SAMPLE DESIGN FOR THE 1990 CENSUS POST ENUMERATION SURVEY

By
H. Woltman, N. Alberti, and C. Moriarity

U. S. BUREAU OF CENSUS

I. Introduction

The 1990 Census Post Enumeration Survey (PES) will be used to measure the net undercoverage of the 1990 Census. In this paper, we discuss the issues involved in designing this survey and provide an overview of the proposed sample design. Design issues that are discussed include the choice of the sampling unit, stratification, sample allocation and the resulting precision of dual systems estimates (DSE) of of population for various geographic areas and demographic subgroups of the population.

To measure the net undercoverage, one needs two samples. A sample of people who should have been counted is used to measure omissions. This is called the population or "P" sample. One also needs a sample of enumerated persons to measure duplicates and other types of erroneous enumerations. This is called the enumeration or "E" sample. The joint implementation of these two samples constitutes the PES.

The 1990 PES is designed to minimize problems encountered in 1980 without creating major new ones. A specially designed sample survey will be used rather than using the sample already selected for another survey such as the Current Population Survey (CPS), which was used for the 1980 Census Post Program (PEP). In addition, the P and E samples will overlap. That is, the same blocks are included in the E-sample as the P-sample [1]. The two samples are used to form an estimateof population using the dual system estimator [2].

A basic design decision made very early in the design cycle was that the primary sampling unit will be the block (or clusters of blocks combined to form a unit of minimum size). The sample will consist of about 5000 blocks (block clusters) totaling approximately 150,000 housing units.

II. PES POSTSTRATA

The PES sample is designed to provide sufficient precision for the dual system estimates of total population for the PES poststrata. The term "poststrata" is used to denote the finest level of detail for which direct PES estimates will be produced - i.e.,dual system estimates of the population. The poststrata are defined by characteristics of the persons enumerated in the PES and are as homogeneous as possible with respect to the Census undercount mechanisms. They were first constructed by Isaki et al to evaluate several synthetic estimates of total population [3]. The variables used to define the poststrata were Census Division, size and type of place, race, origin and overall size (population) of the poststrata. Subsequently, the poststrata were further partitioned by age, sex and in some cases tenure owner, renter). The final poststrata then consist of some 1332 population subgroups defined by geography (Census Division and place and size), race/origin, in some cases tenure, age and sex. The race/origin categories used are black, non-black, Hispanic and all other (non-black and a non-Hispanic; the age categories are 0-9, 10-19, 20-29, 30-44, 45-64 and 65+.The type and size of place typology consists of six categories:1/

1. Central Cities (CC's) in Major PMSA's - New York, Chicago, Detroit, Houston, Dallas, Ft. Worth - Arlington, and Los Angeles - Long Beach.
2. Other CC's in MSA's/PMSA's with at least one CC with a population of 250,000 or more.
3. CC's in MSA's/PMSA's without a CC of population 250,000 or more.
4. Areas outside of the CC's in MSA's/PMSA's in categories 1, 2 or 3 above.
5. Non-MSA/PMSA incorporated places with a population of 10,000 or more and
6. Non-MSA/PMSA areas that are not in an incorporated place or were in an incorporated place with a population of less than 10,000.

III. Sampling Strata

Given the objective of producing DSE's of the population for the poststrata defined above, the essence of the stratification problem is to develop sampling strata that correspond to the 110 poststrata defined by geography race,origin, and tenure as closely as possible. The crossclassifition of Division and the place type and size categories above yields 54 geographic areas that can serve as major sampling strata. The next step involved creating sampling strata within these aereas containing high concentrations of the race/origin/tenure groups corresponding to the poststrata defined by these variables for that same geographic area.

The geographic units used for this additional level of stratification were 1980 Census tracts and block numbering areas (BNAs) where they existed, and places and county remainders where neither tracts nor BNAs existed in 1980. Three steps were required to assign each geographic unit to a race/origin/tenure sampling strata: (geographic units are subsequently referred to as "tracts")

1. Assignment to a Census Division
2. Assignment to one of the 54 geographic areas within the Census

Division, and

3. Assignment to a race/origin/tenure stratum within a geographic area.

The first two steps are based on geographic location and are straightforward. The third step essentially involved combining, where possible, "tracts" with high concentrations of the race/tenure poststrata of interest in each of 54 major geographic areas. For this purpose, 1980 Census counts of occupied housing units by the race of the householder and tenure were used and the collection of geographic units having more than forty percent of one or more of the race/origin/tenure minority black or non-black Hispanic) groups were identified.

