DISTRIBUTION OF POVERTY IN CENSUS BLOCK GROUPS (BGs) AND IMPLICATIONS FOR SAMPLE DESIGN

Joseph Waksberg, Westat, Inc.
Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Sample design, rare populations, low income population

1. Statement of Problem

The sponsors of many Government surveys have a particular interest in analyses of low income households as well as the total population and would like to oversample such households. This, in fact, is done in number of national studies e.g., the Census Bureau's SIPP and the Department of Agriculture's Food Consumption Survey. There have been other surveys in which the sponsors had similar interests but were prevented by the high cost of screening enough sample cases to obtain sufficient low income persons for the required analyses. This paper describes research relating to the efficiency of certain methods of oversampling.

The simplest and most direct method of oversampling is to start off with an initial sample that is much larger than the sample size desired, screen the entire sample, retain all low income households, and subsample the others. However, screening such a large sample is quite expensive.

Another method of oversampling is to identify geographic areas with high concentrations of low-income persons and oversample households within these areas. It has been shown that two conditions are necessary for geographic oversampling to be effective: a large part of the target population should live in these areas; and second, the target population should be a substantial proportion of the total population in the identified areas (Waksberg, 1973, and Kalton, Anderson, 1986). As part of more general research for the 1995 revision of the National Health Interview Survey sample design, Westat examined data from the 1990 Census to see whether the distribution of low income persons satisfies the conditions required to make this method efficient. The Census summary tapes provided the data used for conclusions on sampling efficiency. An earlier analysis of 1990 Census data for minority populations showed that the procedure was quite effective for oversampling blacks and Hispanics, but not very useful for Asian-Americans or Native Americans (Massey, Judkins, Waksberg, 1993).

2. The 1990 Census Distribution of Poverty

1990 Census block group tabulations of the number of low-income persons were prepared for the entire U.S. showing the distribution of low-income persons according to the percentage of low-income persons in the block group. Since Government programs targeted to the low incoming population do not all use the same definition of low income, three alternative definitions were used: persons with income below the poverty level, below 125 percent of poverty, and below 150 percent of poverty. The STF-3 file does not contain data on 125 or 150 percent of poverty broken down by race or ethnicity. Data used for separate analyses of blacks, Hispanics, and persons who are neither black nor Hispanic are therefore restricted to those below the poverty level.

Table 1 shows the 1990 distribution of the low income population by block groups classified according to the proportion of low income population in the BG. The BG's in each of the classes depends on the definition of low income. Thus, for the first column, the classification refers to block groups in which under 5 percent of the population is below poverty, 5-10 percent is under poverty, etc. In the next column, the classification refers to BG's in which under 5 percent of the population is below 125 percent poverty, etc. Similarly, in the third column, the classification refers to persons below 150 percent of poverty. The figures shown in the table are the percentages of low-income persons in each class.

Table 1. Percent distribution of low income population by percent of low income in BG, for three alternative definitions of low income, 1990

Percent of BG population that is low income	Low income = under poverty	Low income = under 125% poverty	Low income = under 150% poverty
Total	100.0%	100.0%	100.0%
< 5%	5.8	3.2	1.8
5-9.9	12.3	8.3	5.7
10-19.9	24.8	21.0	16.8
20-29.9	19.8	20.2	19.2
30-39.9	14.3	15.9	17.0
40-49.9	10.0	12.2	13.7
50% +	13.0	19.3	25.7

Source: Tabulation by Westat of 1990 Census STF-3 file.

Table 1 shows a rather flat distribution of low income among the classes for all three definitions. The concentrations are a little greater for persons under 150 percent than for the other two definitions but even for this group it is not very great. As can be seen, with this definition, only about 25 percent of the poor live in BG's where 50 percent or more of the population is poor. The comparable percentages are 19 percent for persons below 125 percent of poverty and only 13

percent for persons below 100 percent of poverty. Such distributions imply that oversampling households in the strata with relatively high percentages of low-income persons will not be much better than oversampling and screening the entire sampling frame. However, we will look into this more quantitatively a little later.

