BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/07

Report: 3

Census Adjustment Based on Stratification

by

Beverley Causey
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Kirk M. Wolter

Report completed: February 19, 1987

Report issued: February 19, 1987

Bevereley Causey
February 19, 1987

Report 3

Census Adjustment Based on Stratification

In Report 2 we have given breakeven variances and c.v.'s for different loss functions, corresponding to uncertainty in estimated true population total. There we considered the U.S. divided into 51 areas: D.C. and the 50 states. Here we likewise consider only the 6 New England states ("Division 1": Maine, NH, VT, MASS, CT, RI). For these states we do calculations for each of 3 artificial populations considered by Isaki, Diffendal and Schultz (IDS) in their studies of small-area estimation. For each of these 3 populations, as discussed in IDS's SRD Technical Report 87-02, "Report on Statistical Synthetic Estimation for Small Areas", there are artificial census and true population counts, based on the experienced census counts and substitutions.

Beyond merely considering the 6 states as wholes, we primarily want, in this report, to look at breakeven results based on the IDS "sampling strata" for these 6 states. Our investigation for sampling strata is of interest because the strata have been constructed so that, within sampling strata, undercount rates should tend to be relatively constant. In forming these sampling strata, IDS used the 6-state set of ED's, 14,199 in number. Each ED fell into one of the following 4 strata:

(1) Part of a central city, size 50,000 or more, with (for the ED) the combined proportion of black plus non-black Hispanic less than 6%. (2368 ED's; actual census count 1,929,094)

(2) Part of an SMSA but not of a central city as in (1), with the combined proportion of (1) less than 7%. (6030 ED's; actual census count 5,659,669)

(3) Other SMSA ED's with combined proportion at least 6% for (1) and 7% for (2). (2001 ED's; actual census count 1,747,143)

(4) ED's not belonging to an SMSA. (3800 ED's; actual census count 2,849,415).

We will give the following results from Report 2:

(a) C_2, the starting-point c.v.: $|Y/T-1|$, with Y = census and T = true total.

(b) B, the nonnegative term which appears in the equation $C_{3A} = C_2 \pm B$ as in (3A) of Report 2:

$$B = 1-1/(\sum p_i^2/r_i) \qquad (1)$$

with p_i the proportion of census total population that belongs to area i, and r_i the same proportion for true population. (Recall that for C_{3A} we subtract B for Y<T, as is typical, and add it for Y>T.) As discussed in Report 2, the value of B corresponds to differentials among area undercount (U/C) rates. Here, B is of particular interest for our sampling strata because, with these strata constructed so as to make U/C rates more uniform, we conjecture that we should obtain B relatively close to 0.

(c) The value of C_{3A} as given above.

(d) The value of C_3, obtained from (3) in report 2:

$$C_3 = [(Y-T)^2 - 2T(T-Y)B]^{1/2}/T. \qquad (2)$$

As in report 2 we would regard C_3 as based on the most suitable loss function, and thus as our official best breakeven c.v. Once again we show empirically that differences between C_3 and C_{3A} are very minor. Thus we may make use of C_{3A} and its simple decomposition into C_2, based on overall U/C rate, and B, based on differentials among area U/C rates.

Table 1 gives results for each sampling stratum, with ED's viewed as the areas of interest (cf. Report 1, where for the entire country we viewed D.C. and the 50 states as the areas of interest). Table 2 gives similar results, with state-pieces (within a stratum) viewed as the areas of interest. All states are represented in all 4 strata, except for VT in the black-Hispanic stratum 3. To provide a point of comparison for Table 2, we provide results for the 6 states as wholes (without regard to stratification or ED's) in Table 3. For each table there are 3 artificial populations as well as, for Tables 1-2, the 4 strata. In all cases, figures are expressed as a percent.

Remarks concerning these tables:

(1) All figures are rounded to the nearest decimal as shown.

