Heterogeneity and Census Adjustment for the Intercensal Base

D. Freedman and K. Wachter

Abstract. Current techniques for census adjustment involve the "synthetic assumption" that undercount rates are constant within "poststrata" across geographical areas. A poststratum is a subgroup of people with given demographic characteristics; poststrata are chosen to minimize heterogeneity in undercount rates. This paper will use 1990 census data to assess the synthetic assumption. We find that heterogeneity within poststrata is quite large, with a corresponding impact on local undercount rates estimated by the synthetic method. Thus, any comparison of error rates between the census and adjusted counts should take heterogeneity into account.

Key words and phrases: Census adjustment, synthetic method, loss function analysis, small area estimation.

1. INTRODUCTION

The decennial census is used to apportion Congress and to allocate tax moneys. Over the course of a decade, intercensal (more correctly, "postcensal") population estimates are developed for states and other geographical areas. These estimates are used to allocate tax moneys in noncensus years. In effect, census figures are rolled forward using data on births, deaths and migration patterns.

The census is generally thought to suffer from a small undercount, which is differential by race, ethnicity and area. The Census Bureau proposed to adjust the 1990 census to correct the undercount. They divided the population into "poststrata," which are relatively broad demographic subgroups thought to be more or less homogeneous with respect to undercount rates. Adjustment factors for the various poststrata were estimated from a special sample survey done after the census, called the Post Enumeration Survey, or PES. For more information and references, see Fay (1992a), Hogan (1993), Freedman (1991), Freedman et al. (1993) or Freedman, Wachter, Cutler and Klein (1994).

The Census Bureau's proposal to adjust the census by these methods was rejected by the parent agency (Department of Commerce, 1991b). Subsequently, the bureau considered a somewhat different procedure for adjusting the census as a base for the intercensal estimates. Among other things, the poststratification was changed (Bureau of the Census, 1992a, b).

To adjust the census and to compare the errors associated with adjusted and unadjusted counts, the Census Bureau makes the "synthetic assumption" that undercount rates are constant within poststrata across geographical areas. This paper will use 1990 census data to assess the synthetic assumption, focusing on the bureau's revised poststratification for the intercensal base. We find that heterogeneity within poststrata is quite large.

Any comparison of error rates in the raw and adjusted census counts should take into account the error due to heterogeneity. Otherwise, the comparison may be quite biased against the census. The bureau's "loss function analysis," which attempts to provide unbiased estimates of risk, appears to suffer from this problem and is unconvincing for that reason among others (Bureau of the Census, 1992b; Freedman, Wachter, Cutler and Klein, 1994; Wachter and Speed, 1991). The impact of heterogeneity on risk estimates is discussed in section 5.

Before proceeding, we outline the "synthetic method" used to adjust state counts, indicating where the synthetic assumption comes into play. California is the lead example. On census day (April 1, 1990), members of 147 different poststrata were resident in that state.

D. Freedman is Professor of Statistics, and K. Wachter is Professor of Demography and Statistics, Department of Statistics, University of California, Berkeley, California 94720. *Freedman and Wachter testified for the Department of Justice in City of New York et al. v. Department of Commerce et al., a lawsuit in which New York sought to compel census adjustment.*

One poststratum, code #301, consisted of blacks age 0–17 living in owner-occupied housing in rural areas anywhere in the United States. The census counted a nationwide total of 764,400 such persons. The "adjustment factor" for poststratum #301 was estimated from PES data as 1.058. In other words, the nationwide population in poststratum #301 as of census day was estimated as

$$1.058 \times 764{,}400 = 808{,}735,$$

rather than the census count of 764,400.

The adjustment factor of 1.058 was estimated using PES data from all across the country, California to Maine. However, to adjust California by the synthetic method, the multiplier of 1.058 is applied to the 4,260 members of poststratum #301 who were resident in California on April 1, 1990, according to the census. In other words, the number of blacks age 0–17 living in owner-occupied housing in rural areas in California on April 1, 1990, is estimated as

$$1.058 \times 4260 = 4{,}507,$$

rather than the census count of 4,260.

A similar procedure is applied to each of the 147 poststrata with members resident in California, and the adjusted population of California is obtained by summing the 147 products. (The PES sampling frame covered about 98% of the population; the "residual population" is not covered by the counts above and has to be added in separately.)

Each poststratum has its own adjustment factor, estimated from PES data covering many states. However, this factor is applied to the members of the poststratum resident in California. That is where the synthetic assumption enters: undercount rates are assumed to be constant within poststrata across states. Our object is to assess the degree to which the synthetic assumption holds.

