

9408
#196

OVERSAMPLING IN PANEL SURVEYS

Rajendra P. Singh, Rita J. Petroni, Tiwanda M. Allen
U.S. Bureau of the Census

Presented at the American Statistical Association Meetings, August 1994.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

OVERSAMPLING IN PANEL SURVEYS

I. INTRODUCTION

Survey statisticians use oversampling to reduce variances of key statistics of a target sub-population. Oversampling accomplishes this by increasing the sample size of the target sub-population disproportionately.

Survey designers use a number of different oversampling approaches. One approach requires forming two sampling strata -- one with a higher concentration of the target population for the oversampling; and the other with a lower concentration. The sample is selected at a higher rate in the higher concentration stratum and at a lower rate in the lower concentration stratum when the total sample size is fixed (Waksberg, 1973). The approach can be generalized to more than two strata. The Survey of Income and Program Participation (SIPP) in the post-1990 Census redesign, and the National Health Interview Survey (NHIS) in the post-1980 Census redesign used this approach to oversample selected population groups. For details see Huggins, et.al. (1991), Mazur (1983), and Massey et.al (1989). The efficiency of this approach depends on the success in appropriately classifying units into high and low concentration strata.

In a second approach, survey designers screen the population to identify the oversample group. Screening is done prior to or at the time of the actual interview for survey data collection. Prior screening is done using earlier survey data, administrative records or by conducting a telephone or personal screening interview. Those identified as having the target characteristics are retained with certainty and others are retained at a lower rate. The U.S. Census Bureau uses this approach for the Current Population Survey March sample to supplement the Hispanic population (Waite, 1993). The U.S. Department of Health and Human Services used it to supplement Blacks, Hispanics, the poor and near poor, the elderly and persons with functional limitations (Cohen et.al, 1987) for the National Medical Expenditure Survey (NMES). The success of this approach depends on the success of screening in identifying the target population.

In a third approach, survey designers first select a sample at a higher rate in higher concentration areas, and then screen to identify target population cases from the sample selected in the higher concentration areas. Target groups are retained at higher rates than other groups. This approach combines positive aspects of the above two approaches. The U.S. Census Bureau used this approach to oversample in the post-1990 Census NHIS redesigned sample. (See Judkins et. al. 1994). The success of this approach depends on the success in identifying high concentration areas and screening the desired oversample group correctly.

In panel surveys, analysts may consider the following two types of analyses:

- Analysis of the first interview cohort over time,
- Analysis of the oversample group data at different time intervals.

For the first type of analysis, oversampling issues are similar to one-time (cross-sectional) surveys. However, for the second type of analysis, the issue is not only succeeding in oversampling for the first interview, but also maintaining the oversampling group in subsequent interviews conducted over the life of the panel. In this paper, we will discuss issues that one should consider before deciding to oversample in a panel survey for the second type of analysis. We do not present an exhaustive list of the issues but present some general issues for survey designers' consideration.

Section II. presents special features of panel surveys and their implications for oversampling. Sections III. and IV. present results related to these issues. Section V. presents a summary and conclusions.

II. SOME SPECIAL FEATURES OF PANEL SURVEYS

In panel surveys, we conduct multiple interviews on selected sampling units over a period of time. The number of interviews, time between interviews, and period over which these interviews are conducted varies by survey due to differing survey objectives. For example, data collection for the SIPP occurs every four months, eight times over a 29-month period (Jabine, et. al. 1990). For the Panel Survey of Income Dynamics (PSID) it has occurred once every year for over the last 25 years (Duncan and Hill, 1989). For NMES it occurs every 3 to 4 months, four times over a 15 month period (Cohen et.al. 1987).

As the panel ages, some of the characteristics observed on sampling units during the first interview change (sampling units will refer to persons or group of persons for the rest of the paper). Sometimes these changes occur over a short period of time. Generally, the time between interviews and the length of the panel significantly affect the number of these changes (transitions). Obviously, more changes will occur if the panel is longer. On the other hand, some characteristics (such as race and sex) remain unchanged.

Before continuing this discussion, we define the following terms that are commonly used in analyzing panel surveys' data.

Transition: When a sample unit changes from one state, say "A", of economic and labor condition to another state, say "B", we have a transition from "A" to "B".

Spell: The transition from any state "A" to another state "B" ends a spell of state "A" and begins a spell of state "B".

Spell Length: The length of time between the start of state "A" and start of state "B".

Over a given period of time, more transitions means more and shorter spells and vice versa. Transitions have a direct effect on spells. Also, the length of a spell has a direct effect on the number of transitions.

