Appendix
# Source and Accuracy of Estimates

## Source of Data

Most of these estimates come from data obtained in June of 1985 in the Current Population Survey (CPS). The Bureau of the Census conducts the survey every month, although this report uses only June data for its estimates. The June survey uses two sets of questions, the basic CPS and the supplement.

The data in *Living Arrangements of Young Adults Living Independently: Evidence from the Luxembourg Income Study* are not covered in this source and accuracy statement. For further information please see the sources listed.

**Basic CPS.** The basic CPS collects primarily labor force data about the civilian noninstitutional population. Interviewers ask questions concerning labor force participation of each household member 14 years old and over in every sample household.

The present CPS sample was selected from the 1980 decennial census files with coverage in all 50 states and the District of Columbia. The sample is continually updated to account for new residential construction. It is located in 729 areas comprising 1,973 counties, independent cities, and minor civil divisions. About 59,500 occupied housing units are eligible for interview every month. Interviewers are unable to obtain interviews at about 2,500 of these units because the occupants are not found at home after repeated calls or are unavailable for some other reason.

Since the introduction of the CPS, the Bureau of the Census has redesigned the CPS sample several times to improve the quality and reliability of the data and to satisfy changing data needs. The most recent changes were completely implemented in July 1985.

**June supplement.** In addition to the basic CPS questions, interviewers asked supplementary questions in June about marriage, divorce, widowhood, and remarriage of women.

**Estimation procedure.** This survey's estimation procedure inflates weighted sample results to independent estimates of the civilian noninstitutional population of the United States by age, sex, race and Hispanic/non-Hispanic categories. The independent estimates were based on statistics from the 1980 decennial census; statistics on births, deaths, immigration and emigration; and statistics on the size of the Armed Forces.

## Accuracy of Estimates

Since the CPS estimates come from a sample, they may differ from figures from a complete census using the same questionnaires, instructions, and enumerators. A sample survey estimate has two possible types of error: sampling and nonsampling. The accuracy of an estimate depends on both types of error, but the full extent of the nonsampling error is unknown. Consequently, one should be particularly careful when interpreting results based on a relatively small number of cases or on small differences between estimates. The standard errors for CPS estimates primarily indicate the magnitude of sampling error. They also partially measure the effect of some nonsampling errors in responses and enumeration, but do not measure systematic biases in the data. (Bias is the average over all possible samples of the differences between the sample estimates and the desired value.)

**Nonsampling variability.** Nonsampling errors can be attributed to many sources. These sources include the inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, respondents' inability or unwillingness to provide correct information or to recall information, errors made in data collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, and failure to represent all units with the sample (undercoverage).

CPS undercoverage results from missed housing units and missed persons within sample households. Compared to the level of the 1980 decennial census, overall CPS undercoverage is about 7 percent. CPS undercoverage varies with age, sex, and race. Generally, undercoverage is larger for males than for females and larger for Blacks and other races combined than for Whites. As described previously, ratio estimation to independent age-sex-race-Hispanic population controls partially corrects for the bias due to undercoverage. However, biases exist in the estimates to the extent that missed persons in missed households or missed persons in interviewed households have different characteristics from those of interviewed persons in the same age-sex-race-Hispanic group. Furthermore, the independent population controls have not been adjusted for undercoverage in the 1980 census.

For additional information on nonsampling error including the possible impact on CPS data when known, refer to Statistical Policy Working Paper 3, *An Error Profile: Employment as Measured by the Current Population Survey,* Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, 1978; and Technical Paper 40, *The Current Population Survey: Design and Methodology,* Bureau of the Census, U.S. Department of Commerce.

**Comparability of data.** Data obtained from the CPS and other sources are not entirely comparable. This results from differences in interviewer training and experience and in differing survey processes. This is an example of nonsampling variability not reflected in the standard errors. Use caution when comparing results from different sources.

**Note when using small estimates.** Summary measures (such as medians and percentage distributions) are shown only when the base is 75,000 or greater. Because of the large standard errors involved, summary measures would probably not reveal useful information when computed on a smaller base. However, estimated numbers are shown even though the relative standard errors of these numbers are larger than those for corresponding percentages. These smaller estimates permit combinations of the categories to suit data users' needs. Take care in the interpretation of small differences. For instance, even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

**Sampling variability.** Sampling variability is variation that occurred by chance because a sample was surveyed rather than the entire population. Standard errors, as calculated by methods described later in "Standard errors and their use," are primarily measures of sampling variability, although they may include some nonsampling error.

