BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/05

# Report 2: Census Adjustment Based on an Uncertain Population Total

by

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



Recommended by: Kirk M. Wolter

Report Completed: November 12, 1986

Report Issued: January 29, 1987

## Report 2: Census Adjustment Based on an Uncertain Population Total

This report extends our Report 1: A Study of Whether Census Adjustment is Worthwhile. In that report we considered an across-the-board ratio adjustment using a known population total, T. Here we consider the same adjustment except that we do not know T perfectly. We only have an estimator of it, which we call $\hat{T}$.

We retain most of the notation in Report 1: $t_i$ is true population for area i, $y_i$ is unadjusted census count, Y is total unadjusted census count. Whereas in Report 1 we considered the adjustment $a_i = p_i T$ we now consider

$$\hat{a}_i = p_i \hat{T} \tag{1}$$

with $p_i = y_i/Y$. We want to decide whether $\hat{a}_i$ is in general closer to $t_i$ than is $y_i$. As criteria for this decision we use the four loss functions $f_k(\underline{x})$, k=1,2,3,4, in (2) of Report 1. There we compared $f_k(\underline{a})$ and $f_k(\underline{y})$. Here the vector $\hat{\underline{a}}$ depends on $\hat{T}$, which we view as a random variable. Thus we now compare $E(f_k(\hat{\underline{a}}))$, which we call $f_k^*$, against $f_k(\underline{y})$. We will be led, at the end of the report, to recommend a fairly simple and easily interpreted "criterion 3A," in (3A). But for now we want to look at all 4 criteria.

Suppose that $\hat{T}$ is an unbiased estimator of T, with known variance W. We then have

$$f_1^* = f_1(\underline{a}) + W(\sum p_i^2) \tag{2a}$$

$$f_2^* = f_2(\underline{a}) + W \sum p_i^2/y_i \tag{2b}$$

$$f_3^* = f_3(\underline{a}) + W \sum p_i^2/t_i. \tag{2c}$$

For $f_4^*$ we presume, additionally, that $\hat{T}$ is normally distributed. This presumption makes sense if $\hat{T}$ is based on a large sample. For

$W>0$ we are able to show that

$$f_4^* = \sum c_i[d_i(2\phi(d_i)-1) + 2g(d_i)] \qquad (2d)$$

with $c_i = p_i W^{1/2}$, $d_i = |p_i T - t_i|/c_i$, $\phi$ the c.d.f. for $N(0,1)$, and $g$ the density for $N(0,1)$.

The larger $W$ is, the larger each $f_k^*$ becomes. We will compute the value of $W$($>0$, typically) for which $f_k^*$ and $f_k(\underline{y})$ are equal. If this breakeven $W$ is distinctly larger than the anticipated value of $W$, then according to criterion $k$ we do better to use $\hat{a}_i$ in preference to $y_i$. As in Report 1, we will use the 1980 Post Enumeration Project (PEP) in our investigation. Before going to this investigation, however, we consider bias in $\hat{T}$.

Above, we presumed $E(\hat{T}) = T$. A more complete model is $E(\hat{T}) = T+B$, with $B$ possibly nonzero; thus each $f_k^*$ depends on $W$ as well as $B$. We no longer can talk about a breakeven value for $W$, except with reference to a particular value of $B$. Results thus become hard to interpret. However, I think it is best to view $\hat{T}$ as unbiased. If we sense that $\hat{T}$ might be biased, we can use a bias correction, as we think appropriate. Then, $W$ can be viewed as the sum total of sampling error, uncertainty in making the bias correction, etc.

We view $W$ in this manner, with $\hat{T}$ unbiased, in the rest of this report; we are now ready to discuss our investigation. As values for $t_i$ and $T$ we use PEP estimates as we did in Report 1; for each of 12 PEP sets we compute a breakeven value of $W$ for each of our 4 loss functions. For $k=1,2,3$ the form of (2a-c) permits easy computation; for $k=4$ we use a binary search.

We give results in terms of the coefficient of variation (c.v.) $C = W^{1/2}/T$, expressed as a percent. Let $C_k$ be the breakeven c.v. corresponding to criterion $k$. For the 12 PEP sets we have values of $C_k$ as follows:

| PEP Set | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 |
|---|---|---|---|---|
| 2-8 | 1.592 | 1.155 | 1.134 | 1.229 |
| 2-9 | 2.116 | 1.577 | 1.552 | 1.662 |
| 2-20 | 2.438 | 1.896 | 1.869 | 2.031 |
| 3-8 | 1.403 | 1.003 | 0.982 | 1.033 |
| 3-9 | 1.927 | 1.426 | 1.401 | 1.442 |
| 3-20 | 2.251 | 1.745 | 1.718 | 1.810 |
| 5-8 | 1.950 | 1.738 | 1.717 | 1.902 |
| 5-9 | 2.467 | 2.156 | 2.131 | 2.336 |
| 10-8 | 0.430 | 0.309 | 0.296 | 0.291 |
| 14-8 | 0.788 | 0.916 | 0.931 | 0.896 |
| 14-9 | 0.145 | 0.495 | 0.511 | 0.494 |
| 14-20 | - | 0.173 | 0.189 | 0.131 |

(For $C_1$ and 14-20 the breakeven W is negative, corresponding to the fact $f_1(\underline{y})<f_1(\underline{a})$. That is, according to criterion 1 and 14-20 we do better not to adjust even if we know T exactly.) Here our areas, for which census counts are to be adjusted, are the 50 states plus DC.

