UNITED NATIONS
ECONOMIC
AND
SOCIAL COUNCIL

Distr.
LIMITED

E/CN.14/SM/4
E/CN.14/CPH/8
22 March 1968

Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA
Seminar on Sampling Methods
Addis Ababa, 3-14 June 1968

ECONOMIC COMMISSION FOR AFRICA
Seminar on Organization and Conduct
of Censuses of Population and Housing
Addis Ababa, 17-29 June 1968

USE OF SAMPLING IN POPULATION AND HOUSING CENSUSES

Prepared by the
Statistical Office of the United Nations

M68-620

Distr.
LIMITED

ST/STAT/23
22 March 1968

Original: ENGLISH

STATISTICAL OFFICE
OF THE UNITED NATIONS

USE OF SAMPLING IN POPULATION AND HOUSING CENSUSES

TABLE OF CONTENTS

I N T R O D U C T I O N

1. Sampling has a role in population and housing censuses both as an integral part of the planning, execution, analysis and evaluation of the census and through the use of the census as a sampling frame for subsequent sample inquiries. The census may be the only frame for subsequent sampling in many countries which have neither good maps, village lists, lists of living quarters, a register of population, nor a register of households. The two aspects of the use of sampling in population and housing censuses are discussed in Parts One and Two below.

PART ONE. SAMPLING AS AN INTEGRAL PART OF THE CENSUS

2. The rapidly growing needs for extensive and reliable demographic data in most countries have made the modern census of population an undertaking of great magnitude. Censuses are expensive and if all such needs were to be met through complete enumeration, the cost, time and organization required would probably place a serious strain on the resources available for census-taking in a particular country. To lower the cost and trouble of census-taking and at the same time to obtain the maximum amount of data of the best possible quality, sampling methods can be successfully introduced in conjunction with census work proper, thus enabling countries in different stages of statistical development to embark on census programmes adapted to their respective needs and capacities.

3. Depending on the types of problem to be tackled, a country may consider applying sampling methods in one or more of the following phases of a population census: (a) tests of census procedures; (b) enumeration of items in addition to those for which universal coverage is required; (c) post-enumeration field checks; (d) quality control of data processing; (e) tabulation of provisional results; and (f) tabulation of additional data and data required for special studies.

4. The discussion which immediately follows summarizes the major advantages and disadvantages of employing sampling methods before, during or after a census of population. The conditions which determine the nature and extent of such applications are also included in the discussion.

A. Advantages and disadvantages of sampling and conditions of acceptable sample operations

1. Advantages of sampling

a. Reduction in cost

5. A sample operation, as a rule, costs considerably less than a complete enumeration. Sampling, through savings in costs, makes it possible to introduce needed improvements in the census without jeopardizing the quality of census data. Inasmuch as some expenses are incurred in devising appropriate sample designs and in implementing the schemes, it is true in most cases that cost per unit is higher for a sample operation than for a complete enumeration. However, the number of units included in a sample is usually such a small proportion of the total number of units in the population that the total cost of an operation limited to a sample is only a fraction of the total cost of the same type of operation carried out for all units.

b. Shorter time-lag

6. In most countries, including those where tabulating machines are available, the long interval of time between actual enumeration and complete publication of census results impairs their usefulness. If, instead, only a sample of returns is processed and used for tabulation, the advance estimates of basic census items thus obtained would be of current validity. A carefully executed sample operation may make results available so rapidly that, in effect, it measures changes while these are taking place, and in time to be used as bases for administrative decisions. The population may be undergoing such changes that a complete enumeration, at the time when the data so obtained are tabulated and ready for use, may actually present a less accurate picture of the population than would be presented by a sample.

7. The use of sampling methods to obtain advance results is not without its drawbacks. These are discussed later.

c. Lessened burden of response

8. The burden placed on the public by responding to census questions is alleviated when only a sample of the census units is enumerated for supplementary questions.

d. Lessened demand on manpower and other census resources

9. Again this advantage is pertinent when sampling is used for broadening the scope of the census. If complete enumeration is used to obtain the basic census information and supplementary questions are asked of a sample of population, the census authority finds its task lighter in terms of reduced volume of units of information to be coded and tabulated, with the result that less office space and equipment are needed.

10. If a number of census questions are asked on a sample basis, the field staff needed is considerably less than if answers to all questions were elicited from everyone in the population. The result is reduced demand on manpower. Recruitment and training of personnel are thereby made easier and quicker.

e. Higher quality of work and appraisal of reliability

11. Since the use of sampling eases the task of recruiting personnel, it is possible to make a careful selection of high grade enumerators and office assistants. Such selection of staff may eventually lead to better quality of results quite possibly offsetting the sampling errors. This advantage assumes particular significance when the nature of census information sought is such that if obtained through complete enumeration the likelihood of errors of enumeration and processing is high.

12. One of the most important features of modern sampling is that the precision of the results of a scientifically planned sample inquiry can be determined in advance, in the sense that for each figure to be estimated one can state the probable limits of error of estimation. This is possible provided information on variances is available - from a similar earlier enquiry or other sources. This feature enables one

to design a sample survey in accordance with the precision required for the uses that are to be made of the data, or with the precision possible within the range of permissible costs. In other words, a sample can be devised which will yield results of maximum precision at a given cost.

13. Reliance can be placed on results obtained through the use of sampling methods, not just because these results may be consistent with census and other available statistics, but rather because the sampling methods provide some idea of the sampling error in the estimate. In other words, we could have an idea of the extent to which we are likely to be in error on an average in estimating the population parameters from the sample.

14. In deciding on the nature of the census, and the extent to which its scope should be broadened, it is important to bear in mind that in many situations it is possible to obtain usable information through sampling where a complete enumeration would fail on account of gross inaccuracies of responses and because of incomplete coverage. Such information may profitably be obtained at the same time as the census for more efficient estimation of population values from the sample.

15. The precision of a properly designed sample survey can be determined, after the survey is completed, by a statistical examination of the returns. The final measure of precision is independent of census data obtained otherwise and is also independent of any assumption made in the planning of the sample concerning the magnitude of population variates. That is to say, information from some previous census (such as total population, the number of dwellings units, literacy) may have been found very useful at the planning stage, particularly if this prior information is available for small areas. This information may have been used for assigning probabilities to these areas. The drawing of the sample and preparation of the final estimates are so executed that these probabilities cancel out. Obsolescence in the prior information is thus not carried over into the final results.

f. Refinement of methods

16. An essential feature of planning a sample survey is the attempt to minimize sampling errors, which in turn forces attention on minimizing non-sampling errors, viz., response errors, errors arising from incomplete samples, faulty procedures of estimation, errors arising from inadequate preparation of the questionnaire, defective field and office procedures and faulty analysis of data. If sample data are carefully interpreted and compared with data obtained by complete counts, inadequacies are often disclosed in complete counts as well as in samples. The ultimate benefit of the use of sampling in census work may be a refinement in both the sampling procedure and in the methods of complete enumeration and better interpretation of data.

