Appendix B. Source and Reliability of the Estimates

SOURCE OF DATA

Most of the estimates in this report are based on data collected in March of 1970, 1975, 1978, and 1979 from the Current Population Survey (CPS) of the Bureau of the Census. Some estimates are based on data obtained in earlier years. 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 every sample household. In addition, supplementary questions are asked each March about various population characteristics. In order to obtain more reliable data for the Spanish-origin population, the March CPS sample was enlarged to include all households from the previous November sample which contained at least one person of Spanish origin. This results in almost doubling the number of sample persons of Spanish origin. For this report, persons in the Armed Forces living off post or with their families on post are included. Brief descriptions of the sources and the procedures by which the data were obtained are presented below.

The present CPS sample was initially selected from the 1970 census files and is continuously updated to reflect new construction where possible. (See section, "Nonsampling Variability.") The monthly CPS sample is spread over 614 areas with coverage in each of the 50 States and the District of Columbia. The CPS sample areas are comprised of 1,113 counties, independent cities, and minor civil divisions in the nation.

The estimation procedure used for the monthly CPS data involves the inflation of weighted sample results to independent estimates of the civilian noninstitutional population of the United States by age, race, and sex. These independent estimates are based on statistics from decennial censuses; statistics on births, deaths, immigration, and emigration; and statistics on the strength of the Armed Forces. The estimation procedure for the data in the report also involves a further adjustment so that husband and wife of a household receive the same weight.

RELIABILITY OF THE ESTIMATES

Since the estimates in this report 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

Description of the Current Population Survey for March Supplement

Time period	Number of sample area[1]	Households eligible		Housing units visited, not eligible[2]
		Interviewed	Not interviewed	
March 1979	614	55,000	3,000	10,500
March 1978	614	54,500	3,000	10,500
March 1977[3]	614	55,500	2,500	10,000
March 1973 to March 1976	461	46,500	2,500	8,500
March 1972	449	45,000	2,000	8,000
March 1967 to March 1971	449	48,000	2,000	8,500
March 1964 to March 1966	357	33,500	1,500	6,000
March 1962[4]	333	33,500	1,500	6,000
March 1957 to March 1959	330	33,500	1,500	6,000
March 1947, March 1950, and March 1952	68	21,000	500-1,000	3,000-3,500

[1]Beginning in May 1956, these areas were chosen to provide coverage in each State and the District of Columbia.

[2]These are housing units which were visited, but were found to be vacant or otherwise not eligible for interview.

[3]A supplementary sample of housing units in 24 States and the District of Columbia was incorporated with the monthly CPS to produce March 1977 data.

[4]Three sample areas were added in 1960 to represent Alaska and Hawaii after statehood.

errors provided for this report primarily indicate the magnitude of the sampling error. They also 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 to provide correct information on the part of respondents, inability to recall information, mistakes made in 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).

Undercoverage in the CPS results from missed housing units and missed persons within sample households. Overall undercoverage, as compared to the level of the decennial census, is about 5 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 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 the estimates to the extent that missed persons in missed households or missed persons in interviewed households have different characteristics than interviewed persons in the age-sex-race group. Further, the independent population controls used have not been adjusted for undercoverage in the 1970 census, which was estimated at 2.5 percent of the population, with differentials by age, sex, and race similar to those observed in CPS.

A coverage improvement sample was included in computing the estimates beginning in October 1978 in order to provide coverage of mobile homes and new construction housing units that previously had no chance for selection in the CPS sample. This sample is composed of approximately 450 sample household units which represent 237,000 occupied mobile homes and 600,000 new construction units. These new construction units are composed of those units where building permits were issued prior to January 1970 and construction was not completed by the time of the 1970 Census (i.e., April 1970). The extent of other sources of undercoverage of housing units is unknown but believed to be small. The inclusion of this coverage improvement sample in the CPS does not have a significant effect on the estimates.

Use of metropolitan and nonmetropolitan data. In using metropolitan and nonmetropolitan data, particular care should be exercised in comparing estimates from 1977 and later years to each other and to those from earlier years. Methodological and sample design changes have occurred in these recent years resulting in relatively large differences in the metropolitan and nonmetropolitan area estimates.

Decennial Census of Population. The 1940 census data shown in this report are based on a 100-percent sample and 1960 and 1970 census data are based on 5-percent samples from the census. Data obtained from the CPS and the census are not strictly comparable. This is due in a large part to differences in interviewer training and experience and in different survey processes. This is an additional component of error not reflected in the standard error tables. Therefore, caution should be used in comparing results between these different sources.

