APPENDIX H.
Adjustment of the Population Count

CONTENTS

INTRODUCTION

Since the first census in 1790, there has always been an emphasis on obtaining as complete a count as possible. Throughout the history of census taking in the United States, improvements have constantly been made in the census taking process, not only for efficiency, but also for improved quality. In keeping with this history, a major improvement (called adjustment) was used for the first time in the 1990 census. For the 1990 census, the census tabulations shown in this report are based on the actual census enumeration but modified (adjusted) to reflect best estimates of people missed in the original enumeration.

The adjustment to the basic census count was based on a large sample survey which was used to measure the net undercoverage in the basic 1990 census count. This survey was called the Post Enumeration Survey, or PES. Based on the results of this survey, an estimate was made of people missed in the original census count. This process is called dual-system estimation. Then, using a statistical model, these estimates of undercount were applied to all levels of census geography.

Since the tabulations in this report are basic census counts adjusted based on a sample survey, they are subject to some error which you would not expect in basic census tabulations. This appendix presents a discussion of the PES sample design and the statistical concepts that underlay the adjustment methodology as well as a discussion of the errors in the adjusted population counts presented in this report.

SAMPLING PROCEDURE

Overview and Scope

The PES was a multi-stage sample. First a sample of blocks was selected. See the following section for a description of the stratification used to select the block sample. Within each block, each housing unit was generally enumerated. For large blocks, the housing units were subsampled. Within each housing unit, we attempted to account for each person by talking to a knowledgeable respondent.

To measure the net undercoverage, two samples were needed. A sample of people who should have been counted was used to measure omissions. This was called the population or "P" sample. A sample of census enumerated persons was used to measure duplicates and other types of erroneous enumerations. This was called the enumeration or "E" sample. The joint implementation of these two samples constituted the PES.

The primary sampling unit for the 1990 PES was the block or block cluster (i.e., a group of blocks) and the same blocks were included in the P-sample and the E-sample. The Census enumerations in the sample blocks constituted the E-sample. For the P-sample, interviewers returned to the sample blocks after the census enumeration. They tried to identify all people living in the blocks at the time of the PES. This operation took place in late June or early July 1990. This interview was completely independent of the original Census enumeration. The interviewer asked for names and characteristics required to match the persons enumerated in the P-sample to those enumerated in the census. Equally important, the interviewer asked where each person was living on Census Day, April 1, 1990. This information was used to search the Census enumerations to see if the same people were indeed counted in the census. Those who were not located in the census constituted P-sample persons identified as omissions. E-sample persons (i.e., Census enumerations) were also matched to the P-sample to identify various types of erroneous enumerations (i.e., persons enumerated more than once).

A few groups were excluded from the PES sampling frame. People living in institutions were excluded, as were military personnel living in barracks. It was unrealistic to attempt an "independent" interview for these groups. People living in remote rural Alaska were also excluded. These people were enumerated in waves from January to March, often by flying into the village, interviewing, and flying out. By June, many of these people have left the village for remote fishing sites. Finally, the population defined by the Street/Shelter operation, "S-night" was excluded. It was unrealistic to expect to interview these people in June and then to match them at the April 1, location.

Sampling Strata

The Census Bureau has been studying people missed in censuses for many decades. There is some evidence that undercount is differential by certain geographic areas and by certain age-sex-racial groups. Thus, the adjustment methodology was designed to produce undercount estimates by groups (called post-strata) that we judged to be correlated with undercount.

Given the objective of producing estimates of the population for the post-strata, sampling strata were developed so as to correspond to the post-strata (defined by all variables except age and sex) as closely as possible.