2. Disadvantages of sampling

a. Possible danger of delay and higher cost of final results

17. It is mentioned above that one of the advantages of sampling is that it makes possible the saving of cost and time required for tabulating final results. However, when only a sample of census returns is processed and tabulated, these advance results have later to be integrated with those tabulated for the non-sample units to give the final results. These operations may well increase the total cost and time of tabulation beyond what these would be without the special tabulation of advance results.

b. Burden of response

18. Although asking supplementary questions of only a sample of population may ease the burden on the responding public as a whole, if the same enumerators ask both the main as well as the supplementary questions, it may overburden the enumerators and may result in a deterioration of the quality of enumeration of both the main and the supplementary topics.

19. Furthermore, asking questions which are not asked of all persons may raise legal, administrative and even political issues insofar as census information is required under statute and under penalty. In fact, if only a few persons are to be interviewed for supplementary questions, this may well be misunderstood as discriminatory.

20. Sometimes the totality of supplementary questions is divided among different samples of the population but in this case there is one consideration which operates as a restraint on the aim to ease the burden of response. This is the need for cross-tabulation of several sample factors; such cross-tabulation can be done only for those persons in the population who are asked all of the supplementary questions.

c. Limitation of coverage

21. If information of a certain type is needed for every person in the population, a sample is not adequate. It is generally not possible to obtain sufficiently accurate estimates for small administrative areas through sampling. When population counts of such areas have statutory force, the statute usually states that the populations should be determined by a census, i.e., enumeration must be complete. Moreover, if detailed cross-tabulations are required for regional studies, a complete enumeration is the only way to achieve it. It is necessary to point out that increasing the sample size to obtain cross-tabulation may not offer the real benefits of sampling, since beyond a certain sample size, complete enumeration is less expensive and more efficient. The last argument has special force in the case of countries whose total populations are small.

d. Sampling errors

22. Estimated results, especially preliminary estimates, based on samples are subject to sampling errors in addition to whatever errors may be present in data based on a complete operation and this inaccuracy limits the use of these estimates for certain purposes.

23. It unfortunately happens in applications of sampling to demographic data, in which variability in the size of sampling units tends to be large, that, even using the most efficient methods, the sampling error of estimation often tends to be too large. This consideration is especially important in connexion with the use of sampling methods for estimating change. Intercensal changes in certain items (e.g., proportionate age distributions) are small and it may be difficult to ensure that the sampling error for the detailed breakdowns desired is of a lower order of magnitude than the actual change.

24. Although a good sampling design allows sampling errors to be estimated, such sampling errors are not free from limitations. First, their computation is somewhat laborious. In certain complex sampling designs, even when short-cut methods are used, the sampling error calculations become very involved. Secondly, they are difficult to present in a form intelligible to the layman, since they represent only the spread of the probability distribution of the error due to sampling. In other words, the sampling error computation provides probable limits to the magnitude of the error due to sampling; contrary to the impression conveyed to the layman by the expression "estimated sampling error", the computation does not attempt to estimate the error due to sampling, in the sense of the actual discrepancy between the sample value and the population value. The sampling errors are therefore less useful to the layman than he is sometimes led to suppose.

e. Public confidence

25. Politicians, administrators and the general public, under the mistaken impression that complete censuses are necessarily more reliable than any other type of inquiry, may lack conviction about the dependability of the results of sample inquiries. Such prejudices can only be overcome by educational publicity.

3. Conditions of acceptable sample operations[1/]

a. Cost and census resources

26. The question of cost in sampling is of crucial significance. Numerous factors govern the cost of a sample and it is essential that these be fully weighed before a decision is made to combine a sample plan with a complete count. One important factor, for instance, is the size and complexity of the sample, which in turn is governed by the objectives of the survey and the procedures which are regarded as most efficient.

27. The degree of heterogeneity of the population to be sampled also has to be borne in mind in calculations of cost. It may be largely urban or largely rural or a mixture. Costs are particularly high if areas are mountainous or sparsely populated with meagre transportation facilities or if there are numerous islands to be covered.

28. Procedures used in sampling human populations vary depending on the availability of lists, maps, prior census information on characteristics of the population for small areas and the number of householders or the number of inhabitants living within the boundaries, since where such information exists the sample operations can be relatively inexpensive and efficient. It can be stated in general that the better the availability of information on population of small areas, the less expensive the sample required to yield the desired precision. In addition, the prior information need not necessarily be very accurate or up-to-date. Inaccuracies in the prior information are not carried into the final estimates of population values. Ancillary information related to sample subjects may be used either at the sampling stage or at the stage of estimation - in the latter case by allowing more accurate estimates based on regression and ratio methods.

29. Sample operations have to be conducted under the direction of a competent statistician who is conversant with both the theory of sampling

---

1/ The discussion in paragraphs 26-41 relates only to purposes (a), (b) and (c) enumerated in paragraph 3.

and with the practical difficulties of carrying out surveys. His advice is indispensable at all stages of the sample procedure, that is from planning of the sample to estimation of population values.

30. Knowledge of sampling facilitates the effective deployment of available census resources, namely equipment, experience and personnel; sampling also furthers the development of new theory required for special procedures and circumstances.

31. Considerable experience and skill are necessary in devising questionnaires and preparing instructions for office and field, particularly under unusual conditions.

32. Field and office staff have to be adequately trained and supervised. It is essential to ensure that the field staff thoroughly understand the field instructions before being assigned to the field. Instructions and records of procedures used in previous censuses are helpful, especially if these were critically evaluated soon after use.

33. Sufficient attention must be paid to tabulation procedures. The methods used for arriving at estimates of population values from the sample depend on the tabulating equipment that is available or on the capabilities of the office staff if the work is done by hand. Furthermore, the choice of the sample design itself is to some extent conditioned by the procedures that are to be used in tabulation. For example, if tabulations are carried out by hand, this may well call for the inclusion in the sample design of uniform sampling fractions over all areas in order to avoid involved computations and possibilities of errors. Moreover, hand tabulation is satisfactory for counting in broad classes, but not suitable for detailed cross-tabulations.

b. Precision

34. A necessary feature of an acceptable sample plan is a statement of the precision desired in the sample estimates. The higher the precision, the larger and/or the more complex the sample, and hence the more expensive.