SAMPLING VARIABILITY. The standard errors given in the following tables are primarily measures of sampling variability, that is, of the variations that occurred by chance because a sample rather than the whole of the population was surveyed. The sample estimate and its estimated standard error enable one to construct interval estimates that include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these surveyed under identical conditions using the same sample design and an estimate and its estimated 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 result 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 result of all possible samples may or may not be contained in any particular computed interval. However, for a particular sample one can say with specified confidence that the average result of all possible samples is included within the constructed interval.

All the statements of comparison appearing in the text are significant at a 1.6 standard error level or better, and most are significant at a level of more than 2.0 standard errors. This means that for most differences cited in the text, the estimated difference is greater than twice the standard error of the difference. Statements of comparison qualified in some way (e.g., by use of the phrase "some evidence") have a level of significance between 1.6 and 2.0 standard errors.

Note when using small estimates. Summary measures (such as medians and percent distributions) are shown in the report only 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 those combinations of the categories which 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 and could be prepared at a moderate cost, 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 errors.

The figures presented in tables B-1 through B-4 provide approximations to standard errors of various estimates for total, White or Black persons in the total United States for education only. To obtain standard errors for other characteristics, factors from table B-5 must be applied to the standard errors given for education 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 tables may be approximated by interpolation.

Data based on the 1940, 1960 and 1970 censuses. Sampling errors of all sample data from the 1960, and Census of Population for 1940, 1960 and 1970 in this report are small enough to be desregarded. However, these standard errors may be found in the appropriate volumes.

Standard errors of estimated numbers. The approximate standard error, σ_x, of an estimated number shown in this report can be obtained by use of the formula

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

In this formula f is the appropriate factor from table B-5 and σ is the standard error for total or White persons in table B-1 or the standard error for Black persons in table B-2.

Standard errors of estimated percentages. The reliability of an estimated percentage, computed by using sample data for

Table B-1. Generalized Standard Errors for Estimated Numbers of Persons—Total or White

(68 chances out of 100. Numbers in thousands)

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	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
20	5.7	6.2	6.3	6.4	6.4	6.4	6.4	6.4	6.4	6.4
30	6.6	7.4	7.6	7.8	7.8	7.8	7.9	7.9	7.9	7.9
40	7.0	8.3	8.7	8.9	9.0	9.1	9.1	9.1	9.1	9.1
50	7.2	9.1	9.6	9.9	10.1	10.1	10.1	10.1	10.2	10.2
75	6.2	10.4	11.5	12.0	12.3	12.3	12.4	12.4	12.4	12.4
100	-	11.1	12.9	13.6	14.1	14.2	14.3	14.3	14.4	14.4
200	-	9.1	15.7	18.2	19.5	19.9	20.1	20.2	20.3	20.3
300	-	-	15.7	20.8	23.3	24.1	24.5	24.7	24.8	24.8
400	-	-	12.9	22.3	26.3	27.6	28.2	28.5	28.6	28.7
500	-	-	-	22.7	28.7	30.5	31.3	31.8	32.0	32.0
750	-	-	-	19.7	32.9	36.3	37.8	38.8	39.1	39.2
1,000	-	-	-	-	35.2	40.6	43.1	44.5	45.0	45.2
2,000	-	-	-	-	28.7	49.8	57.5	61.6	63.0	63.6
3,000	-	-	-	-	-	49.8	65.8	73.8	76.3	77.5
4,000	-	-	-	-	-	40.6	70.4	83.3	87.2	89.0
5,000	-	-	-	-	-	-	71.8	90.9	96.4	99.0
7,500	-	-	-	-	-	-	62.2	104.1	114.7	119.7
10,000	-	-	-	-	-	-	-	111.3	128.5	136.3
20,000	-	-	-	-	-	-	-	90.9	157.4	181.7
30,000	-	-	-	-	-	-	-	-	157.4	208.2
40,000	-	-	-	-	-	-	-	-	128.5	222.6
50,000	-	-	-	-	-	-	-	-	-	227.2
75,000	-	-	-	-	-	-	-	-	-	196.7
100,000	-	-	-	-	-	-	-	-	-	-

- Represents zero.

Note: 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.

both numerator and denominator, depends on 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:

$$\sigma_{(x,p)} = f\sigma \qquad (2)$$

In this formula f is the appropriate factor from table B-5 and σ is the standard error for total or White persons in table B-3 or the standard error for Black and other races in table B-4.

