Survey Methodology, December 1985
Vol. 11, No. 2, pp. 111–118
Statistics Canada

Sampling Microfilmed Manuscript Census Returns

D.R. BELLHOUSE[1]

ABSTRACT

In the first part of the paper a review of the historical literature concerning microfilmed manuscript census records is given. Several types of sampling designs have been used ranging in complexity from cluster and stratified random sampling to stratified two-stage cluster sampling. In the second part, a method is given to create a public use sample tape of the 1881 Census of Canada. This work was part of a pilot project for Public Archives of Canada and was carried out by the Social Science Computing Laboratory of the University of Western Ontario. The pilot project was designed to determine the merit and technical and economic feasibility of developing machine readable products from microfilm copies of the 1881 Census of Canada.

KEY WORDS: Computerized random sampling; Microfilmed records; Multi-stage designs; Public use samples; Stratification.

1. INTRODUCTION

To write a history of any person or people the historian must rely on the applicable source material. Many historians today seek to write a history of the *common man*. In this area of historical research the source material may include items such as census returns, land records, and business directories. This paper focuses on the use of census returns as a source material. The major problem with using census data is that there is a large mass of it. For an historian with a reasonable research budget there is not enough money, time or manpower to sift through all the census returns. The solution is to take a random sample of the returns. Most census returns available to the historian are microfilm copies of the returns. In Canada this includes the colonial censuses of 1841, 1851, and 1861 and the Census of Canada for 1871 and 1881. The problem then becomes one of finding the appropriate design to sample returns from the microfilm copies.

In section 2 of the paper a review of sampling techniques that have been used by historians is given. The use of sampling techniques by historians has been very uneven. Some applications have been very good; the use of a particular technique was well thought out and applied. At the other end of the spectrum other historians appear to have used overly complex designs when it was not necessary. A complex design could lead to design effects much different from 1 which, in turn, could lead to problems in the analysis of the data. See, for example, Rao and Scott (1981) and Holt *et al.* (1980) for discussions concerning categorical data analysis and Scott and Holt (1982) for regression analysis. One other problem with many of the surveys reviewed here is that there is insufficient discussion in the survey report to ascertain the reasons why a particular design was chosen.

In section 3 of the paper a method is given to sample the returns of the 1881 Census of Canada for the purpose of creating a public use sample tape. The work was carried out as part of a project for Public Archives of Canada. The contract for the research was awarded to the Social Science Computing Laboratory of the University of Western Ontario. A description of the sampling design is given here; a complete report of the project is found in Mitchell *et al.* (1982). In some ways the design is similar to the ones used for creating public

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B9.

use sample tapes for the 1971 and 1976 Censuses of Canada. The sampling designs are all based on stratification; however, in the case of the 1881 Census, stratification could only be carried out on a geographical basis.

2. HISTORICAL REVIEW

The sampling literature for historical census documents may be categorized by the type of sampling method that was used. The order of categorization followed here will be in approximately increasing complexity of the sampling design.

2.1 Cluster Sampling

Ornstein and Darroch (1978) have given a simple cost efficient method of sampling and linking census records over time. The heart of the scheme is to form clusters of surnames and then to sample clusters. The clusters are defined by the first letter of the surname. If the same clusters are sampled over various censuses then an individual who appears in more than one census will be in the chosen sample. This reduces the number of cases to be examined for linkage purposes and hence reduces the cost. This design is particularly useful for historical studies of migration or historical changes over time.

2.2 Stratified Sampling

In all of the designs considered here that used stratification, no attempt was made to use optimal allocation. This was because prior knowledge of the variation within strata was not available to any of the researchers. To obtain the required information would have increased the cost of each project substantially.

Hammarberg (1971) used a type of two-phase or double sampling technique in an attempt to decrease the bias incurred by sampling from an incomplete set of records. The records, sampled at the second phase, were business directories for nine counties in Indiana. In the first phase of sampling, he sampled from an assumed complete record set, the 1870 United States Census. The sampling method was stratified random sampling with proportional allocation so that the sample is self-weighting. The strata were the nine counties. Two aspects of this study recur in subsequent historical sampling studies. The strata are geographical areas and the sample is self-weighting.

