Appendix C. Source and Reliability of Estimates

SOURCE OF DATA

Most of the estimates in this report are based on data collected in October 1981 and 1980 from the Current Population Survey (CPS) conducted by the Bureau of the Census. The monthly CPS deals mainly with labor force data for the civilian noninstitutional population. Questions relating to labor force participation are asked about each member 14 years old and over in each sample household. In addition, supplementary questions regarding school enrollment are asked about all eligible household members 3 years old and over each October.

The present CPS sample was initially selected from the 1970 census file and is updated continuously to reflect new construction where possible. The October 1981 and 1980 samples were spread over 629 areas comprising 1,133 counties, independent cities, and minor civil divisions with coverage in each of the 50 States and the District of Columbia. The October 1981 sample is composed of approximately 60,000 occupied households that are eligible for interview. Of this number, about 2,000 occupied units were visited, but interviews were not obtained because the occupants were not found at home after repeated calls or were unavailable for some other reason. The following table provides a description of some aspects of the CPS sample designs in use during the referenced data-collection period.

Table C-1. **Description of the October Current Population Survey: October 1947 to 1981**

Time period	Sample areas	Housing units eligible	
		Interviewed	Not interviewed
October 1981	629	58,000	2,000
October 1980	629	63,500	2,500
October 1977, 1978, and 1979	614	54,000	2,500
October 1972 to October 1976	461	45,000	2,000
October 1969, 1970, and 1971	449	48,000	2,000
October 1965	357	33,500	1,500
October 1960	333	33,500	1,500
October 1955	230	21,000	500 to 1,000
October 1947 and October 1950	68	21,000	500 to 1,000

The estimation procedure used in this survey involved the inflation of the weighted sample results to independent estimates of the total civilian noninstitutional population of the United States by age, race, and sex. These independent estimates were based on statistics from decennial censuses; statistics on births, deaths, immigration and emigration; and statistics on the strength of the Armed Forces. The independent population estimates used to obtain data for 1981 in this report are based on the 1980 Decennial Census, except for some of the data in table 1 which were obtained using independent population estimates based on the 1970 Decennial Census. Any display of 1981 data using 1970 census based population estimates is so noted. The data in the detailed tables 17 through 32 for 1980 are derived from independent estimates based on the 1970 Decennial Census. Also, in earlier reports in this series (P-20), data for 1972 through 1979 were obtained using independent population estimates based on the 1970 Decennial Census. Estimates for earlier years were based on earlier censuses. For more details on this change, see the section in appendix B entitled **Introduction of 1980 Census Population Controls.**

RELIABILITY OF THE ESTIMATES

Since the CPS estimates are based on a sample, they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaires, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey—sampling and nonsampling. The standard errors provided in this report primarily indicate the magnitude of the sampling error. Also, they partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The full extent of nonsampling error is unknown. Consequently, particular care should be exercised in the interpretation of figures based on a relatively small number of cases or on small differences between estimates.

Nonsampling Variability—As in any survey work, the results are subject to errors of response and nonreporting in addition to sampling variability. Nonsampling errors can be attributed to many sources, e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness of respondents to provide correct information, inability to recall information, errors made in collection (such as in recording or coding data), errors made in processing the data, errors made in estimating values for missing data, and failure to represent all sample households and all persons within sample households (undercoverage).

Undercoverage in the CPS results from missed housing units and missed persons within sample households. Overall undercoverage, as compared to the level of the 1980 Decennial Census, is about 7 percent. It is known that 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. Ratio estimation to independent age-sex-race population controls, as described previously, partially corrects for the biases due to survey undercoverage. However, biases exist in estimates to the extent that missed persons in missed households or missed persons in interviewed households have different characteristics than interviewed persons in the same age-sex-race group. Further, the independent population controls used have not been adjusted for undercoverage in the decennial census.

Comparability With Earlier CPS Data—In using metropolitan and nonmetropolitan data, caution should be used in comparing estimates for 1977 and 1978 to each other or to any other years. Methodological and sample design changes occurred in these years resulting in relative large differences in the metropolitan and nonmetropolitan area estimates. However, estimates for 1979 and later are comparable as are estimates for 1976 and earlier.

