UNITED NATIONS
ECONOMIC
AND
SOCIAL COUNCIL

GENERAL

E/CN.3/Sub.1/27
29 August 1950

ORIGINAL: ENGLISH

STATISTICAL COMMISSION
SUB-COMMISSION ON STATISTICAL SAMPLING
Fourth session
Lake Success, 5 September 1950

SAMPLING METHODS FOR ESTIMATING DISTRIBUTION BY SIZE OF INDIVIDUAL AND FAMILY INCOME

Paper prepared by the Secretariat in connexion with item 9 of the provisional agenda

1. Interest in statistical investigations of the distribution by size of individual and family income has greatly increased in recent years. The applications are numerous and they vary widely. Therefore it is not easy to describe in a few words the importance of this type of study and to summarize the purposes for which the results have been used. When the first studies of income distribution were made late in the nineteenth century (by Pareto and others), it was felt that they could throw light on certain problems related to the existing economic organization of society. In the early days, these studies were more of an academic nature. They were not yet needed for practical purposes of economic policy, and therefore there was less pressing need for accurate statistical methods of observation or for very up-to-date data. With the growing number of applications in the field of economic analysis, more modern methods of measurement had to be introduced.

2. Up to the present, the need has been felt for statistical information on the inequality of the income distribution. Inter-temporal and inter-country comparisons have also become important. Changes over time of the inequality of the income distribution may be related to fluctuations in economic conditions. Various studies have been devoted to this subject. Measures of economic policy may be aimed at reducing the inequality of the income distribution. This has therefore led to an increased need for information in this field, including

information on the distribution of incomes after taxation. Statistics of the distribution by size of incomes have been used also to determine the number of persons whose incomes remain below a level considered necessary for the maintenance of a decent standard of living; they have been used in studies on housing needs, and for similar purposes.

3. Information on the distribution by size of individual and family incomes is a basic source of information for many purposes of economic analysis. It has been used in studies on national income and consumers' expenditure. Research on the "consumption function", that is the relationship between total expenditure on various consumers' goods and income, is not possible without information on the frequency distribution of incomes. Similarly, the study of the relationship between saving and income (the "propensity to save") for the economy as a whole and for various population groups requires data on the distribution of incomes. The study of various problems of economic policy which countries faced in the postwar period of reconstruction and recovery required information on the size distribution of incomes. The statistics are also used in relation to the study of taxation problems. The number of applications may be expected to increase in the near future rather than to decrease. For several of the purposes indicated, information on the distribution of family incomes is more relevant than information on the distribution of individual incomes. This is due to the fact that for many items of consumers' expenditure the family - and not the individual - is the relevant spending unit. In general, however, information on the frequency distribution of individual incomes is more readily available, as it is obtained in relation to government administrative functions mainly in the field of income tax. Statistics of the distribution by size of family incomes must be obtained through special inquiries, and so far only a few countries have been able to carry out such investigations.

4. The following countries have published regularly, or for certain years only, statistics of the distribution by size of individual incomes:

/Australia

Australia
Canada
Ceylon
Czechoslovakia
Denmark
France
India
Indonesia
Ireland
Netherlands
New Zealand
Sweden
Union of South Africa
United Kingdom
United States

The material is usually compiled on the basis of income tax returns. So far only two countries have made attempt to compile income tax statistics through the application of sampling procedures.

5. Among the problems arising in the presentation of income tax statistics are the following. The fiscal income concept may differ from the income concept required for purposes of economic analysis. For example, the fiscal income concept may include capital gains and the stipulations for the application of depreciation allowances may be different from criteria prescribed by economic theory. Consequently, adjustments may have to be applied to the income distributions derived from tax returns. The choice of the recipient unit is another problem. Sometimes returns of husband and wife are joint whereas for certain purposes separate returns may be required. Usually income tax laws grant certain deduction for dependents, life insurance premiums paid etc. Adjustments may be required to present the income statistics before these deductions. Adjustments for under-reporting or no-reporting are in general very difficult to make. To obtain a complete picture of the income distribution, the income tax statistics have to be supplemented by information on the incomes below the tax exemption limit, which in general are not included in the fiscal statistics. Finally, the results of a tabulation of incomes by size may have to be reconciled with independent estimates of total personal income derived from national income studies. Other problems are of a purely technical nature. In those countries where the tax on wages and salaries is levied at the source, difficulties may be experienced in combining wage and salary statistics with income tax statistics. There is also a problem in the timing of the tax data. For the majority of incomes assessed in income tax,

/the data

the data usually refer to the previous calender year but this may not be true for certain categories of tax payers. Difficulties may also be experienced with regard to returns that are overdue, or which have been subject to revision by the tax authorities, legal procedures, etc.

