CONTENTS

CHAPTER 11. Sampling and Estimation

(Page numbers here omit the chapter prefix, 11– , which appears as part of the number of individual pages. The prefix indicates the location of the material in the final consolidated edition of the **Procedural History**.)

Chapter 11. SAMPLING AND ESTIMATION

SAMPLE DESIGN

Basic Sample

The basic sample for the 1970 Census of Population and Housing was a 20-percent sample selected from the census listing of housing units and individuals in group quarters. For persons living in housing units at the time of the census, the housing unit--including all its occupants--was the sampling unit; for persons in large group quarters (15 or more persons), the sampling unit was the person. Small group quarters were sampled in the same manner as housing units.

In nonmail areas--i.e., those areas enumerated by "conventional" personal visits--the enumerator canvassed his assigned area and listed all housing units in an address register sequentially in the prescribed 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 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--i.e., those where the mail-out/mail-back enumeration procedure was used--the list of housing units was prepared prior to Census Day either by a computerized operation in the Bureau's headquarters (see chapter 3) or by listing the units prior to the census in a process similar to that used in nonmail areas (see also chapter 5). Where the computerized list was used, the sample was designated systematically on the computer. In mail areas where addresses were prelisted, every fifth housing unit on these lists, after a random start, was predesignated to be in the sample.

In large group quarters, regardless of whether the quarters were located in a mail or a nonmail census area, all persons were listed in a sample selection book and every fifth person was selected for the sample. (For further details, see chapter 5.)

Subsamples

The 20-percent sample was subdivided into a 15-percent and a 5-percent sample: Every fourth 20-percent sample unit (or sample person in large group quarters) was designated as a member of the 5-percent sample and the remaining sample units became the 15-percent sample. Two different types of sample questionnaires were used, one for the 5-percent and one for the 15-percent sample units. Some sample population and housing questions appeared on both the 5-percent and 15-percent questionnaires, and these made up the 20-percent sample items. Other questions appeared only on the 15-percent or on the 5-percent questionnaires. (For specific questions and sampling rates, see chapter 15.)

Comparison with 1960 Sampling Plan

The 1970 sampling procedure in nonmail areas was similar to the 1960 procedure except that in 1970 the sample lines in the listing books (address registers) were predesignated. Thus, as in 1960, the enumerators were aware of the identity of the sample units as they recorded them in their listing books. However, in the mail areas in 1970, the designation of the sample housing units was not in the enumerator's hands, as the selected questionnaire already had been mailed to the household, and the enumerator's function was to ensure the completion of the questionnaire.

As in 1970, the sample for the 1960 Censuses of Population and Housing was selected from the listing of all housing units and group quarters. In 1960 a 25-percent sample was used rather than a 20-percent sample. The enumerator was instructed to assign a key letter (A, B, C, or D) sequentially to each housing unit in the order in which he first visited the units. The key letter was assigned at the first visit whether or not the interview was completed. Each housing unit assigned the key letter "A" was designated as a sample unit, and all its occupants were included in the sample. In group quarters the sample consisted of a systematic selection of one in four persons in the order they had been listed.

In order to collect as much information as possible on housing characteristics without placing undue burden on the respondents, in 1960 the 25-percent housing sample was divided into a 5-percent and a 20-percent sample. Certain housing questions were asked only at 5 percent of the housing units, others at only 20 percent of the housing units, and some at the full 25-percent sample of housing units. The same set of population questions was asked at all sample housing units. (The subdivision of the full 1970 sample differed from the 1960 method in that for 1970 both the population and housing 20-percent samples were subsampled prior to enumeration, as described in the section on "Subsamples" above.)

The majority of the tabulations for the 1960 Census of Population were from either the complete count or the 25-percent sample. For large areas, a subsample of one-fifth of the original 25-percent schedules was selected and used for some tabulations in order to reduce costs. The subsample was selected on the computer, using a stratified systematic sample design

comprising 38 strata. For 1970, the data collected were tabulated without further subsampling.