A number of changes were made in data collection and estimates procedures in 1980 and 1981. In 1980 the major change was the use of the "householder" concept instead of the traditional "head" concept. The other major change occurred in 1981 in the estimation procedure. Also, due to this later change, caution should be used when comparing estimates for 1981 and later, which reflect 1980 census-based population controls, to those for 1971 through 1980, which reflect 1970 census-based population controls. The change in population controls had relatively little impact on summary measures such as means and medians, but it did have a significant impact on levels. For example, use of 1980-based population controls resulted in about a 2-percent increase in the civilian noninstitutional population and in the number of families and households. Thus, estimates of levels for 1981 and later may differ from those for 1980 and earlier 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.

Sampling Variability—The standard errors given in tables C-2 through C-5 are primarily measures of sampling variability, that is, of the variation that occurred by chance because a sample rather than the entire population was surveyed. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples of the same size with a known probability. For example, if all possible samples of the same size were selected, each of these being 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:

1. Approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average results of all possible samples.
2. 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.
3. Approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples of the same size is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples of the same size is included in the confidence interval.

Also, standard errors may be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common types of hypotheses are: (1) the population parameters are identical, versus (2) they are different. An example of this would be comparing the percent of adults who were high school graduates in 1981 to those in 1971. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the parameters are different when, in fact, they are identical. All statements of comparison in the text have passed a hypothesis test at the 0.10 level of significance or better, and most have passed a hypothesis test at the 0.05 level of significance or better. This means that, for most differences cited in the text, the estimated difference between parameters is greater than twice the standard error of the difference. For the other differences mentioned, the estimated difference between parameters is between 1.6 and 2.0 times the standard error of the difference. When this is the case, the statement of comparison will be qualified in some way; e.g., by the use of the phrase "some evidence."

Using Small Estimates—Summary measures (such as means, medians, and percent distributions) are shown when the base is 75,000 or greater. Because of the large standard errors involved, there is little chance that summary measures would reveal useful information when computed on a smaller base. Estimated numbers are shown, however, even though the relative standard errors of these numbers are larger than those for the corresponding percentages. These smaller estimates are provided primarily to permit such combinations of the categories as serve each user's needs.

Standard Error Tables and Their Use—In order to derive standard errors that would be applicable to a large number of estimates, a number of approximations were required. Therefore, instead of providing an individual standard error for each estimate, generalized sets of standard errors are provided for various types of characteristics. As a result the sets of standard errors provided give an indication of the order of magnitude of the standard error of an estimate rather than the precise standard error.

The figures presented in tables C-2 through C-5 are approximations to standard errors of various school enrollment estimates for persons in the United States. Estimated standard errors for specific characteristics cannot be obtained from tables C-2 through C-5 without the use of the factors in table C-6. The factors in table C-6 must be applied to the generalized standard errors in order to adjust for the combined effect of sample design and estimating procedure on the value of the characteristic. Standard errors for intermediate values not shown in the generalized tables of standard errors may be approximated by linear interpolation.

Two parameters (denoted "a" and "b") are used to calculate standard errors for each type of characteristic; they are also presented in table C-6. These parameters were used to calculate the factors in table C-6. They also may be used to directly calculate the standard errors for estimated numbers and percentages. Methods for direct computation are given in the following sections. The standard errors in tables C-2 through C-5 were calculated using the "b" parameters in table C-6; however, the "a" parameters used for tables C-2 and C-3 were revised to reflect the total persons in the age group.

Standard Errors of Estimated Numbers—The approximate standard error σ_x, of an estimated number can be obtained in two ways. It may be obtained by use of the formula

$$\sigma_x = f \cdot \sigma \qquad (1)$$

where "f" is the appropriate factor from table C-6 and σ is the standard error on the estimate obtained by interpolation from tables C-2 or C-3. Alternatively, standard errors may be approximated by using formula (2). The use of formula (1) will provide more accurate results when the number of persons in the age group is relatively small, otherwise the following formula will provide more accurate standard errors.

$$\sigma_x = \sqrt{ax^2 + bx} \qquad (2)$$

Here x is the size of the estimate and "a" and "b" are the parameters in table C-6 associated with the particular type of characteristic. When an estimate involves two different categories use the "a" and "b" parameters corresponding to the category with the larger "b" parameter.

