OVERSAMPLING IN PANEL SURVEYS

Rajendra P. Singh, Rita J. Petroni, Tiwanda M. Allen*
Rajendra P. Singh, Bureau of the Census, Washington, DC 20233

Key Words: Efficiency, Stability, Strata, Subpopulation, Variance

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 occured 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 recipiency. (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.
- k_B' = rate of target characteristic for the group without auxiliary characteristics of component 2 for design B.
- k_C = 1 = rate of target characteristic for the target group of component 2 for design C.
- k_C' = 0 = rate of target characteristic for the non-target group of component 2 for design C.
- p_C = proportion of component 2 from the target group for design C.
- p_r = proportion retaining the characteristic after one year.
- p_o = proportion obtaining the characteristic after one year.
- p_r' = $1-p_r$
- p_o' = $1-p_o$
- a = proportion of total sample cases from component 1
- b = (1-a) = proportion of total sample cases from component 2
- c = proportion of component 2 sample cases which have the auxiliary characteristics

C. Results

Table 1 presents proportion of persons with the target characteristic over a 4-year period under different transition and target characteristic rates. For example, rows 2 and 3 of column 3 represent sample design C where the transition rate for the target characteristics is assumed to be 50% over a year period. An assumption about the proportion obtaining the characteristics after one year is 2%, i.e., transition to the target characteristic among the total population is assumed 2%. We also assumed that a = .839, b = (1-a) = .161, c = .603. The last four rows show that even though design C increased sample for the target characteristics by 40% during the first year compared to design A, the increase was lost by the fourth year.

Column 2 of the table presents corresponding results for design B. The initial gain was only 20%. It shows that the proportion of sample cases with the target characteristic remained constant at 6% (i.e. no loss in oversampling) over four years. This is because we assumed auxiliary variables and their correlations with target characteristics are stable over time.

Table 1 shows deterioration in sample size for each simulation for design C. The amount of deterioration partly depends on the assumptions about the two transition rates -- P_r and P_o.

Table 2 presents similar results from additional simulations for different combinations of these two transition rates for design B and C. For example, column 3 shows for design C the proportion of the sample population with the characteristic when the proportion obtaining the characteristics is 0.01, much lower than the proportion (.10) of retaining it. The proportion of the sample with the characteristic after one year drops to 0.09 which is lower than for design B. If the retention rate is low (i.e., high transition rate and short spell length) the proportion of the target group will drop in the sample. On the other hand, if the retention rate is high, the proportion of the target group in the sample will remain high and it may even increase. The results also depend on the proportion of the sample from the target group and from the other group. If only a small part is from the non-target group, it takes large P_o to retain k_1 at population level.

IV. EXAMPLES FROM SIPP

We used 1990 panel SIPP data to prepare these examples primarily because its design had the oversampling feature of Design B in Section III. The following is a brief description of the oversampling design for the SIPP 1990 panel.

A. Design of the 1990 SIPP Oversample Panel

The Census Bureau introduced a panel of 23,600 households which included an oversample of the low income population. Instead of screening for low income, the Bureau used demographic characteristics of those who were occupying the sample housing units during February - May 1989 as auxiliary variables. These characteristics are: Black (BLK), Hispanic (HIS),

and female headed with no spouse present living with relatives (FHNSP) households. Such households tend to have higher poverty rates than the general population. (King, 1990.) Table 3 presents sample size by various sample components.

We expected this oversampling to reduce variances of low income and related estimates and increase other variances compared to the regular SIPP of the same size.

B. Results

Allen, et. al. (1993) compared oversampling and non-oversampling designs to show how transitions affected variances of selected characteristics in SIPP. They found that oversampling based on auxiliary variables that were stable over time performed better than oversampling based on household's low income status in SIPP over the life of the panel. In terms of sample size, they found that due to transitions from low income, only 61% of the households with low income status in wave one had the same status in wave 8 of the 1990 panel. The corresponding percentage when stable auxiliary variables were used was 67%--about 10% higher than when oversampling was done based screening on low income status of households. Thus, oversampling based on stable auxiliary variables with higher correlation with low income retained more sample after 8 interviews over about 2 1/2 years.

We use some examples from the SIPP data to show initial gains in variances based on the oversampling design as compared to the regular SIPP design. In this paper, we compare the variances of low-income, program participation and other (such as labor force, high income, etc.) variables for the regular SIPP design with the above SIPP oversample design. We used a replication method to compute variances for the first quarter of 1990 for both designs. For the regular design, we computed variances for the 1990 panel component and then adjusted to the sample size of the oversample design. Below is a brief summary of our results.

