# THE EFFECT OF STRATIFICATION WITH DIFFERENTIAL SAMPLING RATES ON ATTRIBUTES OF SUBSETS OF THE POPULATION

Joseph Waksberg, Westat, Inc.

This paper discusses the strategy to follow in stratification when one is interested in the attributes of subsets of the population, and the subsets cannot be isolated from the general population, in advance of the sampling. It pulls together the theory relating to this problem, provides data for several practical examples, and discusses some of the implications.

An example can best illustrate how the sampling issues arise. Suppose one is interested in attributes of a specific subgroup of the population, e.g., of negroes, low-income families, preschool age children, etc., but the only frames available for sampling comprise the total population and the subsets cannot be determined except as part of the interviewing procedure. A common strategy is to use geographic stratification, classifying such areas as tracts or Census EDs by the proportion of their populations in the specified subgroups. Census data may be used for stratification or more current local knowledge, if that is available.

More specifically, this paper explores the reduction in sampling variance that is possible when: (a) the population is divided into two strata in such a way that one stratum has a considerably higher proportion of the subset of interest than the other stratum, and (b) a higher sampling rate is used in the stratum with the greater concentration. Further, the paper is restricted to situations in which the following conditions apply:

(1) The stratum with the higher concentration contains less than half of the total population.

(2) Simple random sampling is used with a rate small enough so that the finite population correction factor is trivial.

(3) Most of the discussion relates to cases where the population variances in the subset are the same in both strata.

The first condition is fairly minor since, when it does not apply, only trivial gains in the variance are usually possible. In most cases, the second condition should also lead to only a minor loss of generality. The third condition is more troublesome. Situations exist in which the variances can be expected to be different. A re-examination of the major results would have to be made under such conditions, since it is difficult to state general principles when this occurs.

## I. Notation and Fundamental Relations

Assume the population is divided into two strata.

$N_1$ or $N_2$ = population of stratum 1 or stratum 2.

$N_2 = vN_1$, where $v \geq 1$.

$t_1$ or $t_2$ = proportion of stratum 1 or 2 in specified subgroup.

$t_1 = ut_2$, where $u \geq 1$.

$\sigma^2$ = population variance of a statistic within the subgroup, identical in the two strata.

$r_1$ or $r_2$ = sampling rate in stratum 1 or stratum 2.

$r_1 = kr_2$, where $k \geq 1$.

Compare two sampling plans:

A: Uniform sampling rate in the two strata, rate = r.

B: Use of $r_1$ in stratum 1 and $r_2$ in stratum 2, with

$$r(N_1 + N_2) = r_1N_1 + r_2N_2$$

so that the total sample sizes are identical in both plans.

Using the usual approximations to the variance, and assuming the finite correction factors are trivial, the variance of sample means can be expressed as:

$$\sigma^2 \text{ (plan A)} \equiv \sigma_A^2 \doteq \frac{\sigma^2}{r(t_1N_1 + t_2N_2)} = \frac{\sigma^2}{rt_2N_1(u + v)}$$

$$\sigma^2 \text{ (plan B)} \equiv \sigma_B^2 \doteq \frac{\sigma^2}{r_2(t_1N_1 + t_2N_2)^2}\left[\frac{t_1N_1}{k} + t_2N_2\right]$$

$$= \frac{\sigma^2(\frac{u}{k} + v)}{r_2t_2N_1(u + v)^2}$$

$$\sigma_B^2 / \sigma_A^2 \doteq \frac{k + v}{k(1 + v)} \cdot \frac{u + kv}{u + v}$$

## II. Condition for $\sigma_B^2 < \sigma_A^2$

$\sigma_B^2 / \sigma_B^2 < 1$ when $k < u$ -- that is, oversampling in stratum 1 will reduce the variance provided that the extent of oversampling is less than u, the ratio of the concentration of the subset in stratum 1 to stratum 2.

## III. Minimum Value of $\sigma_B^2/\sigma_A^2$ for a Given Set of Values of u and v

For a given set of values of u and v, the optimum value of $k = \sqrt{u}$.

For this value of k, $\sigma_B^2/\sigma_A^2$ is equal to

$$\frac{(\sqrt{u} + v)^2}{(1 + v)(u + v)} .$$

Table 2 shows the size of this ratio for selected values of u and v.

## IV. Minimum Value of $\sigma_B^2/\sigma_A^2$ for a Fixed Value of u

For a given value of u, the minimum value of $\sigma_B^2/\sigma_A^2$ occurs when $k = v = \sqrt{u}$.