(2) Especially for artificial populations AP2 and AP3, there are some minor inconsistencies between results based on summation over ED's and those based on state pieces, because of different computer roundoffs. We have planned the computations so that the essence of our results is not affected. One might observe, however, that C_2 is algebraically the same for tables 1 and 2 but that, for example, in stratum 3 for AP3 we have C_2 = 2.9129 in table 1 and = 2.9157 in table 2.

(3) Among the 3 artificial populations AP2 and AP3 have been considered the most relevant, with AP3 somewhat preferred.

Some conclusions are as follows:

(a) Differences among values for B may be interpreted as in Report 2. We start with C_2, which compares closeness of Y to T against (expected) closeness of estimated $\hat{T}$ to true T. Then, B is an adjustment which corresponds to heterogeneity, with respect to U/C rate, within strata for Tables 1 and 2 (and within all of New England for Table 3). The greater the heterogeneity, in

general, the larger the value of B will be. In Tables 1 and 2 we might regard departure of B from 0 as a measure of the failure of stratification to achieve complete homogeneity of ED's and state pieces within strata.

(b) Much as in report 2 the differences between C_{3A} and C_3 seem relatively minor, in fact, often negligible. Thus C_{3A} and its interpretation seem useful in practice.

(c) In tables 1-2 all magnitudes are especially large for stratum 3, simply because that stratum naturally represents the black-Hispanic neighborhoods with high U/C rates.

(d) Initially it was thought, as above, that B would be lower for table 2, for states with stratification, than for table 3, for states without stratification, where U/C rates are less uniform. From strata 1, 2 and 4 there is some evidence of such a pattern, although the behavior of stratum 3 prevents a completely uniform result. Another consideration, possibly, is that B is made larger because of the breaking of states into 6 pieces through stratification as for Table 2. On the other hand, B is much larger in Table 1 than in Table 2. Perhaps this pattern might be anticipated, because the ED's of Table 1 are smaller and more heterogeneous than the state pieces of Table 2.

For all 3 tables note that C_2, besides being a breakeven c.v., is the U/C rate (in percent) for the cohort under consideration in each line. As illustration, for stratum 3 the true overall U/C rate -- equivalently, the value of C_2 -- is 2.9%. Using the criterion C_{3A}, a ratio adjustment of ED's will be closer to the truth than the original census counts if our estimator, $\hat{T}$, of T has a c.v. of less that 2.60%.

Table 1
ED's Within Sampling Strata

Stratum	Art. Pop.	C_2	B	C_{3A}	C_3
1	1	.8745	.0489	.8255	.8241
1	2	.6568	.0228	.6340	.6336
1	3	.4346	.0104	.4242	.4241
2	1	.6587	.0144	.6442	.6441
2	2	.5087	.0067	.5020	.5020
2	3	.3294	.0020	.3274	.3274
3	1	2.0167	.1980	1.8186	1.8078
3	2	2.6782	.2738	2.4044	2.3888
3	3	2.9129	.3119	2.6010	2.5822
4	1	.7756	.0187	.7569	.7567
4	2	.6001	.0102	.5899	.5898
4	3	.3927	.0046	.3882	.3882

Table 2
State Pieces Within Sampling Strata

Stratum	Art. Pop.	C_2	B	C_{3A}	C_3
1	1	.8745	.0003	.8742	.8742
1	2	.6568	.0002	.6566	.6566
1	3	.4346	.0001	.4346	.4346
2	1	.6588	.0001	.6587	.6587
2	2	.5088	.00003	.5088	.5088
2	3	.3295	.00001	.3295	.3295
3	1	2.0187	.0116	2.0071	2.0071
3	2	2.6808	.0114	2.6694	2.6694
3	3	2.9157	.0110	2.9047	2.9046
4	1	.7757	.0002	.7755	.7755
4	2	.6002	.0001	.6001	.6001
4	3	.3928	.00003	.3928	.3928

Table 3
State Without Regard to Strata

Art. Pop.	C_2	B	C_{3A}	C_3
1	.9159	.0003	.9149	.9149
2	.8713	.0004	.8710	.8710
3	.7406	.0004	.7402	.7402