An Overview of Primary Sampling Units (PSUs) in Multi-Stage Samples for Demographic Surveys[1]

Padraic Murphy
U.S. Census Bureau, Washington, DC 20233-9200

Abstract

This paper will present an overview of historical and current issues in the definition, stratification, and selection of Primary Sampling Units (PSUs) in large demographic surveys. By PSUs, we mean the large clusters (usually geographic areas such as counties) that are the sampling frame in the first stage of a multi-stage sample design. While simple in concept, the details of defining, stratifying, and selecting PSUs can prove to be surprisingly complex. We look at developments pertaining to PSUs (as used by the U.S. Census Bureau) over the past half-century, and describe some current problems. The issues discussed include (1) constraints on PSU size and boundaries, (2) choosing "building blocks" for PSUs, (3) methodology for and objectives of PSU stratification, (4) coordination among multiple surveys, and (5) coordination with preceding designs (maximizing or minimizing overlap.)

Key Words: Sample Design, Stratified Sampling, History of Surveys, Multistage Sampling

1. Introduction

In 1802, the French mathematician Laplace persuaded the French government to use a sample survey to estimate the population of France as of September 22 of that year. The method used was one he had demonstrated earlier in estimating the 1782 population of France; they first took a sample of small administrative districts known as communes, counted the total population y of each sample commune, and then used the known number of births in both the communes and the nation – x and X, respectively – to calculate the ratio estimator $Y = X(y/x)$. A similar method was used by John Graunt to calculate the population of England in 1662. (Wright, 2001)

Ninety-three years later, in 1895, the Norwegian statistician A.N. Kiaer proposed what he called a "representative method" for social and economic surveys, in which a "miniature" of the population is selected by first selecting large aggregate units such as districts or towns, and then systematically selecting sub-units within the selected aggregates. (Wright, 2001)

Both of these examples show that the idea of estimating population characteristics by selecting a sample in multiple stages – where one first takes a sample of large units and then either a census (in Laplace's method) or a sub-sample (in Kiaer's method) within each of the sampled large units – has been around for a long time.

But one finds the true genesis of complex multi-stage sample designs in the 1943 paper by Hansen and Hurwitz. It is these sample designs that made it possible for the U.S. government to begin conducting large-scale demographic surveys with acceptable levels of precision and at a reasonable cost. (Wright, 2001)

In March 1940, the Works Project Administration began conducting the monthly Sample Survey of Unemployment. That survey is known today as the Current Population Survey (CPS) and is currently the shared responsibility of the U.S. Census Bureau and the Bureau of Labor Statistics. The CPS is probably the best-known large-scale demographic survey using a complex multi-stage sample design, and is often cited as a model for other demographic surveys. For that reason I will refer frequently to CPS in this paper.

Some international surveys that have used complex designs with two or more stages include:

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

- India's National Sample Survey (Jain, 1969)
- Australia's Labour Force Survey (Australian Bureau of Statistics, 1997)
- Finland's Income Distribution Survey (EUROSTAT, 2004)

In the U.S., the Census Bureau conducts complex surveys on behalf of several other federal agencies. These surveys include the National Health Interview Survey (NHIS), the Consumer Expenditure Quarterly Interview Survey (CEQ), the Consumer Expenditure Diary Survey (CED), the National Crime Victimization Survey (NCVS), and the American Housing Survey (AHS). The Census Bureau itself sponsors the Survey of Income and Program Participation (SIPP).

Following every decennial United States census of population and housing – last conducted in the year 2000, and to be carried out again in 2010 – the Census Bureau uses the census data to update the sample designs for the ongoing demographic surveys mentioned above. This effort is called the Sample Redesign, and it provides an opportunity to introduce improvements and innovations not just in sample design, but also in the overall design of demographic surveys. The 2010 Sample Redesign, in particular, needs to help surveys adjust to some major changes in the demographic survey environment. These changes include rapidly accelerating growth in the cost of collecting demographic survey data, declining response rates, and the replacement of the decennial census "long form" by the American Community Survey. This paper is part of an effort to evaluate the current status of the Census Bureau demographic surveys and to respond effectively to these challenges.