Hammarberg (1971) also used the classical chi-square test of fit on certain variables to see how well his sample data fit known population distributions from the census reports. In many other studies no attempt was made to check the *representativeness* of the sample.

Soltow (1975) used samples from the 1850, 1860 and 1870 United States Censuses to study wealth in the United States. For each census year he selected a sample from each microfilm reel so that the sample is stratified by reels, an approximate geographical stratification. Soltow's design appears to be a type of systematic sampling. To choose a sample he designated a spot on the screen of the microfilm reader and fed the film through the reader. The feeder arm was given successive half-turns until the manuscript census entry at the designated spot on the screen was acceptable. One criterion for sample unit selection was that the entry had to be male aged twenty years or older. Also, persons "with wealth of $100,000 or more were sampled 40 times more heavily in 1860 than those under $100,000" (p.5) so that the design is not self-weighting. Although it is not stated, the *oversampling* of wealthy people appears to have been done in order to obtain a reasonable number of them for comparison to the less affluent sections of society. Soltow (1975) also compared his sample results to the published distributions but made no statistical tests for goodness-of-fit. He found that the sample data conformed well to the census results in terms of averages and proportions on

various variables. This was true even for variables such as mean wealth, a result which is surprising in view of the oversampling of the more wealthy individuals and since his estimate appears to be the sample average.

In studying the relationship between ethnicity and occupation, Darroch and Ornstein (1980) used a sample of the 1871 Census of Canada. A description of the sampling method is given in Ornstein (1978). For the purposes of both studies it was necessary to *oversample* some ethnic groups so that the design used was not self-weighting. On ignoring the oversampling of certain ethnic groups, the sampling method used was stratified random sampling. The stratification is based on the geographical hierarchical structure of the census records : provinces, districts within provinces, sub-districts, and divisions within sub-districts. The division corresponds to the modern enumeration area. The natural stratification variable seems to be divisions. However, Ornstein (1978) further subdivided divisions into smaller groups which comprise the strata and then sampled two households per stratum. How the further subdivision was made is not given, but Ornstein states that the reason for further stratification is that sampling two units per stratum minimizes the variance of estimates of certain population values. Although it is not stated, it appears that Ornstein (1978) was trying to increase the efficiency of stratification by forming strata within a division as homogeneous as possible. By stratifying in this way the cost to sample was increased. One other aspect of Ornstein's (1978) method is that it was necessary to make at least two passes through the microfilms, the first to obtain the number of households per division and the second to sample the household.

Johnson (1978b) and Graham (1980) obtained a public use sample of the United States Census of 1900. Johnson (1978a) has described some related work in sampling the 1860 Rhode Island Census schedules. The sample was chosen by obtaining random lines on the microfilm, and then by searching for the chosen lines using a microfilm reader with an odometer attachment. Because of the sample selection procedure, the overall sample size is random. A number of criteria are given in Graham (1980, p. 41) for including or excluding sampled lines. The sampling scheme is stratified random sampling with microfilm reels as strata. The stratification is geographically based provided that the contiguous census returns are all grouped in the same microfilm. The advantages of this scheme are that it is operationally efficient and only one pass through the microfilm is needed. Also, it avoids the problem of empty strata or one unit per stratum when the sampling fraction within a stratum is small. One disadvantage is that, since one pass through the data is made, potentially major problems that arise must be dealt with on an ad hoc basis.

2.3 Stratified Cluster Sampling

Bateman and Foust (1974) obtained a sample of farms in the northern United States from the 1860 United States Census. The north was divided into two strata, East and West, and a random sample of rural counties was chosen in each stratum. Within a county one rural township (the cluster) was chosen at random and information was collected on every farm in the township. One reason for clustering appears to be due to cost considerations. The farms were obtained from the census of agriculture schedules and demographic information on the owners or operators was obtained from the census of population schedules. By remaining in the same township the work of matching farms to owners is minimized. Swierenga (1983) has provided a second reason for cluster sampling. He states that township data made it possible to estimate total factor productivity in agriculture and to identify the entire agricultural workforce, including farm laborers not residing in the 12,000 farms included in the sample (p. 793). Since the clusters, townships, were not chosen by probability proportional to size the design was not self-weighting.

