Journal of Official Statistics
Vol. 2. No. 2, 1986, pp. 169–179

Dual System Estimation Using Demographic Analysis Data

Cary T. Isaki and Linda K. Schultz[1]

Abstract : In constructing the dual system estimator used in the Post Enumeration Program (PEP) an assumption of independence between the census and a post enumeration survey was made. There is reason to believe that this assumption may, in fact, not be true. In this paper, assuming demographic analysis is correct, we are able to obtain measures of association between the census and the PEP for each age-race-sex cell at the national level. Using these measures we construct several estimators. The estimators are then compared to the dual system estimator as well as U.S. level demographic analysis age-race-sex cell population numbers.

Key words: Dual system estimator; odds ratio; correlation; demographic analysis.

1. Introduction

There is a great deal of literature on dual system estimation for estimating the size of an animal population. With regard to animal populations the estimation procedure is called capture-recapture. An example of one of the simplest applications of capture-recapture is the estimation of the total number of fish in a pond. The procedure consists of netting fish in a pond, tagging and releasing the fish captured and netting fish in the pond a second time. A number of problems arise when one applies capture-recapture theory to human populations and it is necessary to modify the basic model. A review of some of these modifications can be found in Wolter (1983). Two additional detailed references on the application of dual system estimation are Marks, Seltzer and Krotki (1974) and Krotki (1978).

One of the early applications of dual system estimation to demographic data was presented by Chandrasekaran and Deming (1949). In their paper a formal probability model was presented and their estimation method was applied to the estimation of the number of births and deaths in several Indian villages. Chandrasekaran and Deming used a registration list of vital events and the results of a house-to-house canvass as their two data collection methods and matched events that were obtained by both methods. The authors point out several possible problems in the implementation of their method that has come to be called dual system estimation. One problem is the occurrence of errors in the data, especially matching errors. Another problem is the variation in error rates of reporting vital events among respondent categories. For example, they observed that deaths were

[1] Cary T. Isaki and Linda K. Schultz are mathematical statisticians in the Statistical Research Division, Bureau of the Census, Washington, D.C. 20233. The authors would like to thank the referees and the associate editor for their helpful comments.

recorded with a lower error rate for older age groups. Still another problem is that their estimator assumes independence while the data collection methods may not be statistically independent in that they are likely to record certain vital events and miss others in a systematic manner.

This paper will address the issue of statistical dependence. More specifically, we propose several total population estimators that can be used when one suspects that dependence exists between the two data collection procedures. A discussion of this problem and its effects can be found, for example, in articles by Jabine and Bershad (1968), Greenfield (1975), Greenfield and Tam (1976) and El-Khorazaty and Sen (1975). The correlation model linking the two data collection methods used below was presented by Jabine and Bershad.

2. Dual System Estimator

Let X_{1i} and X_{2i} denote Bernoulli random variables with probabilities of success (capture) P_1 and P_2, respectively for the ith unit in a population of size N. Let $\bar{P}_2 = 1-P_2$ and $\bar{P}_1 = 1-P_1$. We define the correlation between X_{1i} and X_{2i} by

$$\rho(X_1, X_2) = (P_2\bar{P}_2P_1\bar{P}_1)^{-1/2}(E[X_2X_1] - P_1P_2) \tag{1}$$

where $E[\cdot]$ denotes the expectation operator. The probability model we use assumes that any randomly selected unit in the population of size N has a probability P_1 of being observed by method 1, a probability P_2 of being observed by method 2 and that the observations may be correlated. Suppose that two data collection procedures have been applied to N units in the population and the resulting units matched without error. Then the following 2×2 table can be constructed:

		Method 1 observed	Method 1 not observed	
Method 2	observed	M	X	N_2
	not observed	Y	Z	–
		N_1	–	N

(2)

where

M denotes the total matched cases

X denotes the total cases observed, by method 2 but not by method 1

Y denotes the total cases observed, by method 1 but not by method 2

Z denotes the total cases not observed, by either method, and

$N_1 = M+Y$, $N_2 = M + X$ and N denotes the total population (parameter of interest).

