UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 40

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: William R. Bell *(signed 12/31/02)* WRB
Senior Mathematical Statistician for Small Area Estimation

Subject: A.C.E. Revision II: Calculating aggregate data-defined, correct enumeration, and census inclusion rates (for groups that involve aggregation across post-strata)

The attached document discusses how to make aggregate tabulations of A.C.E. Revision II results for the purpose of calculating certain rates related to census coverage for aggregate population groups. While these aggregate rates are not actually used in constructing the A.C.E. Revision II dual system estimates (DSEs), they are useful descriptive statistics for examining approximately how the DSE components differ across aggregate groups.

Calculating aggregate data-defined, correct enumeration, and census inclusion rates (for groups that involve aggregation across post-strata)

William R. Bell

December 31, 2002

This note discusses how to make aggregate tabulations of A.C.E. Revision II results for the purpose of calculating certain rates related to census coverage for aggregate population groups. The purpose of the note is to address questions that were raised regarding how such tabulations should be done given the different post-strata used for the E- and P-samples. The rates considered here are data-defined rates (proportions of census enumerations that are data-defined persons), correct enumeration (CE) rates, and census inclusion rates. Two types of CE rates are considered, one reflecting the proportion of data-defined persons that are correct enumerations, and the other reflecting the proportion of all census enumerations that are correct enumerations. The first is analogous to the usual E-sample CE-rate, the latter we shall call the "census correct enumeration rate."

Data Defined Rates

Let a denote a geographic area or population subgroup of interest. I can think of two versions of a DD-rate for a that could be of interest. One uses the DDs for a computed from direct tabulations, the other uses the synthetically estimated DDs for a. The rate computed with the former I'll call the "actual DD-rate." It makes sense as a descriptive statistic since we know whether all census records are data defined or not. The rate computed with the latter DD tabulation I'll call the "estimated DD-rate." It makes sense since it relates to how we compute DDs in computing the synthetic estimates from the DSEs. Below I give formulas for computing both these rates.

Let t index all person records in any given detailed $h\ell$ post-stratum, including both data-defined persons and non-data-defined persons. To start we need the census tabulations for the intersection of a with the detailed $h\ell$ post-strata (note that $\sum_{t\in(h\ell)\cap a}$ indicates summation over all person records t within a that are in the detailed $h\ell$ post-stratum). We also need the full census tabulation for a.

- $\text{Cen}_{h\ell,a} = \sum_{t\in(h\ell)\cap a} 1$
- $\text{Cen}_{\bullet a} = \sum_{h\ell} \text{Cen}_{h\ell,a} = \sum_{t\in a} 1.$

We compute the actual DDs in the analogous way:

- $\text{DD}^{\text{act}}_{h\ell,a} = \sum_{t\in(h\ell)\cap a} I(t \text{ is data defined})$
- $\text{DD}^{\text{act}}_{\bullet a} = \sum_{h\ell} \text{DD}^{\text{act}}_{h\ell,a} = \sum_{t\in a} I(t \text{ is data defined}).$

We compute estimated DDs for a synthetically:

- $\mathrm{DD}^{\mathrm{est}}_{h\ell,a} = r_{DD,h\ell} \times \mathrm{Cen}_{h\ell,a}$
- $\mathrm{DD}^{\mathrm{est}}_{\bullet a} = \sum_{h\ell} \mathrm{DD}^{\mathrm{est}}_{h\ell,a} = \sum_{h\ell} (r_{DD,h\ell} \times \mathrm{Cen}_{h\ell,a})$

In the above two expressions $r_{DD,h\ell} = \mathrm{DD}^{\mathrm{act}}_{h\ell}/\mathrm{Cen}_{h\ell}$ is the estimated DD-rate for the $h\ell$ post-stratum that is used in the DSE. Notice that if a were simply an $h\ell$ post-stratum then the equation for $\mathrm{DD}^{\mathrm{est}}_{h\ell,a}$ would become $\mathrm{DD}^{\mathrm{est}}_{h\ell} = \mathrm{DD}^{\mathrm{act}}_{h\ell}$. This equivalence of the estimated and actual DD tabulations occurs only for the $h\ell$ post-strata and for their direct aggregates (not for groups that include parts of $h\ell$ post-strata).

The two DD-rates we can compute are as follows:

- Actual DD-rate$_a$ $= \mathrm{DD}^{\mathrm{act}}_{\bullet a}/\mathrm{Cen}_{\bullet a}$
- Estimated DD-rate$_a$ $= \mathrm{DD}^{\mathrm{est}}_{\bullet a}/\mathrm{Cen}_{\bullet a}$.

One final point to note is that for the synthetic calculations that follow we will need the $\mathrm{Cen}_{h\ell,a}$, but we do not really need to retain the detailed $h\ell$ post-stratum tabulations of $\mathrm{DD}^{\mathrm{act}}_{h\ell,a}$ or $\mathrm{DD}^{\mathrm{est}}_{h\ell,a}$.

