from the sample are $\bar{x}$, p, and x', respectively, and are obtained as follows:

$$\bar{x} = \frac{\sum_{i}^{m} x_i}{m} = \frac{x}{m}$$

$$p = m_1/m, \quad q = 1-p$$

and $$x' = M\bar{x} = (1/f)\,x$$

where x_i is the value of the characteristic for the i^{th} sample household;

x is the aggregate or total value of the characteristic for all households in the sample;

m_1 is the number of households having a certain characteristic in the sample;

m is the total number of sample households;

f is the sampling fraction.

The estimates of the variances of the above sample estimates are obtained by the use of the following formulas:

$$s^2_{\bar{x}} = (1-f)\,s^2/m$$

$$s^2_p = (1-f)\,pq/m$$

$$s^2_{x'} = M^2(1-f)\,s^2/m$$

where

$$s^2 = \frac{\sum_{i}^{m} (x_i - \bar{x})^2}{m-1}$$

and M is the total number of households counted in the census.

The estimate p of the proportion of households having certain characteristic may be used conveniently to estimate the number of households M_1 having the characteristic, which is the type of estimate called for in many of the required tabulations for housing data. The sample estimate m'_1 is obtained simply by the use of the relationship

$$m'_1 = pM$$

and the variance is

$$s^2_{m'_1} = M^2(1-f)pq/m$$

2. **The Agriculture Sample.** The agriculture sample is composed of the specified or large farms (all large farms are included in the sample with certainty) and the 20 per cent of all the unspecified farms. Estimates are therefore obtained by adding the value for the specified farms to the estimate for the unspecified farms. For example, the estimate for the total characteristic X for all farms is

$$x' = X_s + x'_u$$

where the subscripts "s" and "u" stand for "specified farms" and "unspecfied farms," respectively, and X_s is the total characteristic for the specified farms. The use of the capital letter X implies complete enumeration;

$$x'_u = (1/f)\sum_i^{n_u} x_i$$

is the estimate of the characteristic for the unspecified farms: x_i is the characteristic for the i^{th} sample farm, f is the sampling fraction and n_u is the number of un-

specified farms in the sample.

The sampling variance of x' may be estimated from the unspecified sample farms by the use of the equation

$$s^2_{x'} = N_u^2 (1-f) s^2_u / n_u$$

where

$$s^2_u = \frac{\sum_i^{n_u} (x_i - \bar{x}_u)^2}{n_u - 1}$$

$$\bar{x}_u = \frac{\sum_i^{n_u} x_i}{n_u},$$

and N_u is the total number of unspecified farms counted in the census.

An estimate of the number of farms n_i having a certain characteristic is given by the equation

$$n_i' = N_{si} + p_u N_u$$

where

N_{si} is the number of specified farms having the characteristic;

P_u is the proportion of households having the charateristic from the sample of unspecified farms.

and the estimate of the variance of n_i is given by the formula

$$s^2_{n_i'} = N_u^2 (1-f) p_u q_u / n_u$$

In modern times there is an increasing demand for more statistical data in all phases of human activity and, as the factors become increasingly dynamic, the situations must often be evaluated and acted upon within a brief period of time. In such cases, total enumeration would be a tardy and lengthy affair. If customary methods were used, this demand for more extensive data, and more rapid and accurate information could not be made by the existing census machinery within the limitation of the available budget, time and manpower. Fortunately, however, the advances attained in the field of sampling during the last few decades have often made it possible to meet the demand for additional and urgent statistical data within the limits of such available resources.

Handbook of Population Census Methods.
United Nations, New York, June 1954.