SAMPLING THEORY AND PROCEDURES

GARRIE J. LOSEE

In July, 1965, at a meeting in Bogotá of members of the Ministry of Health and the Colombian Association of Medical Schools, the Minister of Health selected a single number from a table of random numbers, and thus determined the identity of the 40 sample cities and municipios of the Colombian National Health Survey. This was perhaps the most dramatic moment in a series of new, exciting and challenging events which summed together mark a long step forward in scientific investigation in Latin America.

It is doubtful whether any existing set of statistical information on the health and socioeconomic characteristics of the peoples of a Latin American nation can stand up as well to scientific criticism as can this study. That this is so is only partly attributable to the statistical design that will be described here. It is instead largely attributable to the faithful execution of the design.

Criticism of statistical data can be made on several grounds. Good statistical information should be relevant, reliable, accurate and specific. Statistical sampling has developed over the past 30 years to a level of high technical sophistication. The procedure is well established for selecting representative samples of the population to be studied by a sample survey, for forming unbiased estimates and for describing the probable reliability of the estimates derived from the survey. If the need is for information on which to base national health or economic planning, the sample should be a representative sample of all the people of the nation; that is, every person should have a chance of being selected in the sample. In that way every characteristic of the popula-

tion will be represented in the results, subject to the limitations of the sample size and, consequently, sampling error. Thus the appropriate proportions of males, females, children, adults, villagers, city dwellers, farmers, unemployed persons and so on are included in the sample.

Information derived from convenient samples such as patients in the hospital in which the researcher works, outpatients of a nearby clinic, the population of several "poor" barrios and a municipio close to a school of public health is frequently relevant in formulating national health policy, but it is not specific to the needs for information on which to base national policy. Information on the extent of internal parasitic infections in Colombia was available prior to the National Health Survey from several such convenient samples. Put together, these sources are at best a patch-quilt of information that could lead to serious consequences if the information were used to formulate national or regional policy of government.

Few purely technical problems were encountered in designing the sample of the Colombian National Health Survey. On the other hand many practical problems were encountered in implementing the survey design, but on the whole these were satisfactorily resolved. In the practice of good sample design no unique method exists for selecting a sample for a particular purpose, although given a certain set of conditions the choice of sampling methods is limited to a few that offer the best *a priori* possibility of providing sufficiently reliable estimates for the purposes of the study at the least possible cost. In the case of national household interview surveys this nearly always leads to the use of multistage cluster samples with stratification of sampling units at the first and often subsequent stages of selection. The use of these techniques in the Colombian National Health Survey is amply described in the accompanying text. Whenever applicable, experience accumulated in developing the present design of the United States National Health Survey was transferred and adapted to Colombia.

At this point it may be well to describe some of the problems faced in 1963, when planning the sample design, which are peculiar to the survey to be conducted in Colombia, their solutions and some of the features of the design that distinguish it from others. First, Colombia is a geographically large country containing vast stretches of sparsely populated territory. Second, at the time when the sample was selected a population census had not been conducted since 1951, and that one with questionable accuracy, although a national population census was conducted in 1964. The accuracy of the 1951 census was question-

able because someone remarked, after it was proposed to use its results in designing the current survey, that the 1951 census was not good —and he said that he knew this because he had directed the census operations. Third, because of the terrain, transportation, except by air, is extremely difficult between many sections of the country and in some areas communities are relatively isolated from their neighbors.

In the United States, information is obtained by the National Center for Health Statistics from two separate surveys, the Health Household Interview Survey and the Health Examination Survey. Due to differing cost and variance configurations, the designs of the two surveys are quite different. The Health Household Interview Survey with low unit costs has a large sample of households (about 40,000 annually) in a large sample of places (over 300) throughout the country. On the other hand, for the Health Examination Survey, a relatively small number of adults was examined (6,500) in a small sample of places (42) because unit costs were high.

The first basic decision made in the design of the Colombian National Health Survey was that the primary focus of attention would be on the collection of information from both interviews and examinations to provide a composite picture of an individual's health status, both as he knew it and as it was known through a clinical examination, for a sample of individuals representative of the Colombian population. Drawing largely on the United States experience this decision implied a small sample of places and persons. Since no additional travel costs for interviewers and supervisors was involved, it was possible to provide greater reliability for disability and medical resource utilization data at little extra cost by collecting interview data only for a larger sample of persons than would be both interviewed and examined in each sample place. The design therefore called for interviewing persons in a sample of about 240 households in each of 40 places, either a city or a municipio. A subsample of one out of every ten persons interviewed was selected to receive a clinical examination.

To overcome the limitations described earlier the following features became part of the survey design. Since the intendancies and comisarias accounted for nearly half of the land area of Colombia, but only 1.3 per cent of the population, these extraterritorial zones were excluded from the study. Their exclusion does not create any bias problem since, if included, even extreme variations from the characteristics of the bulk of the population would not have noticeable effect on national averages.

Costs involved in travel, both to the areas sampled and within these areas were minimized in the survey design by having a concentrated sample in a small number of places. The sample places were restricted in size to include only a single city or part of a city and in rural areas usually a single municipio.

Increased efficiency of design, that is, greater reliability per unit of investment, can usually be gained at the sample selection stage when the proportion of total population contained in each place is accurately known. Believing that the 1951 Population Census figures no longer represented the current proportion of population in many places, the population estimates of the Ministry of Health based on adjusted population projections were used in assigning proportions of total population to places in the selection. The resulting unbiased estimates of health characteristics were later adjusted according to a ratio estimation technique to conform with 1964 Population Census totals, not known at the selection stage (1964), but available in 1967 at the estimation stage.

Other technical features of the design worthy of note were: 1. the use of controlled selection in selecting sample places; 2. the provision of flexibility in sample size by the division of the total sample into two balanced random subsamples, each having 20 sample places; 3. the use of 1964 census maps and population counts for sample places in selecting clusters of about ten sample households; 4. a simple but efficient method for estimating sampling errors; and 5. a continued awareness of measurement errors as evidenced by a systematic reinterview program, standardized training of interviewers and examining staff and standardized control procedures throughout the data collection and data processing stages.