**Standard errors and their use.** A number of approximations are required to derive, at a moderate cost, standard errors applicable to all the estimates. Instead of providing an individual standard error for each estimate, generalized sets of standard errors are provided for various types of characteristics. Thus, the tables show levels of magnitude of standard errors rather than the precise standard errors.

The sample estimate and its standard error enable one to construct a confidence interval, a range that would include the average result of all possible samples with a known probability. For example, if all possible samples were surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

A particular confidence interval may or may not contain the average estimate derived from all possible samples. However, one can say with specified confidence that the interval includes the average estimate calculated from all possible samples.

Some statements may contain estimates followed by a number in parentheses. This number can be added to and subtracted from the estimate to calculate upper and lower bounds of the 90-percent confidence interval. For example, if a statement contains the phrase "grew by 1.7 percent (±1.0)," the 90-percent confidence interval for the estimate, 1.7 percent, is 0.7 percent to 2.7 percent.

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis is that the population parameters are different. An example of this would be comparing the number of White women divorced after their first marriage to the number of Black women divorced after their first marriage.

Tests may be performed at various levels of significance, where a significance level is the probability of concluding that the characteristics are different when, in fact, they are the same. All statements of comparison in the text have passed a hypothesis test at the 0.10 level of significance or better. This means that the absolute value of the estimated difference between characteristics is greater than or equal to 1.6 times the standard error of the difference.

**Standard errors of estimated numbers.** There are two ways to compute the approximate standard error, $s_x$, of an estimated number shown in this report. The first uses the formula

$$s_x = fs$$

where f is a factor from table A-3, and s is the standard error of the estimate obtained by interpolation from table A-1. The second method uses formula (2), from which the standard errors in table A-1 were calculated. This formula will provide more accurate results than formula (1).

$$s_x\sqrt{ax^2 + bx}$$

Here x is the size of the estimate and a and b are the parameters in table A-3 associated with the particular type of characteristic. When calculating standard errors for numbers from cross-tabulations involving different characteristics, use the factor or set of parameters for the characteristic which will give the largest standard error.

Table A-1.
**Standard Errors of Estimated Numbers**

(In thousands)

| Estimated number | Standard error | Estimated number | Standard error |
|---|---|---|---|
| 25 ........ | 10 | 5,000 ..... | 138 |
| 50 ........ | 14 | 7,500 ..... | 168 |
| 75 ........ | 17 | 10,000 .... | 193 |
| 100 ....... | 20 | 15,000 .... | 233 |
| 250 ....... | 31 | 25,000 .... | 293 |
| 500 ....... | 44 | 50,000 .... | 385 |
| 750 ....... | 54 | 75,000 .... | 432 |
| 1,000 ...... | 62 | 100,000 ... | 449 |
| 2,500 ...... | 98 | 125,000 ... | 439 |

**Illustration.** Table A shows that in June 1985, there were 11,367,000 women widowed after their first marriage. Using formula (1), the appropriate factor from table A-3, and a standard error obtained by interpolation from table A-1, the approximate standard error is (1.0)(204,000) = 204,000.

Using the second method, formula (2), with a = -0.000019 and b = 3,918 from table A-3, the estimate of the standard error is

$$\sqrt{-0.000019 x 11,367,000^2 + 3,918 x 11,367,000} = 205,000.$$

The 90-percent confidence interval for the number of women widowed after their first marriage is from 11,039,000 to 11,695,000 (i.e., 11,367,000 ± 1.6 x 205,000).

Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples.