The dual system estimator, d_1, is defined to be

$$d_1 = M^{-1}N_1N_2 \text{ or, equivalently,} \\ = M + X + Y + M^{-1}XY. \tag{3}$$

Under the probability model in (1), the expected value of d_1 including second order terms is

$$E[d_1] = [P_2P_1 + \rho g]^{-1}NP_1P_2 + [P_2P_1 + \rho g]^{-2}[\rho g(\rho g - P_1\bar{P}_2 - P_2\bar{P}_1) + g^2], \tag{4}$$

where $g = (P_2\bar{P}_2P_1\bar{P}_1)^{1/2}$.

In the case of independence ($\rho=0$), (4) reduces to

$$E[d_1] = N + (P_2P_1)^{-1}\bar{P}_2\bar{P}_1. \tag{5}$$

In either (4) or (5), with P_2 and P_1 likely to exceed 0.7, the second term in (4) with non-zero ρ or the second term in (5) with $\rho=0$, are negligible. If the correlation is zero (or equivalently that X_2 and X_1 are independent), d_1 is unbiased given the approximations used. It is easy to see that in terms of bias it is better to use data capture methods with small ρ and large capture probabilities. From (4) the sign of ρ dictates whether d_1 is negatively or positively biased. When ρ is positive, d_1 has a negative bias and vice versa. Given two separate dual system estimation methods differing only in their correlations (say, $0 < \rho_1 < \rho_2$), it can be shown that to first order approximations, the absolute value of the bias of the method with ρ_2 will exceed the absolute value of the bias of the other method.

3. The PEP Dual System Estimator

As part of a research program conducted during the 1980 Census of Population, a dual system estimation method was utilized under the Post Enumeration Program (PEP). The PEP was developed to estimate the total population of states and large metropolitan areas as well as the population of certain minority groups for the entire U.S. In this formulation, the census was used as one data collection method while the other method was a household sample (April and August Current Population Survey panels). The PEP was an extensive survey effort and the details of its planning, data collection and editing, imputation and estimation phases can be found in Cowan and Bettin (1982) and Fay and Cowan (1983). In particular, the treatment of nonrespondents and the determination of respondents' match or non-match status led to a dozen separate estimates of total population for a given area. In what follows, we use one of these data sets termed PEP 3–8. The PEP 3–8 data set consists of April Current Population Survey respondents only. Nonrespondents and refusals are not used in the estimation. Incomplete survey cases are imputed using completed cases obtained from follow-ups. In addition, we examined a sample of census respondents to estimate (and adjust) gross overcounts arising from duplication of coverage, curbstoning, etc. A special methodology using a post office review of incomplete cases was used to impute for noninterviews of census reinterview cases.

Certain modifications in the development of the dual system estimator described in the previous section need clarification. We exclude from our discussion those members of the total population who are institutionalized. By institutionalized we mean those persons who are in prisons, mental hospitals, homes for the aged, etc. In most applications of the dual system procedure, methods 1 and 2 are complete enumerations of the entire population or of clusters of large sample areas. This is not the case in the present application where method 1 coverage of the population used a subsample of household segments selected from within primary sampling units for re-enumeration. The method 2 enumeration was the actual census. This modification of method 1 resulted in the type of anomoly caused by the use of such a sampling scheme. That is, some of the observed cell values in (2) may be negative. The cell values M and Y in (2) are obtained by first matching method 1 respondents' reported information to the method 2 listing of persons and applying sampling weights. This matching process was done in one direction only. We obtained an estimate of X by subtracting the estimate of M from N_2. In practice, it was possible for X to be negative.

In some cases, persons listed under method 2 had been imputed or fabricated so that a true match did not exist. To solve this problem, a sample of method 2 responses were reinterviewed. Based on the reinterview sample, an estimate of the imputed and fabricated cases was obtained and subtracted from the original census count to obtain N_2.

In other cases, information from both methods 1 and 2 had been collected, but this information proved inadequate for matching purposes. In these cases it was necessary to impute a match or nonmatch determination.

Because it was felt that the capture probabilities, P_1 and P_2, varied by age, race, sex and state, the estimation of total population for the U.S. was obtained by using d_1 to estimate the total in each age by race by sex by state domain. The totals for each age, race, sex domain were then summed over states to arrive at a U.S. level total population figure for each domain.