For previous work on synthetic adjustment, with a review of the literature, see Wachter and Freedman (1994). For discussion by Census Bureau staff, see Fay and Thompson (1993) or Kim, Blodgett and Zaslavsky (1993). For a discussion of adjustment from other perspectives, see Breiman (1994), Choldin (1994), Citro and Cohen (1985), Ericksen and Kadane (1985), Ericksen, Kadane and Tukey (1989), Ericksen, Estrada, Tukey and Wolter (1991), Fay (1992a), Freedman (1991), Freedman and Navidi (1986, 1992), Hengartner and Speed (1993), Hogan (1993), Hogan and Wolter (1988), Mitroff, Mason and Barabba (1983), Schenker (1993), Schirm (1991), Schirm and Preston (1987), Singh (1992), Steffey (1993), Steffey and Bradburn (1994), Wolter (1986a, 1991), Wolter and Causey (1991) and Ylvisaker (1991). Choldin (1994) reviews work on the 1990 census, and is generally sympathetic to adjustment. On the year 2000, see Steffey (1993) or Steffey and Bradburn (1994), who support current plans for integrating the census, coverage measurement, and adjustment; components would not be made available for separate analysis: hence the description as a "one-number census."

TABLE 1
Proxy variables considered by the Bureau of the Census (1992b)

Substitution rate: the percentage of persons in each group whose entire census records were imputed. (For about 1% of the population, the census determines the number of persons in a household, e.g., by interviewing a neighbor, but has no personal information about the occupants. Then occupants are chosen at random from nearby houses, and their personal characteristics are imputed to the occupants of the target household.)
Allocation rate: the percentage of persons in each group with at least one out of six critical items in the census record "allocated," that is, imputed.
Multiunit housing rate: the percentage of persons in each group who were living in multiunit housing (such as apartment buildings).
Nonmailback rate: the percentage of persons in each group who failed to mail back their census questionnaire. (The denominator is the number to whom forms were mailed.)
Mobility rate: the percentage of persons in each group who were living at a different address five years ago.
Poverty rate: the percentage of persons in each group whose incomes were below the poverty line.

2. RESULTS

Proxy Variables

Of course, direct information on variability of undercount rates within poststrata is hard to come by. Although the PES sampled roughly 400,000 persons, there are 357 poststrata in the scheme for adjusting the postcensal base; the average sample size is about 1,000 persons per poststratum. Crossing the sample with the 50 states produces nugatory sample sizes. To investigate the synthetic assumption, the Census Bureau turned to proxy variables, that is, variables thought to resemble undercount rates in terms of heterogeneity. A "variable" is a numerical characteristic of population groups, with values determined from census data. The relevant groups are the poststrata, the states and the poststratum × state cells. We will consider the six proxies shown in Table 1. The first four are based on census short-form data. The last two come from the census long-form sample. (The denominator for long-form variables consists of population sizes estimated from the long-form sample.)

There is no sampling variability in the first four proxies, and essentially none in the last two, since the census long-form sample is so large. Of course, there is sampling variability in estimated under-

TABLE 2
Age–sex groups defined by the Bureau of the Census (1992b) for adjusting the 1990 census base; intercensal estimates

	Male	Female
0–17	1	
18–29	2	4
30–49	3	5
50+	6	7

count rates, since these are estimated from a sample survey.

The undercount rates are what everybody cares about, but the PES is too thin on the ground to make any very strong case about homogeneity in such rates. That is why the Bureau turned to the proxies. The great advantage of the proxies is complete (or virtually complete) data. The drawback is that the proxies may not behave like undercount rates, in terms of heterogeneity.

Poststrata

The poststrata considered in Bureau of the Census (1992b) for adjusting the intercensals are shown in Tables 2 and 3. Basically, there are 7 age-sex groups shown in Table 2, crossed with the 51 "poststrata groups" (PSG's) shown in Table 3. Thus, there are $51 \times 7 = 357$ poststrata. These are identified by two- or three-digit code numbers. Poststratum #11 is poststratum group 1 crossed with age-group 1; poststratum #517 is PSG 51 crossed with age-group 7. These poststrata are also referred to by sequence number: code #11 is the first in sequence, and code #517 is 357th in sequence.

Age-sex group 1 (top of Table 2) consists of males and females ages 0–17. Age-sex group 7 (bottom right) consists of females age 50 and over.

Poststratum group 51 (at the bottom of Table 3) consists of American Indians on reservations anywhere in the United States; PSG 30 (in the middle of the table) consists of blacks living in owner-occupied housing in rural areas anywhere in the United States; PSG 1 (at the top left) consists of non-Hispanic whites living in owner-occupied housing in large urban areas in the Northeast region of the United States. (The Northeast census region consists of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey and Pennsylvania.)

