Probability of Duplicated Selection and Its Effect on Nonresponse, Bias and Variance

Jay Jong-Ik Kim[1]
U.S. Bureau of the Census
Suitland, MD U.S.A.
jay.jong.ik.kim@census.gov

The U.S. Bureau of the Census redesigns the demographic surveys after every decennial census, i.e., every ten years. Prior to the 2000 redesign, the Bureau reduced respondent burden by making it a rule that no housing unit would be selected by more than one survey. Thus, the Bureau coordinated its sample selection operations among many surveys under its perview and "unduplicated" the housing units if they were in more than one survey. In the sampling process each survey would get its turn. For example, Survey of Income and Program Participation would select its sample first, Current Population Survey would go next, etc. The sample selection was computerized in the address frame and the frame software was designed so that housing units which were already in a survey were ignored by the following surveys in their sample selection. In the area frame, the person who selected the sample made sure that no unit was selected in more than one survey.

The National Health Interview Survey (NHIS) is the only survey considered that has an all area frame design. It uses the area frame even in large urban areas which are normally covered by an address frame in the other surveys. If a block in an urban area was selected in the NHIS, then the other surveys also placed that block in the area frame to maintain unduplication. This resulted in an increase in cost for sample selection for all of the non-NHIS surveys in the redesign and an increase in variances for some surveys. For the 2000 redesign, we conducted an investigation into the possibility of allowing duplicated selection, its effect on the nonresponse rate and sample estimate.

When a household is selected and interviewed for more than one survey, the household might not cooperate with the interviewer for the second survey and the third survey, etc. This phenomenon might be more acute when the time periods of the interviews for different surveys are not spaced far apart. On the other hand, if the interviews are spaced apart by more than a certain time period, say two years, it might not affect the respondents' attitude as much toward subsequent interviews.

In this paper, we will show, by concentrating on six major demographic surveys that the Bureau conducts, how often a household can be selected in more than one survey in selected PSU's, given the sampling intervals and the number of units selected for use for ten years, and show the impact of duplicated selection on the increase in the nonresponse rate, on the bias and the variance of the sample mean under a statistical model. Under a model we will also provide the number of units needed to make up for cases lost to refusal because of duplicated selection, if we wish to maintain the same number of interviews without duplicated selection.

---

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

These will be dealt with first from the perspective of duplicated selection between the NHIS and another demographic survey such as the Current Population Survey. We will also touch on the triplicated selection (i.e., a housing unit selected in three surveys) or higher when we let a household have a chance to be in more than two surveys. Our investigation will be restricted to duplication among six major surveys. They are Current Population Survey (CPS), National Crime and Victimization Survey (NCVS), American Housing Survey - Metropolitan Sample (AHS-MS), Survey of Income and Program Participation (SIPP), Consumer Expenditure Surveys (CE) and NHIS.

Sampling for the demographic surveys is performed in two stages. In the first stage, primary sampling units (PSU's) are selected. PSU's are generally counties or groups of counties. The PSU's which are selected with certainty are called self-representing (SR) PSU's. Other PSU's are called non-self-representing (NSR) PSU's. NSR PSU's are selected with probability proportional to some measure of PSU size. In the second stage, within each selected PSU, housing units are selected. Housing units, which are listed in order and sorted by certain criteria, are selected by
systematic sampling. Roughly speaking, when a unit is hit, a certain number of consecutive units, called a "string," is selected for use for ten years until the next redesign. The size of the string is called "string length." The string includes "reserved units" which are selected and kept for use when a need arises.

Using the sampling intervals (SI's) and string lengths, we can compute the probability that a housing unit is selected in both NHIS and another demographic survey. The typical probabilities for SR PSU's are:

NHIS with	CPS	NCVS	AHS-MS	SIPP	CE
Duplication Prob	.000300	.000127	.000064	.000343	.000464

For CPS the duplication probabilities will vary from state to state as the SI varies from state to state. For AHS-MS and CE the probabilities vary from city to city, since the SI varies from city to city. For surveys which do not require state/city estimates, the SI does not change between states/cities and the duplication probabilities remain the same across the states/cities. The highest duplication probability is observed between NHIS and CE, which is .0464%.

The Bureau allows duplicated selection with the American Community Survey (ACS). We identified households which were in both ACS and CPS. The number of housing units which were in both surveys was small. Even if the scale is very limited, there was no clear sign that the CPS non-response rate increased because some of the housing units were also in ACS. In the 2000 redesign, the Bureau and survey sponsors decided to allow duplicated selection with NHIS

Le Census Bureau produit des plans d=échantillonnage nouveaux, pour toutes ses enquêtes démographiques, après le recensement décennal. Traditionnellement, pour réduire le fardeau du répondant, les plans sont conçus en sorte que les répondants ne fassent jamais partie de plus qu=une enquête a la fois. Par contre, le *National Health Interview Survey +, qui est une enquête indépendante, utilise un plan d=échantillonnage base sure les surfaces couvertes, plutôt que sur une liste d=adresses centrale. Cela complique la tache de design pour les enquêteurs du Census Bureau, en augmentant les coûts de sélection de l=échantillonnage, et en augmentant la variance. Apres une plus longue réflexion

sur les conséquences d=avoir des répondants doubles, le Census Bureau a finalement décide de permettre les répondants de chevaucher ses enquêtes démographiques et le * National Health Interview Survey.