4. Alternative Estimators Under Dual Systems

We now illustrate how the demographic analysis estimates of the noninstitutional population by age, race and sex can be used to construct alternative estimators of total population for areas. The demographic age-race-sex estimates, provided by Passel and Robinson (1984), include 3.5 million illegal aliens. With the assumed 3.5 million illegal aliens, the demographic analysis total population is approximately equal to that from PEP 3–8. While it is recognized that in practice the demographic estimates are subject to error, for our purposes we assume, nevertheless, that the estimates are indeed adequate.

The following description illustrates how the demographic analysis estimates are used in the construction of the alternative estimators. Recall the age-race-sex domain layout as in (2) (although N and Z are not known) is available for each state and the District of Columbia. If j is a subscript denoting the jth state, a single dual system estimator for a specific age-race-sex cell is

$$d_0 = (\sum_{j=1}^{51} M_j)^{-1} \sum_{j=1}^{51} N_{1j} \sum_{j=1}^{51} N_{2j}$$
$$= M^{-1} N_1 N_2 .$$

If it is assumed that P_1 and P_2 do not vary among states and ρ_j is the correlation between methods for the jth state then the approximate expectation of d_0 is

$$E[d_0] = [P_2 P_1 + g\bar{\rho}]^{-1} N_{DA} P_2 P_1 , \qquad (6)$$

where $N_{DA} = \sum_{j=1}^{51} N_{DAj}$ and

$$\bar{\rho} = \sum_{j=1}^{51} \frac{N_{DAj}}{N_{DA}} \rho_j .$$

N_{DA} is the demographic analysis count and $\bar{\rho}$ is a weighted average of state correlations both for a specific age-race-sex cell.

Given that nearly 20 percent of the population change their residences from one year to the next, a feasible method to provide state estimates of population size via demographic analysis (age-race-sex domains, i.e., N_{DAj}) is not available. While birth and death records are available on an annual basis for states, updated population registers, such as those existing for provinces in Sweden, through which population movement can be measured are nonexistent.

Because the individual values, N_{DAj}, are not currently available we estimate $\bar{\rho}$ for each age, race, sex cell by replacing the left hand side of (6) with d_0 and estimating P_1 and P_2 by setting

$$P_1 = \frac{N_1}{N_{DA}} \text{ and } P_2 = \frac{N_2}{N_{DA}} \text{ for a given age, race,}$$

sex cell. This is appropriate because under the model, the expected values of the marginal totals represent the total number of units obtained through the methods.

Since we assume that N_{DA} in (6), obtained via demographic analysis, is correct we can obtain an estimate of ρ, called $\hat{\bar{\rho}}$, for each age, race, sex cell using

$$\hat{\bar{\rho}} = (gd_o)^{-1} [P_1 P_2 (N_{DA} - d_o)].$$

This expression is algebraically equal to
$[(M+X)(M+Y)(X+Z)(Y+Z)]^{-1/2} (MZ-XY)$
where Z is defined as $N_{DA} - M - X - Y$. Substituting this expression for Z (since Z is an unknown quantity) we have another expression for $\hat{\bar{\rho}}$ =

$$[(M+X)(M+Y)(N_{DA}-X-M)(N_{DA}-Y-M)]^{-1/2} \times [M(N_{DA}-M-X-Y)-XY].$$

In calculating the $\hat{\bar{\rho}}$'s for the different age, race, sex cells we discovered that in several cases $N_{DA} < M+X+Y$, implying $Z<0$. Obviously this should not occur. There could be several different explanations for this. First, we have assumed that the matching done between the PEP and the census is perfect, that no errors have occurred in this process. This is certainly not a realistic assumption. It is possible that what we classified as X and Y really should have been classified as a match. Had they been correctly matched we would have observed an increase in M and a decrease in both X and Y. Age misreporting would be one cause of nonmatches. For example, if one classified oneself as 34 in the census and 35 in the PEP, a match might not occur.

Another possible reason for negative Z estimates could be the assumed number of illegal aliens in the demographic analysis estimates. We have assumed the number of illegal aliens to be 3.5 million. Whether there are 3.5 million illegal aliens or not is, of course, unknown. An underestimate for the number of illegal aliens or an incorrect distribution of illegal aliens to the age, race, sex cells could also contribute to the negative estimates in the fourth cell for some of the age, race, sex cells. (There is the possibility of new legislation that would legalize the status of illegal persons in the U.S. and simultaneously place strict sanctions on employers of future illegal aliens. It is felt that such new legislation would minimize the need to speculate on the size of the illegal population.)

