A PROBABILITY MODEL FOR CENSUS ADJUSTMENT

by D. A. Freedman, P. B. Stark and K. W. Wachter
Department of Statistics
University of California, Berkeley, CA 94720

Technical Report No. 557
Prepared for *Mathematical Population Studies* 31 July 2000

ABSTRACT

The census can be adjusted using capture-recapture techniques: capture in the census, recapture in a special Post Enumeration Survey (PES) done after the census. The population is estimated using the Dual System Estimator (DSE). Estimates are made separately for demographic groups called post strata; adjustment factors are then applied to these demographic groups within small geographic areas. We offer a probability model for this process, in which several sources of error can be distinguished. In this model, correlation bias arises from behavioral differences between persons counted in the census and persons missed by the census. The first group may on the whole be more likely to respond to the PES: if so, the DSE will be systematically too low, and that is an example of correlation bias. Correlation bias is distinguished from heterogeneity, which occurs if the census has a higher capture rate in some geographic areas than others. Finally, ratio estimator bias and variance are considered. The objective is to clarify the probabilistic foundations of the DSE, and the definitions of certain terms widely used in discussing that estimator.

1. INTRODUCTION

The US census suffers from a small undercount, which is differential by race and ethnicity. There have been proposals to adjust the census using capture-recapture techniques. Capture is in the census, recapture is in a special sample survey—the Post Enumeration Survey, or PES—done after the census. Undercount rates are estimated for several hundred demographic groups called post strata: one post stratum might consist of non-Hispanic Asian male renters age 30–49, across the whole U.S. Then, small areas—blocks, tracts, cities, states—are adjusted synthetically, assuming that undercount rates are determined by demography not geography: in other words, undercount rates are taken to be constant within post strata across areas. For more details, see Hogan (1993), Freedman and Wachter (1994), or Brown et al. (1999). For institutional background, see Skerry (2000). The Supreme Court has ruled against census adjustment for apportioning Congress. However, adjustment is likely to to be used for redistricting, and allocation of tax funds. Government acronyms are in constant flux: the 1980 analog of the PES was PEP (Post Enumeration Program). For 2000, what was formerly ICM (Integrated Coverage Measurement) is now ACE—Accuracy and Coverage Evaluation Survey.

There are familiar probability models for capture-recapture in animal populations. Extensions to the census context present special complications, even at the conceptual level. For example, blocks are sampled rather than individuals. Furthermore, erroneous enumerations—people counted in the census in error—must be estimated, as well as gross omissions (people omitted from the census). Previous modeling efforts will be reviewed in Section 3 below, but these models do not seem to handle the complications: for instance, erroneous enumerations are not represented. Moreover, even for gross omissions, the models depend on rather tenuous assumptions about individual behavior.

Here, we propose a simple new model for the census adjustment process, which does handle some of the major complications. The stochastic element is generated by the controlled randomness in sampling blocks for the PES. No artificial behavioral assumptions are needed, and the block structure is built into the model from the beginning. Erroneous enumerations are represented, as well as gross omissions. This paper focuses on stating the model. However, a brief summary of empirical results for 1990 is given in Section 7.

Estimated undercount rates for post strata suffer from random error (sampling error when drawing blocks into the PES). This error can be quantified in fairly standard ways; see Section 6. Adjustment factors developed for each post stratum are nonlinear functions of the data, which leads to ratio estimator bias, also quantified in Section 6. Of more practical concern are the biases that arise from operational errors of one kind or another (for instance, respondents may give incorrect census-day address information). These errors will be set aside until Section 7.

There are people who are inaccessible, for one reason or another, both to the census and to the adjustment. These people create another important source of bias—difficult to measure directly—called “correlation bias” (Section 4). Finally, undercount rates cannot really be constant within post strata. This leads to an error called “heterogeneity,” not in the post stratum adjustment factors but in the adjusted counts for small areas (Section 5). Correlation bias, heterogeneity, and ratio estimator bias are sometimes hard to distinguish in the context of census adjustment. For instance, compare Ericksen, Fienberg, and Kadane (1994) with Freedman and Wachter (1994). Our model permits a clear separation of these errors.