35. It was mentioned earlier that in a scientifically designed sample plan, anticipated sampling errors can be computed fairly closely in advance provided there is some prior knowledge of variances. Where this condition is met, the size and type of sample may be planned in relation to the precision desired and the permissible cost. It happens in the case of certain developing countries, that cost is of less interest than available labour force and, in sampling of human populations, the aim would be to secure specified precision, or even maximum precision, for given labour resources. For successful execution of a scientifically designed sample plan, it is essential that strict selection procedures are followed. The procedures must be such that a probability is assigned to every unit in the sampling population. These probabilities are needed for estimating population values and for calculating the precision of these estimates. A further requirement of good selection procedures is that these must be simple and direct so that deviations from prescribed standards or instructions can be held to a minimum.

36. The problem of specifying the intended precision of sample results and its translation into statistical terms does not present any special difficulty when only one characteristic is to be estimated. It is usual to specify the desired value of the coefficient of variation of the estimate. When, however, more than one characteristic is to be estimated, as is generally the case in most applications of sampling to censuses of population, the problem of specifying in advance the accuracy required in the different estimates needs more scrutiny. One method sometimes adopted is to specify that the percentage errors of individual estimates should not exceed certain limits. In this case, however, there is a certain amount of arbitrariness in the choice of the permissible margins of error, and it is therefore important to examine, with a view to reducing costs, how the permissible set of intended percentage errors could be altered without seriously impairing the value of the estimates. This difficulty may be partially removed if, instead of setting upper limits to the desired precisions of the separate estimates, a suitably

defined over-all measure of precision is employed to specify the reliability aimed at.

37. An over-all measure of precision is needed for another important reason. The entries in various cells of a sample table (e.g., age-sex-marital status table) are not mutually independent and as a result the percentage standard errors of the separate estimates may not adequately reflect the accuracy of the sample table as a whole. In any case, even if the percentage standard errors in the various cells are known, it is difficult to comprehend fairly accurately the degree of over-all accuracy of the sample table.

c. Sampling frame and sample units

38. The selection of sample units has to be made from some kind of a complete list of these units, namely the sampling frame. The sampling frame may be a list of small areas, structures, households, people or groups of households. Before a sample selection is undertaken it is necessary to ensure that the sampling frame is free from such defects as inaccuracy, incompleteness, duplication, inadequacy and obsolescence. Emphasis should be laid at this point on the importance of bringing the frame up to the date of the sample inquiry, if such an inquiry is conducted following the census. Otherwise the inquiry results would be biased.

39. If a multi-stage sample design is decided on, the specification of the sample unit to be used at any stage depends not only on what is desirable but on the maps and information which are available from previous censuses. Also, at any stage, sub-frames are required only for the sample units selected at the previous stage.

40. Because basic census conditions differ in different countries, it is difficult to lay down any general rules concerning the choice of sample units employed at various stages of a multi-stage sample design. In most African countries it may be assumed that any sample field operation in relation to a census is likely to be based on a single

stage sample of areas. Such area units should preferably be approximately equal in population and an estimate of the population of each unit should preferably be available. Census enumeration areas are often convenient units for this purpose.

41. This topic is discussed in greater detail in the final section of this paper in connexion with "The census as a basis for subsequent sample inquiries."

B. Tests of census procedures

42. In planning a census, preliminary decisions are required in respect of alternative procedures in various aspects of the design and conduct of the operation. These decisions can then be tested by conducting a sample inquiry, and in most instances the optimum solution can be found with a high degree of precision. Such procedures facilitate a more desirable allocation of available census resources than is possible otherwise and thus enable a country to undertake a census programme adapted to its needs and capacities.

43. Though testing of every aspect of the census plan is desirable in principle, samples of particular interest in Africa are provided by tests relating to (a) questionnaire content and format (errors of response in residence status, age, literacy, wording and arrangement of questions, etc.); (b) size of field staff and related costs (performance rates in field interviews, transport requirements); (c) field organization and control.

44. The nature and extent of census testing depend on the experience which is available from previous censuses or other sources. In countries which expect to expand the scope of their censuses rather substantially in relation to their previous censuses extensive pre-testing may be necessary. This may warrant a broad experimental approach to the entire census plan. Sampling, since it offers a scientific basis for such an approach, can make a valuable contribution towards providing reliable

answers to a number of questions with regard to the plan. In particular, when previous information is not available, it is often worthwhile to organize a pilot survey to obtain an estimate of variability of the important characteristics under study.

45. In housing censuses, the general lack of prior housing statistics makes it difficult to assess in advance the variability of the statistical material to be investigated and also the quality of the interviewers. Therefore, there is usually a need to obtain an idea of these factors by means of an intensive pilot survey.

46. When carrying out census tests, it is well to remember that one does not always have to use probability sampling. Non-random or purposive samples may be used where, for testing the effectiveness of a projected course of action, it is not necessary to base the final decision on any quantitative measures derived from data obtained by means of the test, and particularly where the biases introduced are not of very great significance for the problems investigated. On the other hand, when quantitative measures are needed for comparing efficiencies of different fields procedures (for instance, in examining the anticipated response errors arising from enumerating the de facto as opposed to the de jure population or from different systems of obtaining reports of age) probability sampling must be used.

C. Enumeration of items in addition to those for which universal coverage is required

47. As mentioned earlier, the rapidly growing needs in a number of countries for extensive and reliable demographic and housing data have made sampling methods a very desirable adjunct of any complete census. Sampling has been used in several countries for broadening the scope of the census by asking a number of questions of only a sample of population. Modern experience in the use of sampling techniques has confirmed that it is not necessary to gather all demographic and housing information on a complete basis; the use of sampling actually saves a good deal of time and money and furthermore, under certain circumstances, the sampling approach alone ensures data of acceptable accuracy.

48. An important limiting consideration (already briefly mentioned) which should be borne in mind is the following: Complete counts of the population are taken for granted in most decennial censuses. Basic legislation in some countries requires that the size of the population be obtained as a complete count. In addition to the legal requirements, the uses to which population counts are put are so vital that even trivial errors can have important effects on policy and administration. Most importantly, population counts are often used for legislative apportionment. In addition, during the decade between censuses, large amounts of central (or federal) and provincial funds are distributed on the basis of the population distributions.

49. In analyzing the effects of using samples, it is wrong to consider complete counts as being without error and differences between complete counts and samples as reflecting differences between truth and approximation. It is now generally recognized that in the field of economic and social statistics perfection is unattainable. Statistics on these subjects can be obtained only with various degrees of approximation because of the complexity of the definitions involved and nature of the data-collection process. Furthermore, census data are normally used for a period of ten or more years, and the users either treat census distributions as if they were constant over time or extrapolate the results, frequently on the basis of limited information. Superimposed on this is the fact that census data are frequently included in analyses in which the census definitions are not exactly what the users want, but come close to it, and the census is the only source of related information.

50. Response to supplementary questions may be obtained by one of two different methods. Either the same enumerators may ask the supplementary questions or especially trained enumerators may be engaged for enumerating the sample.