Illustration of the use of tables of standard errors. Table 1 of this report shows that in March 1979 there were 58,986,000 men 25 years old and over. At that time, an estimated 5,049,000 of them had completed elementary school and had not completed at least 1 year of high school. Table B-1 shows the standard error of an estimate of this size to be approximately 97,000. In this situation no factors have to be applied to the standard error, (i.e. f is equal to 1.0). The 68-percent confidence interval as shown by the data is from 4,952,000 to 5,146,000. 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 68 percent of all possible samples. Similarly, we could conclude with 95-percent confidence that the average estimate derived from all possible samples lies within the interval from 4,855,000 to 5,243,000, (i.e. 5,049,000 ± (2 x 97,000)).

Table 8 shows that in March 1979, 75.3 percent of the 12,961,000 Californians 25 years old and over were high school graduates. Table B-3 indicates the standard error on this percentage to be 0.6 percent. Applying the appropriate factor from table B-5 and using formula (2), the approximate standard error is 1.0 x 0.6 = 0.6 percent. Consequently, the 68-percent confidence interval as shown by these data is from 74.7 percent to 75.9 percent and the 95-percent confidence interval is from 74.1 percent to 76.5 percent.

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

$$\sigma_{(x-y)} \doteq \sqrt{\sigma_x^2 + \sigma_y^2} \qquad (3)$$

where σ_x and σ_y are the standard errors of the estimates x and y; the estimate can be of numbers, percents, medians, etc. This will represent the actual standard error quite accurately for the difference between two estimates of the same characteristic 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.

Illustrations of the computation of the standard error of a difference. Table 8 of this report also shows that in March 1979, 68.8 percent of the 10,436,000 persons in the State of New York 25 years old and over were high school graduates.

Table B-2. Generalized Standard Errors for Estimated Numbers of Persons—Black and Other Races

(68 chances out of 100. Numbers in thousands)

Estimated number of persons	Total persons in age group						
	100	250	500	1,000	2,500	5,000	10,000
10	5.0	5.2	5.2	5.3	5.3	5.3	5.3
20	6.7	7.2	7.3	7.4	7.4	7.5	7.5
30	7.7	8.6	8.9	9.0	9.1	9.1	9.1
40	8.2	9.7	10.1	10.4	10.5	10.5	10.5
50	8.4	10.6	11.2	11.5	11.7	11.8	11.8
75	7.2	12.1	13.3	13.9	14.3	14.4	14.4
100	-	12.9	14.9	15.9	16.4	16.5	16.6
200	-	10.6	18.3	21.1	22.7	23.2	23.4
300	-	-	18.3	24.2	27.1	28.1	28.5
400	-	-	14.9	25.9	30.6	32.1	32.7
500	-	-	-	26.4	33.4	35.4	36.4
750	-	-	-	22.9	38.3	42.2	44.0
1,000	-	-	-	-	40.9	47.3	50.1
2,000	-	-	-	-	33.4	57.9	66.8
3,000	-	-	-	-	-	57.9	76.6
4,000	-	-	-	-	-	47.3	81.9
5,000	-	-	-	-	-	-	83.5
7,500	-	-	-	-	-	-	72.4
10,000	-	-	-	-	-	-	-

- Represents zero.

Note: 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.

Thus, the apparent difference between California and New York in the percentage of people 25 years old and over who were high school graduates is 6.5 percent. The standard error (σ_x) of 75.3 percent is 0.7, as shown above. Table B-3 and the factor from table B-5 show the standard error (σ_y) of 68.8 percent with a base of 10,436,000 to be approximately 0.6 = 0.6 x 1.0 percent. Therefore, using formula (3), the standard error of the difference of 6.5 percent is about

$$.8 \doteq \sqrt{(0.6)^2 + (0.6)^2}$$

This means that the 68-percent confidence interval about the difference is from 5.7 percent to 7.3 percent, and the 95-percent confidence interval is from 4.9 percent to 8.1 percent. Since this interval does not contain zero, we can conclude with 95-percent confidence that there exists a difference between the percentage of high school graduates of California and New York who were 25 years old and over.

Standard error of a median. The sampling variability of an estimated median depends upon the form of the distribution as well as on the size of its base. An approximate method for measuring the reliability of a median is to determine an interval about the estimated median, such that there is a stated degree of confidence that the median based on a complete census lies within the interval. The following procedure may be used to estimate the 68-percent confidence limits of a median based on sample data.

1. Determine, using table B-3 or B-4 and the appropriate factors, the standard error on a 50-percent characteristic;
2. add to and subtract from 50 percent the standard error determined in step (1);
3. using the distribution of the characteristic, calculate the 68-percent confidence interval by finding the values corresponding to the two points established in step (2).