There is a minimal unit of census geography called a “block.” (In built-up areas, this is typically a city block.) The census attempts to list all the “usual residents” of each block, as of census day. In each block, some number of persons are counted in error by the census; these are Erroneous Enumerations (EEs). A typical EE is a duplicate record; another typical example is a person whose address was erroneously assigned to the block. Let c_{ie} be the number of EEs in block i. Gross omissions will be discussed below.

We consider the problem of adjusting a geographical area comprising N blocks, indexed by $i = 1, \ldots, N$. After the census is taken, n out of N blocks are chosen for the PES. Denote the sample by S. We will assume the sample blocks are chosen at random without replacement. Of course, our model can also be extended to cover stratified block samples, or unequal sampling weights. For simplicity, we consider only one post stratum. Extending the model to have many post strata, multiple sampling strata, and variable weights would increase the realism, but the exposition would be cumbersome.

The PES makes an independent listing of the residents of the sample blocks, again as of census day. In each block, there are four kinds of residents (the true numbers are shown in parentheses):

- matches, found both by the census and by the PES (c_{im});
- found by the census but missed by the PES (c_{i1});
- missed by the census but found by the PES (c_{ig});
- missed both by the census and by the PES (c_{i0}).

As noted above, the census count for each block also includes

- erroneous enumerations (c_{ie}).

The census count for block i is

$$c_i = c_{ie} + c_{im} + c_{i1}. \tag{1}$$

The true number of residents of block i is

$$\begin{aligned} t_i &= c_{im} + c_{i1} + c_{ig} + c_{i0} \\ &= c_i - c_{ie} + c_{ig} + c_{i0}, \end{aligned} \tag{2}$$

with $c_{ig} + c_{i0}$ representing the number of persons missed by the census in block i. The census overcount in block i is c_{ie}; the undercount is $c_{ig} + c_{i0}$; the net undercount is $c_{ig} + c_{i0} - c_{ie}$. The PES cannot directly determine c_{i0}: these people are missed both by the census and by the PES. The total number of such people, $C_0 = \sum_i c_{i0}$, is necessarily somewhat uncertain, as discussed in Section 2.

Typically, undercounts exceed overcounts. For instance, in the US in 1980 and 1990, the net national undercount was in the range 1% to 2%. Some demographic groups and some geographical areas are more heavily impacted than others. That is why the undercount has become a policy issue. For more discussion of results and statistical questions, see Brown et al. (1999); on the policy questions, see Skerry (2000).

The census determines c_i, but not its components. To focus on the probabilistic issues, we will assume that the PES can accurately determine c_{ie}, c_{im}, c_{i1}, and c_{ig} in each sample block, by matching PES records against census records. Persons who move between the census and the PES create special problems for the fieldwork, assumed away by our model. Likewise, there are some persons counted in the census with insufficient information for matching; this too is ignored. Thus, operational difficulties are not considered here. (Any precise description of the PES fieldwork, the matching, and the potential sources of error would cover many hundreds of pages; there is by now a huge literature on evaluations: Section 7 has a brief review, and citations.)

2. VISUALIZING THE MODEL

To visualize our model, imagine a PES taken in every block in the country, just as the census is taken in every block. Unlike the census and the real PES, this hypothetical PES is assumed free of error—although it may miss some people. Thus, c_{ie}, c_{ig}, and c_{i1} are determined. Almost by definition, the number of people c_{i0} missed both by the census and by the PES remains unobserved. The country-wide PES results are hidden, but we choose our random sample of blocks and uncover the PES data in the sample blocks.

Capture-recapture methods cannot provide any scientific estimates of population size unless one of the samples is generated by some truly random mechanism. With our framework, randomness enters through the sampling. There is no need to treat capture in the census or the PES as stochastic processes with assumed random behavior by persons. No imaginary ensemble of replicate censuses is needed. The country-wide PES is, of course, a construct: given that construct, the model is straightforward. Such constructs are frequently used to analyze samples and experiments; see, for instance, Freedman, Pisani, and Purves (1998, chap. 27), which has further citations to the literature.

