Chapter 11

EARL E. HOUSEMAN
Statistical Standards
Division, USDA

Some Comments on Sampling

THIS PAPER consists of a brief discussion of three separate but related topics: specification of a statistical population; the design of a sample for a specific study;[1] and a simple means of estimating sampling error.

SPECIFICATION OF THE POPULATION AND PARAMETERS

Rigorous use of modern statistical methods in sampling, estimation, and interpretation of results requires detailed specification of parameters (population-values) to be estimated. This means a complete definition of the population and of the data, and complete specification of procedures. Any thorough and careful interpretation of estimates from a sample, including interpretation of estimates of sampling error, must be with reference to a specific set of conditions, because if any of the conditions are changed the results may change. As this paper is limited primarily to sampling, the discussion of specifications will be primarily in reference to those necessary to design and select a sample. Unfortunately, the definition and specification of data, concepts, coverage, and various conditions are often not as fully developed and clarified as they should be.

Definition of the Statistical Population

To define a population one must define the units of observation and the geographical limits.

The Unit of Observation.

A statistical population, for our purposes, is made up of a finite number of units of observation, a unit of observation being, for example,

[1] "Adjustments in dairy farming in the lake states region." This is a cooperative study involving the Farm Economics Research Division of the Agricultural Research Service, USDA, and the states of Iowa, Illinois, Michigan, Minnesota, and Wisconsin. At the time of this writing there are no specific publication plans.

a farm enterprise for which a questionnaire is to be completed. The choice of definition of an observation unit is arbitrary and may involve a compromise between what is desired conceptually for purposes of the study and the practical problems or difficulties in obtaining accurate data for a unit defined in different ways. A definition of an enterprise usually requires a specification of minimum size as well as composition.

Geographical Limits of the Population.

Many farm surveys are limited to the "open country," as it was defined for purposes of the master sample of agriculture. Open country is the area remaining after delineation and deletion of incorporated places and unincorporated settlements having a population of more than about 100 persons and a density of more than about 100 persons per square mile. Whether to limit coverage to the open country is a matter of cost and practical considerations. Selecting an area sample from the open country of a few counties is a quick and inexpensive task. Farms in the nonopen country parts can also be sampled, but the so-called master sample materials are not as well suited for that purpose. That, combined with the difficulty of finding and identifying farms in nonopen country areas, is why the coverage for most farm studies of a research nature is limited to the open country. Perhaps the nonopen country territory should be covered as well — at least places having less than about 2,500 inhabitants.

Decisions must also be made on the broader limits of coverage. Should the area covered be a region, a state, a local area, etc.? That question is obviously related to objectives, costs, and making inferences, which are beyond the scope of this paper. However, when a study is limited to a particular type of operation, such as sugar beet production, census statistics for counties and minor civil divisions can be used to help define the limits of the statistical population to be sampled. For example, for a study of a local sugar beet producing area the statistical population might be defined as a group of minor civil divisions that account for about 90 percent of the production in that area. Should the area be defined to include 95 percent of the population or is 70 percent good enough? Elimination of the peripheral areas where the beet farms are of low density might reduce costs appreciably but how much will the purposes of study be impaired?

With respect to the matter of uncertainty in the making of decisions from survey results, it is clear that definitional or specification errors, as well as sampling errors, response errors, tabulation errors, etc., are a part of the total uncertainty or error picture. A definitional error is the result of defining, for example, the population, a class of the population, or a variable in a way that differs from the corresponding situations about which decisions are made. One would like to have the definitions and data specifications made to serve ideally the ultimate uses, but practical compromises must be made which means the exercise of

judgment on what definitional errors to tolerate. This problem of definitional errors is receiving, and should receive, increased attention by statisticians and subject matter specialists since it is an important problem area in the improvement of research technique.

Tabulation Plans

In addition to the definition of a statistical population there should be a clear understanding about analysis or tabulation plans before a sampler makes final recommendations on sample size and design.