Sampling variability could also be a factor in the negative estimates. Recall that $N_2 = C{-}II{-}EE$ where C is the census total, II is the total number of persons imputed by the census who could not be matched to a method 1 case, and EE is the estimated total number of persons who were erroneously enumerated. Such persons were either fabricated by the enumerator, out of scope to the census (example, persons born after the census date), or coded to incorrect geography, etc. When we have both geographic error and extensive matching, method 1 cases are matched to census coded cases for only a limited part of the census file, i.e., a given geographic area. Under these conditions it was necessary to inflate the EE estimate somewhat for persons whose incorrect census geography made a method 1 match impossible. While this in turn reduces the overall N_2, it also contributes to the negative Z problem by reducing the size of M. We decided to set $Z=0$ for the age, race, sex cells in which the estimate of Z arrived at by subtraction was negative. The resulting $\hat{\bar{\rho}}$ are presented in Table 1 on the following page along with the estimated P_1 and P_2 for race (Black, Non-Black) by sex by age (five year age groups 0 to 64). For two cells, Black male 10 to 14 and Black female 60 to 64 the estimated capture probabilities were close to or exceeded 1.

To estimate $\hat{\bar{\rho}}$, as described above, we assumed that the response probabilities P_1 and P_2 did not vary from state to state. In practice, P_1 and P_2 are likely to vary among states. We made this assumption in order to obtain a crude estimate of $\bar{\rho}$. In constructing alternative dual system estimators, we assumed that each age-race-sex combination had a corresponding $\bar{\rho}$ that was the same for every state. We allowed P_1 and P_2 to vary from state to state and assumed that P_1 and P_2 varied within the limits of the correlation model implicit in (1). This approach is different from a direct application of a synthetic estimation procedure applied to method 2 (census) listings where the same adjustment factor (using demographic analysis age-race-sex data at the U.S. level) is applied in every state. We do not expect a simple synthetic estimator derived from the national level to estimate state populations adequately because of the differential undercount rates by race and geography.

Table 1. Estimates of P_1, P_2, Correlation and the Multiplier of the Odds Ratio for Age, Race, Sex Cells at the U.S. Level

	NB-Male				NB-Female				B-Male				B-Female			
Age	P_1	P_2	$\hat{\bar{\rho}}$	$\hat{\Theta}$	P_1	P_2	$\hat{\bar{\rho}}$	$\hat{\Theta}$	P_1	P_2	$\hat{\bar{\rho}}$	$\hat{\Theta}$	P_1	P_2	$\hat{\bar{\rho}}$	$\hat{\Theta}$
<5	.961	.951	–.046	0	.922	.952	–.065	0	.881	.826	.085	1.827	.897	.831	– .151	.007
5–9	.978	.954	–.033	0	.960	.949	.145	5.553	.916	.866	–.079	.288	.936	.873	.024	1.313
10–14	.978	.966	–.028	0	.963	.968	–.036	0	1.000	.918	*		.943	.919	–.073	0
15–19	.957	.950	–.049	0	.948	.959	–.048	0	.921	.918	–.088	0	.924	.927	–.080	0
20–24	.875	.921	–.103	.054	.921	.944	–.071	0	.719	.824	.145	2.182	.782	.885	–.108	.341
25–29	.871	.915	.236	5.454	.932	.945	.104	3.246	.692	.757	.375	5.955	.871	.878	–.013	.884
30–34	.909	.934	.190	5.127	.935	.958	.072	2.716	.718	.774	.363	5.855	.981	.899	–.047	0
35–39	.898	.917	.364	11.954	.948	.947	.307	13.682	.709	.749	.367	5.682	.815	.886	.128	2.372
40–44	.915	.935	.355	13.661	.949	.960	.020	1.476	.715	.779	.486	11.373	.911	.896	.101	2.422
45–49	.926	.935	.450	24.502	.962	.954	.438	34.976	.684	.741	.483	10.253	.878	.875	.299	6.618
50–54	.939	.952	.307	13.484	.937	.953	.453	31.746	.787	.787	.611	25.139	.956	.919	.270	10.675
55–59	.936	.939	.609	66.264	.948	.956	.422	29.331	.845	.869	.003	1.027	.932	.921	.187	5.180
60–64	.908	.935	.442	21.877	.936	.952	.417	25.332	.873	.891	.049	1.520	1.045	.930	*	