Earlier studies in the 1970's and 1980's showed that although only a minority of the poor lived in defined poverty areas, the proportions were very different among race/ethnic groups, with well over half of blacks and Hispanics living in these areas but only about 20 percent of the white poor. The 1990 Census data show that at the block group level, the differences between whites and blacks or Hispanics is even more dramatic. It can be seen in Table 2 that 40 percent of blacks and 32 percent of Hispanics live in block groups in which a majority of the black or Hispanic residents are poor. Only 6 percent of nonblack, nonHispanics live in similar circumstances. Block groups with 30 percent or more poor do not require an inordinate amount of screening. About 75 percent of blacks below poverty live in such block groups and 69 percent of Hispanics; for nonblack nonHispanics, the percentage is only 19 percent, and for the total poverty population, it is 37 percent. The distributions imply that although oversampling is of dubious value for the total poverty population and for whites, it may be useful when the target population is black or Hispanic poverty. As will be seen later, this is confirmed by a more detailed analysis.

Table 2. Percent distribution of persons below the poverty level by race/ethnicity when BG's are classified by percent of the race/ethnic group in the BG below poverty, 1990

Percent of race/ethnic group in BG below poverty	Total (%)	Black (%)	Hispanic (%)	Nonblack and nonHispanic (%)
Total	100.0	100.0	100.0	100.0
< 5%	5.8	0.6	0.6	10.4
5 - 9.9	12.3	2.2	2.4	19.6
10-19.9	24.8	8.8	11.0	32.6
20-29.9	19.8	13.8	17.0	18.1
30-39.9	14.3	17.0	19.3	9.0
40-49.9	10.0	17.3	17.7	4.6
50%+	13.0	40.4	32.0	5.6
Race/ethnic group as percent of total poverty	100.0	26.7	17.3	56.1

Source: Tabulation by Westat of 1990 Census STF-3 file.

A closer look at the distributions reveals that blacks and Hispanics below poverty are concentrated not so much because the poor are concentrated, but because most blacks and Hispanics, poor and nonpoor, live in predominantly black or Hispanic areas. Table 3 compares the distributions of black and Hispanic poor when BG's are separately classified by the percent of the minorities who are poor in the BG, and by the percent blacks or Hispanics in the BG. The distributions are not greatly different, particularly the proportion of the poor in the most dense stratum.

Table 3. Percent distribution of blacks and Hispanics below the poverty level when BG's are classified by percent blacks and Hispanics in the BG and also by percent black and Hispanics below poverty in BG

	Blacks		Hispanics	
Percent in BG	BG's classified by % blacks in poverty in BG	BG's classified by % blacks in BG	BG's classified by % Hispanics in poverty in BG	BG's classified by % Hispanics in BG
< 5%	0.6	4.0	0.6	4.6
5 - 9.9	2.2	3.7	2.4	5.1
10 - 29.9	22.6	13.2	28.0	19.9
30% +	74.7	79.0	69.0	70.3

3. Sample Design Implications of 1990 Distributions

Equation 1 shows the standard formula for optimum allocation of the sample, with a fixed sample size. It assumes that the within stratum population variances are the same in all strata, in most cases a fairly reasonable approximation.

If the total sample size desired is n, then the optimum allocation in the strata is

$$n_i = \frac{P_i/\sqrt{c_i}}{\sum P_i/\sqrt{c_i}} n \qquad (1)$$

where

P_i = the proportion of the low income population in stratum i; and

c_i = the ultimate cost of a single completed case in stratum i, including the cost of screening enough sample units to identify one member of the target population, and interviewing and processing cost.

