Appendix C.—ACCURACY OF THE DATA

SOURCES OF ERROR

Human and mechanical errors occur in any mass statistical operation such as a decennial census. Errors during the data collection phase can include failure to obtain required information from respondents, obtaining incorrect or inconsistent information, and recording information in the wrong place or incorrectly. Errors can also occur during the field review of the enumerator's work, the clerical handling of the questionnaires, the manual editing and coding operations, and the various stages of the electronic processing of the material. Careful efforts are made in every census to keep the errors in each step at an acceptably low level. Quality control and check measures are utilized throughout the census operation. As was done for the 1950 and 1960 censuses, evaluative material on many aspects of the 1970 census will be published as soon as the appropriate data are accumulated and analyzed. A major concern in the evaluation work is to ascertain, insofar as possible, the degree of completeness of the count of both population and housing units.

EDITING OF UNACCEPTABLE DATA

The objective of the processing operation is to produce a set of statistics that describes the Nation's housing as accurately and clearly as possible. To meet this objective, certain unacceptable entries were edited.

Whenever information was missing, an allocation procedure was used to assign an entry, thereby eliminating the need for a "not reported" category in the tabulations. The assignment was based on related information reported for the housing unit or on information reported for a similar unit in the neighborhood. For example, if tenure for an occupied unit was omitted but a rental amount was reported, the computer automatically edited tenure to "rented for cash rent." On the other hand, if the unit was reported as rented but the amount of rent was missing, the computer automatically assigned the rent that was reported for the preceding renter-occupied unit.

A similar procedure was used when the information reported for an item was inconsistent with other information reported for the unit. For example, if a housing unit was enumerated as having no piped water but having both a bathtub (or shower) and flush toilet for the exclusive use of the occupants of the unit, the computer edited water supply to "hot and cold water," a category considered to be consistent with the reported bathing and toilet facilities.

Specific tolerances were established for the number of computer allocations that would be permitted. If the number of corrections was beyond tolerance, the questionnaires in which the errors occurred were clerically reviewed. If it was found that the errors resulted from damaged questionnaires, from improper microfilming, from faulty reading by FOSDIC of undamaged questionnaires, or from other types of machine failure, the questionnaires were reprocessed.

ALLOCATION TABLES

The extent of allocations for nonresponses and inconsistencies is shown in tables A-1 to A-3 for the data collected on a 100-percent basis and in tables B-1 to B-3 for the items based on a sample. The base on which the percentage is computed is shown for each item. For most items, the percentages are based on all year-round housing units or occupied housing units. In some instances, the base is a specific group of units as indicated in the tables. Percentages are not shown if the item is not published for the specified areas.

SAMPLE DESIGN

The statistics presented in chapter B, tables 31 to 83, are based on a sample of housing units, with sampling rates of 20 percent, 15 percent, and 5 percent. (The data in chapter A, tables 1 to 30, were collected on a 100-percent basis.) For the sample data collected in the 1970 census, the housing unit, including all its occupants, was the sampling unit; for persons in group quarters identified in advance of the census, it was the person. In non-mail areas, the enumerator canvassed his assigned area and listed all housing units in an address register sequentially in the order in which he first visited the units, whether or not he completed the interview. Every fifth line of the address register was designated as a sample line, and the housing units listed on these lines were included in the sample. Each enumerator was given a random line on which he was to start listing and the order of canvassing was indicated in advance, although the instructions allowed some latitude in the order of visiting addresses. In mail areas, the list of housing units was prepared prior to Census Day either by employing commercial mailing lists corrected through the cooperation of the post office or by listing the units in a process similar to that used in non-mail areas. As in other areas, every fifth housing unit on these lists was designated to be in the sample. In group quarters, all persons were listed and every fifth person was selected for the sample; as indicated in Appendix B, information on the housing characteristics of group quarters was not collected in the census.

This 20-percent sample was subdivided into a 15-percent and a 5-percent sample by designating every fourth 20-percent sample unit as a member of the 5-percent sample. The remaining sample units became the 15-percent sample. Two types of sample questionnaires were used, one for the 5-percent and one for the 15-percent sample units. Some questions were included on both the 5-percent and 15-percent sample forms and therefore appear for a sample of 20 percent of the units in the census. Other items appeared on either the 15-percent or the 5-percent questionnaires. The sample rates for the various items appearing in chapter B are shown in table A.