- Variances of 55% of low-income estimates for total population were lower for the oversample design
- Variances of 60% of program participation estimates for total population were lower for the oversample design
- Variances of 26% of other estimates (such as labor force, and income estimates) for total population were lower for the oversample design

We also analyzed three sets of variances by selected auxiliary variables used for oversampling. We found the following:

Low Income Estimates:

- Variances of 79% of low-income estimates for Blacks were lower for the oversample design
- Variances of 83% of low-income estimates for Hispanics were lower for the oversample design

Program Participation Estimates:

- Variances of 73% of program participation estimates for Blacks were lower for the oversample design
- Variances of 77% of program participation estimates for Hispanics were lower for the oversample design

Other Estimates

- Variances of 70% of other estimates for Blacks were lower for the oversample design
- Variances of 26% of other estimates for Hispanics were lower for the oversample design

These results are summarized in table 4.

V. SUMMARY AND CONCLUSIONS

Transition and spell analyses are important for analysts using panel data. Few or no transitions are not of great interest. However, transitions could have significant effect on the efficiency of the oversampling in a panel survey. Therefore, oversampling of a target group in a panel survey should be thought through very carefully before implementing. We are in no way suggesting that oversampling should be avoided in a panel survey. But, its usefulness should be evaluated in terms of its long term effect on the goals of oversampling. Its usefulness depends on various factors. Some are listed below.

Transition Rates

As stated earlier, transitions could have significant effect on the efficiency of oversampling in a panel survey. The higher the transition rate, the lower the efficiency from oversampling.

Spell Length

Spell length also has an effect on oversampling. Longer spell lengths mean fewer transitions. In general, it means that the loss in efficiency of oversampling will be small. If spell length is small and transitions are occurring between the same two states for the same group of sample units, oversampling may remain effective.

Length of Panel

Even low transition rates could have an adverse effect on the success of oversampling if the life of the panel is very long. Over a longer period of time, these transitions will have a cumulative effect similar to a large number of transitions.

Objective of Oversampling

It is critical to know if oversampling will meet its goal. One should consider the impact of the above factors in addition to how to oversample initially.

There are a number of other factors that we have not discussed which will have an effect on oversampling. For example, the parameters we used in our simulation in Section III. Our primary goal here is to focus survey

designers' attention on the complexity involved in oversampling in panel surveys. There are some possible remedies that one may consider in dealing with inefficiencies of oversampling. Czaika and Schrim (1992) have discussed some options in their paper. Survey designers should evaluate their own situation in making design decisions.

ACKNOWLEDGEMENTS

The authors would like to thank Rameswar Chakrabarty, and J. Kim for reviewing this paper and providing thoughtful comments. They would also like to thank Carol Macauley for her excellent typing and patience through the numerous revisions.

REFERENCES

Allen, T., Petroni, R., Singh, R., (1993), "The Effectiveness of Oversampling Low Income Households in the Survey of Income and Program Participation," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Cohen, S., DiGaetano, R., Waksberg, J., (1987), "Sample Design of the National Medical Expenditure Survey - Household Component," 1987 Proceedings of the Section on Survey Research Methods Section, American Statistical Association.

Czaika, J., Schirm, A., (1992),"Selection and Maintenance of a Highly Stratified Panel Sample," Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys, November, 1992.

Duncan, G., Hill, D., (1989), "Assessing the Quality of Household Panel Data: The Case of the Panel Study of Income Dynamics," Journal of Business and Economics Statistics, October 1989.

Huggins, V., Weller, G., Singh, R., (1991), "Oversampling the Low Income Population in the Survey of Income and Program Participation (SIPP)", Proceedings of the Section on Survey Research Methods, American Statistician Association.

Jabine, T., King, K., Petroni, R., (1990), "SIPP Quality Profile", Bureau of the Census, Department of Commerce.

Judkins, D., Marker, D., Waksberg, J., (1994), "National Health Interview Survey: Research for the 1995 Redesign," Draft Prepared for National Center for Health Statistics, Center for Disease Control, Contract No. 200-889-7021.

King, K., (1990), "SIPP: Restructuring the 1989 and 1990 Panels," Internal Census Bureau Memorandum for Documentation, January 31, 1990.

Massey, J., Moore, T., Parson, V., Tadros W., (1989), " Design and Estimation of the National Health Interview Survey, 1985-1994," National Center for Health Statistics, Vital Health Stat 2(110).

Mazur, C., (1983), "HIS Redesign: Differential Sampling to Achieve a Reduction in Black Subgroup variances", Internal Census Bureau Memorandum for Shapiro from Chakrabarty, June 15, 1983.