When this occurs, the ratio $\sigma_B^2/\sigma_A^2$ is

$$\frac{4\sqrt{u}}{(1 + \sqrt{u})^2} .$$

Table 1 shows this minimum for a range of values of u.

In practical situations, it is not possible to manipulate the value of v. Once u is determined, this automatically fixes v. However, it is useful to be able to examine the minimum variance that can occur under the best possible situation.

## V. Value of $\sigma_B^2/\sigma_A^2$ When the Population Variances Are Not Identical in the Two Strata

If the variances in the two strata are not identical, let

$\sigma_1^2$ or $\sigma_2^2$ = population variance in stratum 1 or stratum 2.

$$\sigma_1^2 = w\sigma_2^2 .$$

In this case:

(1) $$\sigma_B^2/\sigma_A^2 = \frac{(k + v)(uw + kv)}{k(1 + v)(uw + v)} .$$

(2) $\sigma_B^2/\sigma_A^2$ will be less than 1 when $k < uw$.

(3) The minimum value of $\sigma_B^2/\sigma_A^2$ occurs when $k = \sqrt{uw}$ . When this occurs, the value of $\sigma_B^2/\sigma_A^2$ is

$$\frac{(\sqrt{uw} + v)^2}{(1 + v)(uw + v)} .$$

## VI. Discussion

1. The reductions in variance will be fairly small unless the concentration of the subset of the population in the stratum to be oversampled is considerably greater than in the rest of the universe. For example, if the concentration in one stratum is twice as great as the other and the variances are the same within the two strata, at best a four percent reduction in the variance can be attained. If the concentration is four times as great, the maximum reduction is 11 percent, and then only if the ratio of the populations in the two strata turns out to be exactly 2 to 1. More likely, the gains will be in the five to ten percent range. When the concentrations get to be of the order of 10 to 1, then sizable reductions occur.

2. On the basis of the preceding comments, it is possible to assess the value of geographic stratification for many types of statistics. For example, it is unlikely that oversampling for such populations as school age children, women of child bearing age, or older persons would have any important payoff. A cursory examination of tract statistics does not reveal any important differences in age distributions among tracts, except in a trivially few tracts. The best one might expect from a stratification of tracts is probably a factor of two or three in the concentrations. Census ED's would be somewhat better, but not strikingly so.

On the other hand, oversampling to produce Negro statistics could produce useful reductions in the variances, and the same is true of low income households although to a lesser extent. A two-way stratification of high and low Negro concentrations by the Bureau of the Census, using 1960 ED's as the units of stratification and 1960 data to classify the ED's shows that the maximum reductions in variance could be in the range 30-50 percent, depending on how current were the data used for stratification. For statistics on low-income households, poverty areas defined on the basis of 1960 data would have produced a 15 percent reduction in variance, about ten years later. Presumably, if smaller areas such as tracts or ED's had been used, the reduction would have been greater, possible of the order of 20-25 percent.

3. It is somewhat deceptive to use Census data some years after the Census and assume the same efficiency applies. For Negro statistics, for example, the

values of u typically dropped by about half between 1960 and 1967, for ED's classified on the basis of 1960 characteristics, resulting in only about two-thirds of the reduction in variance that might have been expected.

4. This deterioration over time in the effectiveness of stratification for many social and economic characteristics will frequently occur even when one uses what would appear to be better modes of stratification than geographic areas. For example, assume that statistics on low-income families is desired, and it is possible to stratify individual families on the basis of the previous year's income. CPS data on the proportion of families that changed their poverty status between 1964 and 1965, indicates that 31 percent of the 1964 poor were nonpoor in 1965, and eight percent of the nonpoor became poor. The values of u and v are about five and nine. Thus the reduction in variance that would occur with the optimum k is only 24 percent. This is not much better than would result from geographic stratification.

5. Classifying the population into two strata will for most cases provide most of the gains that stratification can produce. It would take a very unusual distribution of the population, for additional strata to reduce the variance much further. This can be seen most easily by starting with a two-way stratification and examining the effect of splitting each stratum further. It is clear from comments made earlier that important gains will occur only if there are sizable differences in the concentrations of the subsets in the two substrata formed from each of the original strata. If the original stratification was reasonably effective, it would be highly unusual for substratifications to produce additional differences in concentration of 5 or 10 to 1, these being differences that are required for important reductions in variance.