Basically, the population is split by race and ethnicity (non-Hispanic white, Hispanic, black, Asian and Pacific Islander, American Indian; the first category contains any residual). There is another split on home ownership, and another on place type (big cities, other urban, rural); there is a final split on census region. However, some of the cells in the resulting cross-tabulation are collapsed, as shown in Table 3. The geographic scope of the poststrata is worth attention.The poststrata for whites are regional. However, most of the minority poststrata are national; for instance, small urban areas are combined from coast to coast.

The Data Set

To describe the data, consider a 357×51 matrix, with rows for poststrata and columns for the states (and Washington, D.C.). There is one matrix showing population counts; the (1, 1) entry gives the number of persons in the first poststratum (code #11) living in Alabama; the (357, 51) entry gives the number of persons in the 357th poststratum (code #517) living in Wyoming. (It is a numerical coincidence that there are 51 areas and 51 PSG's.)

Likewise, there is a matrix for each proxy variable, showing counts by poststratum and state. The (1, 1) entry in the "substitution" matrix gives the number of substituted persons in the first poststratum (code #11) in Alabama, and so forth. There is a matrix showing counts of persons in the mail universe (to whom census forms were mailed), used as denominators for nonmailback rates. Finally, there is a matrix showing the population counts as estimated from the census long-form sample, which are used as denominators for poverty and mobility rates. Counts are based on the PES target population (e.g., the institutional population is excluded, as are persons in the wilds of Alaska). These data were kindly provided by the Census Bureau.

Descriptive Statistics

We begin with some descriptive statistics. Table 4 summarizes data on the undercount, as estimated from the PES corrected for certain biases (Bureau of the Census, 1992b). The overall undercount rate is 1.580%. The root mean square standard error (SE) of state undercount rates is 0.466%; this measures sampling error. The SD of undercount rates across states (corrected for sampling error) is about 0.653%; this measures state-to-state differences in undercount rates.

Table 5 gives descriptive statistics for the proxies. Column 1 shows levels, that is, national rates. Column 2 shows the SD of the rates for the 51 areas (states and Washington D.C.). Thus, 0.899% of census forms were substitutions; the SD of the state rates was 0.408%. The third column shows the SD of the rates for the 357 poststrata. The last column shows the root mean square (r.m.s.) over the 357 poststrata of the within-poststratum-across-state SD. For example, take the first poststra-

TABLE 3
Poststrata groups (PSG's) defined by the Bureau of the Census (1992b) for adjusting the 1990 census base; intercensal estimates

	Northeast	South	Midwest	West
Non-Hispanic white				
Owner				
Urbanized areas 250 K+	1	2	3	4
Other urban	5	6	7	8
Nonurban	9	10	11	12
Nonowner				
Urbanized areas 250 K+	13	14	15	16
Other urban	17	18	19	20
Nonurban	21	22	23	24
Black				
Owner				
Urbanized areas 250 K+	25	26	27	28
Other urban		29		
Nonurban		30		
NonOwner				
Urbanized areas 250 K+	31	32	33	34
Other urban		35		
Nonurban		36		
Nonblack hispanic				
Owner				
Urbanized areas 250 K+	37	38	39	40
Other urban		41		
Nonurban		42		
Nonowner				
Urbanized areas 250 K+	43	44	45	46
Other urban		47		
Nonurban		48		
Asian and Pacific Islander				
Owner		49		
Nonowner		50		
American Indians on reservations		51		

TABLE 4
Summary statistics on undercount rates, in percent, corrected for certain biases in the PES

National rate	1.580%
Root Mean Square SE of rates for 50 states and Washington D.C.	0.466%
SD of rates across 50 states and Washington, D.C.	0.653%

tum, code #11. That poststratum had members resident in six states. In those states, the substitution rates were

0.0056 0.0043 0.0059 0.0056 0.0038 0.0087.

The SD of the six rates is 0.0017. Similarly for the remaining 356 poststrata. The r.m.s. of the 357 SD's is 0.00857 = 0.857%.

Table 5 does contain some good news for the pro-adjustment side: poststrata absorb more variability than do states (column 3 is bigger than column 2). There is also bad news: within-poststratum variability is larger than across-state variability (column 4 is bigger than column 2, except for multiunit housing). Differences between state substitution rates, for example, cannot be explained by differences in demographics; indeed, controlling for the poststrata makes the differences larger rather than smaller. This may seem a bit paradoxical, but see Section 4, which makes the connection with analysis of variance.

If the synthetic assumption held true, the SD's in the last column of Table 5 would be negligible. Instead, they are larger than the state-to-state differences summarized in column 2. Thus, Table 5 confirms what is evident a priori: there is quite a lot of heterogeneity within poststrata across geographical areas. Moreover, Table 5 strongly suggests that the synthetic method is counterproductive at the state level; it might be wiser to adjust each state on the basis of its own PES data.

