United Nations | *Nations Unies*

ECONOMIC AND SOCIAL COUNCIL | **CONSEIL ECONOMIQUE ET SOCIAL**

UNRESTRICTED

E/CN.3/Sub.1/19
30 August 1949

ORIGINAL: ENGLISH

STATISTICAL COMMISSION
SUB-COMMISSION ON STATISTICAL SAMPLING
Third session

SAMPLING IN POPULATION CENSUSES

This note covers only a small part of the total field relevant to the problem of sampling methods in conjunction with complete enumeration, namely, the applications of sampling likely to be suitable for countries with relatively little experience in either census-taking or in sampling.

About thirty countries are planning to take population censuses (complete enumerations) around 1950. Since the usefulness of sampling in conjunction with a complete census has been discussed widely since 1940, an indication of the particular uses of sampling in this connexion will be helpful. It is considered necessary too to indicate the possible pitfalls which may impair the results, and to suggest methods which are suitable for adoption by Governments which have little or no sampling experience.

The successful application of sampling methods in population censuses in certain advanced countries has aroused interest in other countries which only lack trained personnel to organize similar sampling operations. A useful step may be taken, therefore, by suggesting simple methods which may be used by persons with no special experience in sampling. Such suggestions are likely to prove acceptable to the administrators. They may be useful too in demonstrating in a simple way the basic ideas underlying random sampling as compared with sampling by purposive selection. Thirdly, if the sampling based on these suggestions produces accurate results, then the advantages from the point of view of time, costs and organization, will be demonstrated in practice. Fourth, sampling on these lines will provide initial practical training for technicians, and finally, the temptation to try out complicated though attractive sampling schemes will be considerably reduced if the outlines for simple schemes are easily available.

General

There are a number of ways in which sampling can be applied usefully

collecting additional data. Second, as the tabulation of complete returns takes a long time - sometimes years - sampling can be used for obtaining quick preliminary results. Third, an appreciable saving in time and money can be effected by using sampling for cross tabulations in studies of inter-relationships among the characteristics of the population. Finally, sampling methods can also be used for controlling the quality of clerical operations; but in view of the complex technicalities involved it is suggested that no attempts to apply sampling in this particular field be made unless the services of persons well conversant with quality control methods are available. Discussion of this point has therefore been excluded from this paper.

The proper organization of census operations requires much preparatory work. This includes the determination of territorial regions within a country and then the demarcation of smaller civil divisions. Dividing lines between the regions should be properly indicated and a list of the smaller divisions prepared. The divisions may then need to be divided into further sub-sections in one or a number of successive stages, to indicate the operational areas of the field units. The areas in every stage of division should be clearly delimited and complete lists of such areas prepared.

The smallest divided area, for census purposes, will be a unit of area which will be suitable in size for an enumerator to cover within a specified period of time. This will be called in this paper an "enumeration unit". The units will not be equal in size all over the country. In urban areas where the density of population is high, the units will be smaller in area than in rural areas where the density is low. But whatever the size, the boundaries of the units must be clearly defined in order to avoid overlapping. In urban areas the enumeration unit may comprise one or more blocks. In rural areas, the unit may be equal to the size of a village. In the open country, where such a distinction is made, the units will probably conform with areas limited by clear geographical demarcations.

In this paper the following terms will be adopted for the successive groupings of areas. A number of enumeration units will be included in an "enumeration district". A number of enumeration districts will be included in a "county" and a number of counties will be equal to a "civil district". It is likely that very often the areas delimited for census purposes in successive steps will conform with similar successive delimitations of the administrative areas. Possible exceptions will be enumeration units, (probably urban areas), too small

to be administrative areas. The terms, civil district, county, enumeration district, and enumeration unit have been used here as illustrations and for the convenience of reference only and are not suggested for adoption in any country. The method of dividing and demarcating areas will depend entirely on conditions in the country.