Let

k = ratio of the total cost of screening, interviewing and processing a target person to the cost of screening a nontarget person;

r_i = ratio of the total population in the i-th stratum to the target population in that stratum;

c_i is proportional to $k + r_i - 1$;

$$n_i = \frac{P_i/\sqrt{k+r_i-1}}{\sum P_i/\sqrt{k+r_i-1}} n \qquad (2)$$

The key parameters are: the percentage distribution of the target population among strata (P_i), the ratio of interviewing and processing cost per case to

screening cost (k); and the amount of screening in a stratum needed to locate one member of the target population (r_i). Although we have concentrated up to now on the values of P_i, the other two parameters are also important factors in the efficiency of oversampling schemes. Note that the P_i's and r_i's come from the distribution of the target and total population, and we have used the Census as approximations to the current values. The value of k, however, depends on the survey operations, and will vary from survey to survey.

Equation (3) shows the variance of the optimum design under some simplifying assumptions. As for Equation (1), it is assumed that all the within-stratum population variances are equal, and that they are equal to the overall population variance. In practice, the population variances are rarely known prior to the conduct of a survey, and some assumptions are necessary to proceed with a sample design. Also, a term of a lower order of magnitude has been dropped from the expression for the variance. This term has approximately the effect of a finite population correction factor, and is trivial in almost all real population surveys.

The variance of a sample with optimum allocation, when variances are approximately the same in the strata

$$V^2 = \left(\sum P_i\sqrt{c_i}\right)\left(\sum P_i/\sqrt{c_i}\right)s^2/n \qquad (3)$$

(See Hansen, et al., formula 12.4.)

This formula drops a term with a lower order of magnitude. The total cost of stratified sample is

$$c = \sum n_i c_i = \frac{\sum P_i\sqrt{c_i}}{\sum P_i/\sqrt{c_i}}n \qquad (4)$$

Through some straightforward algebra, Equation (3) can be used to derive the variance of a SRS at the same cost as the optimum allocation. With SRS, the cost of an interview is $k + r - 1$ times the screening cost, where r is similar to r_i, but is the ratio for the total population. For the same cost as the stratified sample, SRS will provide a sample size of

$$\frac{\left(\sum P_i\sqrt{c_i}\right)n}{\left(\sum P_i/\sqrt{c_i}\right)(k+r-1)}.$$

When the population variance is approximately the same as the variance within strata, the variance of a sample with SRS is

$$\hat{V}^2 = \frac{(k+r-1)\left(\sum P_i/\sqrt{c_i}\right)s^2}{\left(\sum P_i\sqrt{c_i}\right)n}. \qquad (5)$$

The reduction in variance can be measured by $V^2/\hat{V}^2$. The ratio of the two variances is a measure of the efficiency of the optimum allocation. The lower the ratio, the greater the efficiency from oversampling in the higher density strata.

Chart 1 shows the ratio of the variance of the optimum sample to an SRS at the same cost for statistics relating to the population below poverty. Data are shown for several types of geographic areas. Chart 1 indicates that there appears to be moderate advantages to oversampling when k is under 3 or 4, about a 10 to 15 percent reduction in variances. When k is as large as 10, the gains are very slight, and there is virtually no advantage to oversampling BG's with high levels of poverty when K is 20 or larger. There is very little difference in effectiveness among the three types of geographic areas. An examination of more detailed tables indicates that the effectiveness is about the same for other types of geographic breakdowns, e.g., states, large or small MSA's, etc. Conclusions drawn from this analysis will thus approximately apply to subnational surveys.

The value of k affects the efficiency of using geography for oversampling. The reason is that oversampling in BG's with high proportions of low-income persons trades screenings for interviews since the oversampling reduces the amount of screening necessary to identify the desired sample size, but it results in the application of variable sampling rates which requires a larger number of interviews for a specific level of precision. When the cost of screening a household is close to the interview cost, that is, when k is small, it makes sense to use a procedure that reduces the amount of screening even if more interviews are needed. When screening costs are low relative to interview costs, that is when k is large, it doesn't cost much to screen the additional sample, as required by an equal probability sample, and there is little or no advantage in oversampling selected BG's for which a larger number of interviews for equivalent precision is necessary. It should also be noted that introducing variable sampling rates may result in unexpectedly high variances for statistics on subgroups of the population that tend to be concentrated in the non-low income strata.