In order to evaluate the effectiveness of the stratification and to provide estimates of the resulting coefficient of variation on the DSE's we next developed an approximation to the relvariance of the DSE for a particular poststrata defined by geographic area by race,origin and tenure and the corresponding relvariance formula for the DSE of total population at the major geographic area level. Subsequently the parameters of the relvariance formula were modeled using 1980 PEP data, Isaki et al's artificial population results, and 1986 Test Census PES data. The details follow.

A. Relative Variance of the Dual System Estimate

The DSE estimate of total population for a particular race/origin/tenure poststratum d (denoted by N_d) in one of the major geographic areas (e.g. black renters in CC's in the New York PMSA) is given by

$$N_d = \sum_h \frac{N_{eh}\,(1 - P_{edh})}{(1 - P_{mdh})} \qquad (1)$$

Where

h denotes the sampling stratum

N_{edh} is the unadjusted Census count of population in the d^{th} poststratum, h^{th} sampling stratum

P_{edh} is the estimated erroneous enumeration rate ('E' sample estimate) in the d^{th} poststratum, h^{th} sampling stratum.

P_{mdh} is the estimated proportion of persons in the PES who do not match to the Census (P sample "gross miss rate" estimate) in the d^{th} poststratum, h^{th} sampling stratum.

To derive the relvariance of the DSE, shown in (2) below, we used the Taylor series approximation for the relvariance of a ratio estimate and assumed (1). the P and E sample estimates are uncorrelated and (2). the expected erroneous enumeration and gross miss rates for a particular poststratum are constant across sampling strata. ($E\,P_{mdh} = EP_{md}$ & $E\,P_{edh} = EP_{ed}$ for all h).

$$\mathrm{RelVar}(N_d) = \sum_h \left[\frac{W^2_{dh}}{\alpha_{dh} m_h \bar{H}_h \bar{K}_h}\right] \times \left[(P_{ed}/Q_{ed})\,DEF_{eh} + (P_{md}/Q_{md})\,DEF_{mh}\right] \qquad (2)$$

Where

$Q_{ed} = (1 - P_{ed})$
$Q_{md} = (1 - P_{md})$

$DEFF_{eh}$ is the E sample design effect for the h^{th} stratum

$DEFF_{mh}$ is the corresponding P sample design effect for the h^{th} stratum

W_{dh} is the proportion of the d^{th} poststratum population in the h^{th} stratum

α_{dh} is the proportion of the h^{th} stratum population in the d^{th} poststratum

m_h is the number of sample blocks in the h^{th} stratum

$\bar{H}_h$ is the average number of occupied housing units per block in the h^{th} stratum

$\bar{K}_h$ is the average persons per housing unit in the h^{th} stratum

A violation of assumption (1).is likely to result in an overestimate of the relvariance. Some positive correlation is expected between the 'P' and 'E' sample estimates since the same sample blocks will be used for both.

The DSE estimate (N) and relvariance for one of the 54 major areas is given below. Here it is assumed that the poststratum estimates, N_d, are uncorrelated.

$$N = \sum_d N_d$$

$$\mathrm{Rel\ Var}\ N = \sum_d \left[\frac{N_d}{N}\right]^2 \mathrm{RelVar}\,(N_d) \qquad (3)$$

The P and E sample design effects essentially reflect the within block clustering of missed and erroneously enumerated persons as well as variation in block size.

In order to evaluate the coefficient of variation (CV) of the dual system estimate for a particular poststratum we need values of the parameters that reflect the effectiveness of the "tract" stratification within a major geographic area (reflected by the W_{edh} and α_{dh} terms); we need to model the gross miss and erroneous enumeration rates by poststratum (P_{md} and P_{ed}); we need values for the sampling design effects and finally we need the overall sample size allocated to the major geographic areas and the method for allocating this sample to the sampling strata created within the major geographic areas. In the following sections we discuss our assumptions and results.