Bateman and Foust (1974) also used some tests to check the representativeness of their sample. As in Hammarberg (1971), they applied the chi-square test of fit to compare sample counts to expected population counts. For continuous variables they used the t-test. The estimates of the mean and variance were the *simple* estimates, not based on the sampling design.

2.4 Stratified Two-Stage Sampling

Hammarberg (1977) used a stratified two-stage sampling scheme to sample households in the 1880 census for Utah Territory. The strata are a fairly complicated amalgamation of five geographical regions in Utah, some counties within populous regions and some large towns. Within each stratum, a sample of towns or wards was chosen. Towns which were already strata were included with certainty. Wards are geographical divisions in the Mormon Church similar to parishes in the medieval Christian church. Then a sample of households was taken from the chosen towns or wards. The sample was self-weighting on the household. The rationale for stratifying on geographical areas, given on page 460 is compelling:

> "Because the fundamental organization of the mass of people was conceived geographically, and most institutional records, – both church and secular – were organized to correspond to these areal definitions, a sample of the population on an area-by-area basis is also, in large measure, a sample of the records produced and organized for the population."

McInnis (1977) also used a stratified two-stage sampling design to obtain a sample from the 1861 Canadian Census. He studied the relationship between the number of children per family and the abundance of land in certain areas. He first stratified approximately 300 townships by their dates of settlement. Then he took a sample of townships within strata and samples of farms within townships. His reason for choosing a two-stage sample appears related to cost. A sampled farm was matched to the entry in the agricultural census. It takes less time and hence costs less to sample a few townships and match records for several farms within a township than to stratify on townships and match this record for a small number of farms. The same argument applies to Hammarberg's (1977) work. He was also linking other records to the sampled household.

2.5 Stratified Two-Stage Cluster Sampling

Smith (1978) used a stratified two-stage cluster sampling scheme to study older Americans in the 1900 United States Census. The strata are described as census regions with the counties within these regions as the primary sampling units. The primary sampling units, counties, are chosen with probability proportional to the size of their population. Within a county, several pages of census returns were sampled. Every individual over the age of 50 on each sampled page was recorded. Cluster sampling was necessary since it was too expensive to identify every individual eligible to be sampled. There is also an attempt to compare some sample distributions to the published census results. The statistic used is the standard test statistic for hypotheses on a single proportion although the data are multinomial.

A second stratified two-stage cluster sample known as the Parker-Gallman sample is described in Foust (1968, ch. 2). This sample was drawn from the 1860 Census of the United States to study the cotton growing regions in the South. The strata were 405 Southern *cotton counties*, those counties which produced 1,000 or more 400-pound bales of cotton in the year preceeding Census day. Within a county a systematic random sample of pages from the manuscript census was chosen; with a selected page a block of five farms was chosen

at random, the block being the cluster. Cluster sampling was used because information on a particular farm had to be accumulated from three different census schedules. The matching of the farms in the schedules was described as *very laborious*. Fogel and Engerman (1974, pp. 22-25) have listed several additional samples related to the Parker-Gallman sample. Bode and Ginter (1984) have criticized the content of the sample.

Of the large number of samples reviewed here, the Parker-Gallman sample and the samples drawn by Bateman and Foust (1974) are the two that have been most extensively studied. Swierenga (1983) has reviewed much of the work based on these samples.

