Designs of Samples for Surveys

By Earl E. Houseman

In this discussion the author deals with the most frequent questions that come to him, as statistical consultant in the BAE, relative to designing samples for social science surveys. Although he draws heavily on the recent developments in sampling theory, he presents this material in a non-mathematical language.

WHEN designing a survey, the technical sampler usually tries to achieve the minimum sampling error consistent with a given cost or he tries to obtain a specified sampling error at a minimum cost. This is the principle on which much of the research work on sampling methods has been based. It restricts the choice to samples having a measurable sampling error; this, in practice, generally means choosing some form of random sampling. To apply the principle, costs and sampling errors of alternative methods are needed. But the principle of minimum sampling error per dollar is not entirely satisfactory, for it deals with only one of two major components of error in survey results and furnishes no answer to the question of how much different degrees of accuracy are worth.

Kinds of Errors in Survey Data

For purposes of discussion, it is convenient to classify all errors that can occur in a survey into two major categories: Sampling errors and non-sampling errors. Sampling errors, as visualized here, arise from the fact that a sample is involved instead of a complete census. They are the errors associated with the process of selecting a sample or with determining which sampling units are in the sample. Non-sampling errors arise from failure to get accurate information about each sampling unit in the sample. Reasons may include imperfect memory on the part of those being interviewed, omission by interviewer of some of the specifications of the data which should be included in the questions asked, and deliberate misstatements by those interviewed. Non-sampling errors occur during a census as well as during a survey made by means of a sample. Some writers have mentioned a third source of error—physical fluctuations as found in the case of forecasts. This kind of error is not a genuine survey error, for it may occur even though information on factors included in the survey is obtained without error when the survey is made.

It is common knowledge that non-sampling errors, or response errors, are often large enough to cause serious trouble. Estimation of number of farms is a good example. A small area sample for use in estimating the number of farms in the United States with a sampling error of less than 1 percent can be easily designed, but means have not been found for holding the non-sampling error to a negligible quantity because of the problem of defining a farm and following the definition in the field. Differences in farm counts as large as 10 or 15 percent have been observed between surveys when the sampling standard errors were known to be as low as 2 or 3 percent.

The magnitude of non-sampling errors might increase with an increase in size of sample because of the administrative difficulty of maintaining high standards in the interviewing and keeping adequate control over all phases of the work. In general, the smaller the sample the better one can minimize or control the non-sampling error. Hence, more accurate results might be obtained through sampling than through attempting a complete canvass, particularly when large and complex populations are dealt with.

For illustration, suppose a given sum of money is available for a Statewide survey among farmers. By expending very little effort on the control of response errors, a sample of 4,000 farms could be covered. An alternative might be to spend considerable effort on minimizing response errors, and so cover only 3,000 farms as the sample. Which alternative should be followed? The decision should be based upon knowledge of interviewing and operating conditions, and the sampling error, supplemented by best judgment.

The first might be chosen on the basis of getting more schedules per dollar, but fortunately the notion or assumption that size of sample is the only important factor affecting accuracy is gradually disappearing. Assuming that the samples are equally well designed, the writer would ordinarily recommend the second alternative. The writer usually follows the principle, when resources permit, of designing a sample just large enough so the sampling error is at a satisfactory level; then he emphasizes making every effort to minimize non-sampling error.

More work is needed in the general field of controlling non-sampling errors. The Bureau of the Census has been testing the possibilities of self-enumeration as a means of reducing costs and testing the possibility of getting more accurate information that way, because the respondent is free to fill out the questionnaire when he has time and can consult any records he may have.

Total Error the Crucial Point

Most research in survey methodology during the last 10 years has emphasized methods of reducing sampling errors. Much is known about the statistical efficiencies of alternative sampling schemes. But lack of information on survey costs is a major factor limiting extensive use of the principle of minimum sampling error for a fixed cost. Some progress has been made on the development of methods and principles of controlling non-sampling errors, but nearly all methodological studies have aimed at minimizing sampling errors per dollar rather than minimizing total errors. As work on the control of non-sampling error progresses, consideration should be given to the problem of minimizing total errors—that is, sampling errors plus non-sampling errors—per dollar spent. It might be impossible to develop the principle of minimum total error per dollar in terms of mathematical equations as in the case of minimum sampling error per dollar; nevertheless, it is a sounder practical basis for making operating decisions. Examples could be cited where the optimum sample design is clear under the principle of minimum sampling error per dollar, whereas the design which will lead to a minimum total error is uncertain and probably different.