Waite, J., (1993), "Source and Accuracy Statement for the March 1993 Current Population Survey Microdata File," Internal Census Bureau Memorandum for Turner from Waite, July 6, 1993.

Waksberg, J., (1973), "The Effect of Stratification with Differential Sampling Rates on Subsets of the Population", Proceedings of the Social Statistics, American Statistical Association.

* 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.

Table 1. Proportion of Sample Persons with Target Characteristic

Rate of Target Characteristic for Total Population (k_1)	.05			.10			.15			.20		
Design	A	B	C	A	B	C	A	B	C	A	B	C
Rate of Target Characteristic for group with auxiliary characteristics for Design $B(k_B)$	--	.21	--	--	.46	--	--	.71	--	--	.68	--
Proportion Retaining Characteristic after One Year (p_r)	--	--	.5	--	--	.25	--	--	.5	--	--	.5
Proportion Obtaining Characteristic after One Year (p_o)	--	--	.02	--	--	.25	--	--	.25	--	--	.22
Begining of Year:												
1	.05	.06	.07	.10	.13	.14	.15	.20	.21	.20	.24	.27
2	.05	.06	.06	.10	.13	.12	.15	.20	.19	.20	.24	.23
3	.05	.06	.05	.10	.13	.12	.15	.20	.18	.20	.24	.22
4	.05	.06	.05	.10	.13	.12	.15	.20	.18	.20	.24	.22

Table 2. Proportion of Total Sample with Target Characteristic when the Rate of the Target Characteristic for the Total Population is .10 and k_B = .46 for Varying p_o and p_r.

Start of year	Design B	Design C																			
		p_o																			
		.01		.10			.15			.20			.25			.30			.35	
		p_r		p_r			p_r			p_r			p_r			p_r			p_r	
		.10	.95	.10	.75	.95	.10	.60	.75	.10	.45	.60	.10	.30	.50	.10	.15	.40	.10	.35
1	.13	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14	.14
2	.13	.09	.14	.10	.14	.15	.11	.13	.14	.11	.13	.14	.12	.13	.14	.12	.12	.14	.13	.14
3	.13	.09	.13	.10	.13	.15	.11	.13	.14	.11	.13	.14	.12	.13	.14	.12	.13	.14	.13	.14
4	.13	.09	.13	.10	.13	.16	.11	.13	.14	.11	.13	.14	.12	.13	.14	.12	.13	.14	.13	.14

Table 3. Components of the 1990 SIPP Oversample Panel.

Components of Oversample Panel	Number of Eligible Households
Households in addresses originally to be first interviewed in the 1990 panel.	19,700
Households associated with sample addresses which were to first be interviewed in February through May 1989 (i.e., households originally to be in the 1989 panel[1]) and were at that time headed by a Black, Hispanic, or FHNSP.	2,700
Households in one-ninth of all other 1989[1] panel sample addresses.	1,200

[1] The Census Bureau attempted to interview households in all sample addresses from the 1989 panel in February 1989 through January 1990. After January 1990, we did not interview for the 1989 panel. However, for the 1990 oversample panel, we interviewed the 1989 panel households included in the 1990 oversample panel.

Table 4. Proportion of SIPP Estimates with Higher or Lower Variances Under Oversample Design as Compared to Regular SIPP Design (Variances for the First Quarter, 1990)

Estimates		Proportion of SIPP Estimate with							
		Higher Variance for Oversample				Lower Variance for Oversample			
		Total	<10%	10-20%	>20%	Total	<10%	10-20%	>20%
Low Income Estimates									
Total Pop^n	(224)	.45	.55	.27	.18	.55	.40	.45	.15
Black Pop^n	(42)	.21	.67	0	.33	.79	.39	.36	.24
Hispanic Pop^n	(41)	.17	.43	.14	.43	.83	.41	.38	.21
Program Participation Estimates									
Total Pop^n	(171)	.40	.49	.23	.28	.60	.41	.38	.21
Black Pop^n	(44)	.27	.50	.17	.33	.73	.28	.34	.38
Hispanic Pop^n	(43)	.23	.10	.10	.80	.77	.36	.36	.28
Other Types of Estimates									
Total Pop^n	(156)	.74	.57	.35	.08	.26	.39	.49	.12
Black Pop^n	(33)	.30	1	0	0	.70	.35	.61	.04
Hispanic Pop^n	(34)	.74	.40	.44	.16	.26	.44	.33	.22

NOTE: Total number of estimates examined are given in parenthesis.