Chart 1 shows a wide range of values of k, from 1 to 60. In order to determine whether to apply geographic stratification and oversampling in any particular survey, the value of k appropriate to the survey needs to be estimated. Government surveys vary greatly in the effort and cost of carrying out the necessary measurements on a sample case, and this is reflected in the value of k. For example, the Census Bureau has estimated that for the National Health Interview Survey (NHIS), the interview cost is about three times the screening cost per sample household. It is likely that k will also be approximately three in other surveys involving interviews that normally take about 30 minutes and that have fairly rigorous follow-up rules for both interviewing and screening, such as CPS or the American Housing Survey. At the other extreme is a survey like NHANES that involves not only detailed interviews but also physical examinations for each sample person which frequently takes 3 to 4 hours. The examinations require payments to medical and other highly trained personnel, mobile medical

examination facilities, and cash incentives to each participant. The value of k is probably in the 40 or 60 range for NHANES. There are Government surveys which are semi-longitudinal, such as SIPP or the National Medical Expenditure Survey. For such studies, k is probably about 15 or 20. In any particular survey, it is necessary to give some thought to the likely value of k, before making a decision on whether to stratify and oversample some strata and what the oversampling rates should be.

Chart 2 shows a quite different picture than Chart 1. For the black and Hispanic poor, there is over a 60 percent reduction in variance for low values of k. There are also substantial reductions even for high values of k, amounting to 25 or 30 percent when k is 40. If one wants to target black or Hispanic poor, rather than all persons below the poverty level, substantial reductions in variance are possible. This, of course, is a consequence of the much greater geographic concentrations of black and Hispanic than of other low-income persons.

Chart 3 confirms the conclusions one could have reached from the information presented earlier on the distributions of the low-income blacks and Hispanics. Stratification by concentration of all blacks or Hispanics in the BG is not quite effective as stratification by low income, but it comes fairly close. There is about a 10 percent additional improvement through use of low-income stratification, but both show substantial gains over SRS for all values of k.

The fact that the gains in efficiency from stratification by percent blacks or Hispanics in the BG's are almost as great as stratification by the percent below poverty has a bearing on surveys that wish to oversample both all minorities and those below the poverty level. A single mode of stratification serves both purposes quite well.

4. Qualifications of the Analyses

There are important limitations and qualifications of the data and the resulting conclusions. One is the effect of sampling errors. The income data in the 1990 Census are based on a one-sixth sample. The sample size in a typical block group was a little under 100 households. The classification of blocks according to percentage of low-income persons therefore has a fair amount of fuzziness to it, and many block groups will not be in the categories that Census data assign them, but in neighboring classes.

Another limitation comes from the fact that the Census income distributions reflect the situation in 1990. By mid-decade and later, there will be shifts in the distribution of low-income persons. Information on the magnitude of these shifts does not seem to exist. For statistics on minorities (not restricted to low income persons) the population movements in the course of a decade cut in half the gains from stratification and oversampling. The changes in the low income distribution may not be as great, but they are undoubtedly significant.

Consequently, one should not expect whatever potential gains the 1990 Census analyses seem to show from geographic stratification and oversampling to apply in practice. It is sensible to assume that not more than half the gains will actually occur.

It should also be kept in mind that the analyses in this report relate to surveys in which the principal goal is to prepare statistics on the low-income population. When other population domains are also of interest, the additional requirements for these domains need to be taken into account in the development of the sample design and the conclusions in this report may have to be modified.