2. Collection of additional data by sampling

Nature of the additional data

The number of subjects which Governments like to include in a census questionnaire will vary according to circumstances. Generally speaking, countries which have no elaborate census organization or adequate technical staff will find it practical to collect information on a small number of specially important topics, such as total population, age, sex, marital status, and some economic characteristics. On a broader view, however, there are subjects of considerable demographic interest. The United Nations Population Commission, for instance, prepared the following list of subjects:

1. Total population
2. Sex
3. Age
4. Marital status
5. Place of birth
6. Citizenship
7. Mother tongue
8. Educational characteristics
9. Fertility data
10. Economic characteristics
    - (a) Total economically active and inactive population
    - (b) Occupation, industry, and industrial status
    - (c) Population dependent on various types of economic activities
    - (d) Agricultural population
11. Urban and rural population
12. Households (including relationship to household heads)

The Commission pointed out, however, that statistics on total population, sex, age, marital status and economic characteristics were specially important and that countries with little or no previous experience of census-taking should concentrate on these subjects.

In deciding what questions to include in the questionnaire, a Government may draw a line between those questions which are basic to it and those which are useful but secondary. The basic questions may be included in the census questionnaire and the supplementary ones may

be the subject of a sampling inquiry. There is, again, a category of information such as fertility data, where the separate particulars for the smaller areas of a country are not so important as the overall picture for the larger regions or for the whole country. Sampling may be usefully applied to collect this type of data also.

Sample unit

What is the best sample unit? The choice lies between an individual, a group of individuals, and an enumeration unit, any of which may be taken as the sample unit. In the United States, during the 1940 population census, the individual was taken as the sample unit for additional questions. A sheet of paper containing 40 lines on each side - each line for an individual - with suitable column headings was used as a questionnaire and the sample was selected by designating 2 of the 40 lines on each side as the sample lines. The space for additional particulars for persons entered on the sample lines was provided in the questionnaire.

In this connexion, it should be noted that there is a danger of biased sample selection if the enumerators are not familiar with the ideas underlying the random process of selection. Even if the lines are designated by a random process and the enumerators are instructed to enter the members of households in a certain order, the tendency to put an adult in the sample line instead of a child will probably be found very widespread, because most of the supplementary questions will be relevant to adults only. Where neither intensive training on a large scale nor intensive checking and subsequent correcting are possible, it is recommended that individuals should not be taken as the sample units.

The same general remarks apply to the question of choosing a group of individuals say, a household, as the sample unit. If the initiative of selecting or identifying a household is left with an enumerator, there will be a risk of biased selection. This risk can be minimized by training the enumerators and by intensive field supervision, but organizationally, this will not be possible in many countries.

It is much safer to take the enumeration unit as the sample unit. In this case the enumerator has no selection to make. He will make a complete enumeration of all individuals within the limits of his area and no question of a biased selection arises. The selection of sample enumeration units can be made centrally without involving the enumerator. The adoption of the enumeration unit as the sample unit is, therefore, strongly recommended.

/Sampling

Sampling fraction

The size of sample suitable to obtain the results with accuracy within required confidence limits can be determined, provided some previous information about the dispersion of the sample units in relation to the characteristics under study is available. When the information regarding dispersion is not readily available, it is often worthwhile before undertaking a sample survey on a large scale to organize a pilot survey to obtain, amongst other particulars, an estimate of the standard error of the mean of one of the main characteristics under study. The necessary sample size can then be determined on the basis of (a) the estimated standard error and (b) the limits of accuracy required of the results.

However, in the case of sampling taken in conjunction with a census, the organization of a pilot survey to obtain information for ascertaining sample size is unlikely to be worthwhile.

It will usually be best to adopt a sample size which will be organizationally convenient and which will be regarded as adequate within the available resources. From the point of view of minimizing additional work, a sample of 1 in 20 may be found quite convenient. If separate estimates for smaller areas such as civil districts or even counties are considered as important, then a higher fraction like 1 in 10 may be suitable.

