MEASURING THE IMPACT OF CENSUS ENUMERATION ERRORS ON POPULATION PROJECTIONS

Norfleet W. Rives, Jr., University of Delaware

1. INTRODUCTION

The social science professions are large consumers of demographic statistics generated by the decennial census of population and housing. Although the census is theoretically an enumeration of the population, evidence compiled for the three most recent censuses suggests that tabulation counts are neither accurate nor complete.

Census data are subject basically to two sources of error, excluding errors attributable exclusively to sampling. The first source involves errors of coverage. Coverage errors refer to gross omissions from the census count. According to demographic estimates prepared by the Census Bureau, the 1970 count of population was deficient by approximately 5.3 million persons.1/

The second source of error involves errors of content. Content errors refer to the misreporting of census characteristics, such as age, sex, race, educational attainment, occupation, employment status, and family income. The concept of response error is central to the study of content errors, and several different approaches can be taken to measure response errors.2/ The estimation of response errors, as opposed to their measurement, normally involves the application of either record-match methods or reinterview techniques. During the 1970 Census, for example, the reporting accuracy of major subject items on the population schedule was determined by matching census records from an aggregated 20-percent sample with records compiled from the March 1970 administration of the Current Population Survey.3/

The net effect of coverage and content errors on basic demographic statistics generated by the decennial census is called net census undercount. Percent estimates of net census undercount, classified by age, sex, and race, are shown in Table 1 for the 1970 Census.

Since an increasingly large sector of social science research is devoted to problems of public policy, and since policy analysis frequently requires census or census-related population statistics, it is reasonable to conclude that census enumeration errors may dilute the potential usefulness of social science research on public policy. To document the general dimensions of the problem, this paper investigates the effect of census errors on population projections, a common form of demographic information.

TABLE 1. Percent Estimates of Net Census Undercount, by Age, Sex, and Race: 1970

Age Group	White Female	White Male	Non White Female	Non White Male
0-4	2.0	2.3	9.7	10.3
5-9	2.2	2.4	6.6	7.5
10-14	0.9	1.1	2.7	3.5
15-19	0.5	1.3	2.1	3.1
20-24	1.1	2.5	3.4	8.7
25-29	2.8	4.7	6.6	15.6
30-34	2.0	4.0	3.7	14.4
35-39	0.8	4.1	4.6	17.8
40-44	0.1	3.2	3.5	16.3
45-49	0.5	3.5	5.0	13.3
50-54	-0.3	1.8	3.8	10.1
55-59	1.3	2.1	7.4	10.6
60-64	2.7	2.3	5.6	7.3
65-69	-1.1	-0.2	-11.7	-6.7
70-74	0.4	-0.1	5.8	-0.7
75+	5.9	3.6	16.5	0.3

Note: Base of percents is corrected population. Negative entry indicates an overcount.

Source: J. Siegel, "Estimates of Coverage of the Population by Sex, Race, and Age in the 1970 Census," Demography 11 (1974): 13-15.

2. METHODOLOGY

The computer simulation experiment provides a useful approach for establishing the direction and magnitude of the impact of census errors on projected populations. The particular experiment discussed in this paper was conducted for the total resident population of the United States. The experiment consists of two projections.

The first projection incorporates reported statistics, without adjustments for net census undercount. The projection was made from midyear 1970 to midyear 2020, using reported schedules of fertility and mortality.4/ No allowance was made for net international migration, an assumption commonly made for illustrative projections at the national level. The reported schedules of fertility and mortality were derived from published statistics on population and vital registration. The fertility schedules, consisting of central birth rates specific for sex of child, race of child, and age of mother, were constructed using 1970 registered births and midyear population estimates.5/ The mortality schedules, consisting of survival ratios specific for age, sex, and race, were derived from 1970 United States life tables.6/ The fertility and mortality schedules were held constant during the entire projection period, an arbitrary assumption made for purposes of illustration. The projection period was set at fifty years to provide an adequate time interval for demonstrating the effect of census errors on both short-range and long-range projections.

The second projection incorporates population and vital statistics adjusted for net census undercount. The estimates of net undercount shown in Table 1 were used to inflate (deflate) the reported populations for July 1, 1970, producing corrected age-sex-race distributions. The denominators of the birth rates derived from published statistics were adjusted to account for the effect of census coverage and content errors on the female populations of childbearing age. The numerators of the birth rates were not adjusted, because according to birth registration tests conducted during the period from 1964 to 1968, birth statistics at the national level are not subject to any significant errors.7/ Since the female populations of childbearing age are among the most accurately and completely enumerated by the census, the adjustments to the reported birth rates were relatively small; the average adjustment was less than 3 percent. The 1970 United States life tables were corrected for net census undercount using basically the same procedure applied to the birth rates. The denominators of the central death rates from which the life tables are constructed were inflated (deflated) using the estimates shown in Table 1. Corrected survival ratics were then computed from the life tables incorporating the adjusted death rates.