1.1 Definition of PSUs and USUs

In the first stage of a multi-stage sample design, the sampling frame consists of a number of large aggregate units, each of which contains sub-units. We will call a large first-stage unit a Primary Sampling Unit, or PSU. The investigator selects a probability sample of PSUs, and then proceeds to the second stage of sampling, in which a probability sample of sub-units is selected from within each PSU.

One might even have further stages of sub-sampling from the sub-units, the sub-sub-units, and so on. But eventually one selects the final stage sample. We will call an element selected in the final stage of sampling an Ultimate Sampling Unit, or USU.

1.2 Some Simplifying Assumptions

In order to simplify the discussion we make the following assumptions:

- A PSU is the set of all housing units contained in a well-defined geographic area, such as a county or a group of contiguous counties.
- The USUs are the individual housing units.
- There are only two stages of sampling unless stated otherwise.
- The first-stage (PSU) sample is selected with probability proportional to size within each stratum.
- The PSU measure of size is the 16+ civilian non-institutional population.
- A survey will select one or two PSUs per stratum unless stated otherwise.
- A "survey" will refer to one of the demographic surveys conducted by the U.S. Census Bureau.

2. Motivation: Why Use Stratified Multi-Stage Sampling?

"Stratified multi-stage designs are the commonest of all types of sample designs. This is because they combine, as regards cost and efficiency, the advantages of both stratification and multi-stage sampling." (Som, 1973)

This statement is true of institutional, agricultural, environmental, and other types of sample surveys in addition to demographic surveys. However, we will confine our discussion to demographic surveys.

2.1 No Frame, or Poor Quality Frame, for USUs

Often there is no list of the USUs from which to select the final stage sample. Or, if there is, it may be incomplete or otherwise unsatisfactory. Given cost and time constraints, it is usually not practical to create such a list just for the purposes of a single survey. However, if there is a good list of larger entities (such as counties) that taken together contain all of the USUs, then it may be practical to list USUs in a small sample of these larger entities.

In addition, it is often possible in listing the USUs in the sample PSUs to collect additional information that can be used for second-stage stratification, pps sampling, and ratio or regression estimators. In fact, the ability to collect such ancillary information may be sufficient reason to do a multi-stage sample even when an otherwise complete list of USUs exists.

2.2 High Cost, Poor Efficiency of Single-Stage Design

Even if a complete list of USUs with sufficient ancillary information is available, the costs of hiring and training interviewers, in addition to the travel costs associated with personal visit interviews, may be prohibitive with a single-stage sample design. Especially for sparsely populated counties, it may be necessary to hire and train a large number of interviewers, each of whom may have a very small workload. The problem may be exacerbated if the lack of sufficient work leads to high turnover in interviewers in these locales.

In contrast, a multi-stage design allows one to concentrate the interviewing workload in a small number of sample counties in such a way that every interviewer is fully employed; and the total number of interviewers is smaller. In most cases the total time and distance traveled by interviewers will be reduced as well.

The economic motivation for multi-stage sample designs is illustrated in the following passage:

> For example, the transport and administrative costs of drawing an s.r. sample of schools throughout Scotland (say) could be enormous. We could more economically select agents in 5 local education authorities, who can speedily and cheaply collect the required information on schools in their regions, and the savings in cost for such cluster sampling or multi-stage sampling will be large. It is unlikely that any loss of efficiency of estimation (for similar sizes of sample) could outweigh the economic considerations. Indeed the efficiency loss could be easily remedied by taking a *larger* cluster or multi-stage sample whilst still possibly retaining a substantial cost saving. (Barnett, 1991)