Before discussing sampling design, an important development in schedule construction is worth noting. Schedules have usually been set up with spaces to be filled in under abbreviated headings, statements, or questions. These cannot be read off by an interviewer, so he must frame his own questions. Some of the headings may not contain all of the specifications required. For example, a heading might be "acres in cropland." The definition of acres in cropland is perhaps put in a large manual, the contents of which the interviewer is expected to know thoroughly, so when he asks a farmer for the number of acres in cropland on the farm, he can add all of the specifications as to what is meant by cropland. This type of schedule is sometimes called a reporting form.

A more recent alternative is to design a questionnaire insofar as feasible, so each question is written out, and can be asked as it stands. This type of schedule is called a questionnaire or interview form in contrast to the reporting form. Construction of an interview form requires more work, and careful pre-testing of the form is necessary to success. The interviewers are instructed to ask the questions as they appear on the schedule, to the extent possible, so uniform questioning will be assured and so all of the desired specifications will be included. It is not difficult for interviewers to ask the questions as stated if the questionnaire is properly made up, because that is easier than framing their own questions.

No extensive comparisons between the two approaches have been made. But when supervisors have accompanied interviewers who are using the reporting forms, they have found much variation among interviewers in the way questions are asked and have found frequent omissions of specifications, such as the time period, to which a question relates.

It might be argued that an interview form is unnecessary when the interviewing is done by a few persons who are thoroughly familiar with the subject matter and the specifications of the data. Some critics seem to feel that almost any form which provides space for answers is adequate. But many professional persons (not professional interviewers) are more prone than non-professionals to "bias" questions or answers unconsciously. With their knowledge of the situation, they may be more inclined to talk than to listen. That is, they may be inclined to guess the answers, and frame the questions accordingly. Or they may suggest answers instead of waiting for answers. There is no documentary evidence on this point but it may be important to use the interview form and the principles of interviewing related to it when professionals are to do the interviewing as well as when it is to be done by laymen.

Probability Sampling v. Judgment Sampling

No matter how a sample is designed, judgment usually enters in one way or another. In some cases a selection of individuals to form a sample is completely a matter of judgment; in other cases the selection is entirely by chance, and judgment is confined to the choice of restrictions to be imposed in the design of the sample.

Probability Samples

Samples selected in such a way that each element in the population has a known probability, greater than zero in the mathematical sense, of being in the sample, are known as probability samples. This is in contrast to judgment or non-probability samples wherein the sampling units have unknown probabilities of selection.

A main effort in modern sampling has been to make the best choice of restrictions to be employed in a sampling scheme, without losing the feature of known probabilities of selection or randomness. Probability samples have certain advantages, but it cannot be said that they are always preferable to judgment samples. Much depends upon the degree of precision required and the job to be done.

The advantages of probability samples result from the ability to use probability theory in the planning and evaluation of sample designs. This enables the estimation of sampling errors, the optimum allocation of the sampling units over the population, the comparison of the relative efficiencies of alternative sampling designs, and the determination of sample size needed to attain a predetermined magnitude of sampling error.

The performance of non-random sampling designs, on the other hand, can be measured only by means of empirical tests and comparisons with check data.

In practice, complete coverage of all elements selected in a probability sample is usually not attained. A sample may be a probability sample on paper, but it may be impracticable or impossible to get complete data on every unit of observation which should be in this sample survey. To try to get the last few interviews to attain 100 percent response is probably too costly to be worth while, but if a high degree of completeness is not attained the advantages of a probability sample may be lost.

Random samples are likely to be criticized on the grounds that time is lost in areas which seem to yield little or nothing in terms of the survey objectives. This is particularly true for surveys involving sporadically distributed units of observation; the surveyor may think it better to go down the road choosing the particular kind of farms or households wanted in the study.

Judgment Samples

The role of judgment in the design of samples is now receiving considerable attention. We should distinguish clearly between two different uses of judgment in sampling: (1) Use of judgment in determining which elements are to be in the sample and (2) use of judgment in setting up the "restrictions" to be placed upon the sample. (These restrictions do not apply to the selection of the individual elements which are finally chosen for the sample. The use of judgment in the selection of sampling units can range from no judgment through to pure judgment. It is this use of judgment that is thought of when speaking of probability versus judgment samples, or of random versus hand-picked samples.)

Although judgment samples have been frequently used, the literature on sampling methods contains little on numerical evaluation. Evidence can be given, however, on the general performance of judgment samples.