Summary of Sampling Ratios Actually Obtained

Although the 1970 sampling procedure did not automatically ensure an exact 20-percent sample of persons or housing units in each locality, the sample design was unbiased if carried through according to instructions; and generally for larger areas the deviation from 20 percent was found to be quite small. Biases may have arisen, however, when the enumerator (or lister) failed to follow the listing and sampling instructions exactly. The selection of the sample by the enumerator (or lister) at the time of canvassing created opportunities for biases to occur since the order of listing could be manipulated. For example, the enumerator might have included or excluded from the sample a vacant housing unit or one where there was reason to anticipate difficulty in getting cooperation. Quality control procedures were used in the census process, however; and where there was clear evidence that the sampling procedures were not properly followed, enumerators' assignments were returned to the field for resampling and adjustment. Even in mail areas there was some possibility of bias in multiunit structures with unclear apartment descriptions. Where the postman could not place the questionnaire for a specific apartment into a corresponding mail receptacle, it was possible for the postman or even the tenants themselves to determine who would get the sample questionnaire.

In the United States as a whole, 19.4 percent of the population and 19.6 percent of the housing units tabulated were enumerated on sample questionnaires. 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 1 percent) added to the enumeration from the post-enumeration Post Office check, the special check of vacant units, and various supplemental forms (principally from the "Were You Counted?" campaign). (If these classes are excluded from the bases, the respective proportions become 19.6 and 19.7 percent.)

CHECKING THE SAMPLE

Quality Control of the Sample Selection

In the mail areas, checks on the quality of the sample during the field operations were not established because the sample selection was almost entirely out of the hands of the enumerators. In nonmail areas, however, the enumerators had some control over sample selection, so certain quality control checks were made on the sample selection process. The control on the sampling was done by the crew leader (the enumerator's immediate supervisor) and by the field office. (For details, see chapter 5.)

Telegraphic Clearance of Preliminary Field Counts

At the completion of enumeration in a clearance area (a county, or a city of 50,000 inhabitants or more), the district manager telegraphed the preliminary counts of population and housing units to Bureau headquarters in Suitland, Md. The counts were submitted for every clearance area, and were reviewed in Suitland for consistency with expected counts before the district manager released them as preliminary counts to local officials and newspapers. (Additional information on the clearance operation can be found in chapter 5.)

A check on possible bias in the sample selection was made in conjunction with the clearance operation. The check involved comparing sample counts of population and housing with the respective complete counts. If the difference fell within a predetermined range, the sample was accepted. If not, a field review was required. After this review was initiated for only a few areas, it was found that the sample bias review was too time-consuming to be part of the clearance operation. This check then was converted to a post-censal operation (resampling), which is described below.

Resampling Enumeration Districts (ED's) with Significant Sample Bias

In general, when a household sample is selected, the distribution of sample households by number of persons per household should vary from the corresponding distribution for all households only because of sampling variability. Past experience has shown, however, that occasionally the sample is found to have more households of large (or small) size than would have occurred had the sample selection been completely unbiased. This bias normally appears in relatively few ED's and its effect is more noticeable in data for the smaller areas. Since many characteristics are correlated with household size, it was considered desirable to identify areas where an unrepresentative household size distribution occurred in the sample in order to reduce the effect of this bias. In the 1970 census, two methods were used to reduce sample bias. One method, the more important because it was applied to sample data for all areas, was part of the ratio estimation procedure. Ratio estimates were applied to data for sample households in six household size categories. Each of these household size categories was divided into three household types. This method is described more fully in the section on ratio estimation in this chapter.

The second method was the post-censal resampling operation. This operation was employed only in nonmail areas, since the designation of the sample was under less control in these areas and the improvements realized were expected to be greater for the amount of work expended. The operation itself consisted of identifying counties where there appeared to be a significant overall sample bias, and then within these areas identifying the ED's where the bias seemed most pronounced. Within each of these ED's, a new sample of units was designated and the necessary reenumeration completed. The resampling operation produced changes in the sample data for about 850 ED's located in 222 different counties. (There were approximately 250,000 ED's in the United States.)

The problem counties were originally identified by comparing inflated sample population and housing counts with the respective complete-count totals. Counties having differences beyond a certain tolerance were designated as "problem" areas. All ED's within the county were then reviewed to determine those requiring resampling to bring the county count differences to an acceptable level. The ED's to be designated for resampling were determined by the magnitude of the

difference between five times the sample population and the total population of the ED. ED's were ranked according to the absolute value of this difference (i.e., ignoring the sign of the difference). Assuming resampling the ED would reduce the difference for the ED to zero, a sufficient number of ED's was selected in the order of the ranking to reduce the difference for the county to an acceptable level.