B. Design Parameter Values

1. Stratification Parameters

Based on the stratification scheme described earlier, sampling strata were created within each of the 54 major geographic areas in order to support DSE's for the race/origin/tenure post-strata of interest within these areas. This determined the values of W_{dh} and α_{dh} for each poststratum.

2. Assumed Erroneous Enumeration and Miss Rates

In order to model these parameters for each poststratum we used:

a. Erroneous enumeration rates based on those estimated from the 1980 PEP BY geographic area by race as shown in Table A. below.

b. net miss rates for the 110 poststrata defined by geographic area, race, origin and tenure equal to those derived from Isaki et al's artificial population research. Using this assumed net miss rate and the appropriate 1980 PEP erroneous enumeration rate for the same race/origin/tenure category we solved algebraically to determine the gross miss rate. (i.e. the P_{md} value)

c. Design effects appropriate for a block sample were estimated from PES data collected in the 1986 Los Angeles and Mississippi Test Census sites. The assumed values are shown in Table B. below. Note that in our calculations, we expressed the sample sizes assigned to sampling strata in terms of HU's so that the value in Table B are of the form ($DEFF_{eh}/\bar{K}_h$) and ($DEFF_{mh}/\bar{K}_h$). That is, the design effects that reflect clustering of missed or erroneously enumerated persons within blocks have been adjusted by the average number of persons per housing unit.

TABLE A. ASSUMED ERRONEOUS ENUMERATION RATES
(percent)

GEO AREA[1]/	ALL RACES	BLACK	NON HISPANIC	All OTHER[2]/
1, 2 or 3	3.4	4.2	4.6	2.9
4	3.2	4.3	3.8	3.0
5	3.2	4.9	3.6	3.0
6	4.5	7.1	4.6	4.2

(RACE spans BLACK, NON HISPANIC and All OTHER.)

[1]/ the numbers refer to the type and size of place typology given in paragraph II.
[2]/ nonblack and nonhispanic

TABLE B. 1990 PES SAMPLE DESIGN EFFECTS
(adjusted for average persons per HU)

GEO AREA[1]/	SAMPLING STRATUM	P-SAMPLE	E-SAMPLE
1,2,3 and 4	Black	3.5	1.6
	Non-black Hisp	3.5	1.6
	All other	1.1	1.1
5	All	2.5	1.0
6	All	1.9	1.5

1/ See Table A.

Figure 1. illustrates the final sampling strata counts (occupied HU's) crossed by the required poststrata for Central Cities in the New York PMSA. Also shown are the assumed erroneous enumeration (EE) rate, gross and net miss rates by poststratum as well as the assumed design effects for each sampling stratum. A similar table was generated for each of the 54 major geographic areas. Note, that in some cases there were too few minority HU's (black, non-black Hispanic) within a major geographic area to support additional stratification by race or in other cases by tenure within a minority group.

IV. Sample Allocation

The final step in the sample design process was to investigate and decide upon the method of allocating sample (150,000 Hu's) to each of the 101 sampling strata. The solution was to use a two-step procedure. First, a sample size (in terms of Hu's) is assigned to each of the 54 major geographic areas (to achieve some predetermined criteria at this level) and then this sample is allocated by some "optimum" method to the "tract" level sampling strata created within each major geographic area (if there are any). We discuss the second step first.

A. Allocation to "Tract" Strata within Major Geographic Area

The development of the sample allocation method involved examining the optimum allocation to sampling strata to minimize the CV on the DSE for a particular poststratum [formula (2)] or allocating to minimize the CV on the overall DSE [formula (3)] for the geographic area. Since in most geographic areas there are several poststrata for which estimates are desired, this is essentially a multi-variate optimization problem. The statistics examined are again illustrated in Figure 2 for the Central Cities of the New York PMSA. A description of each row of this illustration follows:

* Rows 1 - 3 show the assumed net miss rate, gross miss rate and erroneous enumeration rate by poststratum (as illustrated earlier in Figure 1.)

* Row 4 shows the CV of the poststratum DSE if the sample was

allocated to minimize the CV for that poststratum. These CV's are then, the minimum that can be achieved given the sampling strata.