3. PUBLIC USE SAMPLES FROM THE 1881 CENSUS OF CANADA

Early in the 1980's Public Archives of Canada obtained *Schedule 1: Nominal Return of the Living* for the 1881 Census of Canada. The returns were microfilmed and currently copies are available in most academic and many public libraries. After producing the microfilm copies, Public Archives of Canada was then interested in producing a machine readable edition of the entire census and/or a machine readable public use samples similar to the public use samples for the censuses of 1971 and 1976 (see Statistics Canada (1975, 1979) for documentation). The Social Science Computing Laboratory of the University of Western Ontario obtained a contract to perform a feasibility study and the author was asked to design a sampling scheme to construct the public use sample. In this section the proposed design is described. A report of the feasibility study is found in Mitchell *et al.* (1982).

Schedule 1 contains information on each individual on age, sex , country of birth, ethnic origin, occupation, marital status, whether or not the person had certain disabilities. The other seven schedules contain information on industry, agriculture, forestry, fishing, and mining. A brief description is found in *Census of Canada 1880-81* Vol. 1, pp. v-xv.

The basic requirements of the public use samples are briefly described. To conform to the 1971 and 1976 public use samples it would be necessary to have two independent samples, one of households and one of individuals. If production of only one sample is economically feasible, however, the first priority is the household sample. The public use sample of the 1900 Census of the United States, described by Johnson (1978b) and Graham (1980) is a sample of households. Moreover, the household appears to be the most important sampling unit desired by historians. On taking another cue from the sample of the 1900 census, a sample size in the order of one hundred thousand individuals for either the individual or the households sample is desirable. For the 1881 Census of Canada this would result in an approximate 2½% sampling fraction in either sample. Finally a stratified sampling design with proportional allocation with geographical areas as strata for both samples is desirable. This conforms to sampling practice so far in the historical literature and ensures a self-weighting design. Within a stratum the units should be chosen by simple random rather than systematic sampling. Although convenient, Johnson (1978a) has maintained that systematic sampling is not appropriate for manuscript census schedules. Neighbours possess similar characteristics and would never be included together in a systematic sample. Historians may be interested in studying those individuals with like characteristics.

Based on these basic requirements the following sampling scheme was proposed for the household sample. The design suggested was stratified random sampling with census divisions (the modern enumeration area) as strata similar to Ornstein (1978) rather than microfilm reels as used by Johnson (1978b) and Graham (1978). The census divisions provide natural geographical strata. In addition, the households are consecutively numbered on the enumerators lists with twenty-five individuals per census manuscript page. Thus, if one preliminary pass is made through the microfilms the number of households in each stratum could be easily obtained. With a 2 - 2.5% sampling fraction and proportional allocation, sample

sizes of smaller than two households are obtained in divisions (strata) with fewer than approximately one hundred households. In these cases the division should be grouped with geographically contiguous strata. Further stratification beyond the division as in Ornstein (1978) seems unnecessary and would substantially add to the sampling costs.

The sampling process can easily be made part of a computing environment. From the point of view of a coder sitting at a computer terminal with a microfilm reader to one side the sampling process is straightforward. When a coder is sampling a division, he merely presses the appropriate keys identifying the division he wants and the number of the first household to be sampled appears on the terminal screen. The coder then moves the microfilm forward to the appropriate household number. Once the data are entered, a *next* key is pressed and the second household number appears. When the final sampled household from that division is obtained, pressing the *next* key will result in an instruction to pick another division to sample. In some situations there may be missing households. For example, one or more of the enumerators sheets containing 25 names may have been lost. In this case, when a coder, in the process of sampling, encounters a missing household, the household is entered as missing and also any other missing household numbers that the coder may notice. The coder then continues sampling to the end of the division. Since at least one household sampled was missing the coder is instructed to rewind the microfilm and to continue sampling in the division. The main feature for the coder in this set-up is that with the exception of missing data situations the coder need only move the microfilm reel forward.

The computing algorithm behind this sampling method utilizes a file containing information about the divisions or division groupings and Bebbington's (1975) algorithm for drawing a simple random sample without replacement. After the initial pass is made through the microfilms a file is created containing the division identifier and the number of households in the division. If the divisions have been grouped then the size of each is recorded. When a coder identifies a division to be sampled the appropriate file entry is examined and the division size is obtained. The required sample size in the division is the division size times the sampling fraction for the whole survey which yields proportional allocation. Then Bebbington's (1975) algorithm makes a sequential choice of sample units from an ordered list, the list here being the ordered household numbers in a division. Each household number is examined in turn and is selected for or rejected from the sample. When a household number is selected the number is printed to the terminal screen and the selection procedure pauses for data entry. The sample numbers selected will be in increasing order so that a forward search only is necessary on the microfilm.