For each ED designated for resampling, the arithmetic totals of population and housing counts in the address registers were first re-added. Discrepancies in counts were investigated and occasionally it was found that it was not necessary to continue the resampling operation in an ED since the problem was only an arithmetic error.

For the ED's not cleared up by correcting arithmetic errors, a new sampling pattern was designated. This operation was accomplished by first transcribing the addresses of the housing units in the address register for the ED in the following order:

1. Short form (nonsample) units
2. 15-percent sample units
3. 5-percent sample units

For the original nonsample units, a new sample was designated by selecting one-fifth of the units to be in the sample; every fourth sample unit in the new sample was designated to be a 5-percent sample unit. Sample questionnaires were addressed for units which appeared in the newly designated sample. These questionnaires were sent to the field office of the area involved, and an enumerator visited the units and obtained the necessary sample information. When these questionnaires were returned, they were substituted for the old nonsample questionnaires. For those units originally designated as sample units, four out of five were selected and converted to nonsample questionnaires by deleting the sample information.

Other Post-Census Adjustments in the Sample

During the initial processing of the 100-percent portion of the census questionnaires, the sample and complete counts of population and housing were obtained by ED. A computer program was designed to review the counts for each ED and to display the ED's for which there appeared to be significant discrepancies between the sample and complete counts. Three general types of problems were identified:

Type 1. ED's in which the discrepancy could be attributed to a bad sample of the group quarters population of the ED.

Type 2. ED's in which the count of sample housing units and sample persons both differed significantly in the same direction from an expected 20-percent sample.

Type 3. ED's in which there was a 20-percent sample of housing units but the sample of persons differed significantly from 20 percent. Presumably these were cases of household-size bias as described above.

All three types of problems showed undersamples as well as oversamples. If no corrective action was taken, the ratio estimation procedure would as a matter of course adjust the weights of the sample cases to compensate. In some instances, however, the combination of ED's in a weighting area would not be in the same tabulation area (see the section on ratio estimation below). The weighting procedure produced consistency between complete counts and weighted sample counts only for the weighting area, and the discrepancy in one ED would affect the weights of the sample in other ED's in the weighting area. To the extent that ED's from such a weighting area appeared in different tabulation areas, the sample tabulations could show considerable discrepancies. For this reason, it was decided to manually correct for the bias in the "worst" ED's.

For Type 1 ED's, the following action was performed: If there was an oversample, no correction was made. The ratio estimation procedure was allowed to give less than the expected weight to the sample persons in group quarters. This could normally be expected to occur since the weighting of the group quarters' population was done separately in the ratio estimation procedure. The ratio estimator, however, could not properly correct severe undersamples since the rules allowed the group quarters ratio estimate category to be combined with other population categories if there were insufficient sample cases. Therefore, for ED's with severe shortages in the group quarters sample, sufficient sample questionnaires were imputed from similar group quarters in the same general geographic area. This operation was performed for 80 ED's in 50 counties.

For Type 2 ED's which had oversamples, the sample questionnaires were manually subsampled down to the proper number prior to microfilming. This subsampling affected 25 ED's in 24 counties. For those Type 2 ED's with undersamples of 10 percent or less, the means of correction depended on the actual sampling rate. If the actual sample represented 8 to 10 percent of the ED, the sample questionnaires for the ED were duplicated (10 ED's). For those ED's where the sample was less than 8 percent, an appropriate number of sample questionnaires were imputed from an ED with similiar characteristics in the same geographic area (20 ED's).

Other than resampling, no additional corrective action was carried out for Type 3 ED's. In these cases, the sample as designated by the enumerator showed evidence of a bias in the size of household. Similar situations were noted in 1960 and many were corrected by a large clerical operation. In 1970 the ratio estimation procedure was designed to control on household size and thus partially compensate for the biases.

During subsequent processing, an additional type of problem arose. Data for some ED's could not be found at the time of 100-percent computer processing. In order to keep the proper population levels, data were imputed for these ED's. By the time the sample was processed, however, the data for these ED's might have been available and the sample characteristics did not necessarily agree with the imputed 100-percent characteristics. Since the ratio estimation scheme produces consistency between 100-percent and sample characteristics, some problems arose. For the more

serious cases, the ratio estimation process was modified to a single cell ratio estimate to total population. This operation affected about 350 ED's in 16 States.