Error in Adjustment Due to Heterogeneity

We consider the root mean squared error in synthetic estimates of state-level rates for the proxies,

TABLE 5
Levels and SD's for proxies, in percent: the first SD is over the 50 states and Washington, D.C.; the second SD is over the 357 poststrata; the last SD is within poststrata across state, r.m.s. over state

Proxy	Level (%)	Standard Deviations (%) Across states	Across poststrata	Within poststrata across states
Substitution	0.899	0.408	0.701	0.857
Allocation	16.034	3.573	7.209	3.922
Multiunit housing	22.145	8.926	29.318.	8.755
Nonmailback	25.613	4.716	12.324	6.491
Mobility	42.738	5.270	48.918	7.949
Poverty	12.918	4.132	13.586	7.616

assuming that poststrata-level rates are given. Substitutions are the lead example. Suppose the substitution rate for each poststratum (the row total of substitutions divided by the row total of population) is known, but the entries in the matrix (the number of substitutions in each poststratum × state cell) are unknown. These entries could be estimated synthetically: multiply the poststratum substitution rate by the population count in the poststratum × state cell. Then add the entries in each column of the matrix to get an estimated number of substitutions in the corresponding state. Finally, divide by the state population to get each state's substitution rate. The objects of interest are these estimated state substitution rates.

In effect, each state's substitution rate has been estimated by the synthetic method, just as the Census Bureau estimates undercount rates. With substitutions, the exact answers are available from census data. Therefore the r.m.s. error of the synthetic method, across the 50 states and Washington, D.C., can be computed. This error is due solely to failures in the synthetic assumption, that is, variations in the substitution rate within poststrata across states. The r.m.s. errors are shown in column 1 of Table 6, in percent. For example, with substitutions, the r.m.s. error is about 0.31%. With allocations, the r.m.s. error is about 1.79%.

At first glance, these errors may seem rather small, supporting the synthetic assumption. However, much depends on the standard of comparison. The sampling error in estimated state undercount rates (r.m.s. SE across the 50 states and Washington, D.C.) is about 0.5%; see Table 4. Thus, error due to heterogeneity in the proxies is of the same order as, or even larger than, sampling error in estimated undercount rates. If heterogeneity in undercount rates is comparable to heterogeneity in the proxies, then heterogeneity is a major source of error in census adjustment by the synthetic method. On the other hand, if the proxies are not comparable to the undercount rate with respect to heterogeneity, they seem to provide little evidence about the degree to which the synthetic assumption holds.

Of course, scale matters, and the variables in Table 6 have different overall levels. For example, the Census Bureau's estimate for the undercount rate (from the Post Enumeration Survey corrected for certain biases) is 1.580% (Table 4). The overall substitution rate is 0.899% (Table 5). Practice at the Bureau suggests scaling the substitution rate by the factor $1.580/0.899 = 1.758$, before doing the calculations. The second column of Table 6 gives results for this "level-scaling." (The arithmetic for substitutions is $1.758 \times 0.31 = 0.54$.) As Table 6 shows, substitutions now have the worst r.m.s. error of the proxies, somewhat larger than the r.m.s. SE of state undercount rates.

There are other ways to scale besides levels. One possibility is to equalize state-to-state variability. The SD of undercount rates across states (corrected for sampling variance) is about 0.653% (Table 4). One could scale substitutions to have the same SD, multiplying by $0.653/0.408 = 1.600$ before doing the calculations. Results for this SD-scaling are shown in the third column of Table 6. (The arithmetic for substitutions is $1.600 \times 0.31 = 0.50$.)

Apparently, there are at least two scales on which the proxies differ from the undercounts: level and SD. Generally, it is not possible to match on both scales; although with substitutions or multiunit housing, you come very close. As Table 6 shows, conclusions depend on scaling. Levels matter, so do variances and so would covariation (common patterns in residuals across states). No scaling is perfect, which implies there may be no good proxies for the undercount. Furthermore, the proxies are all positive, while undercounts may be positive or negative, the latter possibility corresponding to overcounts. Thus, caution is in order. Still, Table 6 suggests that

TABLE 6
Root mean square errors of synthetic estimates, in percent; six proxies for undercount

Proxy	Root Mean Square Error (%)		
	Raw	Level-scaled	SD-scaled
Substitution	0.31	0.54	0.50
Allocation	1.79	0.18	0.33
Multiunit housing	3.41	0.24	0.25
Nonmailback	2.37	0.14	0.33
Mobility	3.53	0.13	0.44
Poverty	3.01	0.37	0.48

error in census adjustments due to heterogeneity is comparable in magnitude to the error caused by sample variability, and perhaps even larger.

