# Statistical Aspects of a Census

**Carol C. House**

This paper focuses on the statistical aspects of a census. It addresses issues such as the coverage, classification, sampling, non-sampling error, post collection processing, weighting and disclosure avoidance. The intent of the paper is to demonstrate that most (if not all) of the statistical issues that are important in conducting a survey are equally germane to conducting a census.

KEY WORDS: census, coverage, non-response, error, frames, imputation, disclosure

## 1 INTRODUCTION[1]

In this paper the author will provide a basic overview of the statistical aspects of planning, conducting and publishing data from a census. The intent of the paper is to demonstrate that most (if not all) of the statistical issues that are important in conducting a survey are equally germane to conducting a census.

In order to establish the scope for this paper, we begin by reviewing some basic definitions. Webster's New Collegiate Dictionary defines a "census" to be "a count of the population and a property evaluation *in early Rome"*. Although particularly appropriate to quote at the *CAESAR* conference, we will want to utilize a broader definition. The International Statistical Institute (ISI) in its Dictionary of Statistical Terms defines a census to be "the complete enumeration of a population or group at a point in time with respect to well-defined characteristics". This definition is more useable. We now look at the term "statistics" to further focus the paper. Again from ISI we find that statistics is the "numerical data relating to an aggregate of individuals; the science of collecting, analyzing and interpreting such data." Together these definitions render a focus for this paper -- those issues germane to the science and/or methodology of collecting, analyzing and interpreting data through what is intended to be a complete enumeration of a population at a point in time with respect to well-defined characteristics. Further, because of the nature of the CAESAR conference, this paper will direct its discussion to agricultural censuses. Important issues include the (sampling) frame, sampling methodology, non-sampling error, processing, weighting, modeling, disclosure avoidance, and data dissemination. This paper touches on each of these issues as appropriate to the paper's focus on censuses of agriculture.

## 2 FRAME

Whether conducting a sample survey or a census, a core component of

[1] This paper was presented at the **Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR), June 5-7, 2001**. Carol House is with the National Agricultural Statistics Service, Research and Development Division, and is the Division Director.

methodology is the sampling frame. The frame usually consists of a listing of population units, but alternatively it might be a structure from which clusters of units can be delineated. For agricultural censuses, the frame is likely to be a business register or a farm register. Alternatively it might be a listing of villages from which individual farm units can be delineated during data collection. The use of an area frame is a third common alternative. Often more than a single frame is used for a census. Papers presented at the Agricultural Statistics 2000 conference highlight the diversity of sampling frames used for agricultural censuses (Sward, et. al.; Kiregyera; David).

There are three basic statistical concerns associated with sampling frames: coverage, classification and duplication. These concerns are equally relevant whether the frame will be used for a census or sampled for a survey.

## 2.1 Coverage

Coverage deals with how well the frame fully delineates all population units. The statistician's goal should be to maximize coverage of the frame and to provide measures of under-coverage. For agricultural censuses, coverage often differs by size of farming operation. Larger farms are covered more completely, and smaller farms less so. Complete coverage of smaller farms is highly problematic, and statistical organizations have used different strategies to deal with this coverage problem.

The Australian Bureau of Statistics (Sward, et. al., 1998) intentionally excludes smaller farms from their business register and census of agriculture. They focus instead on production agriculture, and maintain that their business register has good coverage for that target population. Statistics Canada (Lim, et. al., 2000) has dropped the use of an area frame as part of its census of agriculture, and is conducting research on using various sources of administrative data to improve coverage of its farm register. Kiregyera (1998) reports that a typical agriculture census in Africa will completely enumerate larger operations (identified on some listing), but does not attempt to enumerate completely the smaller operations because of the resources required to do so. Instead they select a sample from a frame of villages or land areas, and delineate small farms within the sampled areas for enumeration. In the United States, the farm register used for the 1997 Census of Agriculture covered 86.3% of all farms, but 96.4% of farms with gross value of sales over $10,000 and 99.5% of the total value of agricultural products. The U.S. uses a separate area sampling frame to measure under-coverage of its farm register, and has published global measures of coverage. They are investigating methodology to model under-coverage as part of the 2002 census and potentially publish more detailed measures of that coverage.

