BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: Census /SRD/RR/88/15

REPORT 5
EVALUATION OF CENSUS RATIO ESTIMATION AND SYNTHETIC ESTIMATION

by

Beverley Causey
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Draft 4/19/88

B. Causey:PE:bev5

Report 5

Evaluation of Census Ratio Estimation and Synthetic Estimation

Abstract

This report supplements 4 earlier reports which evaluated methods for estimating area counts using the original counts from the decennial census. Enhancements to those reports are: (A) synthetic estimation based on only 2 cells (poststrata), leading to a single "breakeven" error variance in estimating a nationwide proportion; (B) breakeven variances based on medians along with those of earlier reports based on expected values; (C) weighted combinations of the two separate artificial sets of population counts considered earlier.

Key words: synthetic estimation, breakeven error variance.

This report enhances 4 earlier reports which evaluated methods for estimating area counts, using the original counts from the decennial census. Our 3 sections are: (1) the setting for this report and the earlier reports; (2) the enhancements of this report; (3) the empirical findings of this report, and their implications as to whether estimated counts are more accurate than the original.

1. THE SETTING

As in earlier reports we have: (A) area census and true counts; (B) across-the-board (ratio) estimation; (C) synthetic estimation; (D) "artificial populations" as true counts; (E) "measures of improvement"; (F) uncertainty in estimation.

1A. Counts. Here we consider a division of the U.S.A. into 50 states plus D.C., and then into 3137 counties or equivalents, as areas of interest.

For area i let y_i be census count and t_i be true count.

1B. Across the Board. Let Y and T be nationwide census and true counts. Our across-the-board ratio estimator for area i is $a_i = y_i T/Y$, based on our external knowledge of T.

1C. Synthetic Estimation. Besides our areas, consider a division of the U.S.A. into cells such as age-race-sex cross-classification as we have used. Suppose that for cell j we know T_j, the true population total. Letting y_{ij} and Y_j be the census counts for area i, cell j and for cell j in entirety, our synthetic estimator for area i is $b_i = \sum_j y_{ij} T_j / Y_j$.

1D. Artificial Populations. The true counts t_i, T and T_j come from an "artificial population" (AP) as constructed by Isaki, Diffendal and Schultz, SRD Technical Report 87-02, "Report on Statistical Synthetic Estimation for Small Areas." Because undercount rate information was not readily available for Hispanics, two AP's were formed: "AP1" is based in essence on a presumption that undercount rate patterns for non-black Hispanics resemble those for non-black non-Hispanics, and AP2 on a presumption that patterns for non-black Hispanics resemble those for blacks. In this report, unlike earlier reports, we consider all 50 states plus D.C.

1E. Measures of Improvement. Let x_i be estimated area count, whether original census itself (y_i) or otherwise. Let $\underline{y}$ be the vector $(y_i, \ldots, y_N)$ for N areas, and let $\underline{x}, \underline{t}$, etc. be vectors likewise. As a measure of closeness of $\underline{x}$ to the true $\underline{t}$ we first consider

$$f_1(\underline{x}) = \sum_{i=1}^{N} |x_i - t_i|. \qquad (1)$$

We may express $f_1(\underline{x})$ as $\sum t_i |x_i/t_i - 1|$: the sum of absolute relative errors weighted by true counts t_i. Likewise we consider

$$f_2(\underline{x}) = \sum (x_i - t_i)^2 / t_i, \qquad (2)$$

the sum of squared relative errors weighted by true counts. Thus we have "MOI 1" and "MOI 2." A third MOI, #3, replaces true t_i by census y_i in the denominator of f_2. We see as a drawback to this f_3 the possible distortion that arises when the denominator y_i is erroneously very close to 0.

We also consider MOI's based on proportions rather than counts. Let g_i be t_i/T, the true proportion that area i represents out of the whole. Likewise, with $X=\sum x_i$, let $w_i = x_i/X$, the estimated proportion that area i represents; and let $p_i=y_i/Y$, the original census proportion. Substituting g_i, w_i and p_i for t_i, x_i and y_i, we replace $f_k(\underline{x})$ by $f'_k(\underline{w})$ for k=1,2,3.