ESTIMATION

The estimation procedure for 1970 census sample data was designed to produce estimates that would have a low mean square error within constraints imposed by cost and practicality of application. In general, the procedure dealt with groups of records within specially defined areas called weighting areas (described below). Within each weighting area, complete counts and sample counts were obtained for various characteristics. For a subset of these characteristics, the sample was weighted to agree with the complete counts for the subset. The resulting inflated sample counts were termed target numbers. The sample records for each subset were assigned integral weights such that the sum of these weights agreed with the target number.

The estimation procedure assigned separate sets of weights to the population sample and to the housing unit sample records for each of the three census samples representing 5 percent, 15 percent, and 20 percent of the persons and housing units in the census. The six sets of weights for these samples were derived by processes which were essentially independent but similar in operation. Three-stage ratio estimators were used for population data and two-stage estimators were used for housing data.

The stages of estimation were performed in a given sequence and then repeated again ("iterated") in the same order. The iteration was performed to cause the target numbers used in all stages of the estimate to converge toward the complete counts. There should therefore be close, though not necessarily exact, agreement between tabulations based on weighted sample counts and the complete counts for the totals used to produce the weights in all stages of the estimate. It has been shown that under certain conditions, continuous iteration will produce weighted sample estimates which equal desired row and column complete counts simultaneously. The justification for this estimator arises as an application in statistical information theory. The actual process used came close to, but did not exactly meet, the conditions for convergence.

Background

In the 1960 census, estimates based on sample data were also derived by the use of a ratio estimation procedure. Each sample record was first classified into a ratio estimate group. There were 44 age, sex, and color groups for persons, and seven groups for housing units by color of occupants, occupancy, and tenure. The complete count for each group was determined and weights were assigned to the sample records to sum to the complete count for the group. It was sometimes necessary to combine groups in order to meet conditions imposed to control the bias usually present in ratio estimation procedures.

Experience with the 1960 estimator suggested that the procedure should incorporate household size in the definition of the ratio estimate groups. However, the number of ratio estimate groups defined by expanding each of the 44 groups by six household size categories could not be used efficiently by an estimator of the type used in 1960, and other estimators were therefore considered.

In choosing the estimator to be used in 1970, the following criteria were considered: The estimator should (1) dampen the effect of any biases that occurred in sample selection, (2) reduce the variance of sample estimates, (3) improve the consistency between complete counts and sample estimates, (4) be economical to execute, and (5) permit reasonably accurate estimates of sampling error to be computed. After the 1960 census, the properties of a number of different ratio estimation procedures were examined. The one chosen is explained in some detail in the following sections.

Definition of Weighting Areas

The estimation procedure operated within groups of ED's defined specifically for this purpose. These groups are called weighting areas. Weighting areas were constructed mechanically on the computer by a process of combining ED's to conform as nearly as possible to the smallest areas for which sample tabulations would be published. A single set of weighting areas was defined for use with both the 15-percent and 20-percent samples. The weighting areas defined for the 5-percent sample were made up of integral combinations of 20-percent weighting areas. The weighting areas defined for the population ratio estimator were also used for housing. The procedure was controlled so that normally a weighting area did not include parts of more than one county; weighting areas never crossed State lines. The weighting areas within which ratio estimates were performed were defined with the following guidelines:

1. The set of weighting areas defined for use with both the 20-percent and 15-percent samples normally comprised geographically contiguous territory within a county such as a tract or, for nontracted areas, a place or minor civil division (MCD). The weighting area contained a population of at least 2,500. Areas defined in this way but with less than 2,500 population were, in general, combined with a contiguous area.

2. Weighting areas for the 5-percent sample were defined by combining contiguous 20-percent sample weighting areas until a combination in excess of a minimum population size was achieved. If all 20-percent sample weighting areas in the combination were in the same MCD and in the same county, this minimum population size was 50,000. If all 20-percent sample weighting areas in the combination were not from the same MCD and county, the minimum population size was 25,000.

Ratio Estimation Groups

The ratio estimation process for persons operated in three stages. The first stage employed 19 household-type groups (the first of which by definition had no persons in it). The second stage used two groups, head of household and not head of household, and the third stage used 24 age-sex-race groups.