## 2.2 Classification

A second basic concern with a sampling frame is whether frame units are accurately classified. The primary classification is whether the unit is, in fact, a member of the target population, and thus should be represented on the frame. For example, in the U.S. there is an official definition of a farm: operations that sold $1,000 or more of agricultural products during the target year, *or would normally sell that much*. The first part of the definition is fairly straightforward, but the second causes considerable difficulty with classification.

Classification is further complicated when a population unit is linked with, or owned by, another business entity. This is an ongoing problem for all business registers. The statistician's goal is to employ reasonable, standardized classification algorithms that are consistent with potential uses of the census data. For example, a large farming operation may be a part of a larger, vertically integrated enterprise which may have holdings under semi-autonomous management in several dispersed geographic areas. Should each geographically dispersed establishment be considered a farm, or should the enterprise be considered a single farm and placed only once on the sampling frame? Another example is when large conglomerates contract with small, independent farmers to raise livestock. The larger firm (contractor) places immature animals with the contractee who raises the animals. The contractor maintains ownership of the livestock, supplies feed and other input expenses, then removes and markets the mature animals. Which is the farm – the contractor, the contractee, or both?

### 2.3 Duplication

A third basic concern with a sampling frame is duplication. There needs to be a one-to-one correspondence between population units and frame units. Duplication occurs when a population unit is represented by more than one frame unit. Similar to misclassification, duplication is an ongoing concern with all business registers. Software is available to match a list against itself to search for potential duplication. This process may eliminate much of the duplication prior to data collection. Often it is important in a census or survey to add questions to the data collection instrument that will assist in a post-collection evaluation of duplication. In its 1997 Census of Agriculture, the U.S. conducted a separate "classification error study" in conjunction with the census. For this study, a sample of census respondents was re-contacted to examine potential misclassification and duplication, and to estimate levels of both.

## 3 SAMPLING

When one initially thinks of a census or complete enumeration, statistical sampling may not seem relevant. However, in the implementation of agricultural censuses throughout the world, a substantial amount of sampling has been employed. David (1998) presents a strong rationale for extensive use of sampling for agricultural censuses, citing specifically those conducted in Nepal and the Philippines. The reader is encouraged to review his paper for more details. This paper does not attempt an intensive discussion of different sampling techniques, but identifies some of the major areas where sampling has (or can be) employed.

Reducing costs is a major reason that statistical organizations have employed sampling in their census processes. We have already discussed how agricultural censuses in Africa, Nepal, and the Philippines have used sampling extensively for smaller farms. Sampling may also be used in quality control and assessment procedures. Examples include: conducting a sample survey of census non-respondents to assist in non-response adjustment; or conducting a specialized follow-up survey of census respondents to more carefully examine potential duplication and classification errors. The U.S. uses a sample survey based on an area frame to conduct a coverage evaluation of its farm register and census. It may be advantageous in a large collection of data to sub-divide the population and use somewhat different questionnaires or collection methodologies on each group. Here again is a role for sampling. For example, in order to reduce overall respondent burden some organizations prepare both aggregated and detailed versions of a census questionnaire and use statistical sampling to assign questionnaire versions to the frame units. Alternatively sampling may facilitate efforts to evaluate the effect of incentives, to use pre-census letters as response inducements, or to examine response rates by different modes of data collection.

## 4 NON-SAMPLING ERROR

Collection of data generates sampling and non-sampling errors. We have already discussed situations in which sampling, and thus sampling error, may be relevant in census data collection. Non-sampling errors are always present, and generally can be expected to increase as the number of contacts and the complexity of questions increases. Since censuses generally have many contacts and fairly involved data collection instruments, one can expect them to generate a fairly high level of non-sampling error. In fact, David (1998) uses expected higher levels of non-sampling error in his rationale for avoiding complete enumeration in censuses of agriculture.