51. An important case of the second method, namely the use of especially trained enumerators for the sample, is the "build-in sample", in which

a sample of areas is covered by a specially trained corps of enumerators at the same time as the census, using a more detailed questionnaire. This operation replaces the census in the sampled areas. This method has certain advantages over the first alternative: namely, (a) it avoids biased sampling by the enumerator, (b) the enumerator's task is kept as simple as possible and (c) since only a small number of enumerators are trained with the additional questions, the quality of work is generally high. The build-in method is known to be cheap (generally more so than the PES method mentioned in paragraph 66). Indeed, the cost advantage of this method could approximately be the same as in the first alternative. Built-in samples do not, however, provide the check available from the PES method, except at a total level (see paragraph 66).

52. Field staff for sampling in connexion with a housing census is often required to compile detailed information from the respondents, and may have to make subjective decisions to place a respondent's answer under the correct category. For this reason, sampling field staff is required to be of high calibre, well trained and preferably with prior experience and/or knowledge with regard to housing statistics. Because housing censuses often take place in conjunction with population censuses, the scope of the housing data to be collected may have to be limited by the quality of the interviewers.

53. The suitability of particular questions for a sample enumeration depends on the precision with which results are needed for small areas, population groups or groups of living quarters and on the enumeration costs involved. It is quite possible that certain information, available for larger groups, may be required for small marginal groups for some future research, the need for which could not be foreseen at the time of the sample enumeration. This consideration limits the usefulness of a sample enumeration to some extent, even though information needed immediately can be provided with adequate precision. Account has also to be taken of the desirability of having results of high precision on basic items for areas of various sizes, as those results serve as

benchmarks for numerous inquiries. In fact, it is widely recognized that a simple complete census, even of imperfect quality, almost always constitutes the best basis for a sample inquiry.

54. The sampling errors of estimates of population values derived from sample enumeration can be kept low if use is made of certain related values obtained independently through complete enumeration made in the same census, since it then becomes possible to use the more accurate regression method or ratio method of estimation.

55. The technique of two-phase sampling[2/] can be effectively used in sample enumeration when the cost and accuracy conditions cannot be simultaneously satisfied for single-phase sampling. Suppose it is desired to assess the rate of literacy by using an objective test, which is to be administered to one individual at a time. This requires a lengthy interview and enumerators must be specially trained in the use of the test. Since the total cost of sample enumeration must be held within the amount allocated for it, the data must be secured from a small sample of the population. However, it may be that within these cost limits it is not possible to select a sample which would in itself yield sufficient accuracy. Now suppose the variate is known to be reasonably highly correlated with a second variate (e.g. 5 or more years attendance at school), which can be determined much more rapidly and at low cost per head. Since a very accurate estimate of the second variate can be secured at relatively small expense, and since for any given value of it, the variance of the original variate will be smaller than it is in the whole population, a more accurate estimate of the original variate may be obtained for the same total expenditure by conducting the sampling in two phases. The first phase is to secure data for the second variate only, from a relatively large sample of the population in order to obtain an accurate estimate of the distribution of this

2/ If certain items of information are collected from the whole of the units of a sample and other items of information from a sub-sample of these units, the procedure is termed two-phase sampling.

variate. At the second phase two methods are available. One method is to divide the first-phase sample, as in stratified sampling, into classes or strata according to the value of the second variate and to draw at random from each of the strata a small sample for the costly intensive interviewing necessary to secure data regarding the first variate. A second method, which is simpler and generally preferred, is to draw the second-phase sub-sample without any stratification and to make use of the first-phase sample only at the estimation stage: this is done by raising by the ratio total of second variate in first-phase sample to total of second variate in second-phase sample. An estimate of the first variate based on either of these methods will be more accurate than one based on an equally expensive unrestricted single-phase sample.

D. Post - enumeration field checks

56. As the scope of the modern census is ever increasing with time, the problem of attaining census data of high quality is becoming both more important and more difficult. In the present state of census practice, no procedure is yet available which can keep the census enumeration entirely free from errors. However, in order that census data can be put to practical use, one should attempt to ensure that departure from accuracy does not exceed reasonable limits.

57. Census enumeration, as it extends to every census unit in the population and may involve numerous questions, cannot be completely checked since the cost of complete re-enumeration would be prohibitive and the burden of response imposed on the public excessive. One way of finding out the types and magnitudes of errors in the census enumeration is the re-enumeration of samples of census units for certain questions, as soon as possible after the census day. Another method but seldom applicable in Africa, is to compare samples of census questionnaires with records known to be reliable, e.g., vital registration documents in some countries.

58. The introduction of modern sampling methods has brought about a change in the attitude towards census taking. Until recently, the aim was to produce census results of the best possible quality, without special attention being given to the question of the level of desirable accuracy and to what extent the added cost of increasing accuracy was justified by the final uses. When post-enumeration sample surveys are used for assessing modern censuses, the census results are evaluated not so much from the view-point of attaining the highest possible accuracy *per se* as they are on the principle of determining the level of accuracy that is optimum by balancing the losses due to errors against the costs of greater accuracy.

59. Post-enumeration field checks serve two principal purposes. The one is to inform users of data about the accuracy attached to census figures and the other is to aid census officials in the improvement of subsequent censuses. Insofar as the checks serve the latter purpose, the task of testing subsequent censuses on an experimental basis is greatly eased.

60. As mentioned earlier under section B, "Census tests", one need use probability sampling only when precise quantitative measures are required for appraising the census data (for instance, in determining the amount of under- or over-enumeration) with the best possible reliability. Non probability or purposive samples may be used where quantitative measures derived from the sample data are not of crucial importance, e.g. for testing the effectiveness of a certain census procedures and particularly where the biases introduced by the re-enumeration sample design are not of great significance for the problems investigated. However, while deciding between a probability or a purposive sample, it is well to keep in view the requirements of future demographic or housing inquiries including the next census, since a probability sample, although it may be more expensive in the short term, may well provide a scientific and continuing basis for future investigations of a related nature.

61. Where estimates derived from sample results are used to determine quantitative inaccuracies in census information, the estimates should be stated along with their sampling errors, so that the two components of deviation (sampling error and bias) can be correctly interpreted. Every attempt should be made to obtain the best value of the sampling error and for this reason the estimates of the pertinent variances and covariances should be carefully evaluated. The usual practice in developed countries is to state in advance the precision to be desired in those estimates which are regarded as the most important post-enumeration checks and then to choose the sample design which would cost least. Sometimes, estimating sampling errors for every sample estimate is excessively expensive; in this case, one can obtain sampling errors for a few important estimates and use these as benchmarks for making approximate estimates of errors affecting other sample data by means of some simple error law.

62. Quality of census data may be adversely affected by the following two factors:

(1) content errors,
(2) errors in coverage.