5. Overall Conclusions

The analyses of the 1990 Census data indicate that for surveys of all low-income persons, only small gains are possible with oversampling, and those only when the cost of screening a household is a substantial part of the cost of a complete interview, say, one-third as great or more. Most of these gains are likely to disappear when the limitations discussed in Section 4 are taken into account. In fact, by the middle of a decade or later, when Census data become seriously outdated, there is a strong possibility that stratification and oversampling will reduce efficiency rather than increase it because of the poor relationship between the census data and what is measured in screening.

Stratification and oversampling is a useful device when the focus of interest is on the black or Hispanic poor. The stratification can be either by low-income blacks or Hispanics or by the proportion of the minorities in the BG. However, the limitations discussed earlier still apply. It is probably prudent to reduce the oversampling rates in the areas with high concentrations of black or Hispanics below the levels indicated by census data as the optimum distribution of the sample Half of the gains shown in Chart 2 is probably the best that can be attained. However, even half the reductions in variance are still important gains in efficiency for all but very high values of k. When k is 40 or greater, it is probably preferable to use SRS rather than an oversampling strategy.

6. Prior Information on the Geographic Distribution of Poverty

Studies in the preceding two decades had also indicated that geography is not a very efficient device for oversampling poverty. Less than 40 percent of persons below the poverty level were reported in the March 1985 CPS as living in Census-defined poverty areas (see Table 4). The distributions varied greatly among the race/ethnic groups. Two-thirds of the blacks and 55 percent of Hispanics below poverty lived in poverty areas, but of the nonblack, nonHispanic poor (56 percent of all poor in 1985), only 22 percent were

in poverty areas. The 1986 distribution was not an isolated phenomenon. The 1976 distributions are not very different from 1985, although they indicate that there may be a trend for increased concentration in poverty areas of the black and Hispanic poor, with an opposite trend for the nonblacks and nonHispanics.

Table 4. Percent of persons below poverty who live in poverty areas[1], by race/ethnicity, 1985 and 1976

Race and poverty status	1985	1976
Total	39.7	42.2
Black	68.5	66.9
Hispanic	55.7	48.0
Nonblack, nonHispanic[2]	21.7	28.1

[1]Poverty areas are defined as census tracts in metropolitan areas and minor civil divisions in nonmetropolitan areas in which 20 percent or more of the population was below the poverty level in the previous census.

[2]Approximate data shown. Nonblack, nonHispanic calculated by subtracting black and Hispanic from total. Since there is a small overlap between black and Hispanic, this slightly understates the number of nonblack, nonHispanics.

Source: Census reports P-60, No. 115 and P-60, No. 154.

The periods chosen for the CPS analysis, 1985 and 1976, are approximately in the midpoints of the two decades. Since Census data are used for the delineation of poverty areas, one would expect that the effectiveness of stratification based on Census data would deteriorate with the passage of time. The midpoints of the decade are thus rough indicators of the average effectiveness over time during the course of each decade.

Part of the reason for the mediocre effectiveness of geographic stratification is that although the concentrations are moderately high in central cities, most of the poverty population lived outside these areas. As can be seen from Table 5, 57 percent of persons below poverty lived outside central cities in 1985. Almost half of the 57 percent lived in suburban areas, but of these, less than 20 percent were in poverty areas. The proportion in poverty areas outside MSA's was higher, 37 percent, but still not enough for oversampling through geographic stratification to be very effective.

The poverty areas in the CPS reports consist of complete census tracts or MCD's. Using smaller areas does provide somewhat better discrimination. It was not possible to obtain tabulations that are completely comparable to Tables 4 and 5 but are based on smaller areas (e.g., BG's and ED's instead of tracts) because the CPS public use tape does not contain such detailed geographic codes. However, in connection with another project, Westat prepared tabulations of the 1970

Table 5. Percent of persons below poverty who live in poverty areas[1], by metropolitan status, 1985

	Distribution by MSA status	Percent of poor in poverty areas
Total	100	39.7
In central cities of MSAs	42.8	55.3
In MSAs, not in central cities	27.5	19.0
Outside MSAs	29.6	36.5

[1]Poverty areas are defined as census tracts in metropolitan areas and minor civil divisions in nonmetropolitan areas in which 20 percent or more of the population was below the poverty level in the previous census.