A 95-percent confidence interval may be determined by finding the values corresponding to 50 percent plus and minus twice the standard error determined in step (1).

Illustration of the computation of a confidence interval for a median. Table 8 of this report shows that the median number of school years completed by Californians 25 years old and over is 12.8.

1. There was a total of 12,961,000 Californians 25 years old and over. Using formula (2) and tables B-3 and B-5 the standard error of a 50-percent characteristic is found to be approximately 0.7 = 1.0 x 0.7 percent.
2. To obtain a 95-percent confidence interval on an estimated median, add to and subtract from 50 percent twice the standard error found in step (1). This yields percent limits of 48.6 and 51.4.
3. From table 8 the percent of Californians 25 years old and over who completed 12 years of school is 33.0 = 75.3 – 42.3 percent and 24.7 percent completed less than 12 years of school. By linear interpolation, the lower limit on the estimate is found to be about

$$12.0 + (13.0 - 12.0)\frac{48.6 - 24.7}{33.0} \doteq 12.7 \text{ years}$$

Similarly, the upper limit may be found by linear interpolation to be about

$$12.0 + (13.0 - 12.0)\frac{51.4 - 24.7}{33.0} \doteq 12.8 \text{ years}$$

Hence, the 95-percent confidence interval about the estimated median is from 12.7 to 12.8 years.[1]

Standard error of a 3-year moving average. To calculate the standard error of 3-year moving average, first find the standard error for the estimate of the middle year using the appropriate table(s). Then divide this number by the square root of 3 (1.7) to obtain the approximate standard error of the average.

[1] Note, the estimated median and the upper limit of the 95-percent confidence interval are identical. This result can be attributed to the rounding of the estimated median (from 12.76 to 12.8) and the upper limit (from 12.81 to 12.8).

Table B-3. Generalized Standard Errors of Estimated Percentages—Total or White Population

(68 chances out of 100)

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	6.2	7.2
250	1.3	2.0	2.7	3.9	4.5
500	0.9	1.4	1.9	2.8	3.2
1,000	0.6	1.0	1.4	2.0	2.3
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	0.13	0.2	0.3	0.4	0.5
50,000	0.09	0.14	0.2	0.3	0.3
100,000	0.06	0.10	0.14	0.2	0.2
150,000	0.05	0.08	0.11	0.2	0.2

Note: 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 B-4. Generalized Standard Errors of Estimated Percentages—Black and Other Races

(68 chances out of 100)

Base of percentage (thousands)	Estimated percentage				
	2 or 98	5 or 95	10 or 90	25 or 75	50
75	2.7	4.2	5.8	8.4	9.6
100	2.3	3.6	5.0	7.2	8.4
250	1.5	2.3	3.2	4.6	5.3
500	1.0	1.6	2.2	3.2	3.7
1,000	0.7	1.2	1.6	2.3	2.6
2,500	0.5	0.7	1.0	1.4	1.7
5,000	0.3	0.5	0.7	1.0	1.2
10,000	0.2	0.4	0.5	0.7	0.8
15,000	0.2	0.3	0.4	0.6	0.7
20,000	0.2	0.3	0.4	0.5	0.6

Note: 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 B-5. Factors to be Applied to Generalized Standard Errors in Tables B-1 Through B-4

Type of characteristic	Total or White	Black[2]	Spanish origin[1]
States:			
California	1.0	1.0	1.1
Florida	1.0	1.0	1.1
Georgia	1.1	1.1	1.1
Illinois	1.0	1.0	1.1
Indiana	1.0	1.0	1.1
Massachusetts	1.0	1.0	1.1
Michigan	1.0	1.0	1.1
Missouri	1.0	1.0	1.1
New Jersey	1.0	1.0	1.1
New York	1.0	1.0	1.1
North Carolina	1.1	1.1	1.1
Ohio	1.0	1.0	1.1
Pennsylvania	1.0	1.0	1.1
Texas	1.1	1.1	1.1
Virginia	1.1	1.1	1.2
Regions:			
Northeast	1.0	1.0	1.1
North Central	1.0	1.0	1.1
South	1.0	1.0	1.1
West	1.0	1.0	1.0
SMSA's	1.1	1.1	1.1
Marital status	1.3	1.3	1.5
Household relationship:			
Head, wife, or primary individual	0.8	0.7	0.8
Child or other relative in primary family, secondary family member, secondary individual, or persons living in group quarters	1.3	1.3	1.5
Education and other characteristics	1.0	1.0	1.1

[1]Apply the factors in this column to tables B-1 and B-3 only.
[2]Apply the factors in this column to tables B-2 and B-4 only.