Standard Errors of Estimated Percentages—The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which this percentage is based. 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. The approximate standard error, $\sigma_{(x,p)}$, of an estimated percentage can be obtained by use of the formula

$$\sigma_{(x,p)} = f \cdot \sigma \qquad (3)$$

In this formula, "f" is the appropriate factor from table C-6 and σ is the standard error on the estimate from table C-4 or C-5. When the numerator and denominator of the percentage are in different categories, use the factor or parameters from table C-6 indicated by the numerator. Alternatively, standard errors may be approximated by using formula (4), from which the standard errors in tables C-4 and C-5 were calculated. Use of this formula will provide more accurate results than use of formula (3).

$$\sigma_{(x,p)} = \sqrt{\frac{b}{x} \cdot p(100-p)} \qquad (4)$$

Here x is the size of the subclass of persons or families and unrelated individuals which is the base of the percentage, p is the percentage ($0 < p < 100$), and b is the parameter in table C-6 associated with the particular type of characteristic in the numerator of the percentage.

Illustration of Standard Error Computations—Table C shows that in 1981 there were 18,266,000 persons 30 to 34 years old. It was estimated that 1,211,000 persons 30 to 34 years old were college students in 1981. From table C-2, interpolation shows the standard error to be about 47,000 for an estimate of this size. Applying the appropriate factor from table C-6, and using formula (1) yields an approximate standard error[1] of 1.0 x (47,000) = 47,000. An approximate 90-percent confidence interval as computed from these data is from 1,135,800 to 1,286,200 (i.e., 1,211,000 ± [1.6 x 47,000]). Similarly, the 95-percent confidence interval is from 1,117,000 to 1,305,000 (i.e., 1,211,000 ± [2 x 47,000]).

Therefore, a conclusion that the average estimate derived from all possible samples of this same size lies within a range computed in this way would be correct for roughly 95 percent of all possible samples.

Table 1 shows that 9.4 percent of the 9,660,000 men 25 to 29 years old were enrolled in college in October 1981. Using formula (4) and the appropriate "b" parameter from table C-6, the approximated standard error of the estimate[2] is:

$$\sqrt{\frac{2014}{9{,}660{,}000} \times 9.4\ (100\text{-}9.4)} \doteq .4$$

Thus, a 95-percent confidence interval for this estimate is from 8.6 to 10.2 (i.e., 9.4 ± (2 x .4).

Standard Error of a Difference—For a difference between two sample estimates, the standard error is approximately equal to

$$\sigma_{(x-y)} \doteq \sqrt{\sigma_x^2 + \sigma_y^2} \qquad (5)$$

where σ_x and σ_y are the standard errors of the estimates x and y; the estimates can be numbers, percent, ratios, etc. This formula approximates the standard error quite accurately for the difference between two estimates of the same characteristics in two different areas, or for the difference between two separate and uncorrelated characteristics in the same area. If, however, there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

Illustration of the Calculation of the Standard Error of a Difference—Table 1 shows that in October 1981 an estimated 9.4 percent of men 25 to 29 years old were enrolled in college as compared to 8.0 percent of women 25 to 29 years old. The approximate standard error of 9.4 percent was shown above to be .4. Similarly, the standard error of 8.0 percent is found to be approximately .4. The apparent difference between these two estimates is 1.4 percent. Thus, by formula (5) the standard error of the estimated difference, 1.4 percent, is:

$$\sqrt{(.4)^2 + (.4)^2} \doteq .6$$

This means that a 95-percent confidence interval about the difference of 1.4 is from .2 to 2.6 percent. Therefore, a conclusion that the average estimate derived from all possible samples of this same size lies within a range computed in this way would be correct for roughly 95 percent of all possible samples. Since this particular confidence interval does not include zero, it can be said with 95-percent confidence that the actual percent of men 25 to 29 years old enrolled in college is greater than the percent of women.

[1] Using formula (2) and table C-6 gives a standard error of approximately

$$\sqrt{-.000025(1{,}211{,}000)^2 + 2014(1{,}211{,}000)} \doteq 49{,}000$$

[2] Using formula (3), interpolation from table C-4, and multiplying by the appropriate factors from table C-6 gives a standard error of the estimate of approximately .4.