1F. Uncertainty in Estimation. For across-the-board (ABD) estimation we empirically get $f_k(\underline{a})<f_k(\underline{y})$, and ascertain that for area counts ABD can provide improvement over original census enumeration (OE). In reality, however, we do not know exactly the true population total T. Suppose that we have an unbiased estimator, $\hat{T}$. Let $\hat{a}_i$ be the ABD area-i estimator, a_i, with $\hat{T}$ replacing T. We may compute V_k such that for $Var(\hat{T})=V_k$, $E(f_k(\hat{\underline{a}}))$ is $f_k(\underline{y})$. (Note that ABD does not change area proportions.) For V_1 we assume a normal distribution for $\hat{T}$.

For synthetic (SYN) estimation we get $f_k(\underline{b})<f_k(\underline{y})$. Likewise for area proportions, with $B=\sum b_i$ and $r_i=b_i/B$, we get $f'_k(\underline{r})<f'_k(\underline{p})$. Thus SYN can provide improvement over OE. In reality, however, we do not exactly know the totals T_j. Moreover, because we have many totals T_j rather than a single T, we must now consider not just a single breakeven variance (BEV) but an entire covariance matrix as in Report 4. Under these conditions the notion of a BEV or BEV's becomes awkward. To escape this situation we introduce the enhancement of Section 2A below.

2. ENHANCEMENTS

Our enhancements are: (A) use of a 2-cell SYN which meaningfully leads us to a single BEV; (B) construction of BEV's based on medians along with, as in Section 1F, those based on expected values; (E) the use of weighted combinations of AP1 and AP2.

2A. Two Cells. For SYN we restrict our attention to (MOI's based on) area proportions which, unlike counts, directly confront an important issue: equity of estimated population figures. Based on the absolute differences between proportions in f'_1, that is, MOI 1, we consider the discrepancies among: (1) SYN based, as in Report 4, on age(5)-race(3)-sex(2), 30 cells in all; (2) SYN based merely on 2 cells: (a) either black or Hispanic, (b) other; and (3) the true proportions g_i. Expressed as a percent of the MOI-1 discrepancy between orignial-census and true proportions, these discrepancies for "artificial population" 2: are:

	1-2	1-3	2-3
States	4.71	60.57	59.94
Counties	4.66	64.50	64.84.

That is, (1) and (2) differ relatively little from each other; and the difference in their departures from (3) is relatively small. We thus view (1) SYN estimators $\hat{b}_i$ using 2 cells as being not appreciably different from (2) those based on 30 cells. Under these circumstances we justifiably evaluate SYN for area proportions for these 2 cells, instead of using all 30 cells.

Let u be "T_1/T," the proportion of the entire U.S. population that is either black or Hispanic or both. For area i let

$$d_i = y_{i2}/Y_2 \text{ and } c_i = y_{i1}/Y_1 - d_i.$$

The SYN estimated proportion for area i is now (regardless of the estimated grand total population size)

$$r_i = d_i + c_i u.$$

In practice we do not know u, but must estimate it. Here, in the manner of Section 1F, we consider a breakeven value of Var($\hat{u}$), with $\hat{u}$ an unbiased estimator of u. With $\hat{r}_i = d_i + c_i\hat{u}$ we can compute BEV's for f_k' and SYN in the same way that we have computed them, based on estimated counts $\hat{a}_i = p_i\hat{T}$, for ABD and f_k. Formulas for these ABD BEV's are straightforward, as given in Report 2. A binary search is used for MOI 1 (for which, also, we now use the minor modification for SYN at the end of Section 2B).

2B. Medians. As in Section 1F, for ABD we solved the equation $E(f_k(\hat{\underline{a}})) = f_k(\underline{y})$ for V_k, the BEV. We may also consider the median of $f_k(\hat{\underline{a}})$, which we denote by D_k ; we solve the equation $D_k = f_k(\underline{y})$ for V_k. That is, what is the value of V_k such that, for $\text{Var}(\hat{T}) = V_k$, there is a 50-50 chance that we will have $f_k(\hat{\underline{a}}) < f_k(\underline{y})$? This value for V_k is not distribution-free. We presume, as we do for MOI 1 anyway, that $\hat{T}$ is normally distributed.

Computation of V_k for ABD is as follows. For SYN and 2 cells as just considered, we will then be able to compute V_k' for $\text{Var}(\hat{u})$ in the same way, using the last paragraph of this section. First we consider k=2 and 3; then k=1. One may want to skip these mechanics and go to Section 2C.