The cross-classification of census division and the place type and size categories yield 54 major geographic areas that serve as major sampling strata The next step involved creating, within these areas, addii tional sampling strata by grouping geographic units with high concentrations of the race-Hispanic origin-tenure groups corresponding to the post-strata for the geographic area For this purpose, 1980 census counts of occupied housing units by tenure and the race-Hispanic origin of the householder were used to determine these strata and the collection of geographic units having more than 40 percent of one or more of the race-Hispanic origin-tenure minority (Black or not-Black Hispanic) groups were identified

After grouping geographic units, a total of 101 sampling strata were defined. For example, three sampling strata were defined for the Middle Atlantic Division central cities in the New York City PMSA. One stratum comprised geographic units with a high proportion of Black householders, another stratum comprised geographic units with a high proportion of not-Black Hispanic householders, and the final stratum contained all other geographic units. in the New York City PMSA. Since each sampling stratum contained a high proportion of a specific race-Hispanic origin group, the precision of estimates for the post-strata could be increased by the "optimum" allocation of sample to sampling stratum as discussed below. Finally, a sampling stratum was created having a large proportion of American Indians. This stratum was defined to include persons living on American Indian reservations and tribal trust lands.

Sample Allocation

The method for allocating sample to sampling strata was a two-step process. First, the sample of 150,000 occupied housing units was allocated to the 54 major geographic areas. This allocation was designed to achieve a constant coefficient of variation (the ratio of the standard error of an estimate to the expected value of the estimate) for dual-system estimates of population for these areas. Second, within each of the 54 geographic areas, sample was allocated to the demographic substrata (i.e., the collection of geographic units discussed above). This step can be viewed as a multivariate optimum allocation problem since there is generally more than one post-stratum of interest within each of the 54 areas. Thus, the allocation could be designed to provide the minimum coefficient of variation on the dual-system estimate for a particular post-stratum, for example, Black renters. However, this results in coefficients of variation for the other post-strata that are substantially greater than their own minimum value. Thus, it was decided to allocate the sample to minimize the coefficient of variation on the overall dual-system estimate (i.e., across all the post-strata).

ESTIMATION

After completing the PES enumeration, the next step was to produce estimates of the total population to compare with the census count to estimate net undercount. First, each PES case was assigned to a post-stratum. (See next section). The post-strata were designed to be correlated with undercount. The intent was for undercount. to be as alike as possible within a post-stratum and as different as possible between post-stratum. Then, within each post-stratum, a dual-system estimate was made of the population. It was a dual-system estimate because it was based on two "systems"-the census and the PES. In effect, each PES case was matched to the census. Most as expected, were found in the census. Some were not. These were assumed to be census misses. A similar match was done for the E-sample to estimate erroneous enumeration in the census. The combination of these two estimates produced a dual-system estimate of total population. This process was done for each post-stratum.

' Within each'post-stratum, the dual-system estimate of total population was compared to the actual census count. The ratio of the two is the adjustment factor. Finally, the adjustment factors were applied, by block, to every basic census count to arrive at adjusted census counts, (See sections on Adjustment Factors and Applying the Factors).

PES Post-Stratification

The PES sample (both P and E) is designed to provide sufficient precision for the dual-system estimates (see next section) of total population for the PES post-strata. The term "post-strata" is used to denote the finest level of detail for which direct PES estimates will be produced; i.e., dual-system estimates of the population. The post-strata are defined by characteristics of the persons enumerated in the PES and are defined so that within post-strata persons are as similar as possible with respect to the underlying causes of Census undercount. The variables used to define the post-strata are Census Region, size and type of area,

race and Hispanic origin. Subsequently, the post-strata were further partitioned by age, sex and in some cases tenure (owner, renter). The final post-strata then consist of some 357 population subgroups defined by geography (urban/non-urban and size), race/Hispanic origin, age, sex, and in some cases tenure. The race/Hispanic origin categories are Black, non-Black Hispanic, all others, and in some instances Asian and Pacific Islanders. The age-sex categories are all 0-17, and 18-29, 30-49, 50 and over crossed by sex. Note that the first age group includes males and females. Type and size of place consist of three categories. The full hierarchy is as follows:
Race (4), Housing Tenure (2), Region (4), and Urbanization (3).

