Appendix Source and Accuracy of Estimates

SOURCE OF DATA

Most estimates in this report come from data obtained from the Current Population Survey (CPS) conducted in June of years 1970 through 1990. The Bureau of the Census conducts the survey every month, although this report uses mostly June data for its estimates. Also, some estimates come from March CPS data. The March and June surveys use two sets of questions, the basic CPS and the supplements.

Basic CPS. The basic CPS collects primarily labor force data about the civilian noninstitutional population. Interviewers ask questions concerning labor force participation about each member 15 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 60,000 occupied housing units are eligible for interview every month. Interviewers are unable to obtain interviews at about 2,600 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. These redesigns have improved the quality and reliability of the data and have satisfied changing data needs. The most recent changes were completely implemented in July 1985.

The following table summarizes changes in the CPS designs for the years for which data appear in this report.

Description of Current Population Survey

Time period	Number of	Housing units eligible ¹		
Time period	sample areas	Inter- viewed	Not inter- viewed	
1990. 1985. 1980. 1975.	729 ² 629/729 629 461 449	57,400 57,000 65,500 46,500 48,000	2,600 2,500 3,000 2,500 2,000	

¹Excludes about 2,500 Hispanic households added from the previous November sample. (See "March Supplement.")

²The CPS was redesigned following the 1980 Decennial Census of Population and Housing. During phase-in of the new design, housing units from the new and old designs were in the sample.

June Supplement. In addition to the basic CPS questions, interviewers asked supplementary questions in June about marriage, divorce, and fertility of women 15 to 65 years old.

March Supplement. In addition to the basic CPS questions, interviewers asked supplementary questions in March about family living arrangements.

To obtain more reliable data for the Hispanic population, the March CPS sample was increased by about 2,500 eligible housing units. These housing units were interviewed the previous November and contained at least one sample person of Hispanic origin. In addition, the sample included persons in the Armed Forces living off post or with their families on post.

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 decennial censuses of population; statistics on births, deaths, immigration, and emigration; and statistics on the size of the Armed Forces.

The independent population estimates used for 1981 to present were based on updates to controls established by the 1980 Decennial Census. Data before 1981 were based on independent population estimates from the most recent decennial census. For more details on the change in independent estimates, see the section entitled "Introduction of 1980 Census Population Controls" in an earlier report (Series P-60, No. 133). The estimation procedure for the March supplement included a further adjustment so the husband and wife of a household received the same weight.

The estimates in this report for 1985 and later also employ a revised survey weighting procedure for persons of Hispanic origin. In previous years, weighted sample results were inflated to independent estimates of the noninstitutional population by age, sex, and race. There was no specific control of the survey estimates for the Hispanic population. Since then, the Bureau of the Census developed independent population controls for the Hispanic population by sex and detailed age groups. Revised weighting procedures incorporate these new controls. The independent population estimates include some, but not all, undocumented immigrants.

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 errors: sampling and nonsampling. The accuracy of an estimate depends on both types of errors, 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. There are several sources of nonsampling errors, including the following:

- · Inability to get information about all sample cases.
- · Definitional difficulties.
- Differences in interpretation of questions.
- Respondents' inability or unwillingness to provide correct information.
- Respondents' inability to recall information.

- Errors made in data collection, such as recording and coding data.
- · Errors made in processing the data.
- · Errors made in estimating values for missing data.
- Failure to represent all units with the sample (undercoverage).

CPS undercoverage results from missed housing units and missed persons within sample households. Compared with 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 caused by 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, 1978.

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.

Caution should also be used when comparing estimates in this report (which reflect 1980 census-based population controls) with estimates for 1980 and earlier years (which reflect 1970 census-based population controls). This change in population controls had relatively little impact on summary measures such as means, medians, and percent distributions. It did have a significant impact on levels. For example, use of 1980-based population controls results in about a 2-percent increase in the civilian noninstitutional population and in the number of families and households. Thus, estimates of levels for data collected in 1981 and later years will differ from those for earlier years by more than what

could be attributed to actual changes in the population. These differences could be disproportionately greater for certain subpopulation groups than for the total population.

Since no independent population control totals for persons of Hispanic origin were used before 1985, compare Hispanic estimates over time cautiously.

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. These estimates may not be reliable for 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 next, are primarily measures of sampling variability, although they may include some nonsampling errors.

Standard Errors and Their Use. A number of approximations are required to derive, at a moderate cost, standard errors applicable to all the estimates in this report. 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.

Table 1 provides standard errors of estimated numbers. Tables 2 and 3 provide standard errors of estimated percentages. Table 4 has standard error parameters for persons, families, households, householders, and unrelated individuals. Table 4 also provides factors to apply to the standard errors in tables 1 through 3.