63. The checks instituted to discover these errors are known as "content check" and "coverage check" respectively.

64. Content errors comprise errors in respondent's answers, biases and errors on the part of the interviewer and those arising from imperfection in the design of the questionnaire. The content errors are usually subjected to critical census tests preceding the actual census by conducting experiments under census conditions, but experience has shown that post-enumeration field checks are necessary to determine the nature and extent of such errors in the census enumeration.

65. Special precautions are generally taken in order to limit re-enumeration errors in the collection of information from the sample selected for the post-enumeration field checks. The check is conducted

as close as possible to the date of the original census, using for the re-enumeration the best available personnel (either the supervisors or the best enumerators from the census). The advantages of short time interval are: there will have been a minimal movement of population, respondents will still have fresh in mind the information as of the date of the census, and the public will be prepared to co-operate with the new inquiry because they have not forgotten the publicity and other activities connected with the census. The best of the supervisors and enumerators from the regular census are selected to serve on the post enumeration field check; they receive intensive training and are remunerated on a basis allowing them ample time to obtain valid replies. When necessary, their work is verified by cross-checks with other elements of information. Objective tests (of literacy for example) are sometimes used instead of questions when it is thought that questioning will not yield reliable information. It is important to use as respondents responsible persons, preferably the head of the household or the individuals about whom the questions are asked, even if this means repeat visits. It is necessary to use a carefully designed questionnaire, with clear instructions to elicit the most exact responses.

66. Normally the use of the post enumeration survey (PES) method, with specially trained enumerators, should be preferred to allowing the regular census enumerators to select their own sample of households as they go along. The PES has the following advantages over the latter: (a) it avoids biased sampling by the enumerator, (b) the enumerator's task is kept as simple as possible, (c) since only a small number of enumerators are trained with the additional questions, the quality of work is generally high and (d) it can combine a case-by-case check of the census with the role of collecting additional information. However, the PES method is expensive (generally more so than the built-in method mentioned in paragraph 51).

67. To accomplish its purpose, a true post-enumeration field check, particularly one instituted to check coverage errors, should meat three separate requirements, namely,

(1) it should be independent of the original census;

(2) it should be representative of the whole country and of all population groups;

(3) it should involve one-to-one matching of records to produce an identical group from each investigation.

68. To achieve independence in the post-enumeration field check the ultimate sampling units for re-enumeration should be clearly defined and operationally convenient geographic areas. The sampling units for re-enumeration should be selected not from a list of housing units, households or names enumerated at the census being evaluated, but from a comprehensive frame consisting of geographic area units, preferably the enumeration districts used in the original census, the probable population-size of which is already known and within which every household will be re-enumerated.

69. Representativeness in respect of the whole geographic area and of all population groups calls for the post-enumeration field check to be carried out in a well designed probability area sample, small enough to hold down the cost but large enough to give at least an evaluation of the enumeration in the country as a whole and in its major civil divisions, since determination of quantitative measures of under- and over-enumeration is one of the important objectives of the procedure. Should there be suspicion of great regional variability in the quality of enumeration, a 100 per cent re-enumeration in the areas suspected of large errors might have to be carried out separately, but results of such intensive checks might then be combined with those of a sample of the rest of the country to evaluate over-all accuracy.

70. The one-to-one matching of census schedules to post-enumeration field check schedules is the essence of the re-interview method of census evaluation. However, since a typical post enumeration sample is likely to run into several hundreds of thousands, a complete one-to-one matching for such a sample is virtually inconceivable. Therefore, it might be

desirable to consider the possibility of such one-to-one matching on a subsampling basis. There are however difficulties in name-to-name matching in any country and in cultures where names lack uniqueness it may be practically impossible. Even identification of living quarters and the household may be difficult in those parts of the world where street names and house numbers are practically non-existent. Nevertheless, since the efficacy of the post-enumeration field check as a remedial measure in census taking rests on the identification of errors and their correction, comparisons must be made on a one-to-one basis to the degree possible. Naturally, comparison can be made between gross numbers enumerated in various parts of the country and among various segments of the population with corresponding persons counted in another survey but the possibility of compensating errors might invalidate some conclusions based on evidence derived in this way. Moreover, one-to-one matching of schedules aids in the analysis of the types of error by various population groups, thereby making it possible to improve future census methodology.

E. Quality control of data processing

71. It was mentioned in the previous section that the quality aspect of census data has of recent years come to the forefront of census problems. This statement is even more valid in the case of mass-processing operations of a census. Whereas the previous discussion centred on detecting the sources and sizes of errors in census enumeration, the present discussion concerns the use of sampling for measuring and controlling the quality of census processing operations. These include the editing of questionnaires, their coding, the recording of data on punched cards (or tape production) and their verification, some computations (rates, etc.) and the tabulation of figures. Even in a country of medium population size, these operations involve millions of questionnaires and cards and a far greater number of columns of figures and perforations.

72. Under appropriate conditions, sample inspections and in particular quality control techniques make it possible to control effectively and economically the quality of census data processing and it is no longer necessary or expedient to locate and correct substantially all processing errors by carrying out each operation a second time by way of verification.

73. The magnitude of processing operations of a modern census and their mass execution make it impossible in practice to achieve complete accuracy in census results. Even if it were possible to eliminate all unambiguous errors after several complete reviews of each operation, there are always certain characteristics which are so difficult to classify that even two experts may not agree on the proper code or enumerator's entry. Such considerations suggest the adoption of statistical methods of the type used for controlling the quality of mass-production in industry. This course results in a much lower census cost than if the control had been performed on a complete basis. Moreover, it is in any case not justified to increase the processing costs of a census beyond a certain point for improving the accuracy of census results since these will still be affected by defects in the enumeration (e.g., incompleteness of enumeration and response errors, etc.) and other errors independent of processing. In fact, if the material coming in for processing is already known to contain serious biases, one should exercise extreme caution in deciding on the level of cost of processing to improve the over-all quality of census results. The case for sampling inspections of the type used in industrial quality control is reinforced by the philosophy of not paying for perfection, while still ensuring the level of accuracy that is optimum by balancing the losses due to errors against the cost of greater accuracy. In this connexion one important consideration may be mentioned as guiding in the decision on the optimum method of quality control. When final tabulations are in view, attention has to be paid to the reliability of figures in a series of detailed cross-tabulations, especially those concerning relatively small areas or small groups of individuals. This demands accuracy in small lots of data and imposes more stringent conditions in the application of statistical methods of quality control than if only larger areas or groups were to be considered.