6. Complete compilations of income tax statistics are costly and they also put a burden upon the fiscal administration. Because of the limitations of the data, as indicated above, complete enumerations may seldom be necessary. Applications of sampling methods may be useful because they save costs and will make it possible to concentrate on improving the quality of the information included in the sample. In one country estimates of the income distribution have recently been obtained through the application of probability-sampling to the tax returns of all people assessed in income tax.[1]

7. Information on individual incomes may also be obtained by sampling procedures in connexion with censuses of population. In Sweden material was gathered on taxable incomes in connexion with the partial population census of 1936 which covered a random sample of about one-fifth of the population. The sample survey of 1940 for this country resulted in a distribution of income recipients by size of personal income, together with other relevant information. In the United States the Census of 1950 was used to obtain data on individual and family incomes through probability-sampling of one-fifth of the total population. Detailed information on the methods of tabulation to be used has not been made available as yet. A discussion of this important example of the application of sampling techniques will therefore have to wait until a later date.

8. Statistics of income tax usually refer to the incomes of individuals with the possible exception of joint returns of the husband and wife. Therefore other methods must be used for obtaining information on the distribution of family income. Sampling procedures offer the appropriate device for obtaining information on these important data. It is not the purpose of this paper to describe in detail the various conceptual problems that have to be considered but

[1] Sample Surveys of Current Interest, document E/CN.3/Sub.1/23, page 14, item 3.

/some may

some may be mentioned briefly. In compiling statistics of family incomes, agreement must be reached on what constitutes a family and on the definition of income to be used. In the studies on Income of Non-farm Families and Individuals, issued by the United States Department of Commerce, Bureau of the Census, the term "family" refers to a group of two or more persons related by blood, marriage, or adoption and residing in the same household. (A "household" is defined as a group of persons living together in a dwelling unit, usually with common housekeeping arrangements, or a person living alone.) Lodgers and servants not related to the head of the household are considered as additional families, and not as part of the head's family. In the Netherlands, "household" and "family" are defined in a similar way, except that households formed by two or more unrelated friends are considered as families, whereas the United States Bureau of the Census considers them as two (or more) "individuals not in a family". Detailed information will also have to be obtained on the composition of the families with respect to age and sex, occupation, number of children of different age groups etc., so that the usefulness of the material for the purposes of economic analysis may be as large as possible.[1]

9. Agreement must be reached on the definition of income to be used. The income definition used in the sample surveys of the United States Bureau of the Census excludes from the net income from operation of farm or ranch, the value of food produced and consumed at home and inventory changes. The net income is overestimated in that depreciation charges are not deducted. The deviations are due not to conceptual discrepancies but to the difficulties of obtaining and of evaluating these items. In general, differences in defining or measuring income may be due to conceptual or statistical differences in treatment, for example of:

(a) Income in kind.

(b) Net rental values of owner-occupied houses.

(c) Imputation of interest to holders of savings accounts or bank deposits.

(d) Premiums for life insurance.

(e) Employees' contributions to social insurance and pension funds.

(f) Allowances of the armed forces and their dependents, including mustering-out and discharge pay, bonuses etc.

[1] At a later stage, the Statistical Commission may wish to formulate recommendations for the compilation and tabulation of statistics of families, classified by size, composition and other characteristics.

/(g) Gifts,

(g) Gifts, inheritances.

(h) Capital gains.

(i) Proceeds from sale of assets and dissaving.

For the purpose of measuring family income all wages, salaries, interest, dividents, rents, profits and entrepreneurial incomes received by family members will be included, and also transfer incomes, such as for example pensions, terminal leave payments, social security payments, annuities, unemployment assistance; but not capital transfers. Items (g), (h) and (i) will have to be excluded.