All PEP data are 3–8

* indicates a "correlation" that cannot be calculated

$$\hat{\bar{\rho}} = \frac{P_1 P_2 [N_{DA} - d_o]}{g(d_o)} \qquad P_1 = \frac{N_1}{N_{DA}} \qquad P_2 = \frac{N_2}{N_{DA}} = \frac{C-II-EE}{N_{DA}}$$

$$\hat{\Theta} = \frac{MZ}{XY} \text{ where } Z = N_{DA}-M-X-Y \qquad g = \sqrt{P_1 \bar{P}_1 P_2 \bar{P}_2} \qquad \bar{P}_1 = 1-P_1$$

$II \equiv$ Number of persons substituted in the census. EE$\equiv$ Estimated number of persons erroneously enumerated in the census.

4.1. *Modified Dual System Estimator*

While the dual system estimator d_1 is nearly unbiased when ρ is zero, there is a possibility of substantial bias when ρ is not zero. As can be seen from Table 1, there is a large number of age-race-sex cells where this is the case. To counter this bias, we constructed the alternative dual system estimator $\hat{d}_1$, that follows. The estimator is

$$\hat{d}_1 = [M - \hat{\rho}(XY)^{1/2}]^{-1} N_1 N_2 ,$$

where it is assumed that $\hat{\rho}$ is known and $|\hat{\rho}| < |\rho|$ where ρ is the actual correlation. For a first order approximation it can be shown that

$$E[\hat{d}_1] = [(P_1 P_2 + \rho g) - \hat{\rho} k]^{-1} N P_2 P_1$$

where

$$k = [(P_1 \bar{P}_2 - \rho g)(P_2 \bar{P}_1 - \rho g)]^{1/2}.$$

Under these assumptions, it can be shown that $\hat{d}_1$ has a smaller bias than d_1.

Our motivation for constructing $\hat{d}_1$ follows from (4) where it can be observed that the denominator of the first term is too large when ρ is positive (and vice versa). If one has knowledge of a lower bound of the positive ρ, then one should use that information along with the sample data to remove this positive term and in this way reduce the bias of the estimator.

4.2. *Greenfield Estimator*

The second estimator is one proposed by Greenfield in a different context but translates in a straight forward manner to the present context. Greenfield proposed an estimator of total population under dual system estimation that estimates the missing cell in (2) by solving the quadratic equation in Z arising from

$$r_Z = [(M+Y)(X+Z)(M+X)(Y+Z)]^{-1/2} \times (MZ - XY) . \quad (7)$$

A rough estimate of r_Z is used to solve the equation. We shall refer to the estimator $d_2 = M + X + Y + \hat{Z}$ as Greenfield's estimator where $\hat{Z}$ is the estimated total number of persons missed by both methods and is obtained by solving (7) with $\hat{\bar{\rho}}$ in place of r_Z.

4.3. *Odds Estimator*

An alternative estimator of Z incorporates the estimated odds ratio $\hat{\Theta} = [XY]^{-1} MZ$. Assuming that the demographic analysis estimates are correct, an odds ratio is computed for each age-race-sex cell at the U.S. level. The odds ratio for a given cell is then used to estimate the component Z_j for the jth state, say, by $\hat{\hat{Z}}_j = M_j^{-1} \hat{\Theta} X_j Y_j$. The estimator of total population is then $d_3 = M + X + Y + \hat{\hat{Z}}$ (omitting the subscript j). The reader will note that assuming knowledge of $\hat{\Theta}$ is equivalent to assuming knowledge of ρ. The estimator d_3 differs from d_1 in that d_3 uses the odds ratio to adjust the estimator of Z when correlation is assumed. The estimator d_3 is discussed in Ericksen and Kadane (1985). Also, d_2 and d_3 (called the "odds" estimator) are different only in the way that Z is estimated. If $\hat{\Theta} = 1$ or equivalently $\hat{\bar{\rho}} = 0$, then $d_2 = d_3 = d_1 = \hat{d}_1$.