74. In order to obtain the best results from applying methods of quality control to census processing operations, it is necessary that certain prerequisites of an organizational nature are fulfilled. These include the availability of well-trained and reliable clerks and machine operators, and the existence of adequate work procedures and discipline and therefore of competent office management. These same conditions constitute requirements for carrying out satisfactorily the quality control operations themselves, which form a succession of exactly defined tasks to be executed rigorously as prescribed. Careful thought should, therefore, be given to deciding to what extent, if at all, statistical methods will be used to control the quality of the processing of a census. This decision is particularly delicate to make when there exist urgent problems of increasing productivity per clerk through improved office organization, procedures and methods of recruitment and remuneration, better work conditions, stability of employment and provision of more and better equipment. These problems are often very acute in the census offices of economically less developed countries. Statistical methods of quality control can contribute to the solution of these problems by providing useful qualitative data concerning the deficiencies to be remedied and the standards of performances to be established and maintained, and also by reducing the amount of census processing work through the elimination of verifications by repetition. But the introduction of these methods can never be a substitute for the solution of the fundamental problems of administrative organization and management mentioned above; these methods should not be relied on before an adequate organization is realized. In particular, it would be disastrous for the processing of a census in a poorly organized office, if the substitution of statistical methods of verification for the traditional ones gave rise, among the census staff, to the obviously unfounded belief that henceforth errors will not matter very much, since they have been recognized as being inevitable and will be controlled in any case by the new scientific methods.

75. Some of the processing errors which are likely to affect results most seriously are eliminated by verifications independent of the

statistical controls. These verifications include the usual simple checks on the numbers of documents processed in order to avoid the loss or misplacement of whole lots, and the mechanical editing of information transferred to punched cards which simultaneously provides control over many types of punching error.

76. Experiments with statistical methods of quality control in census data processing have so far tended to show that it is not practical to use these methods for ensuring, with the usually desired high level of probabilities, that the census results published for any given small population grouping do not contain processing errors exceeding certain limits. This would require inspection rates so high as to make the cost of the inspection excessive, and would raise almost insuperable problems of error record-keeping, tracing and correction. The aim of statistical quality control of census processing is therefore limited primarily to ensuring that the frequency of errors in the various census processing operations is small enough for corrections to be unnecessary, and subsidiarily to avoiding the inclusion in the results of those work lots which are of extremely poor quality. In technical terms, this means that acceptance sampling techniques are to be used as a by-product of process control in order to prevent an unusually poor lot from being absorbed in the operation. These conditions can be met with moderate inspection rates. It is possible with a low sampling fraction to detect the occasional extremely poor unit of work so that it can be set aside and verified completely. The controls bear primarily on the performance of individuals. These result in detecting operators whose error rates are above certain minimum standards and in removing them from a particular type of work (at least until they have been further trained), rather than in rejecting moderately defective work lots for further verification and correction. Each individual's performance has to be tested at the beginning of the operation and thereafter from time to time (e.g., every two weeks) since variations over time may be present. If the performance is found to meet consistently the pre-established standards, the lots of work produced by the individual are assumed to contain few errors, and are accepted without further verification.

77. The experience of various countries in the application of sampling to data processing has pointed to further opportunities for improvement of present plans and introduction of new sample controls. Moreover, the effort to attain better quality of census data in the phase of data processing has brought into focus the necessity of improving the quality of work done during the previous phases. In particular, the need for an effective plan to control the quality of enumeration has been recognized. As has been observed in the previous paragraphs, a considerable amount of progress has been achieved in measuring and controlling the quality of a census after it has taken place, but to date not enough is done about controlling quality at the time of the actual enumeration. Consideration of these problems would lead to the important question of integrating quality control of processing with the quality of work of the previous phases.

F. Tabulation of provisional results

78. A complete national census is a huge undertaking particularly when the housing census and the population census are carried out as a combined operation. Unless sampling methods are introduced, several years may elapse beyond the date to which the census refers before some of the tabulations of the greatest economic and social interest are published. This lag may be expected in countries which are large and which do not take censuses regularly and frequently and also in the economically less developed countries where modern tabulation equipment may be lacking or inefficiently used. This danger is no less serious in countries where modern tabulation equipment is available in large quantities, since most of the increased capacity of such equipment has to be utilized to meet heavy additional demands made by the modern census in those countries.

79. It is natural, therefore, that provisional tabulations should be the most frequent application of sampling to census work and perhaps the first one to be made in many countries. Although it is true that in most countries enumerators prepare summaries referring to the number of living

quarters, the number of persons enumerated and some other basic facts at the time of enumeration, there is nevertheless a definite need for early tabulation of some of the results such as the distribution of the population by age groups or by type of living quarters, employment status, branches of industry, etc., which cannot easily be derived from enumerators' preliminary reports. The need for advance tabulations is evident even in a small country which takes censuses of population regularly. It should be mentioned at this point that advance sample tabulations can be used to the best advantage when they are obtained for fairly large areas and for the country as a whole.

80. The greatest benefit of using sampling for obtaining advance tabulations is speed, since sample results can be accomplished in only a fraction of the time needed to make the same tabulation for all census units. As mentioned earlier, in most countries, including those where tabulating machines are available, the long interval of time between actual enumeration and complete publication of census results impairs their usefulness. If, instead, only a sample of returns is processed and used for tabulation, the advance estimates of the basic census items thus obtained would be of current validity. A carefully executed sample operation may make results available so rapidly that, in effect, it measures changes while these are taking place, and in time to be used as bases for administrative decisions. The population or the housing situation may be undergoing such changes that a complete enumeration, at the time when the data so obtained are tabulated and ready for use, may actually present a less accurate picture of the population than would be presented by a sample.

81. Another point to note about advance tabulation is that comparisons can be made in due course between the provisional figures based on samples and the final ones obtained in the traditional manner. The advance tabulations may also provide a convenient tool for internal checks and analyses. For example, one can use the preliminary sample to determine rates of non-response in the census for various characteristics and to decide on final tabulation specifications.

82. On the other hand, securing advance tabulations through sampling has certain disadvantages. The results tabulated for the sample units have to be integrated with those tabulated for the non-sample units to give the final results. These operations increase the total tabulation time, and its cost, beyond what they would be without the special tabulation of advance results. Precautions are necessary in order to minimize the delay that may be caused to the preparation of the final results. To that end, the census plan should carefully co-ordinate the special tabulation of advance results and the tabulation of final results. Otherwise, the sampling scheme may buy time temporarily at the cost of delay in the main programme, particularly if the possibility of assigning additional staff to deal exclusively with the sample has to be ruled out. Another disadvantage is that, owing to the preliminary character of the results, less than wholehearted public interest may be forthcoming for sample results and as a result the tables may be put to only limited and unimportant uses. Still another shortcoming of sample tabulation is revealed when the advance demographic estimates are used to assess changes from the previous census. Intercensal changes in certain items (e.g., proportionate age distributions) are small and it may be that the sampling error of the estimate for the detailed breakdowns desired is of a lower order of magnitude than the actual change.