The capture-recapture method is often summarized in a 2×2 table, like Table 1; in this table, the erroneous enumerations are set aside. There are three observed cells, corresponding to

(i) matches, that is, people captured in the census and in the PES;
(ii) people captured only in the PES;
(iii) people captured only in the census.

The counts in cells (i)–(iii) in block i are c_{im}, c_{ig}, and c_{i1}, respectively; Table 1 shows the totals.

TABLE 1. The 2 × 2 table for capture-recapture. Totals. Erroneous enumerations are set aside.

		Census	
		In	Out
PES	In	C_m	C_g
	Out	C_1	??

In the real world, the table entries are just four numbers, not necessarily related to each other. The three observed cells need not tell us anything about the fourth cell—the people missed by both systems. The people in the different cells are different, and have behaved differently in response to the census and the PES. Any attempt to use the three observed cells to estimate the fourth, unobserved cell must depend on information not found in the census and the PES.

As explained in Section 4, the DSE estimates the number of people in the fourth cell by proportionality:

$$C_1 C_g / C_m.$$

However, "estimation" is being used in its dry sense, without ordinary-language connotations. Thus, it is technically correct to say that the population of Idaho can be used to estimate the population of Iowa. Indeed, this estimator will have little random error, although bias is a problem. Such estimators do not live up to the promise inherent in the terminology. Likewise, the DSE may not be estimating the fourth cell in any satisfactory way—because correlation bias may be substantial.

The proportionality built into the DSE is called the "Independence Assumption:" somewhat informally, capture in the census and recapture in the PES are assumed independent of each other. Proportionality excludes correlation bias. When correlation bias is positive, the true 2 × 2 table is the sum of a table satisfying the "independence" or "zero correlation bias" assumption, plus a table with zeros in the first three cells and some number of additional people in the fourth cell. The additional group may be termed the "fifth cell." People who do not wish to be found by any government agency would go into the fifth cell. In fisheries applications, the fifth cell might be populated by wily trout.

Let $1 - \pi$ and π be the proportions of the true population accounted for by the independence model and by the fifth cell, respectively. Rudas, Clogg, and Lindsay (1994) define such a π-index for general contingency tables. Naturally, the additive representation is mathematically equivalent to many other models. Like π, the structural parameters in those other models—which allow the fourth cell to be determined from the three observed cells—are not identifiable from the data. Independence is an example of an identifying restriction that determines the fourth cell. Some representations make it easier than others to forget the identifiability problem. The additive representation focuses on the critical issue of identifiability.

3. MODELS FOR INDIVIDUAL BEHAVIOR

The typical capture-recapture model has independent individuals. These individuals have independent responses to two surveys, so an individual can be characterized by two response probabilities. In this literature, "heterogeneity" often means that response probabilities vary across individuals, and the "correlation" in correlation bias then refers to correlated probabilities. Cormack (1968, 1999) reviews capture-recapture models for estimating wildlife populations, and Cormack (1966) considers models with unequal probabilities of recapture. For other reviews, see Alho (1994), Hook and Regal (1995), and Pollock (2000).

To develop a capture-recapture model for census adjustment at the level of individual responses—rather than block totals—requires a model for individual behavior. Each person has a response status in the census: $X = 1$ if the person responds, $X = 0$ if not. Likewise, each person either will respond ($Y = 1$) or will not respond ($Y = 0$) to the PES, and Y is observable if the PES samples that person's block of residence. From this perspective, our treatment takes the X's and Y's as deterministic.

Stochastic models for X and Y have been developed, but seem conjectural. For example, Wolter (1986) proposes a model assuming independence across individuals: that is the "Multinomial Assumption" and "Autonomous Independence" (p. 339). His section 3.1 contemplates, as we do, a block sample. However, our treatment of correlation bias reflects the decision to treat responses as deterministic not random. Compare equation (2.2) in Wolter (1986). Moreover, Wolter does not make our distinction between correlation bias and heterogeneity; nor are EEs explicitly represented in his model.