Thus, as an example, set 3-8 and criterion 2 give us a breakeven c.v. of 1.003, or about 1%. That is, we estimate that if $\hat{T}$ has a relative standard error of 1% as an estimator of T, we are indifferent as to whether to use adjusted $\hat{a}_i$ in preference to unadjusted $y_i$. If the relative error is less than 1%, we would use $\hat{a}_i$. If it is greater, we would use $y_i$. For set 3-8 and criterion 3 we have, at 0.982, a breakeven c.v. barely under 1%.

We now look closely at the formulas for the breakeven variance, $W_k$, corresponding to which we have presented $C_k$ above.

The breakeven $W_2$ is just $(Y-T)^2$. Thus according to criterion 2 we simply compare the two squared errors $E((\hat{T}-T)^2)$ and $(Y-T)^2$. That is, if the error (i.e., variance) in $\hat{T}$ is smaller than the error in Y, then the adjusted $\hat{a}_i$ is preferred to the unadjusted $y_i$.

With $p_i = y_i/Y$ we likewise set $r_i = t_i/T$. The breakeven $W_3$ is

$$W_2 + 2T(T-Y)[1/(\sum p_i^2/r_i)-1]. \quad (3)$$

We have $W_3 = W_2$ if we have either: (1) T=Y, or (2) $r_i = p_i$ for all i (that is, the ratio $y_i/t_i$ is constant). Otherwise, use of a Lagrange multiplier shows that the bracketed term in (3) is negative, and we have $W_3 < W_2$ if T>Y (i.e., if Y is an undercount of the total population). For our first 9 PEP sets, above, the estimated T exceeds Y; accordingly, we have $W_3 < W_2$. Thus as in the above discussed example, for PEP set 3-8 the breakeven c.v. falls from $C_2 = 1.003$ to $C_3 = 0.982$: not a major difference. For T<Y, as for the last 3 PEP sets, we have $W_3>W_2$; but T>Y seems more realistic for areas which are hard to enumerate.

The breakeven $W_1$ is

$$W_2 + 2T(T-Y)[(\sum p_i r_i)/(\sum p_i^2)-1]. \qquad (4)$$

As for $W_3$ we have $W_1 = W_2$ for either T=Y or $r_i = p_i$. Otherwise, our empirical results indicate that for the 50 states plus DC the bracketed term in (4) appears to be, in practice, positive. We have $r_i>p_i$ typically, when $p_i$ is largest and $r_i<p_i$, typically, when $p_i$ is smallest (remember that $\sum p_i = \sum r_i = 1$). That is, the undercount rate is generally higher for the larger states, and as a result, for T>Y, the breakeven W is forced upward. Difference between $W_1$ and $W_2$ appear to exceed those between $W_2$ and $W_3$: e.g., for PEP set 3-8 we have $C_1 = 1.403$ and $C_2 = 1.003$. For groups of areas other than the 50 states and DC we may, of course, have a negative bracketed term in (4), with $W_1 < W_2$ for T>Y.

The breakeven $W_4$ is $W_2\pi/2$ (i.e., $C_4 = C_2(\pi/2)^{1/2}$) for $r_i = p_i$ as opposed to $W_1 = W_3 = W_2$ for $r_i = p_i$. Convex-programming and calculus manipulations show that for $r_i \neq p_i$ we have $W_4<W_2\,\pi/2$. For example, for PEP set 3-8 we have $C_4 = 1.033$ - whereas the value of $C_2(\pi/2)^{1/2}$ is 1.003 x 1.253 = 1.256.

Of the 4 criteria we prefer 4, because it works with absolute values, and 3, because it divides squared differences by the <u>true</u> $t_i$. For both of these, in practice, differential rates of undercount lead to a reduction in breakeven c.v. from what it would be if we had $p_i = r_i$ for all i--equivalently, if we had $y_i/t_i$ constant. Thus we might first consider, based on $y_i/t_i$ constant, the breakeven c.v.'s $|Y/T-1|$ for criterion 3, and $(\pi/2)^{1/2}|Y/T-1|$ for criterion 4. These provide useful starting points in deciding whether or not to adjust. That is, we can compare the c.v. of $\hat{T}$ against these breakeven values in making this decision. But we must make some modification to reflect the fact that $y_i/t_i$ is not constant.