If it is assumed as an illustration, that on the whole, one enumeration unit out of every 20 will be selected as a sample, then the next point to decide will be how these samples should be distributed between different areas. The simplest way is to have a uniform sampling fraction of 1 in 20 in all the regions and in all the civil districts. The working sampling ratio will, however, in that case, vary slightly from district to district because the total of enumeration units in most of the districts will not be exactly divisible by 20 and approximations will have to be made such as 1 unit for a remainder of 10 or over, or none for a remainder of 9 or less. Where there are special reasons for closer study of certain areas (e.g. depressed areas) a higher sampling fraction may be used.

Since, in practice, the sampling fraction is seldom uniform for all areas of the country, the proper method of arriving at the regional or country totals will be to estimate the totals for the civil districts (or even for the counties) first and then to add up the district totals

/to obtain

to obtain totals for the larger areas.

The taking up of a number of areas only for the selection of sample enumeration units within those areas rather than selection of samples from all over the country, is a method which can also be adopted. The adoption of such a method is not, however, advisable unless the reasons for economy are compelling. The picture obtained from such an investigation will be only regional and not national. The picture will be national if the regions are so chosen as to represent the whole country correctly. But, for a correct selection in this case a considerable amount of regional information must be available beforehand. It will not therefore be advisable for countries lacking up-to-date and detailed regional data to organize surveys on the basis of a few selected regions only.

Sampling procedure

The selection of sample units purposively in order to make the sample representative of the aggregate population is not considered a correct method. This method assumes, on the part of the selector, a complete knowledge of the character of the sample units which in practice is far from complete. As a result, this kind of selection by judgement will very often be biased. Arranging all the units of a population into comparatively homogeneous groups on the basis of available information is, however, useful. This is what is known as stratification. For sampling, a part of all the units arranged within a group has still to be selected and the best way of doing so is to select the sample units by a random process where the chance of inclusion of each unit is equal. If each unit has an equal chance of being included in a sample, the variability of all the units within the stratum will be reflected in the sample.

The sample units should invariably be selected at the central census office. Complete lists of census enumeration units for each civil district will be prepared during the preparatory work and from these lists sample units can be selected with the help of random numbers. The required number of sample units for each civil district can be deduced by applying the sampling fraction to the district total.

Another method of sample selection is to select enumeration units from the lists at equal intervals starting with a random number. For instance, if it is decided that the sampling fraction will be 1 in 20 then every 20th enumeration unit will be selected from the lists starting with a number at random. This method of systematic selection has been adopted in many cases because of its simplicity. The method

is considered approximately random but the only risk is that if there is any regular periodicity amongst the sample units whose amplitude is equal to the interval of sample selection, the sample will be biased.

Tabulation

In order to reduce the volume of tabulation work at the central office, some countries follow the practice of tabulating data at the field offices. The office in charge of operating work in a census enumeration district will tabulate the primary data and send the results to the higher office and so on successively to the central office for final tabulation and analysis. The editing of primary data and the tabulation work are likely to vary in standard from one area to another in such a system and on the whole, the quality of the work is likely to suffer.

The present tendency, therefore, is to centralize editing and tabulation in order to have a uniform standard and to obtain a better quality of work. With the increasing use of machines in tabulations this tendency has been further accelerated.

Whatever the form of organization in tabulation - centralized or decentralized - the steps to be taken for estimating total figures from the sample data will be similar. First, the civil district totals, (and if necessary the county figures) will be estimated from the figures of sample enumeration units within the civil district by multiplying by a factor inverse to the working sampling fraction. The resulting civil district figures will then be combined to obtain estimates for the whole country or for the larger regions.

Again, irrespective of whether the system is centralized or not, the separate figures for each sample enumeration unit for the whole country must be available in the central office and measures should be taken to ensure that the field offices supply such separate figures for the sample units. These are the particulars on the basis of which the standard errors of the estimates for the country, larger areas, or civil districts, will be calculated.

There is a tendency to exclude the calculation of standard errors in order to avoid additional computational work. This usually arises from lack of appreciation of the value of standard errors in judging the accuracy of the data, or of the significance of differences between estimates. Since considerations relating to costs are usually the restricting factor, a possible procedure is to calculate the standard errors of one or two important characteristics only and to omit the calculations regarding the rest.