Darroch et al. (1993) propose a Rasch model for X and Y, with a latent variable representing a person's capturability; also see Agresti (1994). Erroneous enumerations are not contemplated. Given the latent variable, capture in the PES and in the census are assumed to be independent, as well as capture in a hypothetical third survey. The model also assumes independence across individuals. For the PES as for the census, only one person responds in each household—on behalf of all members of the household. Thus, recapture by the PES, like capture by the census, must be strongly correlated across members of a household; the independence assumption in Darroch et al. is therefore questionable, as is the corresponding assumption in Wolter (1986). Our model makes no independence assumptions, although it may be somewhat stylized in other respects.

Kadane, Meyer, and Tukey (1999) have an individual-level model with no geography and no EEs; they argue that under certain (unverifiable) conditions, correlation bias will be positive. Independence of capture and recapture is often said to be a critical assumption behind the DSE, with positive correlation bias arising from dependence. See, for instance, Citro and Cohen (1985, pp. 140, 168, 343–45), Steffey and Bradburn (1994, p. 109), or Ericksen et al. (1994). This idea is hard to formalize, although the attempts by Wolter (1986) and Darroch et al. (1993) are interesting.

The sources of undercounts and overcounts in the census seem to be too complex to be represented in any convincing way by a probability model. Thus, we prefer to treat the census as a given data set. The focus then shifts to variations in the probability of selection for the PES. Selection probabilities for the sampling units—the blocks—are surely under good control. It is the responses of individuals that are at issue. Again, these are hard to model, so we treat them too as data. In our setup, the only source of randomness is the sampling of blocks for the PES. A difference in average PES response rates between census hits and census misses leads to correlation bias, while differences in census response rates among geographical areas create heterogeneity.

4. THE DSE AND CORRELATION BIAS

We use capital letters for sums. For example, the total census count is

$$C = \sum_{i=1}^{N} c_i = C_e + C_m + C_1, \tag{3}$$

where $C_e = \sum_{i=1}^{N} c_{ie}$, and so forth. The true population is

$$T = \sum_{i=1}^{N} t_i = C_m + C_1 + C_g + C_0. \tag{4}$$

The Dual System Estimator (DSE) for T is

$$\hat{T} = C\hat{Y}/\hat{X}. \tag{5}$$

Here, C is the census count defined in (3) and $C/\hat{X}$ is the usual capture-recapture estimator. The factor $\hat{Y}$ takes care of the EEs.

The formulas for $\hat{X}$ and $\hat{Y}$ involve sums over blocks i in the PES sample S. Specifically,

$$\hat{X} = \frac{\sum_{i \in S} w c_{im}}{\sum_{i \in S} w(c_{im} + c_{ig})} \tag{6}$$

and

$$\hat{Y} = \frac{\sum_{i \in S} w(c_{im} + c_{i1})}{\sum_{i \in S} w(c_{ie} + c_{im} + c_{i1})}, \tag{7}$$

where $w = N/n$ weights the sample to the population. Intuitively, $\hat{X}$ is the match rate, that is, the estimated rate at which people are found both by the census and by the PES; the numerator is the upweighted sample count of matches, and the denominator is the upweighted sample count of PES records. Similarly, $\hat{Y}$ is the estimated rate of correct enumerations in the census. The numerator is the upweighted sample count of census correct enumerations, and the denominator is the upweighted sample count of census records.

The expected values for the numerator and denominator of $\hat{X}$ are C_m and $C_m + C_g$, respectively. Let the actual numerator and denominator be $1 + \zeta_2$ and $1 + \zeta_1$ times their respective expectations. Thus

$$\hat{X} = \frac{C_m}{C_m + C_g} \frac{1 + \zeta_2}{1 + \zeta_1}. \tag{8}$$

The expected values for the numerator and denominator of $\hat{Y}$ are $C_m + C_1$ and $C = C_e + C_m + C_1$, respectively. Let the actual numerator and denominator be $1 + \eta_2$ and $1 + \eta_1$ times their respective expectations. Thus

$$\hat{Y} = \frac{C_m + C_1}{C_e + C_m + C_1} \frac{1 + \eta_2}{1 + \eta_1}. \tag{9}$$