Stage	Group	Household type
I		Male head without own children under 18
	1	1-person household
	2	2-person household
	3	3-person household
	.	.
	.	.
	6	6-or-more-person households
		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
	19	Group-quarters persons
II	20	Head of household
	21	Not head of household (including persons in group quarters)
III		Male Negro
	22	Age under 5 years
	23	5-13
	24	14-24
	25	25-44
	26	45-64
	27	65 and older
		Male, not Negro
	28-33	Same age groups as for male Negro
		Female Negro
	34-39	Same age groups as for male Negro
		Female, not Negro
	40-45	Same age groups as for male Negro

The ratio estimation process for housing 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 by definition had no persons in it); 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.

Occupied housing units

Stage	Group	Household type
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 households

Occupied housing units--Continued

Stage	Group	Household type
		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
II		Owner-occupied
	19	Negro
	20	Not Negro
		Renter-occupied
	21	Negro
	22	Not Negro

Vacant housing units

Stage	Group	
I	23	Year-round vacant for sale
	24	Year-round vacant for rent
	25	Other vacant

Collapsing Ratio Estimation Groups

Certain criteria had to be met before the estimation procedure was performed within a group. If these criteria were satisfied, the ratios for the group were computed. If these criteria were not satisified, however, the complete counts and sample counts for the group involved were each combined (collapsed) with the counts for other groups by a prescribed procedure until the counts for the combined groups did meet these criteria. The order of collapsing was such as to combine a group with another group that was very similar. The collapsing required for all estimation stages was performed before the estimation procedure was executed.

Two sets of criteria were used. The first set of criteria applied to complete counts and sample counts for the 15-percent sample to determine when collapsing was required. These criteria determined the collapsing of both the 15-percent and 20-percent sample counts so that the same cells were combined for the 15-percent sample and the 20-percent sample. A second set of criteria applied to the counts used for the 5-percent sample only. The conditions that had to be met by the complete and sample counts before a ratio estimate was permitted were as follows:

1. The weighting procedure had to produce weights for the sample cases which were less than a certain maximum. This maximum was 20 for the 15-percent sample and 80 for the 5-percent sample. These criteria applied for both housing and population samples.

2. Each of the complete counts had to equal or exceed a given minimum. The minimums were chosen so that the probability of getting three or more sample cases in a given group would be

about 0.999. For population samples, this probability was assumed to be met when the 100-percent count of population was equal to or exceeded 85 persons for the 15-percent sample or 275 persons for the 5-percent sample. For housing samples, the 100-percent count of housing units had to equal or exceed 70 units for the 15-percent sample or 200 units for the 5-percent sample.

The maximum allowable weight condition applied to the ratio of the complete to sample counts for an entire group. It was possible for the estimation procedure to produce a weight for a specific record exceeding the maximum even though the conditions were met by the group as a whole. Because of a program limitation, however, the maximum weight a 5-percent sample record could have was 127; for the 15- or 20-percent sample it was 31. If a weight exceeding these limitations was generated, the record received the maximum weight that could be carried in the record (127 or 31). In these (infrequent) situations, the inflated sample count did not agree with the 100-percent count. A summary of the weights assigned are shown below for percent of 20-percent-sample persons and housing units. The asterisk (*) indicates less than .01 percent.

Assigned weight	Population	Housing
0	.01	*
1-2	.51	.27
3	4.30	2.68
4	22.46	20.92
5	41.39	48.55
6	21.37	21.38
7	6.36	4.53
8	2.06	1.11
9	.79	.34
10-11	.51	.17
12-13	.14	.03
14-19	.08	.01
20 and over	.01	*

Estimation Procedure for Population

The population estimation procedure for a given weighting area was accomplished by three stages of ratio estimation. The steps in the process are illustrated in figure A. The figure attempts to portray the two matrices of sample and complete counts involved in the estimator for population data (one matrix for household heads and one matrix for other persons).

At the start of the process, each of the interior cells of the matrix is assumed to hold the inflated sample counts for the weighting area. In figure A, S_{ij} represents the inflated sample count for the i^{th} family-type household size category (row) and the j^{th} age-sex-race category (column). These counts are also summarized for the rows and the columns into marginal (group) totals (represented by $S_{i.}$ and $S_{.j}$); the figure describes them as "target numbers" because later steps in the process produce adjusted values in the interior cells (defined as target numbers) which are then summarized and carried in these marginal cells. The 100-percent counts for these marginals are also indicated (represented by T_i and $T_{.j}$).