5. Results

A direct method of comparing the performances of the estimators under consideration is to construct state estimates and compare them to a standard. Unfortunately, neither demographic analysis state estimates nor any other comparable figures exist. As an alternative means of comparison we chose to compute each age-race-sex cell estimate for each state, sum them over all states and compare them to the assumed correct U.S. demographic analysis age-race-sex figure. These comparisons do not directly address the question of the accuracy of the state total population estimates. They do, however, provide an idea of the quality of the estimates

Table 2. Sum of Dual System Estimates over 51 States by Age-Sex-Race Noninstitutional Population Only (Including 3.5 Million Illegals)

Age Interval	N_{DA}	d_1*	$\hat{d}_1$	d_2	d_3
			NB-Male		
<5	7 177 181	7 224 494	7 210 574	7 239 409	7 207 037
5–9	7 351 903	7 360 692	7 353 226	7 368 487	7 352 327
10–14	7 905 470	7 944 195	7 938 615	7 950 018	7 936 677
15–19	9 170 115	9 261 551	9 239 158	9 285 545	9 232 231
20–24	9 384 766	9 498 267	9 379 871	9 635 383	9 385 042
25–29	8 812 969	8 488 801	8 755 621	8 812 501	8 839 187
30–34	7 925 814	7 513 386	7 895 671	7 921 729	7 938 794
35–39	6 418 084	6 190 701	6 321 222	6 411 469	6 431 839
40–44	5 254 361	5 106 221	5 180 140	5 238 516	5 219 883
45–49	4 967 150	4 806 864	4 884 595	4 964 726	4 896 027
50–54	5 130 904	5 044 768	5 101 436	5 132 611	5 163 881
55–59	5 087 079	4 889 467	4 957 681	5 095 045	5 096 681
60–64	4 370 655	4 215 182	4 286 588	4 371 394	4 387 627

Age Interval	N_{DA}	d_1*	$\hat{d}_1$	d_2	d_3
			NB-Female		
<5	6 810 193	6 860 033	6 830 611	6 892 484	6 824 737
5–9	6 981 484	6 932 322	6 963 148	6 970 792	6 972 833
10–14	7 547 988	7 570 617	7 561 444	7 580 230	7 558 541
15–19	8 841 396	9 059 874	9 028 274	9 093 538	9 008 103
20–24	9 177 629	9 378 369	9 317 688	9 445 276	9 302 958
25–29	8 619 688	8 564 625	8 611 225	8 617 838	8 628 558
30–34	7 843 937	7 811 955	7 837 659	7 840 073	7 832 443
35–39	6 399 094	6 292 659	6 361 183	6 395 828	6 409 279
40–44	5 291 067	5 284 436	5 288 585	5 288 688	5 288 534
45–49	5 124 656	5 029 129	5 080 215	5 126 304	5 147 889
50–54	5 492 963	5 353 648	5 419 420	5 488 976	5 500 775
55–59	5 564 250	5 448 894	5 511 279	5 561 916	5 569 276
60–64	4 922 069	4 805 977	4 869 375	4 924 171	4 959 080

* d_1 recomputed after collapsing over states in alphabetic sort because $N_2 - M < 0$.

at another level, namely, a given age-race-sex total at the national level. The results provided in Table 2 enable us to make this type of comparison. We believe that an estimator that exhibits superior performance in nearly all cells of national level totals is also likely to exhibit superior performance in estimating state total populations. Because the cell entries M, X and Y in (2) are sample based estimates, it is possible to obtain negative estimates of Z. We have already discussed several possible explanations for the negative estimates of Z. Setting $Z = 0$ in constructing d_1, d_2 and d_3 was likely to have reduced their biasses.

Table 2 (Cont.). Sum of Dual System Estimates over 51 States by Age-Sex-Race Noninstitutional Population Only (Including 3.5 Million Illegals)

Age Interval	N_{DA}	d_1*	$\hat{d}_1$	d_2	d_3
			B-Male		
<5	1 371 853	1 356 136	1 373 271	1 375 253	1 378 997
5–9	1 348 499	1 361 630	1 348 219	1 376 851	1 347 606
10–14	1 366 854				
15–19	1 465 217	1 528 840	1 511 824	1 548 337	1 506 190
20–24	1 376 208	1 323 201	1 370 250	1 381 367	1 381 462
25–29	1 206 981	1 060 359	1 153 904	1 248 411	1 230 821
30–34	985 140	880 364	956 900	1 025 780	1 026 070
35–39	797 820	708 591	774 234	835 768	836 120
40–44	671 583	572 809	607 633	682 048	658 253
45–49	630 071	525 477	663 458	624 480	637 959
50–54	582 742	498 184	520 397	596 149	560 053
55–59	503 209	503 761	504 010	504 010	504 058
60–64	389 922	385 287	387 284	387 427	387 490