The sample estimate and its standard error enable one to construct a confidence interval. A confidence interval is 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 1.645 standard errors below the estimate to 1.645 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 in the report 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 be used to perform hypothesis testing. This is a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis appearing in this report is that the population parameters are different. An example of this would be comparing Whites with Blacks.

Tests may be performed at various levels of significance. The significance level of a test 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 1.645 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

$$s_x = fs$$
 (1)

Table 1. Standard Errors of Estimated Numbers: Living Arrangements —Total or White

(Numbers in thousands)

Size of estimate	Standard error
25	11
50	15
100	22
250	35
500	49
1,000	69
2,500	109
5,000	153
10,000	213
25,000	322
50,000	417
75,000	461
100,000	467
125,000	438
150,000	364

Note: Apply the square root of the factors in table 4 to the above standard errors for other characteristics.

where f is a factor from table 4, and s is the standard error of the estimate obtained by interpolation from table 1. The second method uses formula (2), from which the standard errors in table 1 were calculated. This formula will provide more accurate results than formula (1).

$$s_{x} = \sqrt{ax^{2} + bx} \tag{2}$$

Here, x is the size of the estimate and a and b are the parameters in table 4 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 that will give the largest standard error.

Illustration

Suppose that 46,503,000 children under 18 years lived with two parents. Use the appropriate parameters from table 4 and formula (2) to get

Number, x	46,503,000
a parameter	-0.000026
b parameter	4,785
Standard error	408,000
90% conf. int.	45,832,000 to
	47,174,000

The standard error is calculated as

$$s_x = \sqrt{-0.000026 \ x \ 46,503,000^2 + 4,785x46,503,000} = 408,000$$

The 90-percent confidence interval is calculated as $46,503,000 \pm 1.645 \times 408,000$.

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.

The alternate calculation of the standard error, using formula (1) with f=1.0 from table 4 and s=404,000 by interpolation from table 1, is

$$s_x = 1.0 \times 404,000 = 404,000$$

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. Estimated 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 4 indicated by the numerator.

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

$$s_{x,p} = fs$$
 (3)

In this formula, f is the appropriate factor from table 4, and s is the standard error of the estimate obtained by interpolation from table 2 or table 3.

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

$$s_{x,p} = \sqrt{(b/x) p (100-p)}$$
 (4)

Here, x is the total number of persons, families, households, or unrelated individuals in the base of the percentage, p is the percentage ($0 \le p \le 100$), and b is the parameter in table 4 associated with the characteristic in the numerator of the percentage.

Illustration

Suppose that of a total of 39,732,000 White children under 18 years, 83.0 percent lived with their biological parents. Use the appropriate parameter from table 4 and formula (4) to get

Percentage, p	83.0
Base, x	39,732,000
b parameter	4,785
Standard error	0.4
90% conf. int.	82.3 to 83.7

The standard error is calculated as

$$s_{x,p} = \sqrt{\frac{4,785}{39,732,000} \times 83.0 \times (100.0 - 83.0)} = 0.4$$

The 90-percent confidence interval for the percentage of White children under 18 years who lived with their biological parents is calculated as $83.0 \pm 1.645 \times 0.4$.

The alternate calculation of the standard error, using formula (3), with f=1.0 from table 4 and s=0.4 by interpolation from table 2, is

$$s_{x,p} = 1.0 \times 0.4 = 0.4$$

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

$$s_{x-y} = \sqrt{s_x^2 + s_y^2}$$
 (5)

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.

Table 2. Standard Errors of Estimated Percentages: Living Arrangements—Total or White

Base of estimated percentage	Estimated percentage						
(thousands)	1 or 99	2 or 98	5 or 95	10 or 90	25 or 75	35 or 65	50
25	4.4	6.1	9.5	13.1	18.9	20.9	21.9
50	3.1	4.3	6.7	9.3	13.4	14.8	15.5
100	2.2	3.1	4.8	6.6	9.5	10.4	10.9
250	1.4	1.9	3.0	4.2	6.0	6.6	6.9
500	1.0	1.4	2.1	2.9	4.2	4.7	4.9
1,000	0.7	1.0	1.5	2.1	3.0	3.3	3.5
2,500	0.4	0.6	1.0	1.3	1.9	2.1	2.2
5,000	0.3	0.4	0.7	0.9	1.3	1.5	1.6
10,000	0.2	0.3	0.5	0.7	1.0	1.0	1.1
25,000	0.14	0.2	0.3	0.4	0.6	0.7	0.7
50,000	0.10	0.14	0.2	0.3	0.4	0.5	0.5
75,000	0.08	0.11	0.2	0.2	0.4	0.4	0.4
100,000	0.07	0.10	0.2	0.2	0.3	0.3	0.4
125,000	0.06	0.09	0.13	0.2	0.3	0.3	0.3
150,000	0.06	0.08	0.12	0.2	0.2	0.3	0.3

Note: Apply the square root of the factors in table 4 to the above standard errors for other characteristics.