One additional comparison may be useful. The interest in adjusting state populations is due to differential undercounts, which vary from state to state. The size of these differentials can be measured by the SD of the state rates. Take substitutions as a proxy. The state-to-state differences are on the order of 0.4% (column 2 in Table 5). The error due to heterogeneity is about 0.3% (column 1 in Table 6). This error is about 75% of the effect of interest. Results for other proxies are similar. Heterogeneity is not trivial.

The Sign Test

Come back to the 357 × 51 matrix of population counts for poststrata by states, and the corresponding matrix of substitutions. In each cell, we have the actual number of substitutions from census records, as well as the number estimated from the synthetic method. For each cell, let

$$\text{residual} = \text{estimate} - \text{actual}.$$

In cells with no population, the residual is 0. Otherwise, the residual will be positive or negative.

A strict interpretation of the synthetic assumption would require the residual to be identically 0. Clearly, the data contradict this strict form of the assumption. A weaker interpretation is that the residuals vary randomly (in some sense) from cell to cell and are more or less symmetric. If that is so, residuals would tend to cancel when adjusting states.

A sign test of the weaker interpretation can be made as follows. For each state, there are cells with nonzero population counts, corresponding to poststrata with members resident in that state. A typical state has 147 nonzero cells, although Alaska has 98; no state has more than 147. (There are 7 age groups in Table 2 and 21 rows in Table 3: $7 \times 21 = 147$.) In each nonzero cell, the weak interpretation of the synthetic assumption suggests that the sign of the residual is random. Thus, for each state, the number of positive residuals should be binomial with success probability 1/2, the number of trials being that state's number of nonzero cells. As a result, nearly half the nonzero residuals should be positive in each state. This prediction can now be compared to the data.

TABLE 7
Empirical SD's, in percent: Six Proxies for undercount; the theoretical SD, based on the synthetic assumption, is about 4.4%

Proxy	Empirical SD (%)
Substitution	24.5
Allocation	24.0
Multiunit housing	26.5
Nonmailback	22.3
Mobility	16.8
Poverty	23.7

For each state, count the number of positive residuals and divide by the number of cells with nonzero population. Then compute the SD of the resulting empirical distribution on [0, 1]. The empirical SD for substitutions is 24.5%; the SD suggested by the binomial model is 4.4%; details are in Section 4. The empirical distribution is fairly uniform, the model distribution is rather concentrated. In other words, the distribution of residuals is quite different from the predictions of the synthetic assumption. Cancellation of residuals is an unlikely assumption.

Results for the other proxies are rather similar. The six empirical SD's are shown in Table 7. The binomial SD's are all about 4.4%, since the number of cells with nonzero population is virtually the same for all populations considered (census, mail universe and long-form sample estimated population).

3. SUMMARY AND CONCLUSIONS

Data on proxy variables suggest substantial failures in the synthetic assumption. The r.m.s. error in

estimated state undercount rates arising from failures in the synthetic assumption seems comparable in magnitude to the r.m.s. error arising from sampling variability. Investigators who compare errors in adjusted and unadjusted census data should take heterogeneity into account; otherwise, the analysis may be quite biased against the census. The Census Bureau's loss function analysis appears to suffer from this problem (Bureau of the Census, 1992b) and is unconvincing for that reason among others.

Postscript

At a briefing on 23 December 1992, the Bureau announced its decision not to adjust the census. The Bureau argued that its "loss function analysis" was robust against heterogeneity; that improvements could be made at the state level; but that, for smaller areas like cities and countries, adjustment was of doubtful value (Bureau of the Census, 1992c, d; 1993).

4. MATHEMATICAL DETAILS

Index the 357 poststrata by i and the 51 areas (states and Washington, D.C.) by j. Let c_{ij} be the census count of members of poststratum i resident in state j on census day. Write subscript "+" for addition over a subscript: thus

$$c_{i+} = \sum_{j=1}^{51} c_{ij}$$

is the ith row sum in the c-matrix, corresponding to the census count of members of poststratum i in the 50 states and Washington, D.C.

Substitutions are the lead example. Let s_{ij} be the number of substitutions in the cell corresponding to poststratum i and area j. The first column in Table 5 reports levels, that is, s_{++}/c_{++}. The second column reports the SD of the state rates, that is, the SD of the 51-vector whose jth entry is s_{+j}/c_{+j}. The third column reports the SD of the poststratum rates, that is, the SD of the 357-vector whose ith entry is s_{i+}/c_{i+}. These statistics depend only on the marginal distributions.

The fourth column in Table 5 depends on interior cells. It reports the root mean square of the 357-vector whose ith entry is the standard deviation of

$$\{s_{ij}/c_{ij}\colon j = 1, \ldots, 51 \text{ and } c_{ij} > 0\}.$$

Allocations and multiunit housing are similar. For nonmailback rates, the denominator is the number of persons in the mail universe, rather than the census count. For mobility and poverty rates, the denominator is the estimated census count from the long-form sample.