83. If sampling has been used as an integral part of a complete enumeration to broaden the scope of the census, by obtaining supplementary information, the same sample of individuals, living quarters, households, farms or other units will also provide a sample for advance tabulations of the census proper. Such a sampling scheme, if it is devised efficiently with a view to securing additional census information by small administrative units, may offer excellent opportunities of conveniently obtaining advance tabulations by the same administrative units. This would not be feasible ordinarily without incurring unduly heavy expense and inconvenience.

84. Even when no sampling has been used in the actual enumeration the sample design for advance tabulations may be comparatively simple because the complete census returns provide the sampling frame and the domain or domains of study. Further benefits which such a simple scheme offers are that sampling may in most cases be carried out at a single stage at little or no extra cost, and there is a resulting automatic geographic stratification by area as the returns are generally filed in this manner. If a multi-stage sample design is used for the purpose of getting preliminary tabulations on a sample basis, it is important to remember that the cost of eliciting supplementary information from a particular sample unit at any stage may be considerably less than that incurred for enumerating the same sample unit in a field survey.

85. In the choice of a suitable sample unit, it is essential that the sample units are extractable physically from the remainder of the returns, so that they can be sent through the operations of processing without interference with the regular processing of the census. The sample units must therefore be the same as the fundamental census units that are ordinarily processed as inseparable entities. In many countries those entities are definable areas known for instance as "enumeration districts or areas"; in some countries enumeration areas form suitable sample units. A sample of enumeration areas drawn in a proper manner is then to be put into the stream of processing for advance tabulations.

86. In determining the size of a sample required for preliminary results, it has to be borne in mind that each increase in the size of the sample gives greater detail and more accurate results at the expense of speed in the analysis, for the latter decreases as the sample increases. In the interest of obtaining quick results, it is necessary that the sample be no larger than that required to give only the main characteristics of population with no more than reasonable accuracy, say a coefficient of variation of up to 5 per cent. Where punched cards are used to analyse a census, the only additional work involved in obtaining preliminary results from a sample is that concerned with the selection and the machine

analysis of the sample. The coding of the sample forms and the punching of the sample cards are a part of the work required for the complete analysis and need not be repeated for the latter.

87. It has been noted earlier that there may be some reluctance on the part of certain sections of the public to accept advance sample tabulations of a census as sufficiently reliable. Therefore, the responsibility of securing the closest and the most economical agreement between sample results on the one hand and census results and other independent data on the other devolves on the statistician. When a sample is scientifically designed and carefully executed, it is true that the disadvantage of the presence of sampling errors in the results is more than offset by the value of the extra information and cross-tabulations which sampling makes possible. If, however, scientifically prescribed standards are not strictly adhered to, discrepancies between sample results and other data may arise from such additional causes as biased sample selection, biased interview method, processing errors and biased estimation procedures.

G. Tabulation of additional data and data required for special studies

88. The principal limitations on completely tabulating all the information collected in a population census are time and money. Consequently the tabulation programmes of most population censuses provide for the complete tabulation of only those elements of demographic and economic information which are regarded as of sufficient importance for small administrative areas. Numerous characteristics are tabulated only on a sample basis. Sampling is thus used to extend the scope of census tabulations and for obtaining data for special demographic studies.

89. There are certain demographic and housing characteristics which are needed only by large areas and for the country as a whole. Sampling enables one to obtain tabulations for large areas with reasonably small sampling errors and at a much reduced cost and in shorter time than needed for tabulations on a complete basis. Even greater savings in

cost and time than in the case of simple tabulations are achieved if cross-tabulations are prepared on a sample basis.

90. However, in the case of a housing census, since one of its purposes is to serve local interests, the feasibility of sampling is determined to some extent by the size of the smallest localities for which separate tabulations are needed. In order to obtain data of acceptable precision for these divisions and localities, it may necessary to use high sampling fractions in view of the great variability in housing statistics. There may be instances where even the use of high sampling fractions fails to give results of acceptable precision; in such cases it would be inadvisable to use sampling methods.

91. The use of sampling for enlarging the scope of census tabulations can be most advantageous if the sample used for this purpose is the same as that used in the case of preparing advance tabulations.

92. The problems of sample design and sample selection in the present case are similar to those encountered in connexion with tabulation of advance results on a sample basis and therefore are not repeated here. However, it should be noted that whereas advance sample tabulations are prepared for meeting urgent needs after a census and are published eventually on a complete basis, here one is concerned with sample tabulations and cross-tabulations which are not intended to be prepared on a complete basis at any later date. This important different highlights the need to exercise the utmost care in designing and executing the sample with a view to attaining the best accuracy within permissible costs. Not only is it necessary to attain the best possible accuracy in the sample estimates, but it is also desirable that measures of precision should be supplied to users along with sample tabulations and cross-tabulations. Since these tabulations will never be followed by more accurate and complete results, the information on the margins of error of the tabulated results will enable the users to get a realistic picture of the demographic phenomena in which they may happen to be interested.

93. There is also the difference in timing of the two operations. In the case of advance sample tabulations, the full advantages of speed are attained by selecting a sample of the returns from a complete census and processing the sample returns in advance, that is, before the remainder of the returns are processed. In the present case, as a rule, a sample of the returns from a complete census is selected only after the general editing has been carried out and in countries using mechanical tabulation these sample tabulations may well be delayed until the cards are punched. If however, the tabulations to be made from the sample required the coding of particular information and its transcription on punched cards, and if in addition such information is not needed for the complete tabulation, these coding and punching operations may be carried out after the sample selection.

97. It was pointed out previously that before a sample selection is undertaken it is necessary to ensure that the sampling frame is free from such defects as:

(1) inaccuracy,
(2) incompleteness,
(3) duplication,
(4) inadequacy,
(5) obsolescence[3/]

98. Inaccuracy in a frame may arise not only from wrong information about the units listed in it or defined by it but also from listing of units which are in actual fact non-existent. If some units of the material are omitted entirely, the frame is called incomplete; if some units of the material are included more than once, the frame is said to be subject to duplication. A frame used for a particular sample survey may sometimes exclude certain parts of the material which the survey is expected to cover; such a frame is called inadequate for the survey under study. A frame, though accurate, complete and free from duplication at the time it was constructed, may no longer be so at the time it is required for use. Such a frame is called obsolete and defects of the first three of the above types may be introduced through the use of an obsolete frame.

3/ The following discussion of these five defects is drawn primarily from F. Yates, Sampling Methods for Censuses and Surveys, 2nd ed. revised and enlarged, New York, 1953, pp. 60-62.