Analysis of changes (transitions) from one state of socio-economic conditions to another state and their causes and effect on other characteristics are of great interest to analysts of panel data. These analyses could serve as a very powerful instrument in explaining socio-economic processes and helping federal agencies in developing and evaluating their policies.

Transitions over time could have significant adverse effect on meeting oversampling objectives in panel surveys. They will also have an adverse effect on the reliability of estimates of the group which was not oversampled. Even if we oversample the desired group superbly for estimates from the initial part of the panel, the gain of oversampling may disappear later in the panel due to transitions. Thus, there could be a direct conflict in oversampling a subgroup and analyzing transition and spell data.

Due to the factors stated above, oversampling in panel surveys has very different issues compared to one-time surveys. These issues revolve around transitions in target characteristics for units in the oversample group. Oversampling in a panel survey will be effective if:

- One can use characteristics of interest for oversampling in screening to select the sample and these characteristics have a high degree of stability over time. Examples of stable characteristics include sex, race, and social security reciprocity. (Time refers to the time of interest for the analysis.)

If variables (characteristics) are not stable, the efficiency of oversampling will decrease over time. Thus, for the direct screening approach to be successful in oversampling over time requires long spells (or few transitions) of the target sub-population relative to the period of analysis.

- One can use auxiliary variables that have very high correlation with the desired oversampling group to select the sample and the auxiliary variables and their correlations with the oversampling group are stable over time. Higher correlation means greater success in oversampling.

If correlation is stable, the initial oversampling (which may or may not be very successful) will be maintained. If auxiliary variables and correlations

are unstable, the success of initial oversampling may decrease as the panel ages.

In the next two Sections, we present examples of various oversampling situations using simulations and data from the 1990 SIPP panel.

III. TARGET POPULATION SAMPLE SIZES OVER TIME

In this Section, we present simulations showing how target population sample sizes are affected by various assumptions about transitions from one stage to another for three alternative designs. We did not simulate variances since for the oversampling design they will change over time. This is because the proportion of differential (larger) weights will change among the groups of interest as transitions occur. In general, we expect variances to increase.

A survey designer needs to determine for his/her case what the important estimates are and compute variances for them.

A. Notations and Assumptions

Before discussing the simulations, we outline the designs, assumptions and notations used.

Design A - self-weighting (equal probability of selecting a sampling unit) panel design with n sample cases

Design B - oversample design with two components. Component 1 is a self-weighting sample. Component 2 is obtained from a second self-weighting sample. For this component, a set of auxiliary characteristics is used for screening. All cases with auxiliary characteristics are selected in sample. In addition, component 2 includes a small proportion of sample from the remaining sample.

The total sample in components 1 and 2 is n .

Design C - modified oversample design. The sample design has two components. Component 1 is a self-weighting sample. Component 2 consists of all cases with target characteristics from another self-weighting sample. The target characteristic (for example, poverty status) is used for screening. Additionally, component 2 includes a small proportion from the remaining sample.

The total sample in components 1 and 2 is n .

Assuming no attrition,

- For designs A and B, the number of cases with the target characteristic remains the same from one year to the next.
- For design C, the number of cases with the target characteristic changes over time since cases originally in the target group may lose the characteristic.

B. Formulae

$$R_A = k_1$$

$$R_B = (a)k_1 + (b)(c)k_B + (b)(1-c)k_B'$$

$$R_{C1} = (a)k_1 + (b)p_C k_C + (b)(1-p_C)k_C'$$

$$= (a)k_1 + (b)p_C = (a)k_1 + d$$

$$R_{C2} = (a)k_1 + p_r[(b)p_C] + p_o[(b)(1-p_C)]$$

$$= (a)k_1 + p_r d + p_o d'$$

$$R_{C3} = (a)k_1 + p_r(p_r d + p_o d') + p_o(p_r' d + p_o' d')$$

$$R_{C4} = (a)k_1 + p_r[p_r(p_r d + p_o d') + p_o(p_r' d + p_o' d')] + p_o[p_r'(p_r d + p_o d') + p_o'(p_r' d + p_o' d')]$$

where

R_A = proportion of sample cases with the target characteristic for design A.

R_B = proportion of sample cases with the target characteristic for design B.

R_{Ci} = proportion of sample cases with the target characteristic for design C at the start of year i , $i = 1, 2, 3, 4$.

k_1 = rate of target characteristic for the total population.

k_B = rate of target characteristic for the group with auxiliary characteristics of component 2 for design B.