The first stage of estimation inflates the sample counts to 100-percent counts of population by household type and household size. This amounts to applying the ratio $\frac{T_{i.}}{S_{i.}}$ to each S_{ij} in the i^{th} row. The ratio is based on the values obtained for household heads and other persons combined, but within an interior cell the values of S_{ij} are treated separately for heads and for others.

The second stage adjusts the target numbers produced in the first stage for household heads and for other persons to the complete counts for these two categories. This involves the ratio $\frac{T_H}{S_H}$ and the new values of S_{ij} for heads produced by the previous estimation stage (the term S_H represents the sum of all new values for heads S_{ij} produced in the previous stage). The ratio $\frac{T_{\bar{H}}}{S_{\bar{H}}}$ and S_{ij} for persons other than heads of households are similarly involved.

The third stage adjusts the target numbers produced from the second stage to 100-percent counts by 24 age, sex, and race categories. This involves applying the ratios $\frac{T_{.j}}{S_{.j}}$ to each value of S_{ij} in the j^{th} column. The ratio is based on the values obtained for household heads and other persons combined, but within an interior cell the values of S_{ij} are treated separately for heads of households and for others.

The three stages of estimation are then repeated again in sequence.

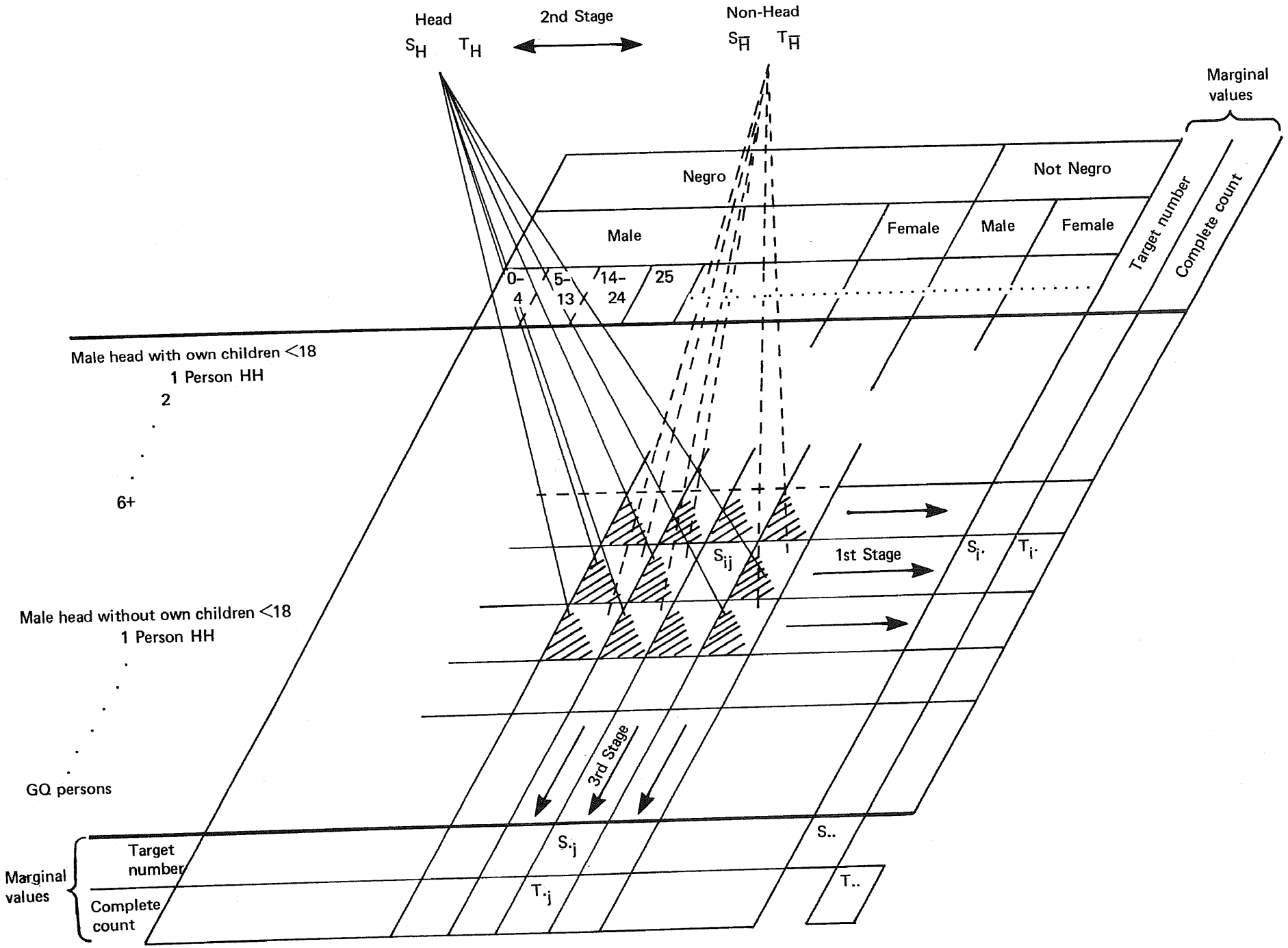
The cross-tabulation of all groups used in the three stages of population ratio estimation produced 864 cells (18 x 2 x 24) of sample counts. One result of the ratio-estimation process was a set of 864 target numbers, one for each of the interior cells in the cross-tabulation. The target number for a cell was used to assign integral weights to all sample records which were members of that cell so that the total of integral weights for sample cases in each cell equaled the target number in the cell. For example, if the target number for a cell was 65 and there were 12 sample cases in the cell, five-twelfths of the sample cases (selected at random) within the cell were assigned weights of 6 and the remaining seven-twelfths were assigned weights of 5.