The random variables ζ_1, ζ_2 in (8) and η_1, η_2 in (9) have mean zero; the randomness is due to sampling variability. By (3),

$$\begin{aligned} \hat{T} &= C \frac{C_m + C_1}{C_e + C_m + C_1} \frac{C_m + C_g}{C_m} \frac{1 + \eta_2}{1 + \eta_1} \frac{1 + \zeta_1}{1 + \zeta_2} \\ &= \frac{(C_m + C_1)(C_m + C_g)}{C_m} \frac{1 + \eta_2}{1 + \eta_1} \frac{1 + \zeta_1}{1 + \zeta_2} \\ &= T^* \frac{1 + \eta_2}{1 + \eta_1} \frac{1 + \zeta_1}{1 + \zeta_2}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} T^* &= (C_m + C_1)(C_m + C_g)/C_m \\ &= C_m + C_1 + C_g + (C_1 C_g / C_m) \\ &= T + (C_1 C_g / C_m) - C_0 \end{aligned} \tag{11}$$

by (4).

In effect, the DSE estimates C_0 as $C_1 C_g / C_m$. The error here is bias. “Estimation” is used in its narrowest technical sense: C_0 is not identifiable from census and PES data without further assumptions, since the people who comprise C_0 are not observed in either system. This has been discussed above, in Section 2.

Definition. Correlation bias is $C_0 - C_1 C_g / C_m$.

The sign is chosen for historical reasons. If correlation bias is positive, C_0 is underestimated, and the DSE will be too low.

Heuristic arguments are often given to suggest that

$$\frac{C_0}{C_g + C_0} > \frac{C_1}{C_m + C_1}. \tag{12}$$

This inequality is equivalent to positive correlation bias. The idea behind (12) is the following. Set aside the EEs. Then $C_g + C_0$ is the number of persons missed by the census, while $C_m + C_1$ is the number found. Moreover, C_0 and C_1 represent PES misses in these two subpopulations (census misses and census hits). What (12) says, then, is that the PES is more likely to miss people among census misses than among census hits. If the inequality in (12) is reversed, correlation bias is negative. There is some evidence for negative correlation bias in 1980: see Fay et al. (1988). If equality holds in (12), there is no correlation bias. Since C_0 is not identifiable, neither is correlation bias.

5. HETEROGENEITY

Census adjustment is done at the block level: the census count for block i is multiplied by the adjustment factor $\hat{\phi} = \hat{T}/C$, where $\hat{T}$ is the DSE defined by (5); thus, $\hat{\phi} = \hat{Y}/\hat{X}$. In other words, to adjust blocks, undercount rates are assumed to be constant (within demographic subgroups) across wide geographic areas. Failures of this constancy assumption are termed "heterogeneity," although "geographical heterogeneity" might be more appropriate.

Definition. Heterogeneity refers to variation in the ratio c_i/t_i of census counts to true counts across geographical areas, indexed by i.

In our model, the difference between correlation bias and heterogeneity can be stated quite sharply.

- Correlation bias refers to a behavioral difference between people missed by the census and people counted in the census: $C_0/(C_g + C_0) \neq C_1/(C_m + C_1)$. Census misses differ from census hits, with respect to the likelihood of responding to the PES.
- Heterogeneity refers to a behavioral difference between residents of different areas: the coverage ratio c_i/t_i varies within demography across geography.

In particular, heterogeneity can be defined by reference to the census and the truth: the PES is not involved. Easy examples can be given to demonstrate correlation bias without heterogeneity, or heterogeneity without correlation bias.

A more complete model would have several post strata, indexed by j. There would be an adjustment factor $\hat{\phi}_j$ for each post stratum, computed as above from PES data on members of the post stratum resident anywhere in the US. Let $c_{i(j)}$ be the census number of residents of block i who are members of post stratum j. This block × post stratum fragment would be adjusted by the factor $\hat{\phi}_j$. The adjusted population in block i would be $\sum_j \hat{\phi}_j c_{i(j)}$, and the adjusted population in any larger area would be obtained by summing over the blocks comprising that area. In practice, block-level counts would be rounded, to avoid tables with fractional entries.