SAMPLE DESIGN

Generally speaking, suitable lists are not available for sampling purposes. Hence, if the principles of probability sampling are to be applied, area sampling is indicated. Much literature is available on sampling, so rather than prepare a general paper it seems more appropriate to use a specific survey as a basis for discussion. Some interest has been expressed in a description of the sample for the study "Adjustments in Dairy Farming in Lake States Dairy Region," so that study will be used.

The statistical population for this study was all commercial farms in economic classes I through V, except specialized poultry, fruit, and truck farms. Each state was divided into regions as indicated in Table 11.1, and each region was treated separately for analysis purposes. Table 11.1 gives some general descriptive information about the population and the sample. Available notes reveal very little about how the sampling rates were determined. However, the matter of setting sampling rates will be briefly discussed later.

A geographically stratified random sample of area segments would

Table 11.1. Some Statistics About the Population and the Sample for the Lake States Dairy Adjustments Study

State	Region	No. of counties	Total No. of segments	No. of farms Class I thru V, 1954	Av. No. of farms per segment	No. of segments selected	Sampling rate	Sampling rate times No. of farms
Minnesota	1	17	11,148	38,384	3.4	99	1/113	340
	2	18	10,224	28,964	2.8	96	1/107	271
Wisconsin	1	21	13,182	44,538	3.3	126	1/105	424
	2	25	17,270	52,825	3.1	132	1/131	403
Michigan	1	5	2,441	6,927	2.4	72	1/34	204
	2	5	5,236	13,013	2.5	64	1/82	159
	3	10	7,318	16,909	2.3	76	1/108	157
	4	13	6,844	14,766	2.2	80	1/95	155
	5	9	7,818	16,640	2.1	84	1/103	162
Iowa	1	11	5,763	20,311	3.5	54	1/107	190
	2	6	3,575	11,944	3.3	54	1/66	181
Illinois	1	8	3,186	10,761	3.4	54	1/59	182

have given a sample of segments well distributed over a region, but because of the small size of the sample the average distance between sample segments would have been large. It did not appear advisable to increase the size of the segments, so a means of introducing some clustering of sample segments was sought. A two stage sample design was indicated. A county was not a suitable primary sampling unit because of the small number of counties in a region. Therefore, except for two regions in Michigan, minor civil divisions were used as primary sampling units. In the two Michigan regions, single stage sampling was used because they were small.

Region 2 in Iowa has been chosen to illustrate how the sample was selected. A sampling rate of 1/66 (see Table 11.1) meant that 54 segments were to be selected. As the sampling plan called for three sample segments in each township (minor civil division), 18 sample townships were needed. The six counties in this region listed in a geographical order, the total number of segments in each county, and the random numbers for designating selected townships are shown in Table 11.2.

Table 11.2. Counties in Iowa Region 2 Surveyed in Lake States Dairy Adjustment Study

County	Total number of segments	Random numbers designating sample townships
Winneshiek	718	45, 243, 441, 639
Allamakee	501	119, 317
Clayton	703	14, 212, 410, 608
Dubuque	547	103, 301, 499
Jones	563	150, 348, 546
Jackson	543	181, 379
Total	3,575	

As three segments were to be selected from each sample township, the sample townships were chosen with probabilities proportional to their numbers of segments. There are various ways to do this but considering the form in which the materials were available, it was actually done as follows: The 3,575 segments may be visualized as a continuous array, ordered geographically within townships and with the townships being in a geographical order within counties. A selection of every 198th segment in the array from a random starting point would give 18 segments and hence 18 corresponding townships. The townships so selected would have probabilities of selection proportional to their numbers of segments. That was the method followed. The starting point, a number selected at random between 1 and 198, was 45. Consequently, 45 is the first random number shown in the table above. The other numbers were obtained by adding 198 successively, but when 198 was added to 639 the result, 837, exceeded the number of segments in the first county. Hence, 718 was subtracted from 837 which gives 119, the first number in the second county, etc.