The following example illustrates how controlling for poststratum increases cross-state variability; contact will be made with analysis of variance. Let x_{ij} be some numerical characteristic of the ijth cell. Let U be a random row index, and let V be a random column index. Consider the random variable

$$X = x_{UV}.$$

Choose U and V with weights proportional to the census counts, so

$$P\{U = i \text{ and } V = j\} = c_{ij}/c_{++},$$

where c_{ij} is the census count in the cell and subscript "+" denotes summation over an index. Take $x_{ij} = s_{ij}$, the number of substitutions in cell (i, j). Now $E\{X \mid V\}$ gives the state substitution rates, and $E(\text{var}\{X \mid U\})$ is the mean across poststrata of the within-poststratum variance; means and variances are weighted by population size. Thus, $E(\text{var}\{X \mid U\})$ is the analog of within-poststratum-across-state variance. However, there is no necessary inequality between $E(\text{var}\{X \mid U\})$ and $\text{var}(E\{X \mid V\})$, although both are less than $\text{var}(X)$:

$$E\Big(\text{var}\{X \mid U\}\Big) + \text{var}\Big(E\{X \mid U\}\Big) = \text{var}\, X.$$

The weighted results may be of some interest (Table 8). For two out of the six proxies, conditioning on poststratum adds variability to the state rates.

It may be useful to put the calculation for Table 6 in algebraic terms. As before, let s_{ij} be the number of substitutions in cell (i, j), and let c_{ij} be the census count. Let $\lambda_i = s_{i+}/c_{i+}$ be the substitution rate for poststratum i. Then the estimated number of substitutions in cell (i, j) is $c_{ij} \times \lambda_i$, the estimated number of substitutions in state j is $\Sigma_i\, c_{ij} \times \lambda_i$, and the estimated substitution rate in state j is

$$(\Sigma_i c_{ij} \times \lambda_i)/(\Sigma_i\, c_{ij}).$$

For the sign test, let n_j be the number of positive cells in state j, that is, the number of poststrata i such that the census count $c_{ij} > 0$. Let ξ_j be binomial, with n_j trials and success probability 1/2, independent in j. Let $\eta_j = \xi_j/n_j$. The number of positive residuals is modeled as ξ_j, according to the weak form of the synthetic assumption. So the empirical distribution for the fraction of positive residuals is modeled as

$$\{\eta_j\colon j = 1, \ldots, 51\}.$$

TABLE 8

Standard deviations for proxies, in percent: the first SD is over the 50 states and Washington, D.C.; the second SD is over the 357 poststrata; the third SD is within poststrata across states, r.m.s. over states; the last SD is taken over all poststrata × state cells; all SD's are weighted by population

	Standard Deviations (%)			
Proxy	Across states	Across poststrata	Within poststrata across states	Across all cells
Substitution	0.26	0.65	0.35	0.74
Allocation	2.37	6.24	1.92	6.53
Multiunit housing	9.44	27.28	5.56	27.84
Nonmailback	4.40	12.89	2.97	13.22
Mobility	5.70	21.67	4.34	22.10
Poverty	3.40	12.31	3.98	12.93

This distribution has empirical variance

$$s^2 = \frac{1}{50}\sum_{j=1}^{51}(\eta_j - \overline{\eta})^2 \quad \text{where } \overline{\eta} = \frac{1}{51}\sum_{j=1}^{51}\eta_j.$$

As usual,

$$E\{s^2\} = 0.25 \times \frac{1}{51}\sum_{j=1}^{51}\frac{1}{n_j} = 0.044 = 4.4\%.$$

The model involves binomials with different numbers of trials. Thus, the empirical distribution is a slightly nonstandard object. However, it seems to be informative.

In principle, skewness in the residuals may arise because some cells are relatively small, and the substitution rate (for instance) is close to 0. Consider, then, a "superpopulation" version of the synthetic assumption. The null hypothesis is that, for each poststratum i, the number of substitutions s_{ij} is binomial, with success probability p_i and number of trials c_{ij}; the numbers are independent across states j; the success probability depends only on the poststratum i. The alternative hypothesis allows the success probability to depend on i and j.

Treating poststrata as independent, the usual likelihood test statistic is $\chi^2 = 2.2 \times 10^5$ on 6,510 degrees of freedom. Taking the 357 poststrata one at a time, the mean number of degrees of freedom is 18 and the mean χ^2 value is 614. The most favorable poststratum (sequence number 258) offers five degrees of freedom and $\chi^2 = 12$, so $P = 3.6\%$. The next most favorable poststratum (256) also has five degrees of freedom but $\chi^2 = 20$ so $P = 0.13\%$. The remaining 355 poststrata all have $P < 0.1\%$. About 20% of the positive cells have fewer than five expected substitutions but censoring such cells does not change the results. The superpopulation idea, whatever its conceptual merits or demerits, is not viable.