Correct Enumeration Rates

Since CEs are identified only for the E-sample cases, and not the whole census, we can only estimate aggregate CE rates synthetically. The synthetic estimate of CEs for area or subgroup a is calculated as follows:

- $\mathrm{CE}_{h\ell,a} = r_{ce,i} \times \mathrm{DD}^{\mathrm{est}}_{h\ell,a} = r_{ce,i} \times r_{DD,h\ell} \times \mathrm{Cen}_{h\ell,a}$
- $\mathrm{CE}_{\bullet a} = \sum_{h\ell} \mathrm{CE}_{h\ell,a} = \sum_{h\ell} (r_{ce,i} \times r_{DD,h\ell} \times \mathrm{Cen}_{h\ell,a})$

where i is the E-sample post-stratum determined by the more detailed post-stratum $h\ell$. The corresponding CE-rate is then

- CE-rate$_a$ $= \mathrm{CE}_{\bullet a}/\mathrm{DD}^{\mathrm{est}}_{\bullet a}$.

This rate reflects the estimated proportion of data-defined persons in area or subgroup a that were correct enumerations. It is analogous to the estimated CE-rates ($r_{ce,i}$) for the E-sample post-strata. In fact, if a is an E-sample post-stratum i, the above rate reduces to $r_{ce,i} \times \mathrm{DD}^{\mathrm{est}}_{\bullet i}/\mathrm{DD}^{\mathrm{est}}_{\bullet i} = r_{ce,i}$.

Note that the denominator of CE-rate$_a$ is the estimate of data defined persons, not the census count. This is because in the DSE the E-sample CE-rates $r_{ce,i}$ are applied to the estimates of data defined persons. We can also calculate the following ratio:

- census CE-rate$_a$ $= \mathrm{CE}_{\bullet a}/\mathrm{Cen}_{\bullet a}$

which can be called the "census correct enumeration rate" (for area or subgroup a). It reflects the estimated proportion of all census enumerations for area or subgroup a that were correct. Since the census count is the sum of data-defined and non-data-defined persons, $\text{Cen}_{\bullet a} \geq \text{DD}^{\text{est}}_{\bullet a}$ and census CE-rate$_a \leq$ CE-rate$_a$. Note that census CE-rate$_a$ reflects for the area or subgroup a the combined effects of the CE- and DD-rates ($r_{ce,i} \times r_{DD,h\ell}$).

Census Inclusion Rates and Undercount Rates

Census inclusion rates estimate the proportion of the true population actually included in the census. The true population for a is estimated by the synthetic estimate for a from the DSEs including any adjustment for correlation bias. Let $\hat{N}$ denote such synthetic estimates (reserving "DSE" for conventional DSEs computed without correlation bias adjustments). Under the "two-group" model for adjusting for correlation bias the $\hat{N}$ are given by

- $\hat{N}_{h\ell,a} = c_k \dfrac{\text{CE}_{h\ell,a}}{r_{m,j}} = \dfrac{c_k}{r_{m,j}} \times r_{ce,i} \times r_{DD,h\ell} \times \text{Cen}_{h\ell,a} = \text{CCF}_{h\ell} \times \text{Cen}_{h\ell,a}$

- $\hat{N}_{\bullet a} = \sum\limits_{h\ell} \hat{N}_{h\ell,a} = \sum\limits_{h\ell} \left(\dfrac{c_k}{r_{m,j}} \times r_{ce,i} \times r_{DD,h\ell} \times \text{Cen}_{h\ell,a} \right) = \sum\limits_{h\ell} (\text{CCF}_{h\ell} \times \text{Cen}_{h\ell,a})$

where j is the P-sample post-stratum determined by $h\ell$ and c_k is the correlation bias adjustment factor for age-race group k (with k also determined by $h\ell$). Note that the c_k are 1 for children and adult females. The coverage correction factors for the $h\ell$ post-strata are:

$$\text{CCF}_{h\ell} = \frac{c_k}{r_{m,j}} \times r_{ce,i} \times r_{DD,h\ell}.$$

The census inclusion rate for a is

- Census inclusion rate$_a$ = $\text{CE}_{\bullet a} / \hat{N}_{\bullet a}$.

This rate uses as its numerator $\text{CE}_{\bullet a}$, because this represents our estimate of the number of correct census inclusions from the true population. Since $c_k \geq 1$ and $r_{m,j} \leq 1$, $\hat{N}_{\bullet a} \geq \text{CE}_{\bullet a}$, and this inclusion rate is ≤ 1. This rate reflects the effects of dividing in the DSEs by the P-sample match rates and then multiplying the DSEs by the correlation bias factors c_k from the two-group model. If a were a P-sample post-stratum that included the full age-sex detail then the census inclusion rate would be simply $r_{m,j}/c_k$. If we were not adjusting for correlation bias in the DSEs (as is the case for children and adult females) then the census inclusion rate for a P-sample post-stratum is simply the match rate for that post-stratum.

If we replace $\text{CE}_{\bullet a}$ in the numerator of the above rate by the census count we get a net census coverage "rate," which would exceed 1 for net overcounts. More familiar is the related census undercount rate for a:

- Census undercount rate$_a$ = $1 - (\text{Cen}_{\bullet a} / \hat{N}_{\bullet a})$

which is negative for overcounts.

Note: Multiplying any of the above rates by 100 expresses them as percentages.