Although the sampling procedure did not automatically insure an exact 20-percent sample of persons or housing units in each locality, the sample design was unbiased if carried through according to instructions; generally for larger areas the deviation from 20 percent was found to be quite small. Biases may have arisen, however, when the enumerator failed to follow his listing and sampling instructions exactly. Quality control procedures were used throughout the census process, and where there was clear evidence that the sampling procedures were not properly followed, some enumerators' assignments were returned to the field for resampling. As shown in table C-1 of the Population Census report PC(1)-C1 for the United States, 19.4 percent of the population and 19.6 percent of the housing units tabulated were enumerated on sample questionnaires. (The PC(1)-C series of State reports shows percentages for each State.) The bases for these percentages included several classes of the population and housing units for which no attempt at sampling was made. These were the relatively small numbers of persons and housing units (in most States, less than one percent) added to the enumeration from the post-census post office check, the various supplemental forms, and the special check of vacant units. (If these classes are excluded from the bases the respective proportions become 19.6 and 19.7 percent.) The ratio estimation pro-

TABLE A. **Sample Rate for Subjects Included in Chapter B.**

Subject	Sample rate (percent)
OCCUPANCY CHARACTERISTICS	
Total housing units	20
Total population	20
Occupied housing units	20
Tenure	20
Race	20
Spanish heritage[1]	15
Population per occupied unit	20
Cooperative or condominium	20
Year moved into unit	15
VACANCY CHARACTERISTICS	
Vacant housing units	20
Homeowner vacancy rate	20
Rental vacancy rate	20
Duration of vacancy	20
UTILIZATION CHARACTERISTICS	
Number of rooms	20
Size of household (persons)	20
Persons per room	20
Bedrooms	5
PLUMBING CHARACTERISTICS	
Plumbing facilities	20
Piped water	20
Flush toilet	20
Bathtub or shower	20
Complete bathrooms	15
Source of water	15
Sewage disposal	15
STRUCTURAL CHARACTERISTICS	
Complete kitchen facilities	20
Access	20
Units in structure	20
Mobile home or trailer	20
Year structure built	20
Basement	20
Elevator in structure	5
EQUIPMENT, FUELS, AND APPLIANCES	
Telephone available	20
Heating equipment	20
Air conditioning	15
Automobiles available	15
Second home	5
Fuels for house heating, water heating, and cooking	5
Clothes washing machine	5
Clothes dryer	5
Dishwasher	5
Home food freezer	5
Television	5
Battery-operated radio	5
FINANCIAL CHARACTERISTICS	
Value	20
Contract rent	20
Gross rent	20

[1] As indicated in the "Introduction," derived figures are not presented if there are fewer than 25 units in the distribution or the base for the 20-percent sample, fewer than 33 units for the 15-percent sample, and fewer than 100 units for the 5-percent sample. However, in the tables for households with heads of Spanish heritage, the minimum base for which derived numbers are shown is determined according to the sample rate of the characteristic shown in this table.

cedure described below adjusts the sample data to reflect these classes of population and housing units.

RATIO ESTIMATION

The statistics based on 1970 census sample data are estimates made through the use of ratio estimation procedures which were applied separately for the 5-, 15-, and 20-percent samples. The first step in carrying through the ratio estimates was to establish the areas within which separate ratios were to be prepared. These are referred to as "weighting areas." For the 15- and 20-percent samples, the weighting areas contained a minimum population size of 2,500. The weighting areas used for the 5-percent ratio estimate were larger areas having a minimum population size of 25,000 and comprising combinations of the weighting areas used for the 15- and 20-percent samples. Weighting areas were established by a mechanical operation on the computer and were defined to conform, as nearly as possible, to areas for which tabulations are produced. Where these areas do not agree (primarily for smaller areas), there may be some differences between complete counts and sample estimates.

