DESIGNING A MULTIPURPOSE NATIONAL HOUSEHOLD PROBABILITY SAMPLE BASED ON THE 1990 CENSUS

Karol P. Krótki, Lorraine Porcellini, Temple University
Karol P. Krótki, Institute for Survey Research, 1601 North Broad Street, Room 502, Philadelphia, PA 19122

Key Words: Sample, frame, Census, issues, innovations

Introduction

The 1990 Census data are now being released and are presenting samplers and survey institutes with the opportunity to update their sample frames. In addition, the country has undergone some important changes in the last decade (e.g., increased presence of minority groups) which offer additional challenges to the sampler.

At the Institute for Survey Research, the process of redesigning the national sample is now underway and involves discussions along several dimensions. One argument centers around the issue of whether the 1980-based national sample should be retained as much as possible or whether a brand-new sample should be drawn.

Another issue is the increased need to survey, on a national basis, specific sub-populations such as women, certain age groups, and racial/ethnic minorities. This last point has become particularly important due to the notable increase in the proportion Hispanic. In addition, there is increased interest on the part of many users to focus on the Asian minorities, which so far have not received special attention.

Finally, the issue of technological advances is inextricably involved with any new sample design. For the first time, Census data will be made available on CD-ROMs offering the potential for developing and implementing the entire sample design on microcomputers. In addition, the TIGER system can be used on PCs to replace that most odious of all sampling operations, map sketching.

Standard issues such as sample size, selection of stages, selection probabilities, and weights will also be discussed in this paper.

I. ISR's 1980 National Sample

ISR's 1980 National Sample entailed a stratified multi-stage area probability sample of the 48 coterminous United States and the District of Columbia. The measures of size (MOS) used in constructing the sampling frame were based primarily on 1985 population projections, i.e., extrapolations of 1970-1980 population growth to 1985.

The construction of primary sampling units (PSUs) was the first step in developing the national sample. A PSU had to satisfy the criterion that at least one interviewer, and preferably two, could be given a full workload on a particular study within one PSU. To minimize variances and to rationalize the interviewer hiring process, each PSU represented about the same population.

The first stage selection of primary sampling units was done using probabilities proportional to size (PPS) samples. Three samples of 50, 100, and 150 PSUs each were selected. In order to maximize cost efficiency, the PSUs of the smaller design were also contained within the larger designs.

When there were sufficient numbers of people living in one location to constitute a PSU, that area was designated as self-representing. This means that the area was included with certainty in each survey. The specific rule was that if an area (group of contiguous counties) included at least 80 percent of a PSU equivalent, it would be designated as self-representing (e.g., 3.76 million for the 50 PSU design). Non-self-representing PSUs were constructed such that each had a MOS not less than 150,000 people. PSUs were first stratified by various demo-socioeconomic characteristics before the sample was drawn.

The second stage of selection involved subselections within PSUs. Subselections were made using PPS to select smaller areas, called secondary sampling units (SSUs). SSUs contained a minimum MOS of 20,000. In the third stage of selection, a single sampling unit called a Listing Area (LA) was selected from each SSU. This selection was also made using PPS. The selected LAs comprised the 1980 "master" frame.

Most ISR household surveys require a much smaller sample size than the "master" frame potentially provides. For each individual

survey, a sample of LAs is subselected at a rate that will provide the required number of interviews.

Once LAs are subselected, trained listers are sent to the field to enumerate households. The fourth stage of selection consists of randomly selecting a sample of households; the fifth stage consists of randomly selecting a sample of respondents from within selected households.

II. Population Changes, 1980-1990

The intercensal growth rate of the U.S. population was the lowest in this decade since the 1930's when it was about 7%. As can be seen in Table 1, the growth is concentrated mainly in the West and South with very few gains elsewhere. Obviously, a larger proportion of any national sample must now be drawn in the South and West.

The population as a whole is aging (Table 2). The median age has increased considerably (from 30.0 to 32.9) in the last intercensal decade and the proportion over 60 has risen from 15.8% to 16.8%. There already is evident a noticeable increase in interest in the elderly on the part of social scientists and this interest is bound to increase as the demographic trend is expected to continue unabated. Surveying the elderly poses particular problems for the survey methodologist since they are typically not clustered geographically. The challenge will be to develop new and efficient ways to locate and survey members of this population.

The average household size for the U.S. decreased over the period in question from 2.75 in 1980 to 2.63 in 1990 (Tables 3 and 4). Indications are that this trend will continue over the next decade. For the survey methodologist designing a national sampling frame, this means that household samples will have smaller units as the final sampling stage.

As can be seen in Table 5, the marked growth in the Hispanic and Asian populations is a notable feature of the change in population composition between 1980 and 1990. These populations are becoming more important for any survey not only because of their increased size but also for their social and political significance. It is already obvious that an increasing effort is being focused on these ethnic populations. Another subpopulation that requires mentioning are the Native Americans whose relative numbers are increasing but also who are the subject of an increasing number of surveys.

Finally, as shown in Table 6, the new populations of interest are concentrated in certain states of the country. This considerably facilitates the task of finding members of these groups and, potentially, interviewing them.

III. Oversampling Subpopulations

As explained in Section II, there are several subpopulations that are becoming increasingly the subject of social surveys. These include Hispanics, Asians, Native Americans, elderly, women, etc. It is impossible to develop a frame that will satisfy all these requirements. It is our recommendation that a frame be designed to anticipate oversampling for Blacks, Hispanics, Asians, and possibly Native Americans. The reason for this is that these groups can all be efficiently oversampled by limiting the sample to certain areas of the country and it is assumed that these groups will almost certainly increasingly be the target of social science surveys in this decade.