Let $S = \text{Var}^{1/2}(\hat{T})$; we express $\hat{T}$ in the form $T + Sz$ where z is N(0,1). Each term in the summation for $f_2(\hat{\underline{a}})$ is a quadratic polynomial in z. Summing and completing the square, we thus express $f_2(\hat{\underline{a}})$, in the form $\gamma + \beta(\alpha + Sz)^2$ with α, β and γ not involving z. Thus we have

$$P(f_2(\hat{\underline{a}}) < f_2(\underline{y})) = \Phi(H_2/S) - \Phi(H_1/S) \qquad (4)$$

with $H_2 = -\alpha + \delta$, $H_1 = -\alpha - \delta$, and $\delta = [(f_k(\underline{y}) - \alpha)/\beta]^{1/2}$. We then use a binary search to solve for the value of S such that the right side of (4) is .5. Thereby we have a breakeven standard deviation (BESD) for $\hat{T}$ based on the median of $f_2(\hat{\underline{a}})$, with S^2 the BEV. For $f_3(\hat{\underline{a}})$ we do likewise.

With e denoting Sz, we may express $f_1(\hat{\underline{a}})$ as

$$F(e) = \sum p_i |h_i + e| \qquad (5)$$

with $h_i = (a_i - t_i)/p_i$. The function F is a continuous function of e; the constants p_i and h_i do not involve e. Except at the points $e=-h_i$ the function F has a derivative, F'(e), which itself is a nondecreasing step function. We may use a binary search to find a point e_0 such that $F'(e)<0$ for $e<e_0$, and $F'(e)>0$ for $e>e_0$. At e_0, F attains its minimum. Also, we have $f(\infty)=f(-\infty)=\infty$. Having gotten e_0, we thus use a binary search to find $H_1<e_0$ such that $F(H_1) = f_1(\underline{y})$. Likewise we find $H_2>e_0$ such that $F(H_2) = f_1(\underline{y})$. Then, as for MOI's 2 and 3, we use a binary search to find S such that the right side of (4) is .5. Thus we get the BEV for MOI 1.

For SYN the "i" term in F(e) of (5) is $|c_i||h_i + e|$ with c_i as in Section 2A and h_i now equal to $(r_i - g_i)/c_i$, except that for c_i extremely close to 0 the whole term is just $|r_i - g_i|$. (The same modification is used for SYN and the BEV at the end of Section 2A.)

2C. Weighted Populations. Construction of AP2 seems more realistic than AP1. As a compromise between the presumptions made for them (Section 1D), however, we may consider true counts which assign weight W to AP2, 1-W to AP1. In Section 3 we set W=.5, .75 and 1. The mixture based on .75 (in essence 75% resemblance to black, 25% resemblance to white) seems the most realistic.

3. EMPIRICAL RESULTS

Our data appendix shows our essential empirical results:

(a) ST denotes state and CO denotes county.

(b) Following ST and CO, we give the weight W assigned to AP2 as in Section 2C.

(c) Ratios are $f_k(\underline{a})/f_k(\underline{y})$ for ABD and $f_k'(\underline{r})/f_k'(\underline{p})$ for SYN.

(d) For ABD we express the BESD as a c.v.; for SYN we give the BESD itself, not a c.v.

(e) All ratios and BESD's are expressed as percents.

Thus, for example, consider a true population which gives .75 weight to AP2 and .25 weight to AP1 (as seems roughly appropriate). For SYN the ratio of $f_1(\underline{r})$ to $f_1(\underline{p})$ is 64.07%: that is, based on MOI 1, SYN produces only 64.07% of the error, in trying to determine the true state proportions, that OE produces, if we know the overall black-Hispanic proportion, u, exactly. According to MOI 2, ABD is preferable to OE in ascertaining county counts if our c.v. in estimating T is less than 1.594% -- using the expectation of $f_2(\hat{\underline{a}})$. But using the median of $f_2(\hat{\underline{a}})$, and presuming that $\hat{T}$ is normally distributed, the breakeven 1.594% is increased to 2.363%.

Such an increase in BESD is a consequence of the extreme skewness in the distribution of $f_2(\hat{\underline{a}})$, based on the squaring of a normal variate. These results may not be very robust against departures from normality; for medians we are inclined to focus much more on the BESD for MOI 1, only 1.760% for ABD and counties, which does not involve squaring. For expectations, the results for MOI 2 (and 3) do offer the appeal of being distribution-free: tentatively, we have computations showing that BESD for MOI 1 for expectations can be quite affected by non-normality.