Table 11.3. Partial List of Count Units in Clayton County, Iowa

Township	Count unit	No. of segments	Cumulative no. of segments
1	1	2	2
	2	3	5
	3	3	8
	4	2	10
	5	2	12
	6	3	15
	7	4	19
	8	3	22
	9	1	23
	10	2	25
	11	2	27
	12	4	31
	13	3	34
	14	2	36
2	1	2	38
	2	3	41
	3	2	43
	4	2	45
	---------- etc. ----------		

There is for each county, as part of the so-called master sample materials, a listing of "count units."[2] Part of the listing for Clayton County is reproduced in Table 11.3. Note that the first random number for Clayton County was 14 which falls in the 6th count unit in the first township. This count unit was divided into three segments and one was selected at random for the sample. As the sampling plan called for three sample segments in each selected township, two additional segments were selected at random within the first township. That, in essence, is the way the sample was designed and selected.

The dairy adjustments study presented a sampling problem that is common to many such studies. How does one manage the situation when information on the size of the statistical population is insufficient to provide a satisfactory basis for setting sampling rates? Selecting a sample and finding half or twice the desired number of sample farms may present a number of difficulties. For a local survey, it is possible to design a sample so the field work may be terminated after approximately the desired number of questionnaires have been completed. This may be done in various ways without loss of the basic principles of probability sampling. Perhaps the simplest procedure is to select an unrestricted random sample of segments and number the segments in the order selected. The segments would be enumerated in the order numbered until the desired number of schedules is obtained. Of course,

[2] A count unit is a group of one or more segments. For a description of a count unit and the master sample materials, see Houseman, Earl E., and Reed, T. J., "Application of probability area sampling to farm surveys," Agr. Handbook. 67, 1954.

administratively, if one knew that he had to cover at least 15 segments he would enumerate the first 15 segments in whatever order was the most efficient and then proceed to 16, 17, etc.

Such a sample would lack stratification, but that can be provided for. Suppose a total of 80 segments was to be selected. One could set up, for example, eight strata and impose a restriction, without introducing bias, such that the sample segments numbered 1 through 8 would constitute a stratified random sample of eight segments, one from each of the eight strata. The same would be true for sample segments numbered 9 through 16, etc. The main point being made is that, particularly for a local survey, there are ways and means of keeping a sample statistically efficient and sound but at the same time have a plan that can be successfully administered and a plan that provides for termination of field work when a given number of schedules have been completed. Provision for making call-backs can and should be included in the plans.

Another kind of problem occurs when, for example, three types of farms A, B, and C are to be compared but their proportions in the population are:

A	10 percent
B	30 percent
C	60 percent

It is possible, with complication, to design a sample so that approximately the same number of farms of the three types would be enumerated. Suppose a sample of 50 farms of each type is desired and that the average size of segment is four farms, considering the three types A, B, and C. On that basis the required number of segments for each type, ignoring call-backs, refusals, etc., would be:

A	125 segments
B	42 segments
C	21 segments

Three samples, X, Y, and Z, could be set up:

X	21 segments	enumerate all three types
Y	21 segments	enumerate only types A and B
Z	83 segments	enumerate only type A

Because of the problem and cost of getting 50 farms of type A one might decide to reduce the size of the sample of type A. On the other hand, because of the low frequency of type A farms, one might decide to make the segments in the "Z" sample 3 times the average size. That is, the "Z" sample could be 26 segments averaging 12 farms.

An alternative to the above approach would be:

1. Use a sample of 125 segments.
2. Canvass all segments and contact each farm, ascertain its type, and list on a special form designed to provide a separate listing of each type.
3. Certain lines on the special form would be checked and farms falling on those lines would be included in the sample. All lines for the listing of type A would be checked. One-third of the lines for type B farms and one-sixth of the lines for the type C farms would be checked. Thus, the interviewing and listing could proceed simultaneously.

Actually, for this alternative approach one would probably use larger segments, perhaps 63 segments averaging eight farms instead of 125 segments averaging four farms.