The sampling frame would thus include, in addition to the basic national sample, several oversample PSUs for the above-identified four racial/ethnic groups.

In tandem with this development it will be necessary to focus on new techniques for cost-effectively sampling rare populations such as the homeless, Ph.Ds, etc. However, it is unlikely that one overall strategy will serve all such populations. Each one is unique with respect to its identifiability and clustering and as such requires its own solution.

IV. Primary Sampling Units (PSUs)

The main question here is whether to retain or not to retain existing PSUs. On the one hand, the arguments in favor of retention are based on huge investments that have been made in staffing and training a field staff in the current 100 PSU sample.

On the other hand, about 60% of the PSUs are self-representing and therefore would remain in the sample under any design. It is also

important to point out that there is sometimes a 50% turnover in the interviewing staff from one survey to the next. Thus the nature of the investment in developing a field staff is not always so permanent that it would be an argument against changing the PSUs. Finally, it should be recognized that response burden would decrease by selecting new PSUs, especially in rural areas.

It has been the practice at ISR and at other organizations to use the county as the PSU, mainly out of practical considerations. Counties on average are about the right size and data are readily available at the county level. However, counties vary greatly in size, both with respect to population and area. Furthermore, with the availability of computerized geocoding, it has become feasible to create relatively equal-sized PSUs by combining smaller units such as tracts or block numbering areas (BNAs), block-groups, and even blocks (Table 7). For the first time, the Census has divided the rural areas entirely into blocks thus enabling such an operation to be performed throughout the entire country.

V. Technological Innovations

One of the most exciting developments for survey methodologists is the phenomenal progress in microcomputer technology. Data from the 1990 Census are now available down to the block-level on CD-ROMs for under $2,000. The data can be read on any moderately powerful PC, thus putting the Census at the fingertips of most reasonably sophisticated researchers. The manipulation of census data has become routine, cheap, convenient, and fast.

For the survey methodologist, these advances translate into a more efficient sample design in that it is now possible to try to test alternative designs. For example, the process of stratification used to be implemented manually with the help of thousands of index cards.

Stratification can now be carried out with programs that will readily permit the quick checking of the efficiency of the final product(i.e., homogeneity of the strata).

Another area in which microcomputers have facilitated the work of the sampler is geocoding. The so-called TIGER file, developed and made available by the Census Bureau, consists of geographical coordinates for all streets, blocks, bridges, and other geographical features in the U.S. The file is available at a reasonable cost and, even though the programming is not straightforward, it is not unmanageable.

Availability of this tool allows the sampler to dispose of manual sketching and, instead, to generate all listing maps by computer. The quality of the maps, based on early reports from organizations that have used this technology, is quite high. It remains to be seen how systems will develop to maintain the TIGER files updated.

VI. Conclusion

This paper has discussed a few issues that are relevant to the redesign of a national sample frame. The question that occurs at this point is whether samplers should continue relying on such an approach. Most survey organizations have a national frame in place and update it every ten years with the release of census data.

There have been several developments that might call this practice into question. It seems that the national sample frame might be receiving less attention for the following reasons:

- The increased use of telephone interviewing
- The increased requirement for oversampling certain subpopulations
- The increased request for geographically-specific samples

However, until a way is found to draw national household area samples quickly and efficiently, it is likely that survey organizations will continue to maintain area sample frames and update them whenever census data are made available.

Table 1
Population Growth by Region

Region	Growth Rate
West	22.3%
South	13.4%
Northeast	3.4%
Midwest	1.4%
All Regions	9.8%

Source: Bureau of the Census, 1990 Census Profile, Population Trends and Congressional Apportionment.

Table 2
Age Changes in Population

Age	1980	1990
5 - 20 Years	33.8%	30.3%
21 - 59 years	50.4%	52.9%
60 and over	15.8%	16.8%
Median Age	30.0 yrs	32.9 yrs

Source: Bureau of the Census, STF1A.

Table 3
Changes in Average Household Size by Region

Region	1980	1990
West	2.68	2.72
South & Northeast	2.77	2.61
Midwest	2.76	2.60
All Regions	2.75	2.63

Source: Bureau of the Census, STF1A.

Table 4
Changes in Average Household Size by Race/Ethnicity

Race/Ethnicity	1980	1990
White	2.71	2.58
Black	3.01	2.87
Hispanic	3.46	3.44
Overall	2.75	2.63

Source: Bureau of the Census, STF1A.

Table 5
Population Growth by Race/Ethnicity

Race/Ethnicity	1980	1990	%CHANGE
White	83.1	80.3	6.0
Black	11.7	12.1	13.2
Hispanic	6.4	9.0	53.0
Asian	1.5	2.9	107.8
American Indian	0.6	0.8	37.9
Other	3.0	3.9	45.1

Source: Bureau of the Census, CB91-215.

Table 6
Percent of Asians and Hispanics
Living in California, Florida, and Texas

Race/Ethnicity	1980	1990
Asian		
CA	35.8	39.1
Hispanic		
CA	31.1	34.4
FL	5.9	7.0
TX	20.4	19.4
Total	57.4	60.8

Source: Bureau of the Census, CB91-100.

Table 7
Geographic Subdivisions
1990 Census

METROPOLITAN	NON-METROPOLITAN
Metropolitan Statistical Area (Population size varies)	**County** (Population size varies)
Minor Civil Division or Census County Division (Population size varies)	**Minor Civil Division or Census County Division** (Population size varies)
Place (Population size varies)	**Place** (Population size varies)
Census Tract (Avg. Population = 4,000)	**Census Tract/BNA** (Avg. Population < 4,000)
Block Group (Avg. Population = 1,000)	**Block Group** (Avg. Population = 1,000)
Block (Avg. Population = 85)	**Block** (Avg. Population = 30)