> "… [a census produces] higher non-sampling error which is not necessarily less than the total error in a sample enumeration. What is not said often enough is that, on account of their sizes, complete enumeration CA's [censuses of agriculture] use different, less expensive and less accurate data collection methods than those employed in the intercensal surveys."

Two categories of non-sampling error are response error and error due to non-response.

### 4.1 Response Error

The literature (Groves; Lyberg, et. al.) is fairly rich in discussions of various components of this type of error. Self-enumeration methods can be more susceptible to certain kinds of response errors, which could be mitigated, if interviewer collection were employed. Censuses, because of their large size, are often carried out through self-enumeration procedures. The Office of National Statistics in Britain (Eldridge, et. al; 2000) has begun to employ cognitive interviewing techniques for establishment surveys much the same as they have traditionally employed for household surveys. They conclude that the "… use of focus groups and in-depth interviews to explore the meaning of terms and to gain insight into the backgrounds and perspectives of potential respondents can be very valuable …" They further conclude regarding self-administered collection that "…layout, graphics, instructions, definitions, routing etc. need testing." Kiregyera (1998) additionally focuses readers' attention on particular difficulties that are encountered when collecting information from farmers in developing countries. These include the "failure of holders to provide accurate estimates of crop area and production … attributed to many causes including lack of knowledge about the size of fields and standard measurement units, or unwillingness to report correctly for a number of reasons (e.g. taboos, fear of taxation, etc.)."

The statistician's role is fourfold: to understand the "total error" profile of the census, to develop data collection instruments and procedures that minimize total error, to identify and correct errors during post collection processing, and to provide, to the extent reasonable, measures of the important components of error.

## 4.2 Non-Response

The statistician's role in addressing non-response is very similar to his/her role in addressing response error: to understand the reasons for non-response, to develop data collection procedures that will maximize response, to provide measures of non-response error, and to impute or otherwise adjust for those errors.

Organizations employ a variety of strategies to maximize response. These include publicity, pre-collection contacts, and incentives. Some switch data collection modes between waves of collection to achieve higher response rates. Others are developing procedures that allow them to target non-response follow-up to those establishments which are most likely to significantly impact the estimates. (McKenzie, 2000)

A simple method for adjusting for unit non-response in sample surveys, is to modify the sampling weights so that respondent weights are increased to account for non-respondents. The assumption in this process is that the respondents and non-respondents have similar characteristics. Most often, the re-weighting is done within strata to strengthen the basis for this assumption. A parallel process can be used for censuses. Weight groups can be developed so that population units within groups are expected to be similar in relationship to important data items. All respondents in a weight group may be given a positive weight, or donor respondents may be identified to receive a positive weight. Weight adjustment for item non-response, although possible, quickly becomes complex as it creates a different weight for each item.

Imputation is widely used to address missing data, particularly that due to item non-response. Entire record imputation is also an appropriate method of addressing unit non-response. Manual imputation of missing data is a fairly widespread practice in data collection activities. Many survey organizations have been moving toward more automated imputation methods because of concerns about consistency and costs associated with manual imputation, and to improve the ability to measure the impact of imputation. Automating processes like imputation are particularly important for censuses because of the volume of records that must be processed.

Yost et. al. (2000) identify five categories of automated imputations: i) deterministic imputation – where only one correct value exists (such as the missing sum at the bottom of a column of numbers; ii) model-based imputation – use of averages, medians, ratios, regression estimates, etc. to impute a value; iii) deck imputation – a donor questionnaire is used to supply the missing value; iv) mixed imputation – more than one method used; and v) the use of expert systems. Many systems make imputations based on a specified hierarchy of methods. Each item on the questionnaire is resolved according to its own hierarchy of approaches, the next being automatically tried when the previous method has failed. A nearest neighbor approach based on spatial "nearness" may make more sense for a census, where there is a greater density of responses, than it would in a more sparsely distributed sample survey.