The three categories for urbanization are:

Urbanized areas with population greater than 250,000
Other urbanized areas
Non-urban or rural areas

In addition, final post-strata were formed for American Indians living on American Indian reservations and tribal trust lands by the same age and sex categories

Dual-System Estimation

To get estimates of the total population, a dual-system estimator (DSE) was used. The DSE is written as:

$$DSE = \frac{N_p\,(CEN - SUB - EE)}{M} \qquad (1)$$

where

N_p = PES population estimate (from the P-sample)
CEN = unadjusted Census count
SUB = number of census whole-person substitutions (i.e., the assignment of a full set of characteristics for a person)
EE = estimate of the number of erroneous enumerations (from the E-sample).

and

M = estimate of the number of persons matched between the census and the PES populations.

A separate DSE was calculated for each post-stratum.

[1]See appendix A for MSA and PMSA definitions.

Applying the Adjustment Factors

An adjustment factor was calculated for each post-stratum as the ratio of the DSE, as described above, to the unadjusted Census count.

These final PES adjustment factors were used to compute the adjusted population by post-stratum for any block by multiplying the known census count by the adjustment factor for the post-stratum. For example, if the adjustment factor for males, age 0-17, not-Black-not-Hispanic, owners, living in Central cities of small MSA's, in the Mid-Atlantic Division was 1.02, then, for every 100 such people counted in the Census, two new people were added. Very few blocks will be so large as to have 100 people in each post-strata. If a block had 25 such people, multiplying by the adjustment factor results in the need to add 1/2 person. To accomplish this, one person was added one half the time. If there were no people with those characteristics living in the block, none were ever added.

ERRORS IN THE DATA

Type of Error

Whereas the census counts have been adjusted based upon the results of the PES, and the adjustment factors were derived from a sample, the adjusted figures in this publication may differ somewhat from the results which would have been obtained if all housing units, persons within those housing units, and persons living in group quarters had been included in the PES sample. The adjusted census counts would also vary if other samples of persons, housing units, and persons within housing units had been selected in the PES sample. The standard error of a survey estimate, such as a PES estimate, is a measure of the variation among the estimates from all the possible samples and thus is a

measure of the precision with which an estimate from a particular sample approximates the average result among all possible samples. The standard error of an adjusted census count is a function of the standard error of the adjustment factor and the size of the unadjusted count. The adjusted census count and its estimated standard error permit the construction of interval estimates with prescribed confidence that the interval includes the average result from all possible samples. The method of. calculating standard errors and confidence intervals is described in the following section, Calculation of Standard Errors.

In addition to the variability which arises from the sampling procedures, the unadjusted counts and the estimates calculated from the PES results are subject to non-sampling error, Non-sampling errors may be introduced during each of the many complex operations used to collect and process census and PES data For example, for the PES, operations such as matching or interviewing may introduce error into the data.

A more detailed discussion of the sources of non-sampling error in the census counts is given in the section on "Control of Non-sampling Error" in appe ix C. This component of error could introduce serious bias in the data and the total error could increase dramatically over that which would result purely from sampling.

Non-sampling error may affect the data in two ways. Errors that are introduced randomly will increase the variability of the data and should therefore be reflected in the standard error. Errors that tend to be consistent in one direction will make both the PES estimates and the unadjusted census counts biased in that direction. For example, if respondents consistently tend to underreport their age, their age distribution will be skewed towards the lower age categories. Then the resulting adjusted count of persons by age category will be below the actual figures. Such biases are not reflected in the standard error.

The error component of the regression model used to smooth the PES sample based adjustment factors is included in the variance of the adjustment factors used to generalize the coefficients of variation of the adjusted counts. Thus, this component of non-sampling error is reflected in the standard errors derived from the coefficients 'of variation shown in table A.

Calculation of Standard Errors

Total-Table H in this appendix contains the information necessary to calculate the standard errors of the adjusted census figures contained in this report.

Table H is a table of generalized coefficients of variation (CV's). The CV is the ratio of the standard error to the adjusted census count. To estimate the standard error of an adjusted census figure, you need only multiply the adjusted count from the publication times the highest logical generalized CV from table H. For example, for the estimated number of White males age 18 or above, use the highest CV among the appropriate two age groups (2044 or 45 years and over) for White males.