Estimation Procedure for Housing

The housing estimation procedure operated independently of population estimation. Also the estimation procedure for occupied housing units was independent of the procedure for vacant housing units.

Weights were assigned to sample occupied housing unit records by a two-stage ratio-estimation procedure. The first stage employed ratio estimates by household size within the three household types defined in the same way as for the population estimates. The second

Figure A. Matrices Used in Population Estimation



stage was a ratio estimate by race and tenure. The two estimation steps were then repeated in sequence again. Weights were assigned to sample vacant housing unit records by a single-stage ratio-estimation procedure performed separately within three classes of vacant housing units.

The ratio-estimation process for housing produced a set of 71 target numbers (17 x 4 + 3), one for each of the cells in the cross-tabulation. For each cell, integral weights based on these target numbers were assigned to all sample records which were members of the cell.

Description of the Diary Printouts

Two diaries were produced to record the results of the ratio estimation procedure. The first, called the "ratio diary," displayed summary measures for weighting areas. The other, called the "weighting diary" was produced at the time weights were assigned to the sample records. Certain checks were made on the sample at both stages.

For each weighting area, and for each of the three samples, the ratio diary displayed the complete counts, the unweighted sample counts, and the weighted sample counts for total population, occupied housing units, and vacant units. These counts were also summarized for each work unit processed on a computer run. In addition, the diary identified the ED's making up the weighting areas and the number of groups collapsed at each stage of estimation. Flags were also shown for each weighting area which failed any of various tolerance checks. For example, one flag indicated any weighting area where the ratio of total population to unweighted 20-percent sample population was less than 4 or greater than 6.

The weighting diary identified the ED's making up the weighting area, and showed the distribution of the weights assigned to sample records. The distribution was shown separately for the 20-, 15-, and 5-percent population and housing samples, and a summary distribution of the weights was provided covering all weighting areas processed on a computer run. Flags were also shown for any weighting area where the assigned weights were unusually large or small.

The flags from both diaries were examined and used to isolate areas where there might have been a processing error. Resulting corrections are described in the section on checking the sample (see p. 2 above).

SAMPLING VARIABILITY

Statistics based on a sample almost always differ somewhat from figures that would have been obtained if a complete census had been taken using the same questionnaires, instructions and enumerators. Sample results are also subject to the same response, reporting, and processing errors which would be present in data from a complete census.

In order that sample statistics from the census be properly interpreted, a statement on their reliability appears in census publications. The estimates of reliability reflect sampling error and the effect of the estimation process but do not reflect the full effect of response or processing variance, or any effect of bias arising in collection, processing, or estimation.

Presenting Sampling Errors

A major concern in the choice of a method of presenting sampling errors arose from the number of statistics produced. To compute and show the sampling error for each published characteristic in each tabulation area would have been costly and time consuming, as well as doubling the number of pages needed to present the results in published volumes. It was decided, therefore, to group the individual census items into homogeneous classes and show in the publications the average of the sampling errors for the items in each class.