One of the best experiments designed to make comparisons between random and judgment samples was conducted by Cochran and Watson.[1] The measurements were on shoot heights of wheat plants, for use in crop-weather studies. They were taken by 12 observers who made judgment selections. Essential points in the findings were: (1) As expected, the biases differed significantly from one observer to another, (2) the average bias for all 12 observers was positive by about 8 per-

[1] COCHRAN, W. G., and WATSON, D. J. AN EXPERIMENT ON OBSERVER'S BIAS IN THE SELECTION OF SHOOT HEIGHTS. Empire Jour. Expt. Agr. 4:69–76 1936.

cent, (3) the average bias varied from plot to plot, and (4) although the observers had been instructed to make the samples representative of all heights of the wheat plants, the variance within the samples was less than the population variance.

In general, the writer's experience conforms with the following two paragraphs written by Yates [2] in 1935:

"The ideal which is aimed at in sampling is to make the sample as representative as possible, so that measurements or observations on it can be taken as virtually equivalent to similar measurements on the whole population. The fact that this ideal is in the mind of the sampler when taking the sample naturally influences his selection if he has any freedom of choice. Most samplers when selecting a representative sample will deliberately reject abnormal material, or if they feel that the sample should be representative of the abnormal as well as the normal will deliberately balance up the different categories of abnormality.

"Unfortunately the sampler's claims to be able to select 'a representative sample' by personal judgment are largely unfounded, and his selection is in fact subject to all sorts of biases, psychological and physical. To avoid these biases and to provide an estimate of the representativeness of the sample, i. e. of the 'sampling error', more rigorous processes of selection have been devised."

It is reasonable to expect that judgment samples will usually be selective in the direction of what the mind regards as the more important elements or in the direction of the less obscure elements. It is also reasonable to expect in general that the variability would be less within judgment samples than within the whole population; if so, this would mean that distributions based on judgment samples might be seriously distorted. In addition to the Cochran and Watson experiment, evidence supporting the two expectations is found in criticisms of well-designed, self-weighted samples. Many comments on stratified random samples are to the effect that in a certain study the sample contains too many households in low-income areas of a city, or the sample contains too many unimportant agricultural counties. Actually, the sample of the given city may have been stratified by income areas, assuring proportionate representation from each.

Intuitively, for some objectives, it makes sense to sample the more important elements at a higher rate. A methodological investigation might indicate that the efficiency of a sample is increased by varying the sampling rate according to some measure of importance of the sampling units. With the aid of sampling theory the question of how much to vary the sampling rate can be answered. Moreover, from knowledge of the different sampling rates, "unbiased" estimates can be made. On the other hand, judgment samples have unknown degrees of selectivity.

Because it is commonly recognized that judgment samples are likely to be distorted, the data are often not treated by computing simple averages. Certain of the sample data might be checked, for example with Census data, then classified and weighted with the hope of removing any appreciable biases. Such treatment of the data should, in general but not always, improve the results, but it may also give a false sense of reliability.

Judgment samples naturally look good to the person who made the selection, but since a sample cannot be accurately appraised by its looks, judgment samplers are likely to have erroneous impressions of the accuracy of their samples.

Comparative Accuracy of Probability and Judgment Samples

Although it is not possible to calculate sampling errors by formula for judgment samples, a simple experiment can be designed to show for various sample sizes how the error in judgment samples compares with that in random samples. For example, consider a population of 100 stones having various weights or a population comprised of 100 counties. By selecting repeated samples for each of several given sample sizes, a relation can be established between the average error (standard error of the sample averages from the population average) and sample size in the case of both random samples and judgment samples. The relationship for the random samples can be plotted, of course, by means of a formula without actually selecting repeated samples. In the case of judgment samples the "error" conceived of here is the standard error of the means of judgment samples of a given size about the population mean, each judgment sample involving a different individual.

When extremely small samples are to be used (samples of about two or three elements) better samples are likely to be obtained through judg-

[2] Yates, F. SOME EXAMPLES OF BIASED SAMPLING. Ann. Eugenics 6:202–213 1935.

ment selection, but as the size of the sample increases the random sample can be expected to yield more accurate results, on the average. There is little documentary evidence to indicate the sample size at which the random sample becomes superior. This will probably vary considerably from one situation to another and with the skill of the sampler. But there seems to be no doubt about this general comparison of the accuracy of judgment and of random samples.