The principles of probability-area sampling can be readily adapted to a wide range of conditions, but to do so successfully, an experienced sampler must work closely with the subject matter specialists. Otherwise, misunderstandings may develop. The sample may not be best for the objectives, the sample may not be properly used in the field, the data may not be properly weighted if weighting is required, etc.

ESTIMATION OF SAMPLING ERROR

Insufficient attention has been given to obtaining estimates of sampling error for interpreting results and planning studies in the future. Therefore, reference is made to a simple means of estimating the sampling error for many selected items, even though the sample design and estimation procedure may be rather involved. In essence, the sample is designed as a composite of several equivalent samples, perhaps eight or ten, set up in such a way that separate estimates may be made from each. The variability among these estimates provides a valid estimate of the sampling error.

As a simple case, consider a sample of 96 segments. Twelve equal sized strata could be formed and eight segments selected at random from each. Such a sample would be equivalent to a composite of eight samples, each being a stratified random sample of 12 segments, one from each stratum. The eight equivalent samples could be separately identified in the sampling operation or they could be established after data collection using appropriate randomization procedures. Separate estimates (probably only for selected items) would be made for each sample. Suppose, for example, that the average number of dairy cows per farm was computed for each of the eight samples, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_8$. The variance among these eight averages is $V = \frac{\Sigma(\bar{x}_j - \bar{x})^2}{n-1}$ where n is the number of samples (eight for the case in point) and $\bar{x}$ is the average of the eight means. The estimated variance of the mean of the entire sample is simply $\frac{V}{n}$.

Actually, for the situation just cited, another way of estimating the sampling error might be recommended, especially if an electronic computer is being used. However, the above approach may become a practical necessity when the structure of an estimate and/or the sample design becomes complicated. The point to be noted is that appropriate steps can be taken in the design of a sample so valid estimates of sampling error can be obtained rather expediently. This is important if it means the possibility of getting estimates of sampling error for many or several items. Otherwise, no sampling errors would be available. Moreover, the arithmetic procedure for estimating the sampling error is simple and can be administered by a nonstatistician. A statistician should be consulted, however, regarding the establishment of the "equivalent" samples to make sure that the differences among them will properly reflect various components of sampling error.

ROBERT D. BELL
South Dakota State College

Discussion

I SUSPECT that, as participants in this workshop, most of us would have been disappointed had not the topic of sampling been included as part of our study. A consideration of sampling problems is a key issue faced by the supply analyst. It is appropriate to have a paper on sampling and to focus attention on the statistical means by which observations should be generated to estimate the supply reactions made by producers to changes in product prices.

Houseman has outlined some of the essential components of sampling, organizing his paper around the related topics (1) specification, (2) the design of a sample, and (3) the sampling error. To "specification" may be attributed the translation of the supply response problem into statistical terms, which we do by defining the appropriate population of farms and by defining the parameters of supply reactions to be studied. The sample survey design treated is the Lake States Dairy Adjustment Study, referred to in several other papers presented at this workshop.

Houseman carried out his indicated objectives and gave a fairly good, though brief, discussion of the above three topics. However, his presentation deals largely with sampling of farms in general and is not a well-pointed paper on sampling for supply functions. I find this so-called "area of omission" a basis for criticism.

The logical question at this point is: "What do we expect to find in a pointed paper focusing attention on the statistical means by which observations should be generated to estimate the supply reactions of producers to changes in product prices?" In a pointed discussion, we would expect a "tie up" to be made between the process of sampling and the tools of analysis to be applied to the observations once they have

been generated and collected. Data needs, including the degree of accuracy, may be expected to vary with the tools used, so that in the sampling scheme it is more than questions concerning the use of a priori knowledge about the population and the cost of obtaining data. I suggest this is especially true in supply analysis where the tools are: (1) sometimes normative in nature, (2) sometimes positive, but (3) often of a combination normative-positive type. A distinction has been made elsewhere of the categories of tools for supply analysis; hence, we do not need to repeat it here. Regression analysis, especially if cross-sectional or some combination of cross-sectional and time-series data, is probably both a positive and a normative tool of analysis.