Almost all of the statistics tabulated from the census sample can be characterized as 0-1 variates; that is, the person is assigned the value one if he has the characteristic and zero otherwise. The design of the census sample and the ratio estimation procedure used suggested that the variances would usually have a fairly simple relationship to those arising from a simple random sample of the same size. This led to a decision to present the sampling errors in the form of "factors over random." Thus the design effect, that is, the ratio of the estimate of the variance of the census sample to the variance for a 20-percent simple random sample, was measured for a set of items in a class. The ratios were averaged over the items in the class and the square root of the average was used in determining the standard error for all statistics for the class.

This decision led to the following method of presenting data on sampling errors. Each census volume contains three tables. Two of the tables show the standard errors of a 20-percent simple random sample for 0-1 characteristics. One of the tables applies to estimates of magnitudes, the other one to percentages. In effect, they show the values of

$$\sqrt{(1-n/N)\frac{P(1-P)}{n}} \quad \text{and} \quad \sqrt{N^2(1-n/N)\frac{P(1-P)}{n}}$$

where N is total population, P the proportion of the population with the characteristic and n=N/5. The third table reflects the design effects, that is, it provides adjustment factors to be applied to either of the first two tables. The reader is required to look up the item of interest in the third table. He then multiplies the factor shown in that table by the appropriate standard error from one of the first two tables to obtain an estimate of the standard error of his census statistic.

Estimating the Census Variances

In order to produce the design effects, it was necessary to estimate the variance of the census statistics. Because a complex estimator and a systematic sample of clusters (households) were used, no simple mathematical formula could be derived that would directly estimate the variance from the census sample. The variances of census estimates were therefore approximated by a half-sample replication approach.

The general estimation technique was as follows. Weighting areas within a set of selected States were paired and, within each pair, the census sample was split into halves, each half being a systematic sub-

sample of the households in the full sample. The ratio estimation was performed on each half-sample independently. Then, for any item, a function of the squared difference of the estimates from the two half-samples was used to estimate the variance from the census sample. The variance that would have resulted from a simple random sample of 20 percent of the population was also computed. The ratio of the census variance to this simple random sample variance was computed and averaged over the paired weighting areas within the State.

The choice of States as areas within which design effects should be averaged was influenced by the fact that census data were processed and published by State. A number of considerations applied. First, the standard errors for each State were required almost simultaneously with regular tabulations to be included in the publications which were released on a State-by-State basis. Second, the order in which the States became ready for publication was not known in advance. A final requirement, dictated largely by the cost of the process, was that variance calculations should be confined to as few separate areas as possible. These constraints were met by the following system:

The States were grouped into eight strata, the major consideration being to maximize the homogeneity of the design effects among the States within each stratum. In each stratum, the variance computation programs were applied only to the first State for which census processing was completed. The design effects calculated on this State then formed the basis of the sampling errors published for all States in that stratum. This system resulted in processing approximately 9 percent of the sample records in the United States in the actual computation of design effects.

The variance calculations were made for a set of representative items rather than for all items tabulated in the census. One of the census tabulations, which included both population and housing items, contained 834 cells and appeared to cover a reasonable representation of all items being tabulated. Accordingly, calculation of design effects was restricted to these 834 items. Statistics, chosen from among the 834 cells, were grouped in rational combinations for which similar design effects would be expected based on their predicted variance behavior. Average factors for these groups appear in publications of the sample statistics and comprise broad areas such as age, labor force items, migration items, etc.

BIBLIOGRAPHY

Ireland, C.T., and S. Kullback, "Contingency Tables with Given Marginals." Biometrika, Vol. 55, No. 1, March 1968, pp. 179-188.

Kullback, Solomon. Information Theory and Statistics. New York, Wiley, [1959]. 395 pp.

U.S. Bureau of the Census. 1960 Censuses of Population and Housing: Procedural History. Washington, D.C., U.S. Govt. Print. Off., 1966. 387 pp.

Waksberg, Joseph; Robert Hanson; and Peter Bounpane. "Estimation and Presentation of Sampling Errors for Sampling Data from the 1970 U.S. Census." Paper prepared for the joint meetings of the International Statistical Institute and the International Association of Survey Statisticians, Vienna, Austria, August 20-30, 1973. 11 pp., tables.