Subsampling within blocks may also be contemplated: this would help make contact with the capture-recapture literature for estimating wildlife populations, where the "block" typically consists of an individual animal. In that literature, "heterogeneity" often refers to variation in the probability of recapture (section 3); then, heterogeneity is a feature of the post enumeration survey rather than the census. Formally, individuals could be classified as responders or non-responders to a post enumeration survey; the usual textbook model take all individuals to be responders. When blocks reduce to individuals, our concept of geographic heterogeneity is unlikely to be very useful.

6. RATIO ESTIMATOR BIAS AND VARIANCE

This section is more narrowly technical, focusing on the sampling distribution of the DSE. In view of (10), it suffices to consider

$$\rho = \frac{1+\eta_2}{1+\eta_1}\,\frac{1+\zeta_1}{1+\zeta_2}. \tag{13}$$

By construction, $E(\eta_1) = E(\eta_2) = E(\zeta_1) = E(\zeta_2) = 0$. However, $E(\rho) \neq 1$, because of the nonlinearity.

Definition. Ratio estimator bias is $T^*[E(\rho) - 1]$, where T^* was defined by (11).

Thus, ratio estimator bias refers to the difference between the expected value of the DSE and T^*, while correlation bias refers to the difference between T^* and T, the true population.

Ratio estimator bias and variance can be traced back to variation, not in c_i/t_i but in the components

$$c_{ie}/(c_{im} + c_{i1}) \text{ and } c_{ig}/c_{im}.$$

Indeed, if

$$\frac{c_{im} + c_{i1}}{c_{ie} + c_{im} + c_{i1}} = \frac{C_m + C_1}{C_e + C_m + C_1}$$

and

$$\frac{c_{im}}{c_{im} + c_{ig}} = \frac{C_m}{C_m + C_g}$$

for all i, there is no ratio estimator bias or variance, since $\rho = 1$ for all samples: see (6) and (7).

The balance of this section, which discusses the asymptotics, is mainly expository. We proceed a bit heuristically, using the delta method and writing $\doteq$ for approximate equality. Rigor can be obtained by letting n and N go to infinity, but this seems out of place here. If we keep only linear and quadratic terms,

$$\rho \doteq 1 + \lambda + \kappa,$$

where

$$\lambda = (\zeta_1 - \zeta_2) - (\eta_1 - \eta_2)$$

and

$$\kappa = \zeta_2^2 + \eta_1^2 - \zeta_1\zeta_2 - \eta_1\eta_2 - (\zeta_1 - \zeta_2)(\eta_1 - \eta_2).$$

The asymptotic bias of ρ is $E(\kappa)$ and the asymptotic variance is var λ.

Of course, $E(\kappa)$ and var λ can be computed in terms of second moments of population quantities c_{ie}, c_{im}, c_{i1}, c_{ig}; then means and variances can be estimated from sample quantities. Suppose u_i and v_i are real numbers for each i, with $\sum_{i=1}^N u_i = \sum_{i=1}^N v_i = 0$. Recall that S consists of n indices chosen at random from $\{1, \ldots, N\}$. Let

$$U = \sum_{i \in S} u_i \text{ and } V = \sum_{i \in S} v_i.$$

Then $E(U) = E(V) = 0$ and

$$\text{cov}(U, V) = n \, \frac{N - n}{N - 1} \, \frac{1}{N} \sum_{i=1}^N u_i v_i.$$

To evaluate cov(η_1, η_2), say, choose

$$u_i = \frac{N}{n} \Big(\frac{c_{ie} + c_{im} + c_{i1}}{C_e + C_m + C_1} - \frac{1}{N} \Big)$$

and

$$v_i = \frac{N}{n}\left(\frac{c_{im}+c_{i1}}{C_m+C_1} - \frac{1}{N}\right),$$

so $\eta_1 = \sum_{i\in S} u_i$ and $\eta_2 = \sum_{i\in S} v_i$. Let

$$d = n\,\frac{N-n}{N-1}.$$

Then

$$\begin{aligned}\operatorname{cov}(\eta_1,\eta_2) &= d\,\frac{1}{N}\sum_{i=1}^{N} u_i v_i \\ &= d\Big(\frac{N}{n}\Big)^2\left[\frac{1}{N}\sum_{i=1}^{N}\frac{c_{ie}+c_{im}+c_{i1}}{C_e+C_m+C_1}\,\frac{c_{im}+c_{i1}}{C_m+C_1} - \frac{1}{N^2}\right]\end{aligned}$$