Source: Census report P-60, No. 154.

census that show breakdowns for such small areas. A concise summary is shown in Table 6. Concentration in 1990 is greater than in 1970, but still not great enough for geographic oversampling to be helpful. Most of the persons that would be oversampled in designated block groups will turn out to be above poverty. Oversampling only on the basis of geography thus turned out to be an inefficient way of increasing the sample of poor people in 1970 and the 1990 Census analysis thus confirmed previous information on efficient sampling strategies.

Table 6. Percent distribution of persons below the poverty level by percent poverty in BG, 1970 and 1990

Percent of BG population that is below poverty level	1970	1990
Total	100.0	100.0
Under 10%	26.5	18.1
10-29.9%	45.1	44.6
30% or more	28.4	37.3

Source: Tabulations by Westat of 1970 Census 5th count and 1990 Census STF-3 file.

REFERENCES

Kalton, G. and Anderson, D.W., 1986, "Sampling Rare Populations," *Journal of the Royal Statistical Society*, Series A, Vol. 149, Pt. 1.

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

Massey, J., Judkins, D., and Waksberg. J., 1993, "Collecting Health Data on Minority Populations in a National Survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hansen, M., Hurwitz, W.N., and Madow, W.G., 1977, **Sample Survey Methods and Theory**, Vol. 1, Chapter 5, Section 12, Wiley & Sons, New York and London.

Chart 1. Ratio of variance of optimum sample to simple random sample at same cost

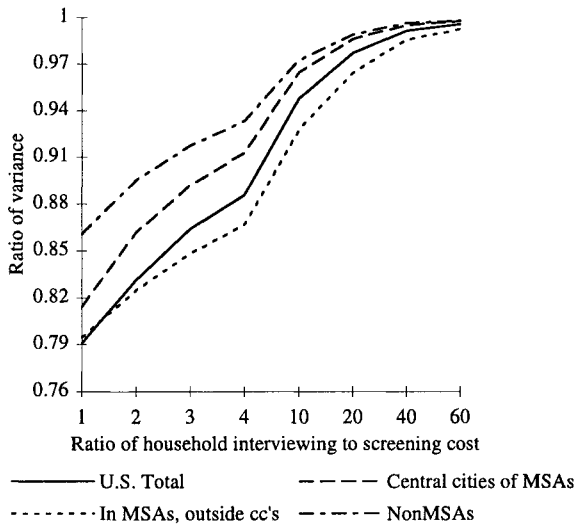

Chart 2. Ratio of variance of optimum sample to SRS at same cost

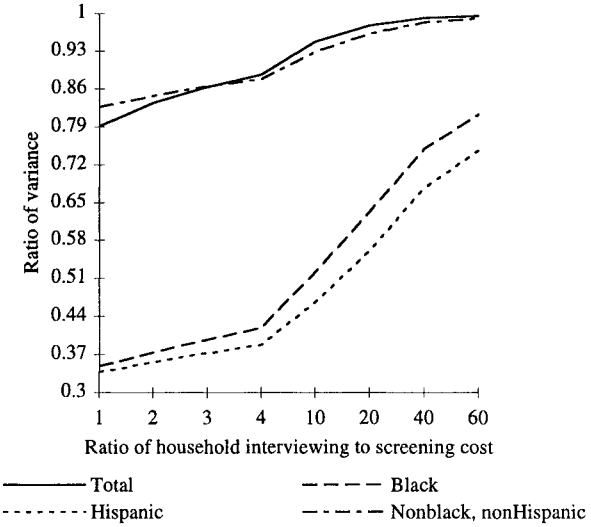

Chart 3. Comparison of ratios of variance when stratification is by percent blacks or Hispanics in poverty and by percent black or Hispanic in BG, 1990