## 5 POST COLLECTION PROCESSING

Post collection processing involves a variety of different activities, several of which (imputation, weighting, etc.) are discussed in other sections of this paper. Here we will briefly address editing and analysis of data. Because of the volume of information associated with a census data collection, it becomes very important to automate as many of these edit and analyses processes as possible. Atkinson and House (2001) address this issue and provide several guiding principles that the National Agricultural Statistical Service is using in building an edit and analysis system for use on the 2002 Census of Agriculture: a) automate as much as possible, minimizing required manual intervention; b) adopt a "less is more" philosophy to editing, creating a leaner edit that focuses on critical data problems; and c) identify problems as early as possible.

Editing and analysis must include the ability to examine individual records for consistency and completeness. This is often referred to as "micro" editing or "input" editing. Consistent with the guiding principles discussed above, the Australian Bureau of Statistics has implemented the use of significance criteria in input editing of agricultural data. (Farwell and Raine, 2000) They contend that "… obtaining a corrected value through clerical action is expensive (particularly if respondent re-contact is involved) and the effort is wasted if the resulting actions have only a minor effect on estimates." They have developed a theoretical framework for this approach.

Editing and analysis must also include the ability to perform macro-level analysis or output editing. These processes examine trends for important subpopulations, compare geographical regions, look at data distributions and search for outliers. Desjardins and Winkler (2000) discuss the importance of using graphical techniques to explore data and conduct outlier and inlier analysis. Atkinson and House concur with these conclusions and further discuss the importance of having the macro-analysis tool integrated effectively with tools for user-defined ad-hoc queries.

## 6 WEIGHTING

When one initially thinks of a census, one thinks of tallying up numbers from a complete enumeration, and publishing that information in a variety of cross tabulations that add to the total. This paper has already discussed a variety of situations in which weighting may be a part of a census process. In this section we focus on the interaction between weighting and the rounding of data values.

Many of the important data items collected in an agricultural census are intrinsically "integral" numbers, making sense only in whole increments (i.e. the number of farms, number of farmers, number of hogs, etc.). For these data, desirable characteristics of the census tabulation is to have integer values at all published levels of disaggregation, and to have those cells sum appropriately to aggregated totals.

The existence of non-integer weights creates non-integer weighted data items. Rounding each of the multiple cell totals creates the situation that they may not add to rounded aggregate totals. This issue can be addressed in one of several ways. In the U.S., the census of agriculture has traditionally employed the technique of rounding weights to integers, and then using these integerized weights. An alternative would be to retain the non-integer weights and round the weighted data to integers. A recent evaluation of census data in the U.S. (Scholetsky, 2000) showed that totals produced using the rounded weighted data values were more precise than the total produced using the integerized weights except for the demographic characteristics, number of farms, and ratio per farm estimates. A drawback to using rounded weighted data values is the complexity these procedures add to storing and processing information.

## 7 MODELING

Modeling can be effective within a census process by improving estimates of small geographic areas and rare subpopulations. Small area statistics is perhaps one of the most important products from a census. However, a number of factors may impact the census' ability to produce high quality statistics at fairly disaggregate levels. The highly skewed distribution of data, which is intrinsic to the structure of modern farming, creates estimation difficulties. For example, many larger operations have production units which cross the political or geographic boundaries used in publication. If data are collected for the large operation and published as if the "whole" farm is contained within a single geographic area, this result will be an over-estimate of agricultural production within that area and a corresponding under-estimate within surrounding areas. Mathematical models may be used effectively to prorate the operation totals to appropriate geographic areas.

Census processes for measuring and adjusting non-response, misclassification, and coverage may produce acceptable aggregate estimates while being inadequate for use at the more disaggregate publication levels. Statistical modeling and smoothing methodology may be used to smooth the measures so that they produce more reasonable disaggregate measures. For example, for the 1997 Census of Agriculture the U.S. provided measures of frame coverage at the state level for farm counts for major subpopulations. They are evaluating several smoothing techniques that, if successful, may allow the 2002 census release to include coverage estimates at the county level instead of just state level, and for production data as well as farm counts.