This general comparison, if accepted, provides a basis for interpreting and answering certain criticisms of random samples. Let us consider, for example, a stratified random sample of 200 counties of the United States. (An unrestricted random sample could be used equally well as an example.) Assume 200 strata based upon type of farming, or some such criterion, with one sample county selected from each stratum. County X, a random selection from stratum A, might be questioned on the grounds that it is unlike any other county in the stratum, or that it is of very little agricultural importance. In other words, the critic regards stratum A as a universe in itself, the problem being to select one county to represent that universe. He does not think of the sample as a whole in terms of how well it represents the United States, or perhaps a region, but in effect incorrectly regards it as 200 individual samples of size 1.

If the purpose is to select one county from stratum A that will give the most accurate representation of that stratum, a hand-picked selection would probably be more satisfactory than a random selection; but no matter what county is selected, it is not an adequate sample of the stratum in any technical statistical sense. A sample should be looked upon as an aggregate. But the question might be asked, "If one hand-picked sample county from each stratum will give the 'best selection' for the strata individually, even though a sample county may not adequately represent a stratum, why does not such a procedure give better results in the aggregate than a random-sampling procedure does?"

The question will be answered with the aid of a hypothetical example. Consider a population of 10 strata with one county to be selected from each. The deviations of sample counties from stratum averages for random and judgment samples might be shown in the tabulation.

	Random selection	*Judgment selection*
	+12	+5
	−10	+8
	+ 5	+6
	−20	+4
	−10	+5
	+15	+3
	+10	+7
	− 4	+6
	− 5	+5
	− 7	+5
Average error	− 1. 4	+5. 4

Stratum by stratum, in this hypothetical example, the judgment selection is usually better; but as with most judgment selections there is a tendency to be too high (or too low), so in the aggregate the average error for the random sample is smaller, unless the samples are exceedingly small. The previous statement is, of course, made with respect to what happens in repeated samplings.

Decisions To Be Made When Designing Samples

Five broad questions that need to be considered when samples are designed are discussed briefly.

1. *Defining the universe and units of observation*

A necessary step in selecting a probability sample is carefully defining the universe. This usually involves (1) establishing its geographic limits and (2) defining the elements that qualify for the study within the area to be covered. A probability sample cannot be selected without a complete listing, or its equivalent, of all sampling units. On the other hand, when judgment samples are used, the universe and the data to be collected are often loosely defined; in fact, a universe may not be defined at all.

Defining the universe may not present problems. All students registered at some college would represent an easily defined universe. But suppose a special survey of cattle feeders is required. Deciding upon the geographical limits of the area to be surveyed and the definition of a cattle feeder may involve difficult and arbitrary decisions. More serious consideration probably should be given to this question when the survey is to estimate the total number of cattle on feed than when the survey does not involve estimating totals, as in a study of production practices.

It is to be remembered that a sample of 1,000 cases might represent the whole of two States about as well as one State. A sample of 1,000 farmers in Iowa and Illinois, for instance, would represent the two States combined with respect to corn production practically as well as a sample of 1,000 farms in Iowa would represent corn production in Iowa, because corn is found on most farms in both States. As this point is not always understood, a survey is sometimes unnecessarily confined to a small area, because of the misbelief that the accuracy of a sample is more closely related to its size in terms of percentage than in terms of absolute numbers. With respect to a general-purpose farm survey, involving many crop and livestock items, a sample of a given number of farms ordinarily would not represent two States as well as one because of the way various items are distributed geographically. It is difficult to give general rules.

For a poultry survey, the specifications may require that a producer, in order to qualify for the survey, must be larger than some arbitrary size. How large should a producer be to qualify? If field contacts are required to ascertain eligibility for the survey, and if it takes only a few minutes to get a schedule from a small producer, it seems feasible to use a very low cut-off point, or no cut-off point at all. Under such circumstances little time would be saved by excluding small producers, and information would be lost on the small fraction of the total which they represent.

The unit of observation is the unit on which separate schedules or data are obtained. In defining a unit of observation two important considerations are utility or meaningfulness as a unit of analysis, and clarity of the definition from the standpoint of field operations. Closely associated with the matter of defining the unit of observation is definition of data to be collected.

2. *Size of sample*

Size of sample is most commonly determined by the funds available. The problem is to make the most efficient use of resources. Sample size will be discussed here, however, under the assumption that funds do not limit the survey seriously.