* Row 5 shows the CV of each poststratum DSE assuming proportional allocation (using total occupied Hu's) to sampling strata.

* Row 6= Row 4/Row 5. This ratio was used to evaluate the effective ness of the stratification by virtue of the high concentrations of the persons in the poststrata of interest.

* Row 7 shows the CV on the poststratum DSE assuming the sample is allocated to minimize the CV on the overall DSE. (for a fixed sample size)

* Rows 8 - 12 show the CV's that occur for each poststratum if the optimum allocation for a particular poststratum was used

* Rows 13 - 15 show the corresponding sample allocation (distribution) to sampling strata.

Note, the CV's shown in this figure are based on a sample size of 3772 HU's.

Based on an analysis of these data for each of the 54 major geographic areas it was decided to allocate sample to minimize the CV on the overall DSE for that area. While this method is not "optimum" for any particular poststratum it still provides better CV's for each poststratum than does proportional allocation except for the all other renter and owner poststrata and is better than using the optimum allocation for any one particular poststratum (e.g., black renters). Specifically, using formula (3), the sample is allocated to sampling strata proportional to

$$[\sum_d (N_d/N)^2 (W^2_{dh}/\alpha_{dh}) \times \{(P_{ed}/Q_{ed}) DEF_{eh} + (P_{md}/Q_{md}) DEF_{mh}\}]^{1/2}$$

where as indicated earlier, the DEFF's have been adjusted for average HU size.

B. <u>Allocation to Major Geographic Area</u>

The final decision made was the method of allocating the sample of 150,000 housing units to the 54 major geographic areas. After discussion, three methods were compared. These were:

1. Assigning an equal sample size of 2778 (150,000/54) HU's to each area

2. Assigning sample to minimize the relative squared error loss function That is, to minimize

$$\sum_i (\hat{N}_i - N_i)^2 / N_i$$

where

$\hat{N}_i$ is the DSE of total population for the i^{th} area

N_i is the true population count for the i^{th} area

The expectation of this function is minimized by allocating sample proportional to $\{\sqrt{N_i} x CV_i\}$ where CV_i is the coefficient of variation for the i^{th} area derived by letting $m_h H_h=1$ in formula (3) above.

3. Assigning sample to achieve an equal coefficient of variation on the DSE estimate of population for each area. To achieve this, sample is allocated proportional to $(CV_i)^2$. The predicted CV for this allocation based on the assumed parameters discussed in Section III is approximately (.007) or 0.7%.

A comparison of the resulting CV's on the DSE of total population under each allocation method indicates that:

* for Type I CC's, the equal CV allocation method gives lower CV's for most areas
* for Type II CC's, the equal CV method yields slightly higher CV's for four areas, while the three methods yield roughly the some CV for three areas.
* for Bal MSA's, the equal CV method yields uniformly higher CV's
* for Non-MSA PL's, the equal CV method yields uniformly lower CV's than the Rel Sqrd Error method.
* for Other Areas, the equal CV method yields uniformly lower CV's then the equal sample size method and yields equal or lower CV's, except for two areas, than the Rel Sqrd Error method.

Here

* Type I CC's include those in categories 1 and 2 described earlier in section II,
* Type II CC's are those in category 3,
* Bal MSA's are those in category 4,
* Non-MSA places are those in category 5
* Other Areas are those in category 6.

Table C. below provides the distribution of the 150,000 HU sample to those same five groups of areas. The major differences between the methods are

(1). the equal CV plan allocates somewhat more sample to the Type I CC areas, compared to the other two methods. (28.6% vs. 22.2 % and 26.4%)

(2). the allocation to Type II CC areas is roughly the same for all three methods. (16.7%, 14.1% and 13.9%)

(3). the equal CV plan substantially reduces the sample allocated to the BAL of MSA's group com pared to the other two methods. (18.1% vs. 27.8% and 28.6%)

(4). the equal CV plan allocates somewhat more sample to the NonMSA PL's and Other Areas groups compared to the other two methods. (39.5% vs. 33.4% and 30.9%)

On the basis of these comparisons, it was decided to allocate sample to the 54 major geographic areas using the equal CV plan since this method allocates more sample to TYPE I CC areas and areas out-of MSA's. These areas are hypothesized to have greater potential for undercount and erroneous enumeration.

