OASI Sampling Methods

In 1937, the first year of operation of the old-age and survivors insurance program, nearly 33 million workers received wage credits in covered employments. In 1938, some 32 million workers received such credits. These figures on coverage and other data on the 1937 and 1938 earnings of workers and their age, sex, race, and State of employment were obtained from a 100-percent tabulation of the wage records under the program. With the 1939 data, however, the Bureau of Old-Age and Survivors Insurance began to use samples to obtain its statistics on workers. A description of the sampling method used is given in the following pages.[1]

For the tabulation of the 1939 data, a 20-percent sample was used. For subsequent years, as more confidence was placed in sampling as an economical means of deriving accurate statistics, the sample sizes for tabulating data on workers were progressively reduced. For 1940–42, samples ranging from 4 percent to 1 percent were used; for 1943 and 1944, use was made of 3-percent and 1-percent samples. Since 1944, reliance has been placed almost exclusively on a 1-percent sample and, for special purposes, on a sample of 0.02 percent.

Sampling developments in tabulating data on beneficiaries under the program are comparable in some respects with those in compiling statistics on workers. For 1940 and 1941—the first years that monthly benefit payments were paid—data were compiled on a 100-percent basis. Since 1941, most of the beneficiary tabulations have been made on a 20-percent sample basis and the rest on a 100-percent basis. Because of the relatively small size of the universe and the many detailed breakdowns, this sample has not been further reduced.

In dealing with a universe of account numbers approaching 100 million, the use of samples is essential both because it is economical and because it makes available a quick and flexible system by which to obtain accurate and up-to-date information on the characteristics of the covered population and the operations of the program. While there are many possible ways of sampling from the social security records, the method adopted was geared to the wage record keeping system and to related administrative operations, in order to yield the required data at the lowest possible cost. Since all individuals under the program are identified by a 9-digit account number, and all records are in numerical order, "digital sampling" was adopted as the most economical way to select samples of account-number holders and beneficiaries.

Description of Universe

Basic statistics are compiled and analyzed in the Bureau for four universes. The largest is the working population in covered employments. Exclusive of new coverage, this universe includes about 82.4 million individuals who have earned some wage credits under the program at any time during the years 1937–50. It is obviously far larger than the size of the total labor force as of any given week (an average of 58 million in 1950) or the total number of persons who were in some kind of civilian work at any time during the year (about 73 million in 1950).

The second universe is composed of persons who have received new account numbers; the tabulations give the number assigned in each calendar quarter and year and the age, sex, race, and State distributions. Nearly 900,000 new numbers were issued in the third calendar quarter of 1950, and about 800,000 in the last quarter.

The third universe consists of old-age and survivors insurance beneficiaries. As of June 30, 1950, benefits were being paid to nearly 3 million persons—1.4 million retired persons aged 65 and over, about 450,000 members of the families of these workers, and 1.1 million surviving dependents of nearly 700,000 deceased persons. The total number of accounts involved in this beneficiary universe is 2.1 million. The records for this universe are in the form of family benefit folders, which may represent a single person or two or more persons in a family. The folders are filed under the worker's account number in numerical sequence.

The fourth universe for which statistics are compiled is that of employers subject to the Federal Insurance Contributions Act who report each quarter the taxable wages of their employees. During 1950 there were 3.5 million such reporting employers. The statistical tabulations of employer reports have been largely on a 100-percent basis, although plans for more elaborate sampling are under consideration.

Issuance of Account Numbers

The social security account number assigned to each worker under the program serves to identify not only his wage record during his working life but also his retirement claim and benefit records. This number consists of nine digits in three segments. The first three digits designate the geographical area where the social security number was issued. The second segment of two digits designates the specific block of numbers issued in any one area. One hundred groups of numbers—from 00 to 99—can be issued for any area. The third segment (known as the serial) consists of four digits; 10,000 numbers can be issued for each block number in any one area. Each area may have 100 blocks, and 1 million numbers can therefore be issued for a single area. Since numbers are now being issued in 612 areas, including areas for railroad workers, 612 million numbers may be issued without the addition of more areas. The numbers in some areas will, of course, be depleted earlier than in others. In theory, it will be possible ultimately to issue 1 billion numbers because there will be 1,000 areas. Assuming that an average of about 2 million new persons will enter the covered labor market each year, the 9-digit account number system can last for several hundred years.

The field offices of the Bureau issue account numbers to individuals as they apply for them. Efforts are made by a screening process to avoid issuing more than one number to a person. Control over the numbers to be issued and the method of issuance is maintained by the Bureau's central office in Baltimore. At present, the account

[1] Summary of a paper delivered in December 1950 before the American Statistical Association by B. J. Mandel, Division of Program Analysis, Bureau of Old-Age and Survivors Insurance.

numbers are released to the field offices in multiples of 500; the area numbers in each shipment are those previously allotted to the State. The field offices must issue numbers consecutively, starting with the lowest number of the series assigned to it.