Table 3. Standard Errors of Estimated Percentages: Women—Total or White

Base of estimated percentage	Estimated percentage						
(thousands)	1 or 99	2 or 98	5 or 95	10 or 90	25 or 75	35 or 65	50
25	2.8	4.0	6.2	8.6	12.3	13.6	14.2
50	2.0	2.8	4.4	6.0	8.7	9.6	10.1
100	1.4	2.0	3.1	4.3	6.2	6.8	7.1
250	0.9	1.3	2.0	2.7	3.9	4.3	4.5
500	0.6	0.9	1.4	1.9	2.8	3.0	3.2
1,000	0.4	0.6	1.0	1.4	2.0	2.2	2.2
2,500	0.3	0.4	0.6	0.8	1.2	1.4	1.4
5,000	0.2	0.3	0.4	0.6	0.9	1.0	1.0
10,000	0.14	0.2	0.3	0.4	0.6	0.7	0.7
25,000	0.09	0.13	0.2	0.3	0.4	0.4	0.4
50,000	0.06	0.09	0.14	0.2	0.3	0.3	0.3

Note: Apply the square root of the factors in table 4 to the above standard errors for other characteristics.

Table 4. Parameters and Factors for the Report: Marriage, Divorce, and Remarriage in the 1990's

Characteristic	Persons			Families		
	a	b	f	а	b	f
Living Arrangements Total or White Black Hispanic. Number of Women	-0.000026 -0.000283 -0.000567	.,	1.00 1.43 1.43	-0.000011 -0.000071 -0.000142	1,899 1,716 1,716	•
Total or White	-0.000038 -0.000279 -0.000280	2,030	1.00 1.00 1.69	(X) (X) (X)	(X) (X) (X)	(X) (X) (X)

X Not applicable.

Note: Apply the factors 0.94, 0.84, 0.73, and 0.73 to the above parameters for data from 1985, 1980, 1975, and 1970.

Illustration		x	У	difference
	Number	32,975,000	2,336,000	30,639,000
Suppose that 32,975,000 White children under 18	a parameter	-0.000026	-0.000283	-
years lived with their biological parents and 2,336,000	b parameter	4,785	6,864	-
Black children under 18 years lived with their biological	Standard error	360,000	120,000	379,000
parents. Use the appropriate parameters from table 4 and formulas (2) and (5) to get	90% conf. int.	32,383,000 to 33,567,000	2,139,000 to 2,533,000	30,016,000 to 31,262,000

The standard error of the difference is calculated as

$$s_{x-y} = \sqrt{360,000^2 + 120,000^2} = 379,000$$

The 90-percent confidence interval around the difference is calculated as $30,639,000\pm1.645$ x 379,000. Since this interval does not contain zero, we can conclude, at the 10-percent significance level, that the number of White children under 18 years who lived with their biological parents is greater than the number of Black children who lived with their biological parents.

Standard Error of a Mean for Grouped Data. The formula used to estimate the standard error of a mean for grouped data is

$$s_x^- = \sqrt{(b/y) s^2} \tag{6}$$

In this formula, y is the size of the base of the distribution and b is a parameter from table 4. The variance, s², is given by the following formula:

$$s^2 = \sum_{i=1}^{c} p_i \bar{x_i} - \bar{x^2}$$
 (7)

where,

x, the mean of the distribution, is estimated by

$$\bar{x} = \sum_{i=1}^{c} p_i \bar{x}_i$$
 (8)

- c is the number of groups; i indicates a specific group, thus taking on values 1 through c.
- p_i is the estimated proportion of households, families or persons whose values, for the characteristic (x-values) being considered, fall in group i.

 x_i is $(Z_{i-1} + Z_i)/2$ where Z_{i-1} and Z_i are the lower and upper interval boundaries, respectively, for group i.

 x_i is assumed to be the most representative value for the characteristic for households, families, and unrelated individuals or persons in group i. Group c is open-ended, i.e., no upper interval boundary exists. For this group the approximate average value is

$$\bar{X}_{c} = \frac{3}{2} Z_{c-1} \tag{9}$$

Standard Error of a Ratio. Certain estimates may be calculated as the ratio of two numbers. The standard error of a ratio, x/y, may be computed using

$$s_{x/y} = \frac{x}{y} \sqrt{\left[\frac{s_x}{x}\right]^2 + \left[\frac{s_y}{y}\right]^2 - 2r \frac{s_x}{x} \frac{s_y}{y}}$$
 (10)

The standard error of the numerator, s_x , and that of the denominator, s_y , may be calculated using formulas described earlier. In formula (10), r represents the correlation between the numerator and the denominator of the estimate.