TABLE C. SAMPLE SIZE DISTRIBUTION FOR ALLOCATION METHODS

AREA	1980 OCC. HU's	EQUAL SAMPLE	MIN REL SQRD ERROR	EQUAL CV
TYPE I CC's	21.5%	22.2%	26.4%	28.6%
TYPE II CC's	12.8	16.7	14.1	13.9
BAL MSA's	42.3	27.8	28.6	18.1
NON-MSA PL's	4.8	16.7	9.6	17.4
OTHER	18.7	16.7	21.3	22.1

V. Summary

The sample for 1990 Census Post Enumeration Survey is designed to provide reliable dual system estimates for poststrata defined by census geography (Division, place type and size), race/origin, tenure, age and sex. The design will use 101 sampling strata defined by Division, place type and size and within these areas by groups of census geographic units (e.g., tracts) containing higher concentrations of the race/origin/tenure groups for which poststrata are defined.

The sample will be allocated to sampling strata in two steps: First, the sample of 150,000 occupied HU's will be allocated to 54 major geographic areas defined by the cross classification of Census Division and six place type and size categories. This allocation is designed to attain a constant coefficient of vari ation (0.7% based on assumed parameters) for dual system estimation of population computed for the 54 areas.

Second, within each of the 54 geographic areas sample will be allocated to the demographic substrata (i.e., the collection of geographic units discussed above). This step can be viewed as a multivariate optimum allocation problem since there is generally more than one post stratum of interest within each of the 54 areas. Thus, the allocation could be designed to provide the minimum coefficient of variation on the DSE for a particular poststratum, for example, black renters. However, this results in coefficients of variation for the other poststrata that are substantially greater than their own minimum value.

As a result, sample will be allocated to minimize the coefficient of variation on the DSE of the total population since this results in coefficients of variation for each of the poststrata that are close to the minimum that could be achieved under optimal allocation targeted at specific poststrata.

REFERENCES

[1]. Childers, D., Diffendal, G., Hogan, H., Schenker, N. and Wolter, K. (1987), The Technical Feasibility Correcting the 1990 Census," Proceedings of the Social Statistics Section, 1987 meeting of the American Statistical Association.

[2] Wolter, Kirk M., (1986), "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-346.

[3]. Isaki, C., Diffendal, G., and Schultz, L. (1986), "Statistical Synthetic Estimates of Undercount for Small Areas," Proceedings of the Second Annual Research Conference, Bureau of the Census, June 1986, pp. 557-567

1/ The general concept of an MSA is one of a large population nucleus together with adjacent communities which have a high degree of social and economic integration with that nucleus. CMSA's are large metropolitan complexes within which individual components, designated as PMSA's, have been defined. MSA's, CMSA's, and PMSA's are defined by the Office of Management and Budget as a standard for Federal agencies in the preparation and publication of statistics relating to metropolitan areas.

FIGURE 1:

MIDDLE ATLANTIC DIVISION (NJ, NY, PA): POSTSTRATA AND SAMPLING STRATA

GEOGRAPHIC CATEGORY: CENTRAL CITIES IN NEW YORK CITY PMSA

	PROPOSED POSTSTRATA: 1980 HU TALLIES BY RACE OF HOUSEHOLDER										TOTAL OCCUPIED HU	
	BLACK RENTER		BLACK OWNER		NONBLACK HISPANIC		ALL OTHER RENTER		ALL OTHER OWNER			
	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	SUM	TOT. %
SAMPLING STRATA:												
BLACK	385780	61.68	91346	14.60	82326	13.16	46931	7.50	19084	3.05	625467	22.28
NONBLACK HISPANIC	53926	19.30	4850	1.74	159597	57.12	50169	17.95	10885	3.90	279427	9.95
ALL OTHER	88300	4.64	14403	0.76	180714	9.50	1141942	60.02	477179	25.08	1902538	67.77
TOTAL	528006	18.81	110599	3.94	422637	15.05	1239042	44.13	507148	18.06	2807432	100.00