10. In the United States sampling techniques have been used for obtaining statistics on the distribution of family incomes by size. Two major projects may be mentioned here. The Current Population Reports on Consumer Incomes, published by the United States Bureau of the Census, provides information on the distribution of income of families and persons in the United States. These data are collected from approximately 25,000 households in 68 sample areas located in 42 states and the District of Columbia. Of the 25,000 schedules, approximately 7 per cent were "noninterview" for which duplicates of other schedules in the sample were substituted. Furthermore, approximately 8 per cent of the schedules in 1948 lacked income information for some person listed on the schedule. Substitutions were not made for these schedules. For a more detailed account of the basic concepts and classifications used in the sampling procedure, reference is made to publications of the Bureau of the Census.[1/]

11. A detailed sampling inquiry of an entirely different nature has been carried out on a continuous basis by the Survey Research Center, University of Michigan, in co-operation with the Federal Reserve Board in Washington, D. C. The purpose of this Survey of Consumer Finances is to obtain information on the current financial status, the recent spending and saving behaviour, the attitudes towards their own financial situation and prospects of the families in the inquiries as well as of those of the country as a whole. The size of the samples is small and it is limited to about 3,500 consumer units. Among the many items concerning the financial status of the reporting families and major items of expenditure, data on the family income figure as one of the basic economic indicators. For the description of the sampling procedure used, reference is made to a summary which is attached to this paper as Appendix I.

[1/] Cf. Current Population Reports, Series P-60, No. 6, 14 February 1950.

/12. The sample

12. The sample survey of families in Rangoon, carried out by the Government of Burma, may also be mentioned here, since it includes questions on the income of the people of Rangoon.[1]

13. The Sub-Commission may wish to review the applications of sampling methods made in obtaining statistics of distribution of individual and family incomes by size. In particular it may wish to recommend:

(a) That the attention of Governments be drawn to the usefulness of sampling procedures in compiling income tax statistics;

(b) The sampling procedure to be used for obtaining statistics on the distribution of family incomes by size.

The Sub-Commission may wish to suggest that the Statistical Commission, in drawing up recommendations for the collection of information on the distribution of individual and family incomes, draw the attention of Governments to the usefulness of sample procedures in this field and to the appropriate methods to be used.

[1] For a summary description of this Sample Survey, reference is made to Sample Surveys of Current Interest, document E/CN.3/Sub.1/23, page 2.

/APPENDIX I

APPENDIX I

Methods of the Survey of Consumer Finances[1]

How the sample is chosen. The sampling procedures of the Surveys of Consumer Finances are based on the principles of probability sampling. They are, however, more complex because of the nature of the sampling problems, as described below; there are also slight departures from the ideal as will be discussed in the paragraph relating to the listing of dwelling units.

There is no list of all the families or spending units in the United States from which a sample could be selected and designated. The establishment of such a list would not be a practical undertaking. Furtheremore, even if a list were available, the individuals selected from it would be so widely dispersed geographically that the cost of interviewing would be very high.

The Surveys of Consumer Finances are desinged on the basis of a work load of about 40 to 50 interviews within each primary area selected (usually a country), and two interviews to a sample block within towns. This procedure, known as "clustering" the sample, reduces the costs of travel and interviewer time in reaching designated respondents. The clustering is intended to achieve the most acceptable compromise between two factors which have opposite effects on the efficiency of sample design: the greater the spread of a sample of given size, the more precisely will it represent the diverse elements of the population; the smaller the spread of the sample, the less the cost per interview.

The sampling procedure used in the Surveys of Consumer Finances is known as multi-stage area sampling. The process of selection has several stages; at each stage the area to be sampled is divided into several parts with clearly designated boundaries, and some of the parts are then selected into the sample according to specified probabilities. First counties are selected; then cities, towns, or townships within the counties; then city blocks in cities and small geographical areas in other places; finally dwelling units within the blocks or areas. Thus by successive selections of areas, individual dwellings are selected and the spending units living in these dwellings are designated for the sample. Despite these complexities the essential qualities of probability sampling are maintained. That is to say, the equivalent of a list representing the population covered by the survey is used at each stage in the process of selecting the sample, and thereby

[1] Source: "Methods of the Survey of Consumer Finances", Federal Reserve Bulletin, July 1950.

/each member

each member drawn into the sample is randomly designated.

Techniques for increasing sample precision. Two major devices are used for increasing sample precision or the likelihood that the sample will have the same characteristics as the total population. One of these is stratification. The other is selection with "probabilities proportional to size".