The ratio estimation process operated in two stages for occupied housing units, and in one stage for vacant units. The first stage for occupied units employed 18 household-type groups (the first of which was empty by definition); the second stage for occupied units used four groups: owner and renter occupied units, by race. The single stage for vacant units employed three groups: year-round vacant for sale, year-round vacant for rent, and other vacant.

Group	
Occupied housing units:	
	STAGE I
	Male Head With Own Children Under 18
1	1-person household
2	2-person household
3	3-person household
.	.
.	.
6	6-or-more-person household
	Male Head Without Own Children Under 18
7-12	1-person to 6-or-more-person households
	Female Head
13-18	1-person to 6-or-more-person households
	STAGE II
	Owner Occupied
19	Negro
20	Not Negro
	Renter Occupied
21	Negro
22	Not Negro
Vacant housing units:	
23	Year-round vacant for sale
24	Year-round vacant for rent
25	Other vacant

At each stage, for each of the occupied housing groups, the ratio of the complete count to the weighted sample count of the housing units in the group was computed and applied to the weight of each sample unit in the group. This operation was performed for each of the 18 groups in the first stage, then for the four groups in the second stage. As a rule, the weighted sample counts within each of the 4 groups in the second stage for occupied units should agree with the complete counts for the weighting areas. Close, although not exact consistency can be expected for the 18 groups in the first stage. Similarly, the weighted sample counts within each of the 3 groups in the single stage for vacant housing units should agree with the complete counts for the weighting area.

There are some exceptions to this general rule, however. As indicated above, there may be differences between the complete counts and sample estimates when the tabulation area is not made up of whole weighting areas. Furthermore, in order to increase the reliability, a separate ratio was not computed in a group whenever certain criteria pertaining to the complete count of housing units and the magnitude of the weight were not met. For example, for the 20-percent sample the complete count of units in a group had to exceed 70 units and the ratio of the complete count to the unweighted sample count could not exceed 20. Where these criteria were not met, groups were combined in a specific order until the conditions were met. Where this occurred, consistency between the weighted sample and the complete counts would apply as indicated above for the combined group but not necessarily for each of the groups in the combination.

Each sample housing unit was assigned an integral weight to avoid the complications involved in rounding in the final tables. If, for example, the final weight for a group was 5.2, one-fifth of the units in the group (selected at random) were assigned a weight of 6 and the remaining four-fifths a weight of 5.

The estimates realize some of the gains in sampling efficiency that would have resulted had the population been stratified into the groups before sampling. The net effect is a reduction in both the sampling error and possible bias of most statistics below what would be obtained by weighting the results of the sample by a uniform factor (e.g., by weighting the 20-percent sample results by a uniform factor of 5). The reduction in sampling error will be trivial for some items and substantial for others. A byproduct of this estimation procedure is that estimates for this sample are, in general, consistent with the complete count for the housing unit groups used in the estimation procedure. A more complete discussion of the technical aspects of these ratio estimates will be presented in a separate report.

SAMPLING VARIABILITY

The estimates from the 20-, 15-, and 5-percent sample tabulations are subject to sampling variability. The standard errors of these estimates can be approximated by using the data in tables B through D. The chances are about 2 out of 3 that the difference (due to sampling variability) between the sample estimate and the figure that would have been obtained from a complete count is less than the standard error. The chances are about 19 out of 20 that the difference is less than twice the standard error and about 99 out of 100 that it is less than 2½ times the standard error. The amount by which the estimated standard error must be multiplied to obtain other odds deemed more appropriate can be found in most statistical textbooks. The sampling errors may be obtained by using the factors shown in table D in conjunction with table B for absolute numbers and in conjunction with table C for percentages. These tables reflect the effect of simple response variance, but not of bias arising in the collection, processing and estimation steps nor of the correlated errors enumerators introduce; estimates of the magnitude of some of these factors in the total error are being evaluated and will be published at a later date.

Table B shows approximate standard errors of estimated numbers for most statistics based on the 20-percent sample. In determining the figures for this table, some aspects of the sample design, the estimation process, and the size of the area over which the data have been compiled are ignored. Table C shows standard errors of most percentages based on the 20-percent sample. Linear interpolation in tables B and C will provide approximate results that are satisfactory for most purposes. Table D provides a factor by which the standard errors shown in tables B or C should be multiplied to adjust for the effect of the sample size (i.e., whether a 15-percent or 5-percent sample) and the effect of the estimation procedure.