Table C-2. Generalized Standard Errors for Estimated Numbers of Persons—All Races or White

(In thousands. For meaning of symbols, see page 126)

Estimated number of persons	Total persons in age group									
	100	250	500	1,000	2,500	5,000	10,000	25,000	50,000	100,000
10	4.3	4.4	4.4	4.5	4.5	4.5	4.5	4.5	4.5	4.5
20	5.7	6.1	6.2	6.3	6.3	6.3	6.3	6.3	6.3	6.3
30	6.5	7.3	7.5	7.7	7.7	7.7	7.8	7.8	7.8	7.8
40	7.0	8.2	8.6	8.8	8.9	8.9	9.0	9.0	9.0	9.0
50	7.1	9.0	9.5	9.8	9.9	10.0	10.0	10.0	10.0	10.0
75	6.1	10.3	11.3	11.8	12.1	12.2	12.2	12.3	12.3	12.3
100	-	11.0	12.7	13.5	13.9	14.0	14.1	14.2	14.2	14.2
200	-	9.0	15.5	18.0	19.3	19.7	19.9	20.0	20.0	20.0
300	-	-	15.5	20.6	23.1	23.8	24.2	24.4	24.5	24.5
400	-	-	12.7	22.0	26.0	27.2	27.8	28.2	28.3	28.3
500	-	-	-	22.4	28.4	30.1	30.9	31.4	31.6	31.7
750	-	-	-	19.4	32.5	35.8	37.4	38.3	38.6	38.7
1,000	-	-	-	-	34.8	40.1	42.6	44.0	44.4	44.7
2,000	-	-	-	-	28.4	49.2	56.8	60.9	62.2	62.8
3,000	-	-	-	-	-	49.2	65.0	72.9	75.4	76.6
4,000	-	-	-	-	-	40.1	69.5	82.3	86.1	87.9
5,000	-	-	-	-	-	-	71.0	89.8	95.2	97.8
7,500	-	-	-	-	-	-	61.5	102.8	113.3	118.2
10,000	-	-	-	-	-	-	-	109.9	126.9	134.6
20,000	-	-	-	-	-	-	-	89.8	155.5	179.5
30,000	-	-	-	-	-	-	-	-	155.5	205.7
40,000	-	-	-	-	-	-	-	-	126.9	219.9
50,000	-	-	-	-	-	-	-	-	-	224.4
75,000	-	-	-	-	-	-	-	-	-	194.3
100,000	-	-	-	-	-	-	-	-	-	-

Note: These values must be multiplied by the appropriate "f" factor in table C-6 to obtain the standard error for a specific characteristic. To estimate the standard errors for the 1956-66 period, multiply these standard errors by 1.23. For years prior to 1956, multiply by 1.5.

Table C-3. Generalized Standard Errors for Estimated Numbers of Persons—Black and Other Races

(In thousands. For meaning of symbols, see page 126)

Estimated number of persons	Total persons in age group						
	100	250	500	1,000	2,500	5,000	10,000
10	4.5	4.7	4.7	4.7	4.7	4.8	4.8
20	6.0	6.5	6.6	6.7	6.7	6.7	6.7
30	6.9	7.7	8.0	8.1	8.2	8.2	8.2
40	7.4	8.7	9.1	9.3	9.4	9.5	9.5
50	7.5	9.5	10.1	10.4	10.5	10.6	10.6
75	6.5	10.9	12.0	12.5	12.8	12.9	13.0
100	-	11.7	13.5	14.3	14.7	14.9	15.0
200	-	9.5	16.5	19.0	20.4	20.9	21.1
300	-	-	16.5	21.8	24.5	25.3	25.7
400	-	-	13.5	23.3	27.6	28.9	29.5
500	-	-	-	23.8	30.1	31.9	32.8
750	-	-	-	20.6	34.5	38.0	39.6
1,000	-	-	-	-	36.9	42.6	45.1
2,000	-	-	-	-	30.1	52.1	60.2
3,000	-	-	-	-	-	52.1	69.0
4,000	-	-	-	-	-	42.6	73.7
5,000	-	-	-	-	-	-	75.2
7,500	-	-	-	-	-	-	65.2
10,000	-	-	-	-	-	-	-

Note: These values must be multiplied by the appropriate "f" factor in table C-6 to obtain the standard error for a specific characteristic. To estimate the standard errors for the 1956-66 period, multiply these standard errors by 1.23. For years prior to 1956, multiply by 1.5.