Stratification. By this device, the population to be sampled is first sorted into several groups (strata) on the basis of relevant social and economic variables. Subsequently units within each of these strata are selected for the sample, thus ensuring that it will more accurately reflect the diversity of the population in regard to those variables. In so far as the variables used in stratification are related to the variables being measured by the survey, the precision of findings is increased. The 12 largest metropolitan areas, each of which contained a million or more inhabitants in 1940, are considered separately from the rest of the country for survey purposes. These 12 areas contain 48 counties and about 30 per cent of the nation's population. The largest is the New York area with about a tenth of the population in 15 counties and the smallest is the Cleveland area in Cuyahoga County, Ohio. Each of the central cities of the 12 metropolitan areas is included in the sample. A sample from a list of the cities, towns, and rural districts in the suburban areas surrounding these central cities is drawn. The sampling of blocks and dwelling units within these cities and towns is similar to that described below for cities and towns outside the metropolitan areas.

Outside the 12 metropolitan areas there are about 3,000 counties, each of which (or sometimes an adjacent group of two or three counties) is a potential primary sampling area. Originally, these counties were sorted into 54 groups (strata) on the basis of the following variables: percentage of 1940 population living in urban places; average per capita savings bond sales in 1943; degree of industrialization as indicated by the proportion of the 1940 working population employed in manufacturing industries; percentages of the 1940 population which were native white; and average size of farm according to the 1940 Census of Agriculture. One primary area per stratum was selected in a random manner from the list of areas for each stratum.

Work has been under way for some time to make it possible to change to a new set of 54 primary areas. In this new grouping, more emphasis is being given to such factors as population concentration and geographic location and, in some

/instances,

instances, primary sampling areas larger than single counties have been established. Moreover, a new technique for controlling the selection of primary areas has been devised, which on tests of several important items has yielded increased precision of results.[1] The shift from the old to the new set of primary areas is being made gradually and to date only 17 of the original 54 primary areas have been replaced by new ones. (The 54 selections and the 12 large metropolitan areas comprise the 66 primary sampling areas of the survey.)

Each of the 54 primary areas is divided into two parts: (1) cities, towns, villages, and unincorporated congested areas, and (2) open country. The areas included in (1) are divided into several subgroups (substrata) and from each subgroup one place is selected for the sample. The entire area of each place selected is divided into blocks (small areas with definite boundaries, usually streets), the blocks are listed and numbered consecutively, and a set of sample blocks scattered through the various parts of the town is selected from the list. A map of the town showing the sample blocks, and a separate "listing sheet" for each sample block, with a sketch of the boundaries, are given to the interviewer, who is instructed to enter on separate lines the complete address (with description where necessary) of every dwelling unit located within its boundaries. From these lists a sample of dwelling units is selected and the interviewer is directed to take interviews at the selected addresses.

The sparsely populated "open country" portions of the primary area are sampled in a slightly different but analogous manner: the entire area is divided into small segments bounded by roads, railroads, streams, township lines etc. These subdivisions are numbered consecutively, and random selection from this listing yields several segments scattered through the various parts of the county. The interviewer is given a county map showing these segments and told to take interviews at each dwelling located inside their boundaries.

[1] Briefly, this technique makes sure that the primary areas selected from the various strata will be better distributed with respect to geographical location and other variables than they would ordinarily be by stratification alone. It involves a co-ordination of the selection of primary areas within the various strata at the same time adhering rigorously to principles of probability sampling. For further details, see the forthcoming article "Controlled Selection - A Technique in Probability Sampling", by Roe Goodman and Leslie Kish in the September 1950 Journal of the American Statistical Association.

/In all the

In all the different stages of molding the sample, the selections are made in the Ann Arbor, Michigan, office in accordance with predetermined probabilities, with the use of tables of random members.