						OVER ALL POSTSTRATA
ASSUMED NET MISS RATES	7.59%	6.08%	5.77%	0.80%	0.64%	3.10%
ASSUMED GROSS MISS RATES	11.5%	10.0%	10.1%	3.68%	3.52%	
ASSUMED E.E. RATES	4.20%	4.20%	4.60%	2.90%	2.90%	

DESIGN EFFECT FOR MISSES- BLACK, NONBLACK HISPANIC SAMPLING STRATA: 3.500 ALL OTHER SAMPLING STRATA: 1.100

DESIGN EFFECT FOR EEs- BLACK, NONBLACK HISPANIC SAMPLING STRATA: 1.600 ALL OTHER SAMPLING STRATA: 1.100

FIGURE 2:

MIDDLE ATLANTIC DIVISION (NJ, NY, PA): POSTSTRATA AND SAMPLING STRATA

GEOGRAPHIC CATEGORY: CENTRAL CITIES IN NEW YORK CITY PMSA

	PROPOSED POSTSTRATA: 1980 HU TALLIES BY RACE OF HOUSEHOLDER										TOTAL OCCUPIED HU	
	BLACK RENTER		BLACK OWNER		NONBLACK HISPANIC		ALL OTHER RENTER		ALL OTHER OWNER			
	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	COUNT	ROW %	SUM	TOT. %
SAMPLING STRATA:												
BLACK	385780	61.68	91346	14.60	82326	13.16	46931	7.50	19084	3.05	625467	22.28
NONBLACK HISPANIC	53926	19.30	4850	1.74	159597	57.12	50169	17.95	10885	3.90	279427	9.95
ALL OTHER	88300	4.64	14403	0.76	180714	9.50	1141942	60.02	477179	25.08	1902538	67.77
TOTAL	528006	18.81	110599	3.94	422637	15.05	1239042	44.13	507148	18.06	2807432	100.00

						OVER ALL POSTSTRATA
1. ASSUMED NET MISS RATES	7.59%	6.08%	5.77%	0.80%	0.64%	3.10%
2. ASSUMED GROSS MISS RATES	11.5%	10.0%	10.1%	3.68%	3.52%	
3. ASSUMED E.E. RATES	4.20%	4.20%	4.60%	2.90%	2.90%	
4. CVs OF INDIVIDUAL OPT.	0.0192	0.0376	0.0211	0.0069	0.0105	
5. CVs OF PROPORTIONAL	0.0257	0.0533	0.0246	0.0071	0.0108	
6. CV RATIO IND.OPT./PROP.	0.7484	0.7055	0.8551	0.9790	0.9726	
						OVER ALL POSTSTRATA
7. CVs OF OVERALL OPT.	0.0209	0.0428	0.0221	0.0080	0.0124	0.0070

CVs THAT WOULD OCCUR IF EACH POSTSTRATUM'S OPTIMAL ALLOCATION (INDIVIDUAL OPT.) WAS ACTUALLY USED:

						OVER ALL POSTSTRATA
8. BLACK RENTER	0.0192	0.0380	0.0244	0.0101	0.0157	0.0076
9. BLACK OWNER	0.0195	0.0376	0.0267	0.0105	0.0163	0.0079
10. NONBLACK HISPANIC	0.0228	0.0477	0.0211	0.0082	0.0127	0.0072
11. ALL OTHER RENTER	0.0314	0.0662	0.0265	0.0069	0.0105	0.0086
12. ALL OTHER OWNER	0.0316	0.0662	0.0285	0.0070	0.0105	0.0088

THE FIGURES BELOW ARE THE SAMPLE ALLOCATION PROPORTIONS (%) TO THE SAMPLING STRATA WHICH MINIMIZE THE POSTSTRATA CVs (INDIVIDUAL OPT.); THE FIGURES IN THE FAR RIGHT COLUMN MINIMIZE THE GEOGRAPHIC AREA CV (OVERALL OPT.).

						OVERALL OPTIMUM
13. BLACK	57.03%	63.45%	28.46%	13.86%	14.03%	37.22%
14. NONBLACK HISPANIC	14.25%	9.77%	26.48%	9.58%	7.08%	15.46%
15. ALL OTHER	28.72%	26.77%	45.06%	76.57%	78.89%	47.32%

SAMPLE SIZE=3772