Table C-4. Generalized Standard Errors of Estimated Percentages—All Races or White

Base of percentage (thousands)	Estimated percentage				
	2 or 98	5 or 95	10 or 90	25 or 75	50
100	2.0	3.1	4.3	6.1	7.1
250	1.3	2.0	2.7	3.9	4.5
500	0.9	1.4	1.9	2.7	3.2
1,000	0.6	1.0	1.3	1.9	2.2
2,500	0.4	0.6	0.9	1.2	1.4
5,000	0.3	0.4	0.6	0.9	1.0
10,000	0.2	0.3	0.4	0.6	0.7
25,000	0.1	0.2	0.3	0.4	0.4
50,000	0.1	0.1	0.2	0.3	0.3
100,000	0.1	0.1	0.1	0.2	0.2
150,000	0.1	0.1	0.1	0.2	0.2

Note: These values must be multiplied by the appropriate "f" factor in table C-6 to obtain the standard error for a specific characteristic. To estimate the standard errors for the 1956-66 period, multiply these standard errors by 1.23. For years prior to 1956, multiply by 1.5.

Table C-5. Generalized Standard Errors of Estimated Percentages—Black and Other Races

Base of percentage (thousands)	Estimated percentages				
	2 or 98	5 or 95	10 or 90	25 or 75	50
75	2.4	3.8	5.2	7.5	8.7
100	2.1	3.3	4.5	6.5	7.5
250	1.3	2.1	2.9	4.1	4.8
500	0.9	1.5	2.0	2.9	3.4
1,000	0.7	1.0	1.4	2.1	2.4
2,500	0.4	0.7	0.9	1.3	1.5
5,000	0.3	0.5	0.6	0.9	1.1
10,000	0.2	0.3	0.5	0.7	0.8
15,000	0.2	0.3	0.4	0.5	0.6
20,000	0.1	0.2	0.3	0.5	0.5

Note: These values must be multiplied by the appropriate "f" factors in table C-6 to obtain the standard error for a specific characteristic. To estimate the standard errors for the 1956-66 period, multiply these standard errors by 1.23. For years prior to 1956, multiply by 1.5.

Table C-6. "a" and "b" Parameters and "f" Factor for Calculating Approximate Standard Errors of Estimated Numbers and and Percentages

(For meaning of symbols, see page 126)

Characteristic	Persons			Families or family heads		
	Parameter			Parameter		
	a	b	f factor	a	b	f factor
Persons enrolled in school, 14 to 34 years old:						
All races or White	-0.000025	2,014	1.0	-0.000010	1,389	0.8
Black	-0.000179	2,265	1.0	-0.000087	1,255	0.7
Spanish origin[1] (levels)	+0.001519	1,856	(X)	-0.000033	2,397	1.1
Spanish origin[1] (percentages only)	(X)	3,374	1.3	(X)	(X)	(X)
Children enrolled in school, 3 to 13 years old, all race groups and Spanish origin	-0.000063	2,350	1.1	(X)	(X)	(X)
Marital status, household, and family:						
All races or White	-0.000017	3,500	1.3	(X)	(X)	(X)
Black	-0.000210	5,020	1.5	(X)	(X)	(X)
Spanish origin[1]	-0.000043	7,469	1.9	(X)	(X)	(X)
Income:						
Persons tabulated by family:						
All races or White	-0.000018	3,770	1.4	(X)	(X)	(X)
Black	-0.000154	4,310	1.4	(X)	(X)	(X)
Spanish origin[1]	-0.000067	10,112	2.2	(X)	(X)	(X)
Residence:						
All races or White	-0.000024	4,253	1.5	-0.000016	2,170	1.0
Black	-0.000308	7,402	1.8	-0.000178	2,561	1.1
Spanish origin[1]	-0.000074	15,027	2.7	-0.000014	2,844	1.2

[1]To compute standard errors for school enrollment levels (number of persons), only use formula (2). To calculate any other Spanish standard errors either use formulas (1) or (2), or use the generalized standard errors from tables C-2 or C-4 in conjunction with formulas (1) or (3).

Note: (a) For nonmetropolitan data cross-tabulated with other data, also apply the factor 1.2 to the "f" factor and the factor 1.5 to the "a" and "b" parameters. (b) Multiply the "a" and "b" parameters by 1.5 for CPS data collected from 1956 to 1966 and by 2.25 for CPS data collected before 1956. (c) For western census region data cross-tabulated with other data, also apply the factor .92 to the "f" factor and .84 to the "b" parameter (for any race or Spanish Origin). Also multiply the "a" parameter by .84 for only the all races or white category.