which can be estimated as

$$d\Big(\frac{N}{n}\Big)^2\left[\frac{1}{n}\sum_{i\in S}\frac{c_{ie}+c_{im}+c_{i1}}{\hat{C}_e+\hat{C}_m+\hat{C}_1}\,\frac{c_{im}+c_{i1}}{\hat{C}_m+\hat{C}_1} - \frac{1}{N^2}\right],$$

with $\hat{C}_e = \sum_{i\in S} wc_{ie}$, and so forth. Other terms may be handled the same way, and we omit further details. For more information on ratio estimators like the DSE, see Kish (1965, pp. 206–210).

7. EMPIRICAL RESULTS FOR 1990

The US census of 1990 counted 248.7 million persons, and the production DSE would have added 5.3 million. Ratio estimator bias and variance are relatively easy to quantify, using just the data from the PES, and were not a major problem in 1990—at least for large population aggregates. For the national net undercount rate of 2.1%, ratio estimator bias was on the order of a tenth of a percentage point, while sampling error was on the order of two tenths of a point: see Mulry and Spencer (1993). For many individual post strata or states, however, sampling error was a concern: see Freedman et al. (1993, 1994).

Heterogeneity is more difficult to measure. In principle, variation in undercount rates within post strata can be measured across geography, but allowance must be made for sampling variation; thus, large samples are needed to carry out the program. For some results, see Alho et al. (1993) or Hengartner and Speed (1993). It may be more practical to use “proxy variables” whose behavior mimics undercount rates; Kim (1991), Freedman and Wachter (1994). The evidence shows that heterogeneity is a large-scale problem, which significantly complicates the tasks of adjusting the census or comparing the accuracy of adjusted and unadjusted counts. This is so even within post strata, and for geographical areas like counties or states; Fay and Thompson (1993), Freedman and Wachter (1994).

Evaluation Followup data can be used to measure error rates in the PES arising from sources like bad census-day address information. As noted above, the production DSE would have added

5.3 million persons to the census count. Of this total, 3.0 to 4.2 million persons would have been added due to errors in the PES rather than errors in the census, with a central estimate of 3.6 million. See Mulry and Spencer (1993), Breiman (1994), or Wachter and Freedman (2000).

Correlation bias can be determined at the national level, by combining results from the PES, the Evaluation Followup, and an independent estimate of the population. Perhaps the most solid such estimate is provided by Demographic Analysis, which is keyed to administrative records. DA suggests adding 4.7 million persons to the 1990 census count. Thus, correlation bias can be estimated as a residual: $4.7 - (5.3 - 3.6) = 3.0$ million.

The object of the PES is to put back into the census some 5 million people who were left out. But these people have to be allocated to the right places—states, cities, counties, congressional districts. Otherwise, adjustment could make the census worse, not better. If the PES itself stands in need of correction, first subtracting 3.6 million people then adding back 3.0 million different people, its reliability for adjusting the census seems questionable. The people who must be added back to the PES do not resemble those who must be taken out, in terms of race, sex, and likely place of residence; Wachter and Freedman (2000). For other perspectives, see Belin and Rolph (1994) or Cohen, White, and Rust (1999).

ACKNOWLEDGMENTS

We thank Richard Cormack (St. Andrews) for many helpful comments, and the William H. Donner Foundation for its financial support. The work was performed while PBS was on appointment as Miller Research Professor in the Miller Institute for Basic Research in Science. KWW and DAF have served as consultants to the Freshpond Institute and as expert witnesses for the government in litigation over proposed adjustments of the 1980 and 1990 censuses.

REFERENCES

Agresti, A. (1994) Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 50: 494–500.

Alho, J. M. (1994) Analysis of sample based capture-recapture experiments. *Journal of Official Statistics* 10: 245–56.

Alho, J. M., Mulry, M. H., Wurdeman, K., and Kim, J. (1993) Estimating heterogeneity in the probabilities of enumeration for Dual System Estimation. *Journal of the American Statistical Association* 88: 1130–36.