Selection with "probabilities proportioned to size". Another step in increasing sample precision is to give each primary area a probability of being chosen proportional to a measure of the number of people it contains. The sampling rates within primary areas are controlled so that each dwelling unit has the desired probability of being selected, regardless of where it is located. This technique, in addition to increasing the sampling precision, contributes to easier administration by making for a relatively stable number of interviews from each type of sampling unit (county, city, or block). Although the measure is usally based on the 1940 population, there is no fixed "quota" of interviews to be taken in any one area. In so far as some sample counties, towns, or blocks have increased in population since 1940, this increase will be reflected, within limits of sampling variability, in a larger sample from those places. For cities with over 50,000 population, the number of dwellings in each block shown in the 1940 Census Block Statistics is used, supplemented by an additional selection from blocks which had no dwellings in 1940.[1/] In smaller places aerial photographs are utilized to obtain a rough count of the dwellings in the blocks. The "Master Sample", from which the listing of the towns and rural congested areas in the sample counties is obtained, also provides the material for selection of segments in the open country areas.[2/]

Oversampling of high-income groups. Another important device used for improving the precision of some of the survey results is the procedure for increasing the number of interviews with people at higher economic levels. This group represents the far end of the highly skewed distributions of income, of amounts saved, and of assets. Because of the concentration of income and saving among a relatively small proportion of the population, information received from a relatively few respondents weighs heavily in the means, aggregates, and

[1/] When 1950 Census Block Statistics become available, these data rather than the 1940 data will be used.

[2/] The Master Sample comprises maps and other materials for the entire country which greatly facilitate the selection procedures involved in area sampling. Developed jointly by Iowa State College, the US Bureau of Agricultural Economics, and the US Bureau of the Census, the material can now be obtained from the Bureau of the Census.

/distributions of

distributions of aggregates collected in the survey. There is great variation in the amounts received, held, invested, and spent by the members of this group. By increasing the number of "wealthy" respondents in the sample, a more reliable representation of this important group is obtained. In the tabulation of results the interviews from the oversampled dwellings receive a proportionately smaller weight so that they appear in their proper proportions in the final results.

The procedure for oversampling must rely on indirect means, because direct identification of dwellings with hihg-income occupants is not usually possible. In the Surveys of Consumer Finances various indirect procedures are used. For cities with populations over 50,000, Census figures give the average rent paid per block; dwellings in high-rent blocks, and also dwellings in high-rent suburbs, are oversampled in the Surveys of Consumer Finances. Also, at the time of listing dwellings in these blocks the interviewers are instructed to indicate whether they think the dwellings are occupied by high-, medium-, or low-income families. Dwellings rated "medium" are sampled at twice the rate of "low" dwellings. Dwellings rated "high" at six times this rate. (In the three surveys prior to 1950, the "high" dwellings were sampled at four times the rate of the "low" and that experience indicated the advisability of greater oversampling). In other cities and towns the dwellings rated high as well as those rated medium are sampled at twice the rate of the lows; in these smaller places the highest rate of oversampling is not applied because it is believed that the additional cost of this procedure is not justified in view of the relatively small proportion of potential high-income respondents.

It should be noted, of course, that this device of oversampling on the basis of subjective ratings does not affect the representativeness of the original sampling procedures. The weight assigned to each interview takes into account the rate of sampling. If some dwellings rated high prove to contain low-income families, this merely increases the number of interviews from low-income families, without adding to their weighted proportion in the final sample, and fails to add interviews from high-income families. Hence, inaccuracies in the subjective ratings reduce the gains in over-all precision accruing from the oversampling procedure; but they do not bias the sample results.

<u>No substitutions in sample</u>. After the dwelling units have been selected each interviewer is given relatively simple instructions with respect to procedure. At each dwelling assigned to him, he is instructed to list the occupants, to identify

/the family units

the family units and the spending units, and to interview the head of every spending unit. Substitutions for non-responses are not allowed because they would not be true substitutes, and because their effect on the over-all procedure might be to render the sample results less accurate.

A high enough sampling rate is taken to obtain approximately the desired number after allowing for losses due to non-response.

Inaccuracies in listing. In the carrying out of field operations there are some departures from specifications. Occasionally some dwellings are overlooked at the time of the listing. A number of these omissions are later discovered and included during the interviewing period. The interviewer may also make a mistake in identifying the boundaries of a sample block or segment. Finally, there is the difficulty of including in the listings all the newly constructed dwellings as they become occupied. Some listings are from one to four months old, at the time interviewing begins. These listings, however, include dwellings under construction, and such dwellings are included in the addresses in the samples. For block listings which are older than that a procedure is used to bring into the sample newer dwellings in these blocks in their proper proportion. This is done by selecting a sample of these blocks for inspection by the interviewer, who locates any new and unlisted dwelling while he is interviewing in the block.