The CV's in table H are listed for three levels of geography: metropolitan areas, non-metropolitan areas, and statewide. CV's are given at each of these three levels for characteristics defined by race/ Hispanic origin, sex, and age and for total adjusted population and count adjustment population.

For example, if one wanted to know the standard error of the statewide adjusted count of White males aged 0-9, then one would multiply the adjusted census count of White males aged 0-9 times the generalized CV for White males aged 0-17 statewide from table H. If one needed to estimate the standard error of the number of female widows within a particular city, then one would multiply the adjusted count of female widows in that city times the highest generalized CV for females 18 years old and over across all of the race/ Hispanic origin categories from table H.

Sums and Differences-The standard errors derived from this table are not necessarily directly applicable to differences between and sums of two sample estimates. The standard error will approximately be equal to the square root of the sum of two individual standard errors squared; that is, for standard errors:

$$Se_i \text{ and } Se_j \text{ of estimates } \hat{X} \text{ and } \hat{Y}$$

$$Se_{(i+j)} = Se_{(i-j)} = \sqrt{(Se_i)^2 + (Se_j)^2}$$

This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated.

Count Adjustment Population-The count adjustment population is defined to be the difference between the adjusted census count and the unadjusted census coumt. The unadjusted census count is not subject to sampling error. Therefore, the standard error of the count adjustment population is equal to the standard error of the adjusted census count.

Ratios-The standard error of the ratio of two adjusted census counts, say X and Y, may be approximated as follows:

$$Se_{(\hat{X}/\hat{Y})} = \frac{\hat{X}}{\hat{Y}} \sqrt{\frac{(Se_{\hat{X}})^2}{\hat{X}^2} + \frac{(Se_{\hat{Y}})^2}{\hat{Y}^2}}$$

where $\hat{X}$, for example could represent the adjusted count of Blacks aged 2040 and $\hat{Y}$ could represent the total adjusted population. (Y could also represent the total adjusted Black population.)

Medians-For the standard error of the median of a characteristic (e.g., median age), it is necessary to examine the distribution from which the median is derived, as the size of the base and the distribution itself affect the standard error. An approximate method is given here. As the first step, compute one-half of the number on which the median is based (refer to this result as (N/ 2)). Treat N/2 as if it were an ordinary estimate and obtain its standard error as instructed above using table A. Compute the desired confidence interval about N/2. Starting with the lowest value of the characteristic, cumulate the frequencies in each category of the characteristic until the sum equals or first exceeds the lower limit of the confidence interval about N/2. By linear interpolation, obtain a value of the characteristic corresponding to this sum. This is the lower limit of the confidence interval about the median. In a similar manner, continue cumulating frequencies until the sum equals or exceeds the count in excess of the upper limit of the interval about N/2. Interpolate as before to obtain the upper limit of the confidence interval for the estimated median. When interpolation is required in the upper open-ended interval of a distribution to obtain a confidence bound, one may use 1.5 times the lower bound of the open-ended confidence interval as the upper bound of the confidence interval.

Confidence Intervals

A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that will contain the average value of the estimated characteristic that results over all possible samples, with a known probability. For example, if all possible samples that could result under the 1990 census PES design were independently selected and surveyed under the same conditions, and if the adjustment factors and its estimated standard error were calculated for each of these samples, then:

1. Approximately 68 percent of the intervals from one estimated standard error below the estimate to one estimated standard error above the estimate would contain the average result from all possible samples;
2. Approximately 90 percent of the intervals from I.645 times the estimated standard error below the estimate to 1.645 times the estimated standard error above the estimate would contain the average result from ail possible samples; and
3. Approximately 95 percent of the intervals from two estimated standard errors below the estimate to two estimated standard errors above the estimate would contain the average result from all possible samples.

The intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively. The average value of the estimated characteristic that could be derived from all possible samples is or is not contained in any particular computed interval. Thus, we c&not make the statement that the average value has a certain probability of falling between the limits of the calculated confidence interval. Rather, one can say with a specified probability or confidence that the calculated confidence interval includes the average estimate from all possible samples.