Sampling collapsed strata or grouped divisions can also be done using this algorithm. Suppose L strata of sizes $N_1, \ldots, N_L$ have been grouped into one stratum of size $N = N_1 + N_2 + \ldots + N_L$. It is necessary only to use the stratum sizes to obtain the sampled household in each stratum. Suppose in the algorithm units $s(1), \ldots, s(n)$ have been chosen for the sample, $1 \leq s(i) \leq N$. If for any $i (i = 1, \ldots, n)$ $N_1 + \ldots + N_{h-1} < s(i) \leq N_1 + \ldots + N_{h-1} + N_h$, where $N_1 + \ldots + N_{h-1} = 0$ for $h = 1$, the unit $s(i)$ is in stratum h and the household number within that stratum is $s(i) - (N_1 + \ldots + N_{h-1})$.

The general sampling algorithm can also be modified to account for missing households. The method described does not require enumerating these missing households prior to sampling. When the sampling of a stratum by Bebbington's (1975) algorithm has been completed, two possibilities arise: no missing households were encountered or some were encountered. In the former situation, there is no problem; the sampling has been completed for that situation. In the second situation, the achieved sample size, say m, is less than the desired size n. To obtain a sample of size n of the existing households, it is necessary to sample $n - m$ additional households. To achieve this, the sampling process for this stratum is started again but a list is created of the sampled and known missing households. Suppose there are M

previously sampled and known missing households ($M \geq n$: a coder may notice and record households that are missing other than those which were chosen for the sample). Define an N-dimensional vector v where the value of the u^{th} entry is u, $v(u) = u$ for $u = 1, \ldots, N$. The u^{th} entry is a pointer to the u^{th} household in a division. Now delete all entries in v corresponding to households on the microfilm which are missing or previously sampled and collapse the vector into an $(N - M)$-dimensional vector w. The values $w(u), u = 1, \ldots, N - M$ will contain the household numbers left to sample. In the algorithm, it is necessary only to restate the population size as $N - M$ and the sample size as $n - m$.

A separate and independent sample of individuals can be easily obtained using the household method of sample selection with slight modifications. The key to the modifications is that the pages of the enumerators lists are numbered with 25 names to a page. In the first pass through the microfilms, it is necessary to find the final page number and the number of lines on the last page of each division. On applying Bebbington's algorithm the computer will print the page and line number of the individual sampled.

This method of sample selection has been programmed and tested by the Social Science Computing Laboratory with positive results. For example, in the feasibility study the percentage of time spent searching for sampled units represented approximately 6% of the total estimated data entry time for the household sample and 18.5% for the individual sample. See Mitchell *et al.* (1982 pp. 20-21).

REFERENCES

BATEMAN, F., and FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.

BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.

BODE, F.A., and GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.

DARROCH, A.G., and ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.

FOGEL, R.W., and ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods.* Boston: Little, Brown and Co.

FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production.* New York: Arno Press.

GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook.* Seattle: Centre for Studies in Demography and Ecology, University of Washington.

HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.

HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880. *Journal of Interdisciplinary History*, 7, 453-476.

HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, 143, 303-320.

JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.

JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.

McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the Past,* (R.D. Lee ed.), New York: Academic Press, 201-227.

MITCHEL, S.P., LINK, D.G., and HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada.* DSS Contract Ser. No. 0SU80-00326.

ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. Unpublished manuscript, York University, Toronto.

ORNSTEIN, M.D., and DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.

RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.

SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870.* New Haven: Yale University Press.

STATISTICS CANADA (1975). *1971 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.

STATISTICS CANADA (1979). *1976 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.

SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.