Although a census may be designed to collect all information from all population units, there are many cases in which circumstances and efficiencies require that census data not stand alone. We have already discussed methodologies in which a separate survey may be used to adjust census numbers for non-response, misclassificaion and/or coverage. Sometimes sources of administrative data are mixed with census data to reduce respondent burden or data collection costs. Most often the administrative data must be modeled to make it more applicable to the census data elements. Alternatively, some census collection procedures utilize a "long" and "short" version of the questionnaire so that all respondents are not asked every question. To combine the data from these questionnaire versions may also require some type of modeling.

## 8 DISCLOSURE AVOIDANCE

The use of disclosure avoidance methodology is critically important in preparing census and survey data for publication. Disclosure avoidance can be very complex for agricultural census publications because of the scope, complexity and size of these undertakings. Disclosure avoidance is made more difficult by the highly skewed nature of the farm population. Data from large, or highly specialized, farming operations are hard to disguise, especially when publishing totals disaggregated to small geographic areas.

Disclosure avoidance is typically accomplished through the suppression of data cells at publication. A primary suppression occurs when a cell in a publication table requires suppressing because the data for the cell violates some rule or rules defined by the statistical agency. Typical rules include:

a) *threshold rule*: the total number of respondents is less than some specified number, i.e. the cell may be suppressed if it had fewer than 20 positive responses.
b) *(n,k) rule*: a small number of respondents constitute a large percentage of the cell's value, for example a (2,60) rule would say to suppress if 2 or fewer responses made up 60 percent or more of the cell's value.
c) *p-percent rule*: if a reported value for any respondent can be estimated within some specified percentage.

Secondary suppression occurs when a cell becomes a disclosure risk from actions taken during the primary suppression routines. These additional cells must be chosen in a way that provide adequate protection to the primary cell and at the same time make the value of the cell mathematically underivable.

Zayatz et. al. (2000) have discussed alternatives to cell suppression. They propose a methodology that adds "noise" to record level data. The approach does not attempt to add noise to each publication cell, but uses a random assignment of multipliers to control the effect of the noise on different types of cells. This results in the noise having the greatest impact on sensitive cells, with little impact on cells that do not require suppression.

## 9 DISSEMINATION

Data products from a census are typically extensive volumes of interconnected tables. The Internet, CD-rom, and other technical tools now provide statistical agencies with exciting options for dissemination of dense pools of information. This paper will discuss several opportunities to provide high quality data products.

The first component of a quality dissemination system is metadata, or data about the data. Dippo (2000) expounds on the importance of providing metadata to users of statistical products and on the components of quality metadata.

> "Powerful tools like databases and the Internet have vastly increased communication and sharing of data among rapidly growing circles of users of many different categories. This development has highlighted the importance of metadata, since easily available data without appropriate metadata could sometimes be more harmful than beneficial."

> "Metadata descriptions go beyond the pure form and contents of data. Metadata are also used to describe administrative facts about data, like who created them, and when. Such metadata may facilitate efficient searching and locating of data. Other types of metadata describe the processes behind the data, how the data were collected and processed, before they were communicated or stored in a database. An operational description of the data collection

> process behind the data (including e.g. questions asked to respondents) is often more useful than an abstract definition of the "ideal" concept behind the data."

The Internet has become a focal point for the spread of information. Web users expect: to have sufficient guidance on use; to be able to find information quickly, even if they do not know precisely what they are looking for; to understand the database organization and naming conventions; and to be able to easily retrieve information once it is found. This implies the need, at a minimum, for high quality web design, searchable databases, and easy to use print and download mechanisms. The next step is to provide tools such as interactive graphical analysis with drill-down capabilities and fully functional interactive query systems. Graphs, charts and tables would be linked, and users could switch between these different representations of information. Finally, there would be links between the census information and databases and websites containing information on agriculture, rural development, and economics.