In accordance with the attributes of W=.75, MOI 1 (unsquared) for ratios and medians, and MOI 2 for means, we would focus on an abridged data set:

	RATIOS		BE-ABD		BE-SYN	
	ABD	SYN	MEAN	MEDIAN	MEAN	MEDIAN
ST	45.05	64.07	1.616	2.347	1.091	1.488
CO	77.33	67.70	1.594	1.760	1.056	1.552.

For example, our 2-cell SYN procedure appears to be preferable if the actual SD of our estimator $\hat{u}$ is distinctly less than 1.091% for states and the use of the mean, less than 1.552% for counties and use of the median. That is, if $\mathrm{Var}^{1/2}(\hat{u})<1.091\%$, our expected error for SYN, according to MOI 2, will be less than the error for OE; and if $\mathrm{Var}^{1/2}(\hat{u})<1.552\%$, with $\hat{u}$ normally distributed, there is a better-than-even chance, according to MOI 1, that our error for SYN will be less than that for OE. Thus a decision as to whether SYN is preferable can readily be based on the anticipated precision of $\hat{u}$ as an estimator of u.

To investigate robustness against non-normality we computed BE values for states and medians with W=.75 based in effect on a standardized normal error term being replaced by (a) standardized χ^2_{60}, and (b) in effect a standardized $-\chi^2_{60}$. For BE-ABD the value 2.347 (for MOI 1) changes to (a) 2.357, (b) 2.344. For BE-SYN the value 1.488 changes to (a) 1.462, (b) 1.519. For counties, for ABD 1.760 changes to (a) 1.827 and for SYN 1.552 changes to (a) 1.518.

Regarding SYN, the original-census nationwide proportion Y_1/Y, of persons that are black or Hispanic or both, is 17.945%. For W=.5, .75 and 1 the true proportion, u, is 18.708, 18.783, and 18.857% respectively. Differences between truth and census are accordingly 0.753, 0.827, and 0.902%. One might conjecture that the BESD for SYN should be roughly equal to this difference; yet we find throughout our full data that the BESD is distinctly larger.

For ABD, on the other hand, we have shown analytically that the BESD for $\hat{T}$ for MOI 3 exactly equals the difference T-Y; for MOI 2 this same BESD is reduced by an amount which corresponds to area-by-area differentials in undercount rate. One might contemplate the use of ABD after having used SYN to alter only proportions but not the grand total. With SYN having been used, the area differentials should be less, and the BESD for MOI 2 thus closer to T-Y, than would be the case for OE. (Conditionality, and dependence between $\hat{u}$ and $\hat{T}$, could be an issue.)

DATA APPENDIX

	RATIOS-ABD			RATIOS-SYN		
	MOI 1	MOI 2	MOI 3	MOI 1	MOI 2	MOI 3
ST,.5	42.83	24.74	24.79	67.70	47.62	47.68
ST,.75	45.05	26.08	26.11	64.07	43.32	43.39
ST,1.	48.02	27.74	27.75	59.94	40.17	40.24
CO,.5	74.54	52.56	52.79	70.98	59.73	60.11
CO,.75	77.33	54.69	55.01	67.70	56.37	56.87
CO,1.	80.46	57.09	57.55	64.81	53.90	54.63

	BE-ABD-MEANS			BE-ABD-MEDIANS		
	MOI 1	MOI 2	MOI 3	MOI 1	MOI 2	MOI 3
ST,.5	1.823	1.617	1.625	2.352	2.397	2.410
ST,.75	1.806	1.616	1.625	2.347	2.396	2.410
ST,1.	1.783	1.615	1.625	2.343	2.395	2.410
CO,.5	1.347	1.596	1.625	1.820	2.367	2.409
CO,.75	1.291	1.594	1.625	1.760	2.363	2.409
CO,1.	1.223	1.591	1.625	1.691	2.359	2.409

	BE-SYN-MEANS			BE-SYN-MEDIANS		
ST,.5	1.007	1.012	1.016	1.367	1.537	1.544
ST,.75	1.118	1.091	1.095	1.488	1.651	1.658
ST,1.	1.245	1.169	1.173	1.619	1.765	1.772
CO,.5	1.067	0.972	0.984	1.455	1.467	1.487
CO,.75	1.155	1.056	1.070	1.552	1.591	1.615
CO,1	1.243	1.138	1.156	1.645	1.712	1.742