5. LOSS FUNCTION ANALYSIS

Heterogeneity is appreciable, adding substantial uncertainty to estimated undercounts for states and smaller areas. That is our message so far. Now there is another topic: the impact of heterogeneity on what the Census Bureau calls "loss function analysis", that is, estimation of risks for the census and adjustment.

The Bureau's loss function analysis is done on population shares rather than rates. To keep the focus on heterogeneity, we ignore problems created by bias and sampling error in the Post Enumeration Survey. Then risk is just total squared error in population shares. (For background, see Freedman, Wachter, Cutler and Klein, 1994.)

In the presence of heterogeneity, estimated risks for the census and for adjustment are severely biased. Hence, the risk difference may be biased. The sign of the bias in the risk difference can go either way; cancellation is also a possibility. Bias in estimated risks is the topic of this section.

To address heterogeneity, the Bureau uses "artificial population analysis," creating hypothetical true and census populations from data on proxies. Artificial population counts have to be generated for each poststratum × state cell. There are many ways to do this, but our approach is quite straightforward. The census goes in as itself. Next, each proxy variable corresponds to a set of poststratum × state cell counts, for instance, the number of substitutions in each cell. For the "true population," we use

$$\text{census} + \lambda \times \text{proxy}.$$

The scale factor λ is chosen to make the overall undercount rate in the artificial population match 1.6%, the rate estimated for the census of 1990:

$$\lambda \times \frac{\text{total of proxy cell counts}}{\text{total of census cell counts}} \approx 1.6\%.$$

TABLE 9

Loss function analysis using the proxies: risks have been multiplied by 10^8; scale factors are chosen to match the level of undercount. The "true risks" are computed using data on proxies in each poststratum × state cell. The "estimated risks" are computed using the synthetic assumption. The estimated risk for adjustment is 0, so the estimated risk for the census coincides with the estimated risk difference; DIFF = 8 × SUB − 0.2 × NMB

Proxy	Scale factor	True risks		True risk difference	Estimated risk difference
		Census	Adjustment		
SUB	1.8	51	26	25	33
ALL	0.10	14	3	11.1	10.6
MUH	0.072	223	27	196	138
NMB	0.063	35	6	29	24
MOB	0.038	26	3	23	22
POV	0.12	40	19	21	24
DIFF	**	459	381	77	224

The scale factors are shown in column 1 of Table 9; SUB stands for substitutions, ALL for allocations and so forth.

Of course, the proxies are part of the census population, while omitted persons are disjoint from that population. For present purposes, that may not matter. On the other hand, all methods for generating artificial populations have a basic problem: the hypothetical "true population" may not be in close correspondence to the actual population.

In effect, Table 6 reports the results of artificial population analysis for estimates of adjustment factors, while the present section deals more directly with shares. Wolter and Causey (1991) review Bureau research on artificial population analysis for the 1980 census; also see Bureau of the Census (1992d, 1993) on the 1990 census.

Turn now to the algebra. Recall that c_{ij} is the census count in poststratum i and state j, so the census count in state j is

$$c_{+j} = \sum_{i=1}^{357} c_{ij}.$$

The total census population is

$$c_{++} = \sum_{i=1}^{357}\sum_{j=1}^{51} c_{ij}.$$

The census share for state j is $\text{csh}_j = c_{+j}/c_{++}$.

Substitutions are the lead example; s_{ij} is the number of substitutions in poststratum i and state j; and the "true undercount" t_{ij} in that cell is taken as $t_{ij} = 1.8 \times s_{ij}$. (As noted above, scale factors like the 1.8 are chosen to make the level of the proxy correspond to the level of undercount: $1.8 \times s_{++}/c_{++} \approx 0.016$.)

The "true population share" of state j is

$$\text{tsh}_j = (c_{+j} + t_{+j})/(c_{++} + t_{++}).$$

The total squared error for the census is then

$$\sum_{j=1}^{51}(\text{csh}_j - \text{tsh}_j)^2.$$

This is $51/10^8$, as shown in column 2 of Table 9.

The synthetic method estimates t_{ij} as $c_{ij} \times t_{i+}/c_{i+}$, where t_{i+} is assumed to be known. So the adjusted population for state j is

$$a_{+j} = \sum_{i=1}^{357} c_{ij} \times t_{i+}/c_{i+},$$

and the adjusted share for state j is

$$\text{ash}_j = a_{+j}/a_{++}.$$

The total squared error for adjustment is

$$\sum_{j=1}^{51}(\text{ash}_j - \text{tsh}_j)^2.$$

This is $26/10^8$, as shown in column 3 of Table 9.