## 10. SUMMARY

Conducting a census involves a number of highly complex statistical processes. One must begin with a quality sampling frame, in which errors due to under-coverage, mis-classification and duplication are minimized. There may be opportunities in which statistical sampling will help bring efficiency to the data collection or facilitate quality control measurements. Non-sampling errors will be present, and the design must deal effectively with both response and non-response errors. Post collection processing should allow both micro and macro analysis. Census processing will probably involve weighting and some type of modeling. The dissemination processes should prevent disclosure of respondent data while providing useful access by data users.

## REFERENCES

Atkinson, D., House, C. (2001) A Generalized Edit and Analysis System for Agricultural Data, *Proceedings of the Conference on Agricultural and Environmental Statistical Application in Rome,* International Statistical Institute. The Netherlands.

David, I. (1998) Sampling Strategy for Agricultural Censuses and Surveys in Developing Countries, *Proceedings of Agricultural Statistics 2000, 83-95.* International Statistical Institute. The Netherlands.

DesJardins, D., Winkler, W. (2000) Design of Inlier and Outlier Edits for Business Surveys, *Proceedings of the 2nd International Conference on Establishment Surveys*, 547-556, American Statistical Association, Washington.

Dippo, C. (2000) The Role of Metadata in Statistics, *Proceedings of the 2nd International Conference on Establishment Surveys*, 909-918, American Statistical Association, Washington.

Eldridge, J., Martin, J., White, A. (2000) The Use of Cognitive Methods to Improve Establishment Surveys in Britain, *Proceedings of the 2$^{nd}$ International Conference on Establishment Surveys*, 307-316, American Statistical Association, Washington.

Farwell, K., Raine, M. (2000) Some Current Approaches to Editing in the ABS, *Proceedings of the 2$^{nd}$ International Conference on Establishment Surveys*529-538, American Statistical Association, Washington.

Groves, R. (1989) *Survey Errors and Survey Costs*, John Wiley & Sons, New York.

International Statistical Institute (1990) *A Dictionary of Statistical Terms*, Published for the International Statistical Institute by Longman Scientific & Technical. Essex CM20 2JE England.

Kiregyera, B. (1998) Experiences with Census of Agriculture in Africa, *Proceedings of Agricultural Statistics 2000, 71-82,* International Statistical Institute. The Netherlands.

Lim, A., Miller, M., Morabito, J. (2000) Research Into Improving Frame Coverage for Agricultural Surveys at Statistics Canada, *Proceedings of the 2$^{nd}$ International Conference on Establishment Surveys*, 131-136, American Statistical Association, Washington.

Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N., Trewin. D., Eds. (1997) *Survey Measurement and Process Quality*, John Wiley & Sons, Inc. New York.

McKenzie, R. (2000) A Framework For Priority Contact of Non Respondents, *Proceedings of the 2$^{nd}$ International Conference on Establishment Surveys*, 473-482, American Statistical Association, Washington.

Scholetzky, W. (2000*) Evaluation of Integer weighting for the 1997 Census of Agriculture*, *RD Research Report Number RD-00-01* National Agricultural Statistics Service. U. S. Department of Agriculture. Washington.

Sward, G., Hefferman, G., and Mackay, A. (1998) Experience with Annual Censuses of Agriculture, *Proceedings of Agricultural Statistics 2000*, 59-70, International Statistical Institute. The Netherlands.

*Webster's New Collegiate Dictionary* (1977) G. & C. Merriam Company, Springfield, MA .

Yost, M., Atkinson, D., Miller, J., Parsons, J., Pense, R., Swaim, N. (2000) Developing *A state of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond--An unpublished staff report*, National Agricultural Statistics Service. U. S. Department of Agriculture. Washington.

Zayatz, L., Evans, T., Slanta, J. Using Noise for Disclosure Limitation of Establishment Tabular Data, *Proceedings of the 2$^{nd}$ International Conference on Establishment Surveys*, 877-886, American Statistical Association, Washington.