Age Interval	N_{DA}	d_1*	$\hat{d}_1$	d_2	d_3
			B-Female		
<5	1 340 910	1 374 465	1 338 680	1 420 928	1 340 428
5–9	1 320 079	1 317 182	1 320 066	1 320 152	1 320 586
10–14	1 355 907	1 420 557	1 408 138	1 434 386	1 401 592
15–19	1 498 440	1 553 611	1 539 815	1 569 288	1 538 751
20–24	1 478 350	1 509 103	1 483 206	1 559 190	1 476 049
25–29	1 293 728	1 306 574	1 303 814	1 309 399	1 302 596
30–34	1 053 381	1 076 331	1 072 708	1 080 202	1 071 281
35–39	846 759	824 430	838 726	841 643	842 867
40–44	714 618	704 604	710 954	711 795	712 449
45–49	664 288	639 716	659 301	669 492	679 405
50–54	635 940	625 313	631 059	634 540	637 158
55–59	577 905	569 826	575 853	578 072	583 185
60–64	485 481				

* d_1 recomputed after collapsing over states in alphabetic sort because $N_2 - M < 0$.

For example in the case of d_3 this results in $\hat{\Theta} = 0$ and we do not add a positive term to $M + X + Y$ as would be the case with d_1 (see (3)). In Table 1, $\hat{\rho}$ and $\hat{\Theta}$ are not strictly interpretable as a correlation coefficient and an odds ratio, respectively. Rather, they should be viewed as adjustment factors. Situations where Z is negative are identified in Table 1 where the $\hat{\Theta}$'s are equal to zero.

In addition to the negative Z estimate, it is also possible to obtain negative X estimates for some state age-race-sex cells. As mentioned previously, negative X estimates occurred because X was obtained by subtraction. To obtain the U.S. age-race-sex cells in Table 2 we listed states in alphabetical order (a random sort) and collapsed the cells of adjacent states (or groups of states) until a positive X estimate resulted. The same collapsing was used for all estimates. We then computed the estimators at the grouped level. It is anticipated that a future PEP type application will be based on a block sample design (cluster of fifty households) rather than a segment sample design (four households). We are not likely to get negative X estimates because with a block sample design, the re-enumeration of households for method 1 and the census sample to eliminate curbstoning, census imputes, out of scopes, etc. are done on the same blocks. The reader will recognize that the estimators $\hat{d}_1$, d_2 and d_3 are not defined when X is negative.

As might be expected, the data in Table 2 show that d_1 is not as close (as measured by absolute differences) to the "correct" value as are the other estimators for each age-race-sex cell except when $\hat{\bar{\rho}}$ is near zero. The Greenfield estimator, d_2 unit, does best overall when $\hat{\bar{\rho}}$ is positive but is very poor when $\hat{\bar{\rho}}$ is negative. This is a consequence of the quadratic equation implicit in its construction that ignores the sign of $\hat{\bar{\rho}}$. Estimator $\hat{d}_1$, the modified dual system estimator, is viewed as a compromise between d_1 and d_2 in the sense that it takes account of $\hat{\bar{\rho}}$ (which d_1 ignores) but is cruder than d_2 in the sense that it does not use all of the available information. The "odds" estimator, d_3, on the whole, does not perform as well as d_2 when $\hat{\bar{\rho}}$ is positive, but is the preferred estimator when $\hat{\bar{\rho}}$ is negative or when $\hat{\bar{\rho}}$ is interpreted as an adjustment factor. Exceptions do occur. For example $\hat{d}_1$ performs better than d_2 or d_3 for some cells (B-male, 30 to 34) but overall d_2 and d_3 are to be preferred. The ratio adjustment of d_1 to the known demographic analysis figure by age-race-sex is an option. However, since it was the aim of this work to compare the performance of the "raw" estimates, this was not done.