6. It should be noted that all of the discussion is related to attributes of subsets of the population. If one is interested in estimates of the sizes of the subsets, the same reductions do not apply. In fact, under some circumstances, the optimum sampling rates for the attributes will result in an increase in variance over a uniform sampling rate.

7. Tables 3 and 4 indicate the values of u and v that can be expected for kinds of items for which geographic stratification is most effective -- characteristics of the Negro and low-income population. Table 5 shows the deterioration over time in effectiveness when the population is stratified into poor and nonpoor families.

## APPENDIX

### I. Fundamental Relations

Since

$$N_2 = vN_1$$

$$t_1 = ut_2$$

$$r_1 = kr_2$$

$$\sigma_1^2 = w\sigma_2^2$$

and

$$\sigma_B^2 = \left(\frac{t_1N_1}{t_1N_1 + t_2N_2}\right)^2 \frac{\sigma_1^2}{r_1t_1N_1} + \left(\frac{t_2N_2}{t_1N_1 + t_2N_2}\right)^2 \frac{\sigma_2^2}{r_2t_2N_2},$$

replacing $N_2$, $t_1$, etc. by their values above

$$\sigma_B^2 = \frac{1}{(t_1N_1 + t_2N_2)^2}\left(\frac{\sigma_2^2t_2N_1}{r_2}\right)\left(\frac{wu}{k} + v\right).$$

For plan A, k = 1, and $r_2$ is replaced by r.

Since

$$r_1N_1 + r_2N_2 = rN$$

and

$$N_1 + N_2 = N\ ,$$

replacing $N_2$ by $vN_1$ and $r_1$ by $kr_2$, it follows that

$$r = r_2\,\frac{k + v}{1 + v}\ ,$$

which leads to

$$\sigma_A^2 = \frac{1}{(t_1N_1 + t_2N_2)^2}\,\frac{(\sigma_2^2t_2N_1)(1 + v)}{r_2(k + v)}\,(wu + v)$$

and

$$\sigma_B^2 \Big/ \sigma_A^2 = \frac{(wu + kv)(k + v)}{k(1 + v)(wu + v)}\ . \qquad (1)$$

When the variances in the two strata are equal w = 1, in this case

$$\sigma_B^2 \Big/ \sigma_A^2 = \frac{(u + kv)(k + v)}{k(u + v)(1 + v)}\ . \qquad (2)$$

## II. Optimum Value of k.

Differentiating equation (1) with respect to k and equating to zero, results in

$$k = \sqrt{uw} \ . \qquad (3)$$

With this value of k, equation (1) becomes

$$\text{Minimum } \sigma_B^2 / \sigma_A^2 \equiv \frac{(\sqrt{uw} + v)^2}{(1 + v)(uw + v)} \ . \qquad (4)$$

When the variances in the two strata are equal and w = 1, equations (3) and (4) become

$$k = \sqrt{u} \qquad (5)$$

$$\text{Minimum } \sigma_B^2 / \sigma_A^2 = \frac{(\sqrt{u} + v)^2}{(1 + v)(u + v)} \ . \qquad (6)$$

Table 1. Minimum Value of $\sigma_B^2 / \sigma_B^2$ for Specified Values of u

| u | Optimum value of k & v | Minimum of $\sigma_B^2 / \sigma_A^2$ |
|---|---|---|
| 1 | 1 | 1.00 |
| 2 | 1.4 | .97 |
| 4 | 2 | .89 |
| 9 | 3 | .75 |
| 16 | 4 | .64 |
| 25 | 5 | .55 |
| 49 | 7 | .44 |

Table 2. Minimum Value of $\sigma_B^2 / \sigma_A^2$ for Specified Values of u and v

| u | Optimum value of k | Minimum value of $\sigma_B^2 / \sigma_A^2$ when v = : 1 | 2 | 4 | 6 | 8 | 12 | 16 | 24 | 30 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2 | $\sqrt{2}$ | .96 | .96 | .97 | .98 | .98 | .99 | .99 | .99 | .99 | 1.00 |
| 4 | 2 | .90 | .89 | .90 | .91 | .93 | .94 | .95 | .97 | .97 | .98 |
| 9 | 3 | .80 | .76 | .75 | .77 | .80 | .82 | .85 | .88 | .90 | .93 |
| 16 | 4 | .74 | .67 | .64 | .65 | .67 | .70 | .74 | .78 | .81 | .87 |
| 25 | 5 | .69 | .60 | .56 | .56 | .57 | .60 | .63 | .69 | .72 | .79 |
| 49 | 7 | .64 | .53 | .46 | .44 | .44 | .46 | .48 | .53 | .56 | .64 |
| 100 | 10 | .60 | .47 | .37 | .35 | .34 | .34 | .34 | .37 | .40 | .47 |