Sampling Methods

Twenty percent of the numbers in each shipment to the field offices have the digit 2 or 7 in the first place of the serial. In this way, 20 percent of all account numbers issued contain these digits, and they yield a controlled subuniverse of 20 percent. The account numbers in this subuniverse are generally issued in clusters of 100 (but never in clusters of more than 1,000) for the selection of smaller samples. Statistics on workers for 1939 were tabulated from this subuniverse of stratified clusters of 20 percent of the accounts.

In tabulating employee data for 1940 and subsequent periods, use was made of samples of 4 percent, 3 percent, 1 percent, and 0.02 percent of the universe of established accounts, all selected from the 20-percent subuniverse. The 4-percent sample was obtained by selecting all accounts in the subuniverse having either 0 or 5 in the last place of the serial number. This method included 2 account numbers out of every 10 and provided a stratified, systematic 20-percent sample from the subuniverse of 20 percent, or a 4-percent sample.

Only the first and last digits of the serial number were relied on to get 20-percent and 4-percent samples. For the 3-percent, 1-percent, and 0.02-percent samples, however, it was necessary to go to the next to the last place of the serial number, or the eighth digit of the account number. The 3-percent sample was obtained by splitting the 4-percent sample into two segments of 1 percent and 3 percent. The 1-percent segment was composed of accounts with 2 or 7 in the first place of the serial, and 05, 20, 45, 70, or 95 as the last two digits of the serial. Since the eighth and ninth place of the account number for persons in the 4-percent sample contained 20 possible numbers, selection of five of them provided a fourth of the 4-percent sample, or 1 percent. It should be noted that both high and low numbers were almost equally represented in order to make the sample as representative as possible of the universe. The 3-percent sample, of course, was the residual segment after the 1 percent was sorted out of the 4-percent sample.

To obtain the 0.02-percent sample, the first step was to select from the five groups in the 1-percent sample the group that contained the digits 05 in the last two places. An 0.2-percent sample was thus obtained. Selecting from this segment only the accounts with the digit 5 in the seventh place of the account number yielded one-tenth of the 0.2-percent segment, or a sample of 0.02 percent. This is a stratified, systematic sample, since it consists of every five-thousandth number in the account-number population and is selected proportionately from each area. A great variety of samples of the same or different sizes can be selected by digital selection based on the serial in the account number.

The same system was used to select the 20-percent sample for tabulating most of the data on beneficiaries under the program and the 20-percent sample of persons who received account numbers.

Representativeness of Samples

From the description of the methods of sampling, qualitative conclusions can be drawn about the over-all representativeness of the six account-number and beneficiary samples. The system of sampling assures stratification by geographical area, under which variations in the age, sex, race, earnings, and industrial characteristics of workers are automatically reflected in the over-all samples for the United States. In addition, the national samples may be expected to reflect fairly closely the characteristics of the universe because of the large absolute size of the samples and the fact that the total sample is a combination of many small samples, built from systematically selected numbers that originate from field offices throughout the country.

Quantitative facts collected in the past several years confirm the belief that these samples are highly representative. Unlike most sampling programs, that of the Bureau of Old-Age and Survivors Insurance through its accounting and tabulating system is provided with selected universe and subuniverse totals. Comparison can thus be made of certain common characteristics as determined from the universe data and independent samples drawn from that universe. In addition, common characteristics from independent samples may be compared with one another.

Universe data are available for over-all totals of the accounts established and the beneficiaries; it is therefore possible to check the actual size of the expected subuniverse of 20 percent and the size of the smaller samples taken from the subuniverse. The data show, for example, that the subuniverse of accounts included 19.99 percent of all accounts established as of January 1949. The sample of persons in receipt of benefits as of January 1, 1950, was 19.93 percent of the known universe of beneficiaries at that time.

Some internal comparisons have been made of the samples of accounts showing wage credits. For example, the 3-percent sample of active accounts in 1945 contained 75.02 percent of the active accounts in the 4-percent sample; the 1-percent sample for that year contained 24.98 percent of the accounts in the 4-percent sample; the 0.02-percent sample contained 2.02 percent of the 1-percent sample for 1945.

It has also been possible to compare the actual size of the sample for different groups of workers or beneficiaries with that theoretically expected. For example, the sample of men aged 65 and over who were awarded benefits in 1949 was 19.9 percent of the total; for women aged 65 and over, it was 19.8 percent; in the age group 65–69, it was 20.0 percent for men and 19.6 percent for women. All these differences from the theoretically expected percentages and most of the other differences that were studied were found to fall well within the range of the expected sampling variations.