**Standard errors of estimated percentages.** The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends on the size of the percentage and its base. Esti-

Table A-2.
**Standard Errors of Estimated Percentages**

| Base (in thousands) | 1 or 99 | 2 or 98 | 5 or 95 | 10 or 90 | 25 or 75 | 50 |
|---|---|---|---|---|---|---|
| 25 | 3.9 | 5.5 | 8.6 | 11.9 | 17.1 | 19.8 |
| 50 | 2.8 | 3.9 | 6.1 | 8.4 | 12.1 | 14.0 |
| 75 | 2.3 | 3.2 | 5.0 | 6.9 | 9.9 | 11.4 |
| 100 | 2.0 | 2.8 | 4.3 | 5.9 | 8.6 | 9.9 |
| 250 | 1.2 | 1.8 | 2.7 | 3.8 | 5.4 | 6.3 |
| 500 | 0.9 | 1.2 | 1.9 | 2.7 | 3.8 | 4.4 |
| 750 | 0.7 | 1.0 | 1.6 | 2.2 | 3.1 | 3.6 |
| 1,000 | 0.6 | 0.9 | 1.4 | 1.9 | 2.7 | 3.1 |
| 2,500 | 0.4 | 0.6 | 0.9 | 1.2 | 1.7 | 2.0 |
| 5,000 | 0.3 | 0.4 | 0.6 | 0.8 | 1.2 | 1.4 |
| 7,500 | 0.2 | 0.3 | 0.5 | 0.7 | 1.0 | 1.1 |
| 10,000 | 0.2 | 0.3 | 0.4 | 0.6 | 0.9 | 1.0 |
| 15,000 | 0.2 | 0.2 | 0.4 | 0.5 | 0.7 | 0.8 |
| 25,000 | 0.12 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
| 50,000 | 0.09 | 0.12 | 0.2 | 0.3 | 0.4 | 0.4 |
| 75,000 | 0.07 | 0.10 | 0.2 | 0.2 | 0.3 | 0.4 |
| 100,000 | 0.06 | 0.09 | 0.14 | 0.2 | 0.3 | 0.3 |
| 125,000 | 0.06 | 0.08 | 0.12 | 0.2 | 0.2 | 0.3 |

mated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more. When the numerator and denominator of the percentage are in different categories, use the factor or parameter from table A-3 indicated by the numerator.

The approximate standard error, $s_{x,p}$, of an estimated percentage can be obtained by use of the formula

$$s_{x,p} = fs$$

In this formula, f is the appropriate factor from table A-3, and s is the standard error of the estimate obtained by interpolation from table A-2.

Alternatively, formula (4) will provide more accurate results:

$$s_{x,p} = \sqrt{bp(100 - p)/x}$$

Here x is the total number of persons, families, households, or unrelated individuals in the base of the percentage, p is the percentage ($0 \leq p \leq 100$), and b is the parameter in table A-3 associated with the characteristic in the numerator of the percentage.

**Illustration.** Table A shows that of the 11,367,000 women widowed after their first marriage, 2,573,000 or 22.6 percent, remarried. Using formula (3), the appropriate factor from table A-3 (1.0), and a standard error from table A-2, the approximate standard error is (1.0)(0.8) = 0.8.

Using the alternative method, with b = 3,918, the approximate standard error of 22.6 percent is

$$\sqrt{3{,}918 \times 22.6 \times (100 - 22.6)/11{,}367{,}000} = 0.8$$

This means that the 90-percent confidence interval for the percentage of women widowed after their first marriage who remarried is from 21.3 to 23.9 percent (i.e., 22.6 ± 1.6 x 0.8).

**Standard error of a difference.** The standard error of the difference between two sample estimates is approximately equal to

$$s_{x-y} = \sqrt{s_x^2 + s_y^2}$$

where $s_x$ and $s_y$ are the standard errors of the estimates, x and y. The estimates can be numbers, percentages, ratios, etc. This will represent the actual standard error quite accurately for the difference between estimates of the same characteristic in two different areas, or for the difference between separate and uncorrelated characteristics in the same area. However, if there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

**Illustration.** Table A shows that there were 17,142,000 women divorced after their first marriage. It also shows 11,367,000 women widowed after their first marriage. The apparent difference is 5,775,000. Using formula (2) and the appropriate parameters from table A-3, the approximate standard errors are

$$\sqrt{-0.000019 \times 17{,}142{,}000^2 + 3{,}918 \times 17{,}142{,}000} = 248{,}000,$$

and

$$\sqrt{-0.000019 \times 11{,}367{,}000^2 + 3{,}918 \times 11{,}367{,}000} = 205{,}000.$$

Therefore, using formula (5), the approximate standard error of the estimated difference of 5,775,000 women is

$$\sqrt{248{,}000^2 + 205{,}000^2} = 322{,}000.$$

This means that the 90-percent confidence interval for the difference between the number of women divorced and widowed after their first marriage is from 5,259,800 to 6,290,200. A conclusion that the average estimate of the difference derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples. Since this interval does not contain zero, we can conclude with 90 percent confidence that the number of women divorced after their first marriage is greater than the number of women widowed after their first marriage.