A pointed paper would probably relate itself in someway to this categorization of tools. It might, however, be in the form of some reference to:

1. The sampling scheme when supply response is to be derived from inter-firm production functions.
2. The sampling scheme when supply response is to be estimated through variable price programming.
3. The sampling scheme when supply response is to specifically reflect planned and/or realized reactions of producers rather than normative reactions.

Would not the sampling scheme be different for some two or all three of these situations? We would expect to find a partial answer in a pointed paper.

Supply response based on inter-firm production functions needs a sampling scheme chosen with the requirements of production function analyses in mind. Actual random samples are not the most efficient designs for this purpose. A different kind of sample would be more efficient if designed to select firms for adequate coverage of the variation in inputs within the sample and reduction in multicolinearity in observations of inputs of the sample. It is not an easy task to select farms into a sample for minimization of correlations between inputs. For one thing, economic conclusions derived from production functions, including those related to supply response, can be valid only if the firms in the sample are operating such that marginal productivity decreases continuously and is always less than average productivity. For another, the selection and classification of groups of farms of varying managerial capacity to insure that they are on approximately the same production functions present serious problems. It is possible in relevant situations for the investigator to identify a range in managerial capacity, thereby selecting for study groups of firms of either a rather uniform level or situations of randomly distributed divergencies in this inter-firm capacity.

It can be easily seen that the appropriate sampling scheme when supply response is derived from inter-firm production functions involves judgment at all stages of its empirical application. Such judgment could pay handsome dividends in increasing the reliability of

estimation. In fact, increases in efficiency which can be derived by such sampling over random or other common types of samples should go up substantially as the extent of this a priori knowledge of the universe of farms increases. This need for prior knowledge explains the real reason why supply response derived from inter-firm production function does not have as great a future as some related techniques. Prior knowledge in advance of selecting farms to obtain the necessary data includes: (1) the expansion path along which production is typically expanded by a population of farms operating more or less on the same production function, (2) the divergencies of individual farms from the norm that belongs in the population, and (3) the information to restrict considerations to firms operating under a declining positive marginal productivity curve as the law of diminishing returns requires.

Few if any researchers have such a priori knowledge available, although this information can be more or less approximated in a carefully designed research study. Even if such were available to the researcher, there is likely to be real difficulty in observing a range of proportions in which resources are combined sufficiently for representing particular portions of the production surface and its derived supply curve. In other words, we know how the sample should be drawn with respect to intercorrelations among the input variables. But conditions of farming and methodological issues are such as to reduce the plausibility of finding the kinds of observations we need in an otherwise homogeneous population in respect to resources available and techniques of production.

Supply response based on activity analysis needs a sampling scheme chosen with the requirements of continuous programming in mind. Probability sampling is appropriate for this purpose. This being the case, the sampling scheme should handle by stratification characteristics that affect the slope and elasticity of the firm supply relationship. We usually consider the following characteristics important, in that they often affect not only the slope and elasticity but also the reversibility of supply curves: Size of farm, tenure arrangement, amount of available capital, risk aversion, age of operator, cropland-pasture ratio, managerial ability of operator, and productivity of relevant resources. The most relevant of these should be the ones held constant in the analysis, at least within strata. The appropriate sampling scheme is expected to identify typical farm units possessing the most relevant of characteristics in varying combinations and also allow estimation of appropriate weights on the basis of the combinations' occurrence in the population.

Much of the same sampling procedure is involved when supply response is to reflect planned and/or realized reactions of producers.

In describing the sampling scheme for the Lakes States Dairy Adjustment Study, Houseman is dealing with stratified sampling. His paper deals mainly with a geographically stratified random sample with some clustering of sample segments introduced. This is certainly a practical approach to probability sampling of farms. Perhaps it accomplishes, to some extent, some of the objectives mentioned above.