The numerators of the death rates were not modified, because although a national death registration test has never been conducted, deaths are thought to be as completely registered as births. The only problem with mortality statistics may be the accuracy of reporting age at death, especially for older cohorts where birth certificates are not widely available to confirm date of birth. The presence of content errors in official mortality data could affect the conclusions of this study, particularly if the errors were significant and occurred at younger ages as well as older ages. Since this possibility has been documented with only very limited evidence, however, there is really no defensible basis for adjusting the numerators of the reported death rates.8/ Consequently, for purposes of this study the effect of coverage and content errors on national death statistics is assumed to be negligible.

The second projection was made in precisely the same manner as the first. The estimated 1970 populations, adjusted for net census undercount, were projected to the year 2020, using the corrected schedules of fertility and mortality. Once again, no allowance was made for net international migration, and each vital schedule was held constant during the entire projection period.

3. IMPACT ON PROJECTED TOTAL GROWTH

The simulation experiment was designed to assess the impact of census errors on two major dimensions of the population projection--projected total growth and projected age structure. The findings with respect to projected total growth are presented below, while the findings regarding projected age structure are presented in the following section.

Percent errors between reported and corrected total populations for the projection experiment are shown in Table 2. The projected total population for each time period is subject to two sources of error. The first source involves the misestimation of the initial population, and the second concerns the misestimation of the growth rate. The latter source can be traced to the underlying errors in the reported fertility and mortality schedules. There are several interesting characteristics of the data presented in Table 2. The first, and certainly the most obvious, is the relative magnitude of the discrepancy for each sex-race group within projection periods. Given the date of the projection, one finds that census errors have the greatest impact on the nonwhite male estimates and the least impact on the white female estimates. This reflects, to a large extent, the underlying sex-race structure of census errors.

A second characteristic, and perhaps the most intriguing, is the systematic decline in the percent error for each of the sex-race cohorts over the entire projection period. The discrepancy associated with the nonwhite female population, for example, declines from 4.9 percent in 1970 to 1.1 percent in 2020. Ironically, this means that with respect to total growth, the long-range projections are actually less affected by census errors than the short-range projections. The negative trend in each discrepancy series can be explained in the following manner.

Let P_a be the true 1970 population and r_a, the true annual growth rate. Likewise, let P_c be the reported 1970 population and r_c, the reported growth rate. The proportionate difference between

reported and corrected population estimates k years beyond 1970 is given by the expression

$$(P_a e^{r_a k} - P_c e^{r_c k})/P_a e^{r_a k},$$

where e is the base of the natural logarithms. Dividing both numerator and denominator by $P_a e^{r_a k}$ yields the expression

$$1 - (P_c/P_a)e^{(r_c - r_a)k}$$

The important component in the latter expression is the difference $(r_c - r_a)$. If, and as long as, r_c is greater than r_a, the exponential term will increase over time, causing the proportionate difference between reported and corrected population estimates to diminish. This is precisely what happens for the total population of each sex-race group.

4. IMPACT ON PROJECTED AGE STRUCTURE

The second dimension of the population projection to be investigated using the simulation experiment is projected age structure. Percent errors between reported and corrected age estimates could have been presented in tabular form, but to facilitate the analysis, comparisons were made of entire age distributions using the index of dissimilarity, a common technique for summarizing differential age composition.9/ Indexes of dissimilarity between reported and corrected age distributions are shown in Table 3. With respect to sex and race, the indexes generally reflect the underlying structure of census errors. Given the date of the projection, one finds that census errors produce the greatest distortions in the nonwhite male age distribution and the smallest distortions in the white female age distribution.