**Standard error of a median.** The sampling variability of an estimated median depends on the form of the distribution and the size of the base. One can approximate the reliability of an estimated median by determining a confidence interval about it. (See the section on sampling variability for a general discussion of confidence intervals.)

Estimate the 68-percent confidence limits of a median based on sample data using the following procedure.

1. Determine, using formula (4), the standard error of the estimate of 50 percent from the distribution.

2. Add to and subtract from 50 percent the standard error determined in step 1.

3. Using the distribution of the characteristic, determine upper and lower limits of the 68-percent confidence interval by calculating values corresponding to the two points established in step 2.

   Use the following formula to calculate the upper and lower limits.

$$X_{pN} = \frac{pN - N_1}{N_2 - N_1}(A_2 - A_1) + A_1$$

   where

   $X_{pN}$ = estimated upper and lower bounds for the confidence interval ($0 \leq p \leq 1$). For purposes of calculating the confidence interval, p takes on the values determined in step 2. Note that $X_{pN}$ estimates the median when $p = 0.50$.

   N = for distribution of numbers: the total number of units (persons, households, etc.) for the characteristic in the distribution.

   = for distribution of percentages: the value 1.0.

   p = the values obtained in step 2.

   $A_1$, $A_2$ = the lower and upper bounds, respectively, of the interval containing $X_{pN}$.

   $N_1$, $N_2$ = for distribution of numbers: the estimated number of units (persons, households, etc.) with values of the characteristic greater than or equal to $A_1$ and $A_2$, respectively.

   = for distribution of percentages: the estimated percentage of units (persons, households, etc.) having values of the characteristic greater than or equal to $A_1$ and $A_2$, respectively.

4. Divide the difference between the two points determined in step 3 by two to obtain the standard error of the median.

Table A-3.
**Parameters and Factors for Remarriage Among Women: June 1985**

| Characteristic | a | b | Factor |
|---|---|---|---|
| Total or White . . . | -0.000019 | 3,918 | 1.00 |
| Black . . . . . . . . . . | -0.000235 | 5,620 | 1.43 |
| Hispanic origin. . . | -0.000049 | 9,471 | 2.42 |

**Illustration.** Table B shows the median age at divorce, for women whose first marriage ended in divorce, is 27.7. The base of the distribution from which this median was determined is 17,142,000.

1. Using formula (4), and the appropriate b parameter from table A-3, the standard error of 50 percent on a base of 17,142,000 is

$$\sqrt{3{,}918 \times 50.0 \times (100 - 50.0)/17{,}142{,}000} = 0.8 \text{ percent.}$$

2. To obtain a 68-percent confidence interval on an estimated median, add to and subtract from 50 percent the standard error found in step 1. This yields percentage limits of 49.2 and 50.8.

3. An unpublished table shows that 64.0 percent of women divorced after their first marriage were 25 years of age or older and 39.6 percent were 30 years of age or older. The upper limit of the estimate may be found to be

$$\frac{49.2 - 64.0}{39.6 - 64.0}(30-25) + 25 = 28.0.$$

   Similarly, since the 50.8 percentage falls within the same age category, the lower limit of the estimate may be found to be

$$\frac{50.8 - 64.0}{39.6 - 64.0}(30-25) + 25 = 27.7.$$

   Thus, the 68-percent confidence interval for the median age at divorce of women whose first marriage ended in divorce is from 27.7 to 28.0 years.

4. The standard error of the median is

$$\frac{28.0 - 27.7}{2} = 0.15 \text{ years.}$$

☆ U.S. GOVERNMENT PRINTING OFFICE: 1991 - 287-465 - 814/21340