Advantages of Digital Sampling

The most obvious advantages of the type of digital sampling used by the

Bureau of Old-Age and Survivors Insurance are simplicity and flexibility, since it lends itself readily to yield smaller or larger samples, as the need arises, by sorting on selected digits in the serial number. Second, the cost of selecting samples of different sizes is kept to a minimum because statistical sampling is linked to administrative operations. Third, by selecting smaller samples from the larger samples it is possible to control mechanical or other errors by comparing sample totals with previously tabulated subuniverse or larger sample totals—an important factor both from the standpoint of accuracy and economy. Fourth, this sample is most appropriate for the type of continuous work-history tabulations made in the Bureau, since it automatically yields representation of the changing universe by the addition each year of a sample of new workers from the group receiving new account numbers with the predetermined sample digits. Thus, all persons who in 1950 obtained new account numbers having the serial 2505 or 7505 will automatically be represented in the 0.02-percent sample. A fifth advantage of this type of digital sample is that, because of the automatic identification of the persons involved, it affords a simple method of coordinating informational items for given workers in the sample with those of other agencies operating on the account-number system. It is relatively easy to supplement the old-age and survivors insurance sample with information from the Railroad Retirement Board or the State unemployment insurance agencies, since they also use the 9-digit account number system. Finally, because of the systematic methods of selection and the automatic stratification by area, the sample yields results highly representative of the universe from which it is drawn.

Conclusions

The compilation of statistics under the sampling program is not devoid of problems despite the availability of selected universe and subuniverse data and a simple, reliable sampling scheme. Two problems in particular need further study. One is the need for data to measure the bias introduced in the employee statistics when it is assumed that a sample of accounts to which wage credits have been posted is representative of individual workers. Despite all efforts to avoid issuing more than one number to an individual, it is known that some persons have multiple account numbers and thus may have wages credited under more than one account. The inclusion of these multiple accounts causes some overstatement in the number of workers and some understatement in the amount of average wage credits. A special study is under way to measure the significance of the multiples.

The second problem is the need to measure the extent to which the variance in the statistics on employees, by industry, exceeds that for random samples. Admittedly, some bias was introduced into the employee sample in the early days of the program, when, to alleviate the heavy initial registration load, clusters of account numbers were given to employers for issuance to their employees. While this bias is probably insignificant for broad industry groupings of the data, it is not known how significant it is for more detailed breakdowns. This problem is also being studied.

The probable main developments in the future sampling program may be summarized as follows:

1. The digital sampling system for employees and beneficiaries has proven itself the most feasible. It may be assumed, therefore, that it will be extended to provide data on the characteristics of the new groups of employees covered under old-age and survivors insurance for the first time by the 1950 Amendments to the Social Security Act.

2. The procedure for maintaining a sample of sufficient size for tabulating detailed data and using smaller subsamples from the larger sample for tabulating selected data will be continued because of its flexibility and economy.

3. Sampling of business establishments, up to now restricted to small-scale studies, will become a necessity under the extended program, when about 7–8 million businesses will be required to report. Consequently, it will be necessary to develop a feasible sampling system for use in compiling statistics on employing organizations and their characteristics.

Economic Status of Aged Persons and of Dependent Survivors

Estimates for December 1950 have been made of the number of aged persons, widows under age 65, and paternal orphans under age 18, and of the number with income from employment, social insurance and related programs, and public assistance. Such estimates are prepared semiannually by the Social Security Administration to aid in program planning and for other purposes.

The most significant development in the economic status of these groups between June and December 1950 is the sharp increase in the number of old-age and survivors insurance beneficiaries. The Social Security Act

Table 1.—*Estimated number of persons aged 65 years and over, receiving income from specified source, December 1950*

[In millions]

Source of income	Number of persons		
	Total	Men	Women
Total population aged 65 and over [1]	12.3	5.7	6.6
Employment	3.7	2.3	1.4
Earners	2.8	2.3	.6
Wives of earners	.9	-----	.9
Social insurance and related programs:			
Old-age and survivors insurance	2.6	1.5	.11
Railroad retirement	.3	.2	.1
Federal civil-service retirement	.1	.1	(2)
Veterans' program	.3	.1	.1
Other [3]	.4	.1	.3
Old-age assistance	2.8	1.3	1.5

[1] Total population is preliminary estimate for April 1950 based on a sample of census returns and is subject to change. Includes persons with no income and with income from sources other than those specified. Some persons received income from more than one of the sources listed.

[2] Less than 50,000.

[3] Beneficiaries of Federal retirement programs other than civil service, and of State and local government retirement programs, and the wives of male beneficiaries of programs other than old-age and survivors insurance.

Sources: Total population and earners from Bureau of the Census. Number of persons in receipt of payments under social insurance and related programs and from old-age assistance, reported by administrative agencies, partly estimated. Number of wives of earners and number of wives of male beneficiaries of programs other than old-age and survivors insurance estimated from Census data on marital status.