TABLE 2. Percent Errors between Reported and Corrected Total Populations, by Sex and Race: 1970 and Projections, 1980, 1990, 2000, 2010, and 2020

Sex and Race	1970	1980	1990	2000	2010	2020
White						
Female	1.4	1.2	1.0	0.8	0.4	0.1
Male	2.5	2.2	1.8	1.3	0.7	0.2
Nonwhite						
Female	4.9	4.2	3.9	3.2	2.2	1.1
Male	9.0	7.6	6.4	5.0	3.4	1.9

TABLE 3. Indexes of Dissimilarity between Reported and Corrected Age Distributions, by Sex and Race: 1970 and Projections, 1980, 1990, 2000, 2010 and 2020

Sex and Race	1970	1980	1990	2000	2010	2020
White						
Female	0.4	0.5	0.5	0.6	0.7	0.8
Male	0.5	0.6	0.8	0.9	0.9	0.8
Nonwhite						
Female	1.0	1.4	1.5	1.6	1.7	1.7
Male	2.4	2.8	2.9	2.7	2.4	2.1

Once again, the most interesting characteristic of the data shown in Table 3 concerns the time pattern of each index series. Index estimates rise from lower levels in 1970 to higher levels in subsequent years, and in the case of both male cohorts, the series actually declines again from peak levels to lower levels. This means that while population projections based on reported statistics tend to improve over the entire projection period with respect to projected total growth, they generally tend to deteriorate with respect to projected age structure. Minor rehabilitations in the quality of the projected age data can be observed only for the two male cohorts, as previously noted. The index series for the nonwhite female population may be converging on a reversal of trend, having stabilized at 1.7 between 2010 and 2020, but the white female series shows no apparent sign of improvement.

Interestingly, the index series for each of the component populations will ultimately converge on a relatively low level of discrepancy. This result follows directly from the particular assumptions governing the simulation experiment. Stable population theory guarantees that any projection continuously subject to constant fertility and mortality will converge asymtotically on a unique age distribution, an age distribution completely independent of the initial age composition and determined entirely by the vital schedules.10/ Since the adjustment for net census undercount does not produce radical differences between reported and corrected vital schedules, one would expect the reported and corrected long-range projections for each sex-race group to either exhibit similar age properties or, at least, to be moving very definitely in that direction. Indeed, it would be only if the reported and corrected vital schedules were significantly different that one would observe permanent distortions of projected age structure.

5. SUMMARY AND CONCLUSIONS

This paper has presented evidence that census enumeration errors bias population projections. Improved estimates can be obtained by correcting population projections derived from reported statistics for net census undercount. Discrepancies between reported and corrected projections are generally greater for the nonwhite population, regardless of sex, and the male population, regardless of race. Furthermore, errors associated with projected age structure are somewhat less predictable than errors associated with projected total growth.

There are two general conclusions to be drawn from this research. First, despite the prominence of modern sampling theory and the prevalence of sample surveys, the decennial census continues to exert a pervasive influence on all sectors of public policy, urban planning, and social science research. Second, given the state of census technology, including prospects for the immediate future, and the recurrent problem of public distrust of government data systems, coverage and content errors which measurably distort the accuracy and completeness of census information may very well be a permanent part of the statistical landscape.

FOOTNOTES

1/ See J. S. Siegel, "Estimates of Coverage of the Population by Sex, Race and Age in the 1970 Census," DEMOGRAPHY 11 (1974): 13-15.

2/ See M. Spiegelman, Introduction to Demography, Revised Edition (Cambridge: Harvard University Press, 1968), pp. 55-58.

3/ See U. S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program PHC (E)-11, Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match, 1975.

4/ Population data for the experiment were derived from published sources. See U. S. Bureau of the Census, Current Population Reports, Series P-25, No. 519, "Estimates of the Population of the United States, by Age, Sex, and Race: April 1, 1960 to July 1, 1973," 1974, pp. 29, 79.

5/ See U. S. National Center for Health Statistics, Vital Statistics of the United States, vol. 1, sec. 1, Natality, 1970, Table 1-9, p. 1-10.

6/ See U. S. National Center for Health Statistics, Vital Statistics of the United States, vol. 2, sec. 5, Life Tables, 1970, pp. 5-7 to 5-9.

7/ See U. S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program PHC (E)-2, Test of Birth Registration Completeness: 1964-1968, 1973

8/ See F. Bayo, "Mortality of the Aged," Transactions of the Society of Actuaries 24 (1972): 1-24.

9/ See H. S. Shryock, J. S. Siegel, and Associates, The Methods and Materials of Demography, vol. 1 (Washington: U. S. Government Printing Office, 1973), pp. 231-233. The reader should note that larger values of the index of dissimilarity indicate greater discrepancies between the age distributions being compared.

10/ See. A. J. Coale, The Growth and Structure of Human Populations: A Mathematical Investigation (Princeton: Princeton University Press, 1972), chapter 3.