2.3 Benefits of Stratifying PSUs

If it is possible to define and stratify PSUs in such a way that

- each PSU is internally heterogeneous with respect to the USU values of the variable(s) of interest,
- the PSUs within each first-stage stratum are homogeneous with respect to a PSU-level aggregate value of the variable(s) of interest, and
- the PSU strata are heterogeneous with respect to a stratum-level aggregate of the variable(s) of interest,

then only a small number of PSUs need be selected in each stratum (Som, 1973). In fact, most large-scale surveys select only one or two PSUs per stratum. A multi-stage stratified sample design will often achieve the same level of precision as a simple random sample or a single-stage stratified sample design, but at a significantly lower cost.

3. Three Major Tasks in First Stage Sample Design

Three major tasks in the first stage of a multi-stage stratified sample design are:

- Defining PSUs
- Stratifying PSUs
- Selecting PSUs

Defining PSUs involves deciding how to group the USUs into PSUs. Generally the only feasible way to do this is to make use of geographic entities already defined for other purposes, such as states, counties, municipalities, ZIP code or other postal delivery areas, or Census blocks or tracts. We may refer to these pre-defined entities as “building blocks.”

The PSU definition task is to (1) decide which set of building blocks to use, and (2) determine how best to put the building blocks together to create PSUs. The details of defining PSUs are the subject of section 4.

Stratifying PSUs is the process of grouping the defined PSUs in order to minimize the component of sampling variance due to the first stage of sampling. This will be the topic discussed in section 5.

Selecting PSUs after they have been stratified involves two steps. The first is assigning probabilities of selection to the PSUs. The second is carrying out the random selection of PSUs using the assigned probabilities. In some cases this is straightforward; but in the case of large national repeated surveys that are redesigning their PSU samples it can be more complex. These complexities will be further explored in section 6.

4. Defining PSUs

For large-scale demographic surveys, the first step in defining PSUs is often to identify areas that must be represented in the survey, either because estimates are specifically required for those areas or because they are large enough and different enough from other areas that to leave them out would be likely to bias the survey estimates. These areas are sometimes called "certainty" PSUs; but we will call them "self-representing" or SR PSUs. This is because one may think of an SR PSU as a stratum with only one member; therefore the first stage sample in that stratum represents only itself and is not a proxy for any other PSUs. In keeping with this nomenclature, the other "non-certainty" PSUs are called non-SR, or NSR.

Most surveys conducted by the U.S. Census Bureau define SR PSUs to conform to the delineation of metropolitan areas published by the White House Office of Management and Budget (OMB). Before 2000, these were called Metropolitan Statistical Areas (MSAs), but they are now called Core Based Statistical Areas (CBSAs).

More specifically, SR PSUs are usually determined by designating the *n* largest CBSAs -- either for the whole nation or for each state -- where *n* is chosen independently by each survey. For example, in the 2000 Redesign, the 151 largest CBSAs nationally were designated SR for CPS. (U.S. Census Bureau, 2006)

Once SR PSUs are defined, the next step is to define the NSR PSUs. For many surveys, the over-riding concern in this process is that each NSR PSU have sufficient population and be geographically compact enough that all interviewers working for the survey be fully employed but not overworked. This often translates to defining the PSU so that the expected USU sample size is equivalent to the workload for one interviewer wherever possible. Of course, some NSR counties have large enough populations that they will require more than one interviewer; but these counties are likely to either be single-county PSUs or to be the "core" of a PSU with small-population counties appended to it. In these cases, the objective is to define the PSU so that the workload will be a whole-number multiple of a single interviewer's full load. The challenge is in deciding what to do with small-population NSR counties. There are some counties that are very sparsely populated but too large to be easily covered by one interviewer.