For one type of ratio, the denominator is a count of families or households and the numerator is a count of persons in those families or households with a certain characteristic. If there is at least one person with the characteristic in every family or household, use 0.7 as an estimate of r. An example of this type is the mean number of children per family with children.

For all other types of ratios, r is assumed to be zero. If r is actually positive (negative), then this procedure will provide an overestimate (underestimate) of the standard error of the ratio. Examples of this type are the mean number of children per family and the poverty rate.

NOTE: For estimates expressed as the ratio of x per 100 y or x per 1,000 y, multiply formula (10) by 100 or 1,000, respectively, to obtain the standard error.

Illustration

Suppose the ratio of male movers from abroad, x, to female movers from abroad, y, is 1.28. Use the appropriate parameters from table 4 (7,130 is from another report). The standard error of this ratio is calculated as follows:

	×	у	ratio
Estimate	641,000	501,000	1.28
a parameter	-0.000025	-0.000025	•
b parameter	7,130	7,130	-
Standard error	68,000	60,000	0.20
90% conf. int.	-	-	0.95 to 1.61

Using formula (10) with r=0, the estimate of the standard error is

$$s_{x/y} = \frac{641,000}{501,000} \sqrt{\left[\frac{68,000}{641,000}\right]^2 + \left[\frac{60,000}{501,000}\right]^2} = 0.20$$

The 90-percent confidence interval is calculated as 1.28 \pm 1.645 x 0.20.

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 standard errors and their use 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.
- 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.

$$x_{pN} = \frac{pN - N_1}{N_2 - N_1} (A_2 - A_1) + A_1$$
 (11)

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_{DN} .

 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 A₁ and A₂, respectively.

4. Divide the difference between the two points determined in step 3 by 2 to obtain the standard error of the median.

Illustration

A recent report by the Bureau of the Census¹ shows the following distribution and median income for families in 1989.

Income levels	Families
Total	66,090
Under \$5,000	2,398
\$5,000 to \$9,999	4,141
\$10,000 to \$14,999	5,354
\$15,000 to \$19,999	5,565
\$20,000 to \$24,999	5,461
\$25,000 to \$29,999	5,576
\$30,000 to \$34,999	5,294
\$35,000 to \$39,999	4,959
\$40,000 to \$44,999	4,464
\$45,000 to \$49,999	3,689
\$50,000 to \$54,999	3,545
\$55,000 to \$59,999	2,595
\$60,000 to \$64,999	2,278
\$65,000 to \$69,999	1,839
\$70,000 to \$74,999	1,463
\$75,000 to \$79,999	1,251
\$80,000 to \$84,999	1,036
\$85,000 to \$89,999	774
\$90,000 to \$94,999	695
\$95,000 to \$99,999	518
\$100,000 and over	3,197
Median income(dollars)	\$34,213

- 1. Using formula (4) with b = 2,058, the standard error of 50 percent on a base of 66,090,000 is about 0.3 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 percent limits of 49.7 and 50.3.
- 3. The lower and upper limits for the interval in which the median falls are \$30,000 and \$35,000, respectively.

Then, by addition, the estimated numbers of families with an income greater than or equal to \$30,000 and \$35,000 are 37,597,000 and 32,303,000, respectively.

Using formula (11), the upper limit for the confidence interval of the median is found to be about

$$\frac{0.497 \times 66,090,000 - 37,597,000}{32.303.000 - 37.597,000} \times (35,000 - 30,000) + 30,000 = 34,500$$

Similarly, the lower limit is found to be about

$$\frac{0.503x66,090,000-37,597,000}{32,303.000-37,597,000} \times (35,000-30,000) + 30,000 = 34,100$$

Thus, a 68-percent confidence interval for the median income for families is from \$34,100 to \$34,500.

4. The standard error of the median is, therefore,

$$\frac{34,500-34,100}{2}=200$$

¹U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 168, *Money Income and Poverty Status in the United States:* 1989 (Advance Data from the March 1990 Current Population Survey) U.S. Government Printing Office, Washington, DC, 1990.