Confidence intervals may also be constructed for the ratio, sum, and difference between two adjusted figures, This is done by computing the ratio, sum of or difference between these figures, obtaining the standard error of the ratio, sum or diierence (using the formulas given earlier), and then forming a confidence interval for this estimated ratio, sum or difference as above. One can then say with specified confidence that this interval includes the ratio, sum or difference that would have been obtained by averaging the results from all possible PES samples.

The estimated standard errors given in this report do not include all portions of the variability due to nonsampling error that may be present in the data. The standard errors reflect the effect of simple response variance, but not the effect of correlated errors introduced by enumerators, coders, or other field or processing personnel. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (i.e., 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this publication based on the estimated standard errors. For more information on confidence intervals and nonsampling error, see any standard sampling theory text.

Use of Tables To Compute Standard Errors

Suppose that City A has an adjusted Hispanic age less than 18 population count of 12,000. We wish to determine a 90 percent confidence interval for this figure. We then look at the Table of Generalized Coefficients of Variation for this State, and determine which of the CV's are logical. Two different CV's are logical: the CV's for metropolitan areas for Hispanic origin age O-1 9 for two sex categories. Suppose that the maximum of the CV's is 0.01. We then estimate the standard error as:

$$Se = 12{,}000 \times 0.01 = 120.$$

We would then estimate the 90 percent confidence interval of the adjusted Hispanic origin population count as:

$$[12{,}000 - 1.645(120)] \text{ to } [12{,}000 + 1.645(120)]$$
or
11,803 to 12,197.

One can then say with about 90 percent confidence that this interval includes the value that would have been obtained from averaging the results from all possible samples in the PES.

Suppose that the unadjusted census count of Hispanics less than 18 years old in City A was 11,500. The estimated count adjustment population of Hispanics less than 18 years old in City A is, therefore, 12,000 – 11,000 = 500. The standard error of this estimate is the same as the standard error of the total adjusted census count for this group (120). We would then estimate the 90 percent confidence interval for the count adjustment population of Hispanics less than 18 years old in Cii A as:

$$[500 - 1.645(120)] \text{ to } [500 + 1.645(120)]$$

Of

303 to 697;

The calculation of standard errors and confidence intervals of sums will be illustrated Suppose that the adjusted census count of Hispanics, age less than 18, in County B (in the same State as above, but in a different metropolitan area) is 20,000. The generalized CV is again 0.01. The standard error for the adjusted census count of Hispanics, age less than 18, in County B is then 20,000 x 0.01 = 200. The standard error of the sum of the adjusted census count of Hispanics, age less than 18, in City A and County B is then:

$$\sqrt{120^2 + 200^2} = 233.$$

Suppose that one wanted to know, for example, the ratio of the number of women over the age of 45 years who are of Hispanic origin to the number of women over the age of 45 in some non-metropolitan region of a State. Let's say that the adjusted census counts show that there are 400 Hispanic origin women over 45 years of age ($\hat{X}$) and 4,000 women ($\hat{Y}$) in this age group. The ratio would then be:

$$\hat{R} = \hat{X} \div \hat{Y} = 400 \div 4{,}000 = 0.1$$

We then select the CV for Hispanic origin women aged 45 years and over and the CV for all races, total for women aged 45 years and over for non-metropolitan areas from the Table of Generalized CV's. Let's suppose that these turned out to be 0.004 and 0.010, respectively. We would then estimate the standard error for the adjusted number of Hispanic origin women in this age group as:

$$Se\hat{x} = 400 \times 0.004 = 1.6$$

and the standard error for the adjusted count of women in this age group as:

$$Se\hat{y} = 4{,}000 \times .010 = 40.$$

The standard error for the ratio would then be estimated as:

$$Se_{(\hat{x}/\hat{y})} = \frac{400}{4000}\sqrt{\frac{(1.6)^2}{(400)^2} + \frac{(40)^2}{(4000)^2}} = 0.001$$

The 90 percent confidence interval for the ratio would then be:

$$[0.1 - 1.645(0.001)] \text{ to } [0.1 + 1.645(0.001)]$$

or

,098 to ,102