For the 1990 and 2000 Sample Redesigns, the workload objective was expressed in terms of two constraints: (1) The minimum population for a PSU was 7,500 and (2) the maximum land area for a PSU was 3,000 square miles. (U.S. Census Bureau, 2006)

If possible after workload requirements have been addressed, it may also desirable to define PSUs in such a way that the heterogeneity of USUs with respect to the survey's target characteristics is maximized within each PSU. The assumption here is that the survey is trying to minimize the number of PSU strata formed, and therefore each PSU needs to be a good representative for other PSUs that end up in the same stratum. The reason for minimizing the number of NSR strata is to minimize the number of interviewers needed.

Another consideration is the existence of CBSA delineations for smaller metro areas and micropolitan areas. It may be desirable to define NSR PSUs to conform to these definitions so that data users can relate data from the survey to information from other sources that is summarized at the CBSA level.

The PSUs defined in the preceding sample design will usually satisfy the workload requirements stated above; and it may also be useful for other reasons to retain the previous definitions whenever possible. At the very least, they serve as a starting point for making adjustments.

For a variety of reasons, some surveys want their PSU definitions to be consistent with other surveys' definitions. For example, in the 2000 Sample Redesign SIPP, SCHIP, and NCVS all used the same PSU definitions as CPS. In some cases, the reasons may be statistical; in others, it may just be a matter of convenience.

5. Stratifying PSUs

Once PSUs have been defined, the next phase is to stratify the NSR PSUs. For the purposes of this discussion assume that one PSU will be selected from each stratum. If one also assumes that the total USU (second stage) sample size has been fixed, then the primary objective of PSU stratification is to minimize the between-PSU component of sampling variance. Another way to say this is that we are trying to group PSUs so that each stratum is as homogenous as possible at the PSU level with respect to the survey target variable. This also has the effect of maximizing the variance across strata of the stratum-level aggregates of the target variable.

The approach taken to PSU stratification at the U.S. Census Bureau for surveys such as CPS has been to see it as a disjoint clustering problem. In particular, the disjoint clustering algorithm put forth by Friedman and Rubin (1967) was incorporated into PSU stratification software in successively modified forms in the 1980, 1990, and 2000 Sample Redesigns at the Census Bureau (Friedman and Rubin, 1967; Jewett and Judkins, 1988; Kostanich, Judkins, Singh, and Schautz, 1981; Ludington, 1992.)

The software used in the 2000 Sample Redesign, known as the Stratification Search Program, or SSP, performed other functions in addition to clustering the NSR PSUs, such as calculating workloads and total variance estimates. However, we believe that there may be ways to better satisfy survey requirements and reduce the amount of human effort necessary in the 2010 Redesign; either by creating an entirely new approach or by improving the SSP. In particular, we are examining alternative clustering algorithms, especially those incorporated into software packages such as SAS, to see if these might be better suited to the problem than the modified Friedman-Rubin algorithm used in previous sample redesigns.

Aside from the algorithm used to stratify PSUs, the other important consideration is what data to use. Ideally, one would stratify on the variable of interest itself; but of course this is not possible. Instead, we look for available data that we have reason to believe will be good predictors of the variable of interest. The likely sources of data include:

- The 2010 Census
- Administrative Records
- American Community Survey data
- Historical Survey data
- Census Bureau Population and Housing projections

For the CPS, the key variable of interest is the unemployment rate; the PSU stratification characteristics used in the 2000 Sample Redesign included

- Unemployment rates measured by geography and demographic sub-populations in the 2000 Census
- Average monthly wages in certain industries (from economic survey data, 1990-1998)
- Average number of persons employed monthly in certain industries (from economic survey data, 1990-1998)
- Number of families with female head of household in the 2000 Census

A good deal of research time is spent in every Sample Redesign attempting to select the "best" set of variables to use in PSU stratification (and also in second-stage sorting and stratification of USUs.) In addition to analyzing the direct correlations of single potential stratification variables with proxies of the variable of interest, multivariate regression may also be employed when no single variable seems to do the job. Also, the situation where surveys have more than one key variable must be carefully considered, especially if the key variables would result in widely divergent stratifications in isolation.