The risk difference is the total squared error for census shares minus the total squared error for adjusted shares. This works out to $51 - 26 = 25/10^8$, as shown in column 4 of Table 9. Positive numbers favor adjustment: the census makes larger errors. Of course, in application the true shares tsh_j would not be known. However, we are assuming no bias in the Post Enumeration Survey and no sampling error; so the "adjustment factors" t_{i+}/c_{i+} are known with certainty.

Invoking the synthetic assumption, the error due to adjustment is estimated as 0. On the same basis, the squared error for the census is estimated

as

$$\sum_{j=1}^{51}(\text{csh}_j - \text{ash}_j)^2.$$

This is $33/10^8$, as shown in the last column of Table 9. (The estimated risk for the census coincides with the estimated risk difference.) In this example, the estimated risk difference based on the synthetic assumption is about one-third larger than the true risk difference. In other words, the Bureau's loss function analysis, which rides on the synthetic assumption, overstates the advantages of an adjustment based on the synthetic method. In the presence of heterogeneity, loss function analysis is biased toward adjustment.

Results for other proxies in Table 9 are computed in a similar way. (Here, the denominators for nonmailback rates are the census counts; the denominators for mobility rates and poverty rates are census counts estimated from the long-form sample.) The poverty rate (POV) goes like the substitution rate. For allocations (ALL) and mobility (MOB), however, heterogeneity causes almost no bias in estimated risk differences. For multiunit housing (MUH) and nonmailback rates (NMB), heterogeneity creates a bias against adjustment.

A further complication should be mentioned: the proxies are all positive, but undercounts can be negative—when the corresponding cell has been overcounted. To indicate the possibilities, we use DIFF = 8 × SUB − 0.2 × NMB as a proxy. The bias against the census is remarkable, a factor of 3. $(\text{DIFF}_{++}/c_{++} \approx 2.1\%$, the estimated level of undercount on July 15, 1991; 8 × SUB more or less matches the gross omissions while 0.2 × NMB matches the erroneous enumerations; see Department of Commerce, 1991b.) In short, Table 9 shows that almost anything can happen.

A bit of algebra may make the situation clearer. The bias in the loss function analysis is the difference between the true risk difference and the estimated risk difference. For state j, this is

$$(\text{csh}_j - \text{tsh}_j)^2 - (\text{ash}_j - \text{tsh}_j)^2 - (\text{csh}_j - \text{ash}_j)^2$$
$$= 2 \times (\text{ash}_j - \text{csh}_j) \times (\text{tsh}_j - \text{ash}_j).$$

Thus, loss function analysis is "conservative," that is, biased against adjustment, when adjustment of shares is conservative:

$$\text{ash}_j > \text{csh}_j \quad \text{and} \quad \text{tsh}_j > \text{ash}_j$$

or

$$\text{ash}_j < \text{csh}_j \quad \text{and} \quad \text{tsh}_j < \text{ash}_j.$$

TABLE 10
Correlations of proxy rates with undercount rates; the unit of analysis is either the poststratum or the state

Proxy	Unit of Analysis	
	Poststratum	State
Substitution	0.40	0.28
Allocation	0.0080	0.10
Multiunit housing	0.24	0.0021
Nonmailback	0.43	0.37
Mobility	0.33	0.53
Poverty	0.42	0.44

Otherwise, loss function analysis is biased against the census.

It may be noted that none of the proxies are well correlated with undercounts (Table 10). Allocations and multiunit housing are particularly weak in this respect. Furthermore, the biases (like the risks themselves) are driven by data for only a few of the 51 areas.

A minor inconsistency in our scaling should be noted too: as the denominator for its undercount rate, the bureau uses the estimated true population; our denominator for the proxy rates is the analog of the census population. In principle, these rates should be level-scaled to $1/(1 - 0.0158) \approx 1.605\%$ rather than 1.580%.

Table 9 quantifies bias in estimated risks due to heterogeneity—for the proxies. For undercount rates, however, the bias is unknown. The Census Bureau seems to argue that, for a majority of its proxies, the bias is negligible or runs against adjustment (Bureau of the Census, 1993). However, we think the basic question is still open: are undercounts more like substitutions, or allocations, or multiunit housing, let alone DIFF? The data do not answer this question.

To sum up, in proxies for the undercount rate, heterogeneity is appreciable. This creates substantial extra uncertainty in estimates for states and smaller areas. The same conclusions are likely to hold for undercounts themselves. With respect to loss function analysis, the bias due to heterogeneity may be substantial; the data do not decide the issue.

ACKNOWLEDGMENTS

We thank Richard Cutler (Utah), Terry Speed (U.C. Berkeley) and Amos Tversky (Stanford) for helpful comments. Research partially supported by NSF Grant DMS-92-08677 (to D. Freedman).