To estimate the standard error for a given characteristic, locate the factor in table D for the appropriate characteristic and the sample size used to tabulate the data, and multiply this factor by the standard error found in table B or C. If an item, although collected on one sample basis, has been tabulated for a smaller sample, use the factor appropriate for the smaller sample.

The standard errors estimated from these tables are not directly applicable to differences between two sample estimates. In order to estimate the standard error of a difference, the tables are to be used somewhat differently in the three following situations:

1. For a difference between the sample figure and one based on a complete count (e.g., arising from comparisons between sample statistics for 1970 and complete-count statistics for 1960 or 1950), the standard error is identical with the standard error of the 1970 estimate alone.

2. For a difference between two sample figures (that is, one from 1970 and the other from 1960, or both from the same census year), the standard error is approximately the square root of the sum of the squares of the standard errors of each estimate considered separately. This formula 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. If, however, there is a high positive correlation between the two characteristics, the formula will overestimate the true standard error. The approximate standard error for the 1970 sample figure is derived directly from tables B through D. The standard error of a 25-percent 1960 sample figure may be obtained from the relevant 1960 census report or an approximate value may be obtained by multiplying the appropriate value in table B or C by 0.9.

3. For a difference between two sample estimates, one of which represents a subclass of the other, the tables can be used directly with the difference considered as the sample estimate.

The sampling variability of the medians presented in certain tables (median rooms, median value, median gross rent, etc.) depends on the size of the base and on the distribution on which the median is based. An approximate method for measuring the reliability of an estimated median is to determine an interval about the estimated median such that there is a stated degree of confidence the true median lies within the interval. As the first step in estimating the upper and lower limits of the interval (that is, the confidence limits) about the median, compute one-half the number on which the median is based (designated N/2). From table B, following the method outlined in other parts of this section, compute the standard error of an estimated number equal to N/2. Subtract this standard error from N/2. Cumulate the frequencies (in the table on which the median is based) until the total first exceeds the difference between N/2 and its standard error, and by linear interpolation obtain a value corresponding to this number. In a corresponding manner, add the standard error to N/2, cumulate the frequencies in the table, and obtain a value in the table on which the median is based corresponding to the sum of N/2 and its standard error.

The chances are about 2 out of 3 that the median would lie between these two values. The range for 19 chances out of 20 and for 99 in 100 can be computed in a similar manner by multiplying the standard error by the appropriate factors before subtracting from and adding to one-half the number reporting the characteristics. Interpolation to obtain the values corresponding to these numbers gives the confidence limits for the median.

TABLE B. **Approximate Standard Error of Estimated Number Based on 20-Percent Sample**

(Range of 2 chances out of 3; for factors to be applied see table D and text)

Estimated number[1]	Number of housing units in area[2]								
	1,000	10,000	25,000	100,000	250,000	1,000,000	3,000,000	5,000,000	7,000,000
50	15	15	15	15	15	15	15	15	15
100	20	20	20	20	20	20	20	20	20
250	30	30	30	30	30	30	30	30	30
500	30	45	45	45	45	45	45	45	45
1,000	...	60	60	65	65	65	65	65	65
2,500	...	90	95	100	100	100	100	100	100
5,000	...	100	130	140	140	140	140	140	140
10,000	...	...	150	190	200	200	200	200	200
15,000	...	...	150	230	240	240	240	240	240
25,000	...	...	...	270	300	310	310	320	320
50,000	...	...	...	320	400	440	440	440	450
75,000	...	...	...	270	450	520	540	540	540
100,000	...	...	...	...	490	600	620	630	630

[1] For estimated numbers larger than 100,000, the relative errors are somewhat smaller than for 100,000.

[2] An area is the smallest complete geographic area to which the estimate under consideration pertains. Thus, the area may be the State, city, county, standard metropolitan statistical area, urbanized area, or the urban or rural portion of the State or county. The rural-farm or rural-nonfarm units in the State or county, the Negro-occupied units, etc., do not represent complete areas.