Henceforth we restrict our discussion to criterion 3, largely because computation for criterion 4 has required the additional assumption, not yet fully justified, that $\hat{T}$ has a normal distribution. Thus as a breakeven c.v. our starting point is $|Y/T-1|$, which is $C_2$. As we have seen, departure of the actual $C_3$ from $C_2$ is a consequence of $y_i/t_i$ not being constant. Our table, above, indicated that the departure is small. Expressed as a percent, $|C_2 - C_3|$ never exceeds .027: barely 1/40 of 1%. Thus one might be able to regard departure of $C_3$ from $C_2$ as a secondary matter; but here we regard it as a primary matter. We develop a simple approximate representation for this departure as follows. Consider $W_3$ in (3). Take the square root of it, and consider the 1st-order Taylor expansion for this square root about the point $W_2$. Dividing by T, we have that $C_3$ is equal to approximately

$$C_{3A} = C_2 \pm [1 - 1/(\Sigma\ p_i^2/r_i)]. \qquad (3A)$$

(Relative accuracy of the approximation is greatest when departure of $C_3$ from $C_2$ is smallest.) In (3A) the bracketed term, which we call B, is positive. In regard to the $\pm$ sign we subtract B if $Y<T$: that is, if there is overall undercount as seems typical. We add B if $Y>T$: that is, if there is overcount. Thus we have developed our criterion 3A, against which we compare the c.v. of $\hat{T}$, in deciding whether to make adjustment for a set of areas. It has two components, one ($C_2$) based on the relative difference between Y and T and one (B) based on differentials in undercount rates.

Note what happens when Y is close to T. The value of $C_2$ becomes essentially 0, thus $C_{3A}$ becomes -B for Y<T and +B for Y>T. Thus there is a discontinuity in the value of $C_{3A}$ and an internal inconsistency in our decision rule. However, for Y very close to T the adjustment (i.e., difference between $y_i$ and $\hat{a}_i$) is so small that it does not matter whether we make it or not. Hence we are not disturbed by the discontinuity. If one is disturbed by it, one can just use $C_3$, which is the official exact breakeven c.v. We have introduced $C_{3A}$ only because it is so easy to interpret. Empirical results, as below, show that $C_3$ and $C_{3A}$ are almost the same.

For our 12 PEP sets the departures $C_{3A}$-$C_3$, expressed in percent, always positive, seem inconsequential:

| PEP Set | $C_3$ | $C_{3A}$ | $C_{3A} - C_3$ |
|---|---|---|---|
| 2-8 | 1.134054 | 1.134247 | .000193 |
| 2-9 | 1.552395 | 1.552591 | .000196 |
| 2-20 | 1.869063 | 1.869255 | .000193 |
| 3-8 | 0.981651 | 0.981876 | .000225 |
| 3-9 | 1.400801 | 1.401016 | .000215 |
| 3-20 | 1.717828 | 1.718035 | .000208 |
| 5-8 | 1.716989 | 1.717113 | .000124 |
| 5-9 | 2.131294 | 2.131434 | .000140 |
| 10-8 | 0.296249 | 0.296505 | .000256 |
| 14-8 | 0.931456 | 0.931582 | .000126 |
| 14-9 | 0.510619 | 0.510859 | .000240 |
| 14-20 | 0.189079 | 0.189815 | .000737 |

On this basis we would prefer the easily interpreted $C_{3A}$. For PEP set 3-8, as an example, the difference in breakeven c.v. is only .000225 of 1%, or .00000225.

Using $C_{3A}$, we might look more closely at the bracketed term, B, in (3A). Perhaps some insights can be gotten from special cases. Suppose we have just 2 areas with $r_1$=c, $r_2$=1-c (two population proportions) and $p_1$= c+δ, $p_2$= 1-c-δ (census proportions). Then we have, for δ>0,

$$B=1/[1+c(1-c)/\delta^2], \text{ or } \delta^2/[\delta^2+ c(1-c)]$$

Suppose we have 3 areas with $r_1 = r_2 = r_3 = 1/3$, $p_1 = 1/3 + \delta$, $p_2 = 1/3$, and $p_3 = 1/3 - \delta$. Then we have

$$B=1/[1 + 1/6\delta^2], \text{ or } \delta^2/[\delta^2 + 1/6].$$

(Here, a function of general form $f(\delta)=1/[1+\alpha/\delta^2]$ for constant $\alpha$ has $f(0)=0$, $f(\infty)=1$, $f^1(0)=0$, $f^1(\infty)=0$, $f^1(\delta)>0$, for $\delta>0$, and point of inflection $\delta=\alpha^{1/2}$. If $\delta$ is small, however, f behaves pretty much like the sample quadratic $\delta^2/\alpha$ .)