Table 3. Use of 1960 Census Data for Stratification of E.D.'s for the Nonwhite Population; Effectiveness at Time of Census and Seven Years Later

| Geographic area | Percent nonwhite: In nonwhite stratum | Percent nonwhite: In residual stratum | Enrichment factor U | Percent of total pop.: In nonwhite stratum | Percent of total pop.: In residual stratum | Ratio of residual stratum pop. to nonwhites stratum pop. V | Percentage reduction in variance with optimum k |
|---|---|---|---|---|---|---|---|
| **Data for 1960** | | | | | | | |
| SMSA's 1,000,000+ | | | | | | | |
| N.E. | 69.7 | 2.6 | 27 | 12.1 | 87.9 | 7 | 45 |
| N.C. | 85.6 | 1.5 | 57 | 13.7 | 86.3 | 6 | 62 |
| S | 77.3 | 1.7 | 45 | 25.5 | 74.5 | 3 | 49 |
| W | 47.9 | 1.4 | 34 | 16.4 | 83.6 | 5 | 48 |
| SMSA's 250,000-1,000,000 | 85.3 | 3.0 | 28 | 9.9 | 90.1 | 9 | 44 |
| SMSA's <250,000 | 63.9 | 1.6 | 40 | 13.7 | 86.3 | 6 | 51 |
| Balance | 56.0 | 6.6 | 8 | 8.2 | 91.8 | 11 | 21 |
| **Data for March 1967** | | | | | | | |
| SMSA's 1,000,000+ | | | | | | | |
| N.E. | 81.8 | 5.2 | 16 | 8.7 | 91.3 | 10 | 31 |
| N.C. | 90.5 | 6.7 | 14 | 10.5 | 89.5 | 9 | 28 |
| S | 82.3 | 6.9 | 12 | 19.1 | 80.9 | 4 | 30 |
| W | 68.8 | 4.2 | 16 | 11.8 | 88.2 | 7 | 34 |
| SMSA's 250,000-1,000,000 | 92.6 | 6.4 | 14 | 7.0 | 93.0 | 13 | 25 |
| SMSA's <250,000 | 55.6 | 2.5 | 22 | 10.1 | 89.9 | 9 | 39 |
| Balance | 52.0 | 7.5 | 7 | 7.4 | 92.6 | 13 | 13 |

NOTE: Data based on stratification performed by the U.S. Bureau of the Census for a special survey performed for the O.E.O. Data for the top half of the table are from the 1960 Census; data for the lower half are from the special survey (SEO).

Table 4. Effectiveness of Using Poverty Areas as Strata for Families in Poverty /1

| | |
|---|---|
| Percent in Poverty | |
| In Poverty Areas | 14.8 percent |
| In Non-Poverty Areas | 2.6 percent |
| Enrichment Factor | u = 6 |
| Percent of Total Pop. | |
| In Poverty Areas | 14.4 percent |
| In Non-Poverty Areas | 85.6 percent |
| Ratio of Total Pop. in Non-Poverty to Total Pop. in Poverty Areas | v = 6 |
| Reduction in variance with optimum k | 15 percent |

1 Poverty areas are defined on basis of 1960 Census data, and restricted to SMSA's of over 250,000 population. The population distributions shown are for 1968.

Table 5. Poverty Status in 1964 and 1965 for Matched Families /1

| Classification in 1965 | Classification in 1964 | | |
|---|---|---|---|
| | Total | Poor | Nonpoor |
| Number of Cases (in 000) | | | |
| Total | 43,845[2] | 7,621 | 36,224 |
| Poor | 7,968 | 5,246 | 2,722 |
| Nonpoor | 35,877 | 2,375 | 33,502 |
| Percent Distribution | | | |
| Total | 100 | 100 | 100 |
| Poor | 18 | 69 | 8 |
| Nonpoor | 82 | 31 | 92 |

v = 4.8

u = 8.5

Maximum reduction in variance = 24 percent

1 Source: special tabulations of March 1964 and 1965 CPS records.

2 The number of matched families is less than the total number of families because of births, deaths, migration, and changes of family composition between 1964 and 1965.