Belin, T. R., and Rolph, J. E. (1994) Can We Reach Consensus on Census Adjustment? *Statistical Science* 9: 486–508 (with discussion).

Breiman, L. (1994) The 1991 census adjustment: undercount or bad data? *Statistical Science* 9: 458–537.

Brown, L. D., Eaton, M. L., Freedman, D. A., Klein, S. P., Olshen, R. A., Wachter, K. W., Wells, M. T., and Ylvisaker, D. (1999) Statistical Controversies in Census 2000. *Jurimetrics* 39: 347–75.

Citro, C. F., and Cohen, M. L., eds. (1985) *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D. C.: National Academy Press.

Cohen, M. L., White, A. A., and Rust, K. F., eds. (1999) *Measuring a Changing Nation: Modern Methods for the 2000 Census*. Washington, D.C.: National Academy Press.

Cormack, R. M. (1966) A test for equal catchability. *Biometrics* 22: 330–342.

Cormack, R. M. (1968) The statistics of capture-recapture methods. *Oceanogr. Mar. Biol. Ann. Rev.* 6: 455–506.

Cormack, R. M. (1999) Population size estimation and capture recapture methods. Technical Report, Division of Statistics, St. Andrews University.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* 88: 1137–1148.

Ericksen, E. P., Fienberg, S. P., and Kadane, J. B. (1994) Comment. *Statistical Science* 9: 511–515.

Fay, R. E., Passel, J. S., Robinson, J. G., and Cowan, C. D. (1988) *The Coverage of the Population in the 1980 Census*. Washington, D. C.: U. S. Government Printing Office.

Fay, R. E., and Thompson, J. H. (1993) The 1990 Post Enumeration Survey: statistical lessons, in hindsight. *Proceedings, Bureau of the Census Annual Research Conference*. Washington, D. C.: Bureau of the Census.

Freedman, D. A., Pisani, R., and Purves, R. (1997) *Statistics* (Third Edition). New York: Norton.

Freedman, D. A., Wachter, K., Coster, D., Cutler, R., and Klein, S. (1993) Adjusting the census of 1990: The smoothing model. *Evaluation Review* 17: 371–443.

Freedman, D., Wachter, K., Cutler, R., and Klein, S. (1994) Adjusting the census of 1990: Loss functions. *Evaluation Review* 18: 243–280.

Freedman, D., and Wachter, K. (1994) Heterogeneity and census adjustment for the intercensal base. *Statistical Science* 9: 476–485.

Hengartner, N. and Speed, T. P. (1993) Assessing between-block heterogeneity within poststrata of the 1990 Post-Enumeration Survey. *Journal of the American Statistical Association* 88: 1119–29 (with discussion).

Hogan, H. (1993) The 1990 Post-Enumeration Survey: operations and results. *Journal of the American Statistical Association* 88: 1047–1060.

Hook, E. B., and Regal, R. R. (1995) Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 17: 243–64.

Kadane, J. B., Meyer, M. M., and Tukey, J. W. (1999) Yule's association paradox and ignored stratum heterogeneity in capture-recapture studies. *Journal of the American Statistical Association* 94: 855–9.

Kim, J. (1991) 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. Washington, D.C.: Bureau of the Census.

Kish, L. (1965) *Survey Sampling*. New York: John Wiley and Sons.

Mulry, M., and Spencer, B. (1993) Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association* 88: 1080–91.

Pollock, K. H. (2000) Capture-recapture models. *Journal of the American Statistical Association* 95: 293–96.

Rudas, T., Clogg, C. G., and Lindsay, B. G. (1994) A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B* 56: 623–39.

Skerry, P. (2000) *Counting on the Census*. Washington, D. C.: Brookings.

Steffey, D. L., and Bradburn, N. M., eds. (1994) *Counting People in the Information Age*. Washington, D. C.: National Academy Press.

Wachter, K. and Freedman, D. A. (2000) The fifth cell: correlation bias in U. S. census adjustment. *Evaluation Review* 24: 191–211.

Wolter, K. (1986) Some coverage error models for census data. *Journal of the American Statistical Association* 81: 338–346.