When the results of such sampling investigations are published, it is recommended that the relevant standard errors be also included in the publication. In this connexion, attention is drawn to the recommendations of the Sub-Commission on Sampling on the preparation of sampling survey reports (See Statistical Papers. Series C. No. 1 issued by the Statistical Office of the United Nations).

3. Quick preliminary results by sampling

As separate entries in respect of every individual in a country are made in the census questionnaires, the volume of work involved in tabulation is very large indeed, and the period between the actual enumeration and the publication of the results is usually long. Sampling can be used, and has been used in many countries, to select a part of the total data for tabulation so that the preliminary results on topics of immediate and general interest, such as structure of population in broad age-groups, sex, and certain economic characteristics, may be published in a short time.

Sample unit

The sample unit may be an enumeration unit, a group of individuals, or an individual. Here it is desirable that the individual be taken as the sample unit. The individual is also the unit of enumeration in the census and particulars of sample individuals selected from all parts of the country will give more accurate results than that obtained from larger and less scattered units. Since the processing is done in the central office, the difficulties referred to (page 4) when individual sampling is attempted in the field, do not arise.

Sampling fraction

Since the sampling method is applied in this case for a quick preliminary return and since standard errors of the estimates are of no practical value, the size of the sample or the sampling fraction will depend on the organizational convenience and on the time in which it is planned to publish the figures. Usually, a sampling fraction between 1 in 100 and 1 in 20 will be quite adequate.

Sampling procedure

If the form of a census questionnaire is a sheet containing a number of lines with each line meant for a person, then the sample lines may be selected at regular intervals or by a random process whichever is found convenient. The choosing of a sampling fraction may also be a matter of convenience. If, for instance, a questionnaire, including both sides of the sheet, contains 80 lines, a sampling fraction of 1 in 100 will

/be written

be rather inconvenient for selecting samples. One in 80 will be a more convenient fraction. In such a case a fixed line from each sheet may be selected, if the selection is made systematically; but perhaps the better way would be to vary the line somewhat from sheet to sheet. Say, 4 or 5 random numbers between 1 and 80 may be selected and the lines in the consecutive sheets may be selected as indicated by the random numbers. The whole order of numbers will then be repeated continuously. The use of fewer numbers and their repetition will be more convenient from the point of view of speed than the selection of sample lines on the basis of a larger set of random numbers.

Where the form of a census questionnaire is a small slip each meant for an individual, the sample slips may be selected on the basis of the serial number used in each enumeration unit. Sampling fractions like 1 in 100 or 1 in 50 will be found convenient for selecting samples because the slips with serial numbers ending in 00 or in 50 can be sorted out comparatively easily. The process of random selection is likely to take more time.

Tabulation

Whatever the sampling fraction selected, the working sampling fraction will in practice vary somewhat between different enumeration units because of approximations and minor errors. Hence it will be more accurate to estimate the totals of the small areas first and to combine these figures to obtain the totals for larger areas. The estimating of separate totals for small areas like the enumeration units may be an unnecessarily detailed procedure to obtain preliminary results and, therefore, separate totals for larger areas like civil districts may be estimated directly from the sample results by using appropriate multiplying factors. Such figures may then be combined to obtain regional or country totals.

4. Cross-tabulations by sampling

The census data provide a rich source of information for studies in social and economic fields. The information on inter-relationships between various factors requires more intensive investigation than is usually permitted by the main census tables. Such investigations can be more suitably done on a sample basis. Not only from the point of view of costs but also from that of time and the difficulties of organizing tabulation on a large scale, sampling, in most cases is probably the only practical procedure.

The question of the adequacy of the size of a sample will depend on the nature of the study undertaken and on the available resources. However, the data relating to (a) sample enumeration units or (b) sample persons selected for preliminary tabulation may be the starting basis for such a study. The work of selecting fresh samples from the records of the total population can thus be avoided. Even the volume of data relating to sample enumeration units, or sample persons for preliminary tabulation, may prove to be unnecessarily large for such studies. The data, in such cases, can be further reduced by sampling.