99. Each one of the five types of defect in a frame introduces defects of a different type into the sample procedure. Inaccuracy in the frame, to the extent that it relates to the selected units, will automatically be discovered and corrected in the course of the survey and so the sample results will remain unaffected. In the event that the information contained in an inaccurate frame has been used as a basis for stratification or as supplementary information, the inaccuracy in the frame will not be carried into the final estimates of population values because the actual accuracy attained in the results can be assessed from the results themselves.

100. Incompleteness in the frame is a serious defect for two reasons. First, it cannot be discovered while the survey is in progress and secondly, only units possessing some special characteristics are likely to be omitted from an incomplete frame. As a consequence of such omissions of the survey material, biases will be introduced into the survey results and, particularly in view of the second deficiency, these biases are likely to be very serious. Duplication, on the other hand, does not suffer from the first deficiency mentioned above inasmuch as an examination of the frame will bring to light any units included more than once; the second deficiency, nevertheless, does occur in a frame subject to duplication but this will have no effect on the survey results if prior to the survey the duplicated units are rejected after a careful examination of the frame. A complete examination of the frame, though capable of yielding unbiased results, almost always means a tedious operation and consequently results in increased cost of the survey. If, however, a certain amount of bias can be tolerated because of the possibility of reducing survey costs substantially and if, furthermore, the degree of duplication in the frame is large, a fairly high sampling fraction can help reveal the extent of duplication in the course of the survey itself. When one deals with EAs (Enumeration Areas), however, incompleteness or duplication in the list of EAs is not generally a serious problem, but incomplete coverage or overlapping

on the ground is of great importance and the central problem consists of defining the boundaries of EAs.

101. Inadequacy in the frame is the least troublesome of the five defects. The extent of inadequacy is known merely from the specification of the frame and the scope of the proposed survey. Often, the construction of a subsidiary frame covering the omitted categories can resolve the difficulty.

102. The foregoing observations point to the desirability of a careful investigation of those parts of a population census which are to be used as the sampling frame for a proposed sample survey, since as pointed out earlier, some defects are not at all apparent until a detailed examination has been made. Such an object may be accomplished by examining the relevant administrative machinery of the census and the way in which the census information is kept up-to-date; sometimes the investigation may even entail a certain amount of field work.

103. Although this remedy is appropriate when the frame in question is a population or a housing census already enumerated and processed, a much more effective approach toward avoiding the above mentioned defects would be to give very careful consideration to the problem at the stage of planning the census, even though here one comes up against the difficulty of not knowing the precise nature of the subsequent sample surveys which a country may find it necessary to undertake. The usefulness of a census as a sampling frame is enhanced if in addition to careful planning, the various phases of the census are executed with proper care and if the obstacles to successful execution are fully recognized. Specifically, the census enumeration needs to be very carefully conducted, close attention being paid to non-response; certain important post-enumeration field checks have to be instituted to verify doubtful information; and it is necessary to ensure that editing, coding and punching operations are free from any defects, assuming, of course, that the last-mentioned operations are among the relevant bases for the subsequent sample surveys.

104. In housing censuses, the units of enumeration, namely, living quarters, households and buildings, may sometimes furnish the sampling frame for subsequent sampling inquiries. If data on types of living quarters are available from previous censuses of housing and/or buildings, stratification by type of living quarters could form the basis of an efficient sampling scheme.

105. Methodological factors in census-taking of the type mentioned in the next paragraph have an important bearing on subsequent sample inquiries. These factors, one must remember, are particularly relevant for sample surveys covering the whole population of a country and capable of giving certain broad subdivisions, but not for small administrative areas. The reason is that such surveys present much more difficult sampling problems than local sample surveys covering a particular town or rural area, or a few contrasted towns or rural areas in which no attempt is made to obtain a sample which is fully representative of the country as a whole. Furthermore, surveys of the first type give much greater scope for increase in efficiency by the use of the more elaborate sampling methods. Since results are not required for small areas, administrative or other areas can form the first stage of a multi-stage sample procedure thus enabling the sampling to be concentrated in relatively few areas instead of being spread over the whole country.

106. The following are some of the methodological factors to which reference has been made in the preceding paragraph.

(a) The general framework and geographical subdivisions of the field organization, e.g., regions, towns, communes and enumeration areas.

(b) Methods of enumeration, e.g., whether by the enumerator or the householder method, and whether enumeration is based on households or individuals.

(c) The general plan for processing and tabulating results, and in particular, the codes utilized (territorial units, socio-professional categories, etc.) and the geographical areas for which the results are tabulated and published. These factors have an important influence on later possibilities of stratifying and establishing of sampling fractions.

(d) The choice and composition of the various groups and sub-groups adopted for the tabulation of data; for example, would the needs of subsequent sample inquiries be better served by a de facto or de jure enumeration.

(e) In general terms the choice of the definition of the statistical units, of which the most important are the individual, the household, the dwelling unit and the enumeration area. To secure the maximum comparability between the results of censuses of population, censuses of housing and sample surveys, the definitions of these units should be, as nearly as practicable, the same in the three types of inquiry.

(f) The census as a list, not only of households or persons, but also of manufacturing and distributive establishments for use as a sampling frame for sample inquiries.

107. It is felt necessary to elaborate the first factor concerning the general framework and geographical subdivisions of the field organization since special significance is usually accorded to this factor for multi-stage sample inquiries which may be based on censuses of population to be taken in the near future throughout the world. The first stage sample units are normally in Africa enumeration areas and careful attention must be given to the demarcation of these so that they can provide particularly convenient sample units at subsequent stages. The following considerations ought to be borne

in mind if the desired results are to be achieved:

(a) It is desirable that each EA be as homogeneous as possible; each EA should if possible be entirely urban or entirely rural and the number of EA's of mixed character must be held at a minimum.

(b) The average size of the EA should approach as closely as possible the optimum for efficient sample inquiries based on the EA as sample unit. This optimum depends on the variate with which the survey is concerned. Studies in Africa and other developing regions suggests that for many variates the optimum would be around a few hundred population. It would, however, be optimistic to hope that the size of EAs might be fixed with the needs of sample inquiries primarily in view. Perhaps the most that could be hoped for in the majority of cases would be the avoidance of very large EAs. The existence of even a small number of very large EAs causes considerable inconvenience in the planning of subsequent surveys.

(c) It is desirable that the variance in the size of EAs should be as small as possible. In those cases where the geographic dispersion of the population forces the establishment of EAs with very small numbers of inhabitants, subsequent groupings of these EAs for obtaining sample units of convenient size should be considered.

(d) It is desirable to number the EAs approximately within each district, and the districts within each region, etc., so as to ensure on the one hand that contiguous geographic EAs are given consecutive numbers and on the other that the numbering provides a code indicating a certain characteristic of the EA (e.g., urban or rural character); similar precision should be applied within each EA to the numbering of housing units and to the listing of the persons on the household list.