6. Selecting PSUs

Once PSUs have been stratified, the final phase of the first stage can begin. If this were a single survey being implemented for the first time, the process would be quite straightforward. Each NSR PSU within a stratum would be assigned a probability of selection; often this is proportional to some measure of size within strata, such as total population or total housing units. The assignment of probability would also depend on the number of PSUs to be selected in each stratum (usually one or two) and -- if the number of sample PSUs per stratum is more than one -- whether the sampling is done with or without replacement. Then a PSU sample is selected in each stratum based on the assigned probabilities.

However, the assignment of PSU probabilities of selection generally has an added layer of complexity for large ongoing demographic surveys. Both to reduce cost and to maintain data quality, survey organizations prefer to minimize the number of interviewers that must be replaced when a PSU sample is redesigned. Interviewers are generally hired from within the counties in which they will be conducting interviews, or at least from a neighboring county. When a sample county is dropped and replaced by one not previously in sample, the organization is losing valuable experience and also incurring the costs of hiring and training a new interviewer. Therefore the organization wants to maximize the likelihood that counties in sample in a current PSU design will continue to be in sample in a subsequent design. At the same time, they still want the PSU sample to be a probability sample. Methods for achieving this objective are often called "maximizing overlap." (Ernst, 1999)

There may be situations in which one wants instead to minimize overlap (for example to reduce respondent burden.) More generally, the assignment of probabilities of selection that depend on the results of a different sample design is sometimes called "sample coordination."

The problem of sample coordination was complicated when surveys began to change strata definitions between Redesigns; it used to be that strata were preserved from one Sample Redesign to the next by most of the big surveys (Ernst, 1999.) Another complication is introduced by selecting two PSUs per stratum rather than just one (Kim, Flanagan, and Corteville, 2002.)

One traditional approach to the sample coordination problem is to formulate it as a transportation problem, in the parlance of Operations Research (Causey, Cox, and Ernst, 1985.) However, Matei and Tillé (2006) propose an alternative method called Iterative Proportional Fitting (IPF.) The IPF method has the advantage of requiring less computation than traditional solutions to the transportation problem.

One potentially troublesome issue with maximizing overlap is that if it is done more than once, independence assumptions that make the procedures valid may be violated. There may be bias introduced in the sense that the unconditional probabilities of selection calculated for each PSU are significantly in error (Ernst, 1999.) This, of course, will affect the weights used in calculating estimates. Research will be done to assess this risk and weigh it.

7. PSUs in 2010 and Beyond

7.1 PSU Definitions Issues

As stated in section 3, the first step in defining PSUs involves the use of CBSA delineations. Unfortunately, the anticipated date for final determination of the 2010-based CBSA delineations is too late for the 2010 Sample Redesign if PSU definitions are to be completed in time to allow for implementation of a new CPS sample in April 2013 as currently planned. A similar problem was encountered in the 2000 Sample Redesign, and the solution was to request projections of the CBSA delineations from the Population Division of the U.S. Census Bureau. Our experience with these projections was not without problems. In particular, AHS found it necessary to make major changes to their sample after the initial selection; and this required a considerable amount of effort.

This raises the question of whether to attempt compressing the schedule for defining PSUs and subsequent activities that are dependent on it; and also whether the risk incurred by using CBSA projections is high enough that we should consider postponing the planned implementation of the new sample designs.

The population and area constraints designed to optimize interviewer workloads severely limit the potential for defining PSUs to meet the heterogeneity objective. Given the fact that the American Community Survey (ACS) is now conducting interviews in every U.S. county over a three-year period, it may be possible to relax these constraints. However, this assumes that ACS interviewers would be capable of conducting the other surveys' interviews; this assumption is debatable. Still, in the current climate of constantly rising survey costs, all reasonable alternatives should be carefully examined.

One other consideration that must not be overlooked is that local governments will sometimes re-define legal boundaries in such a way that county definitions are changed. For example, whoever is defining PSUs for the 2010 Sample Redesign will need to account for the following changes that have taken place since the 2000 Census.