The previous discussion naturally suggests the use of a hybrid estimator $d_4 = d_2$ if $\hat{\bar{\rho}} > 0$ and $= d_3$, otherwise. One could refine d_4 by using its components only when $\hat{\bar{\rho}}$ is safely away from zero while using d_1, otherwise. We leave these issues and the construction of estimates of state total population for future research.

One disturbing aspect of the entries of Table 1 that was pointed out by the referees was the lack of smoothness of $\hat{\bar{\rho}}$ for NB-female 40–44 and for B-male 55–59 and 60–64 (smaller $\hat{\bar{\rho}}$'s). We cannot explain this lack of smoothness. Small values of $\hat{\bar{\rho}}$, however, resulted in alternative estimates that were nearly always an improvement over d_1.

6. Summary

The purpose of this work was to explore the means of combining information available from demographic analysis for use in estimating total population via dual frame estimation methods. In the process we have discovered and attempted to remedy several problems relating to the alternative dual system estimators described. Ericksen and Kadane (1985) proposed use of d_3 for the Black population because this population is not believed to contain many illegal persons. However, in such populations, one still experiences negative estimates of Z.

Assessing the accuracy of the estimators by aggregating them for comparison at higher levels is not satisfactory. However, in the absence of other data, it is the best available evaluation criterion. Based on the assumptions that the demographic analysis numbers are correct and that the previously mentioned errors of the post enumeration program are absent, the results of Table 2 indicate that gains in estimation can be obtained. The hybrid estimator, d_4, appears to be the better estimator among those considered. Again, the work assumes that the demographic analysis data are correct. In the future, we intend to produce state estimates of population using the above estimation methods and compare them with existing PEP estimates. Construction of state estimates will involve collapsing age-sex cells and require recomputation of $\hat{\bar{p}}$ and $\hat{\Theta}$ at the U.S. level. An unresolved problem to be faced is determining which state estimate works best in the absence of a standard.

7. References

Chandrasekaran, C. and Deming, W.E. (1949): On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal of the American Statistical Association, 44, pp. 101–115.

Cowan, C.D. and Bettin, P.J. (1982): Estimates and Missing Data Problems in the Post Enumeration Program. Report presented to a special committee of the American Statistical Association on undercount. U.S. Bureau of the Census.

El-Khorazaty, M.N. and Sen, P.K. (1975): The Capture-Mark-Recapture Strategy as a Method for Estimating the Number of Events in a Human Population with Data from Dependent Sources. Joint Statistical Agreement Report to the U.S. Bureau of the Census.

Ericksen, E. and Kadane, J. (1985): Estimating the Population in a Census Year: 1980 and Beyond. Journal of the American Statistical Association, 80, pp. 98–109.

Fay, R.E. and Cowan, C.D. (1983): Missing Data Problems in Coverage Evaluation Studies. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 158–163.

Greenfield, C.C. (1975): On the Estimation of a Missing Cell in a 2×2 Contingency Table. Journal of the Royal Statistical Society, Series A, 138, pp. 51–61.

Greenfield, C.C. and Tam, S.M. (1976): A Simple Approximation for the Upper Limit to the Value of a Missing Cell in a 2×2 Contingency Table. Journal of the Royal Statistical Society, Series A, 139, pp. 96–103.

Jabine, T.B. and Bershad, M.A. (1968): Some Comments on the Chandra Sekar-Deming Technique for the Measurement of Population Change. Paper presented at the CENTO Symposium on Demographic Statistics, Karachi, Pakistan.

Krotki, K. (1978): Developments in Dual System Estimation of Population Size and Growth. The University of Alberta Press, Edmonton.

Marks, E., Seltzer, W., and Krotki, K. (1974): Population Growth Estimation. The Population Council, New York.

Passel, J. and Robinson, G. (1984): Unpublished tabulations, Population Division, U.S. Bureau of the Census.

Passel, J. and Robinson, G. (1984): Revised Estimates of the Coverage of the Population in the 1980 Census Based on Demographic Analysis: A Report on Work in Progress. Paper presented at the American Statistical Association meeting, Philadelphia, PA.

Wolter, K.M. (1983): Coverage Error Models for Census and Survey Data. Bulletin of the International Statistical Institute, L, pp. 415–432.

Received March 1985
Revised April 1986