- A new county, Broomfield, was created in Colorado by combining parts of four existing counties.
- The independent city of Clifton Forge, VA was merged with the surrounding county of Alleghany.

7.2 Other Current PSU Sample Design Issues

The major concern for demographic surveys right now is the accelerating increase in the cost of survey operations. There are areas of inquiry in the theory of sampling that may yield more precise estimates with fewer resources required to gather the data. Some of the potentially fruitful first stage sample design topics for investigation include:

- Evaluating the potential impact of relaxing constraints on PSU size, both positive and negative
- Evaluating alternative building blocks for PSUs
- Using the same variables used for stratification to aid in PSU definition
- Evaluating/improving the algorithm and associated software used for PSU stratification
- Improving the selection and use of PSU stratification data
- Evaluating the IPF procedure for maximizing PSU overlap across time, and comparing it with the procedures used in previous redesigns
- Addressing the issue of potential bias induced by maximizing PSU overlap two or more times in succession

As suggested in the introduction, the history of multi-stage sampling is long and interesting. While it has often proven to be the cheapest and most effective way to gather important information, it is clear that there is still work to be done to improve how we do large demographic surveys.

References

Australian Bureau of Statistics (1997), "Labour Force Survey Sample Design," Information Paper 6269.0

Barnett, V. (1991), *Sample Survey Principles and Methods*, 2nd Edition, New York, Oxford University Press

Causey, B.D., Cox. L.H., and Ernst, L.R. (1985), "Applications of transportation theory to statistical problems," *Journal of the American Statistical Association*, 80, 903-909

EUROSTAT (2004) Labour force survey in the EU, Candidate and EFTA countries: Main characteristics of the national surveys, pp. 33-34

Ernst, L. (1999) "The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results," *Bulletin of the International Statistical Institute, Proceedings* Tome LVII, Book 2, 293-296

Friedman, H.P. and Rubin, J. (1967), "On some invariant criteria for grouping data," Journal of the American Statistical Association, 62, 1159-1178

Hansen, M. H., and Hurwitz, W. N., (1943) "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, Vol. 14, No. 4, pp. 333-362

Jain V.C. (1969) " A Note On One And Two-Stage Sampling From Finite Populations and Effect on Enlargement of PSU's On The Variance in Two-Stage Design," *Sankhya: The Indian Journal of Statistics*, Volume 31, Series B, Pt. 1, 2, pp. 107—112

Jewett, R. S., and Judkins, D.R. (1988) "Multivariate Stratification with Size Constraints," *SIAM Journal on Scientific Computing*, Volume 9, Issue 6, pp. 1091-1097

Kim, J.J., Corteville, D. R., and Flanagan, P.E. (2002). "Maximizing Retention of Primary Sampling Units (PSUs) in a Two-PSU Per Stratum Design," Proceeedings of the Section on Survey Research Methods, American Statistical Association, 1833-1838

Kostanich, D., Judkins, D., Singh, R., and Schautz, M. (1981), "Modification of Friedman-Rubin's clustering algorithm for use in stratified PPS sampling," *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290

Ludington, P. (1992) " Stratification Of Primary Sampling Units For The Current Population Survey Using Computer Intensive Methods," *American Statistical Association Proceedings of the Section on Survey Research Methods*, 752-757

Matei, A. and Tillé, Y. (2006), "Maximal and Minimal Sample Coordination," *Indian Journal of Statistics*, 67, 590-612

Som, R. K., (1973) *A Manual of Sampling Techniques*, London, Heinemann Educational Books Ltd, Chapters 14 and 19

U.S. Census Bureau (2006), *Current Population Survey Design and Methodology Technical Paper 66 October 2006*

Wright, T., (2001) "Selected Moments in the Development of Probability Sampling," *Newsletter of the Survey Research Methods Section*, Issue 13, pp. 1-6