A major determinant of sample size is the degree of break-down wanted in the analysis. Are State summaries adequate? Are tabulations by subareas of a State required? Are tabulations by income groups within subareas needed? An increase in the number of groups for which separate tabulations are wanted requires roughly a proportionate increase in the size of sample. It is advisable to work out skeleton tables or an outline of the analysis in advance of a survey, but this is seldom done. The skeleton tables are useful when developing the questionnaire as well as the sampling design. It helps to avoid the omission of important questions or the inclusion of unimportant questions. It gives the sampler a good picture of what is required of a sampling design to meet the survey objectives and provides a better opportunity for the sampler and the subject-matter specialist to reach a clear understanding on what alternative designs can accomplish satisfactorily.

Another major determinant of sample size is the degree of accuracy required. It cannot be said categorically that a sample estimate with a 6-percent coefficient of variation is no good; that a coefficient of variation of 5 percent is required to bring useful results. When measuring change from one point in time to another, however, if the changes are known to be small (2 or 3 percent) it is clear that a precision job is required, and a sampling design with a 5-percent coefficient of variation is inadequate. Specifying a standard of accuracy for a survey calls for an evaluation of how much various degrees of accuracy are worth.

Other factors affecting the size of sample needed are the variability from one sampling unit to another of the information to be collected, the frequency of occurrence of the items, and the efficiency of the sample design. A larger sample of farms is required when making an estimate of the number of hired workers with a given coefficient of variation than when estimating the number of family workers with the same coefficient of variation.

Many persons still think of size of sample in terms of percentage. Questions commonly asked are, "Do you think a 5-percent sample is large enough; or what percentage of the population should be included in the sample?" Actually, it is the number of cases that is important, not whether the sample covers 1 percent or 5 percent of the total. But the proportion that the sample is of the total population does have an appreciable effect on accuracy when sizable proportions are involved—more than roughly 10 percent.

3. Restrictions to be employed in the sampling design

There are numerous ways of imposing restrictions in a sample design. It is possible to indicate here the nature of only a few of the decisions that may be necessary. No mention will be made of the principle of stratification or the choice of criteria to be used in stratification as the purpose of stratification is well known.

First, should the sample be a single-stage or a double-stage sample? In a single-stage sample there is only one stage of selection; that is, there is no subsampling of sampling units. A simple stratified random sample of small areas of land (sampling units) in a State when the small areas are completely enumerated is a single-stage sample. A common example of double-stage sampling is making a selection of a sample of counties (primary sampling units) which is followed by selecting small areas (secondary sampling units) within the selected counties. Such a double-stage sampling scheme reduces the scatter of a sample and its cost compared with a single-stage sample. But the reduction in scatter means a reduction in the efficiency or accuracy of the sample. With knowledge of appropriate components of cost and of the magnitudes of components of sampling error, the sampler can determine the allocation of the sample which will give the lowest sampling error per dollar; that is, he can decide how many primary sampling units should be selected relative to the amount of sampling within the primary sampling units.

In the case of single-stage sampling, the size of the sampling units, and in the case of double-stage sampling, the sizes of the primary and secondary sampling units, must be determined. The objective of maximum accuracy per dollar is used in making the decision. As an illustration, a single-stage sample of 1,000 farms, in 100 clusters of 10, is less efficient and costs less to use than a sample of 1,000 farms in 200 clusters of 5. By and large, with respect to agricultural items, methodological research indicates that it pays to do what may seem excessive traveling.

In general, the more that is known about the population to be sampled, the more efficient the sample that can be designed. This is somewhat ironical for if enough were known about the population a sample would not be necessary.

4. The non-response problem

What instructions should be given regarding the number of calls to make when no one is at home? Making one call at every farm or household in the sample usually yields interviews with roughly 70 percent of the specified farms or households. Here are three ways of dealing with the remaining 30 percent. (1) Make no effort to get information on the remaining 30 percent. (2) Take a subsample of the remaining 30 percent and make every effort to get an interview with everyone in the subsample. Some theory [3] has been worked out for indicating the optimum degree of subsampling of non-respondents. The same theory applies in the case of a mailed questionnaire when one wishes to take a follow-up sample of non-respondents by personal interview. This scheme requires weighting the data to get unbiased estimates. (3) Use three calls, or enough to get at least a 90-percent coverage. The survey specifications should be set up in such a way that the degree of completeness is known.

A decision on the question of call-backs is again a matter of accuracy and cost. Experience has shown that individuals who are found at home on the first call often differ so much from those who can be found only after two or more calls that the non-response problem cannot be ignored. However, the feasibility of making call-backs is sometimes seriously questioned because of the expense.

Let us examine the costs in terms of schedules per dollar. For purposes of discussion, assume area sampling to get an expected number of farms equal to 100. (The word "calls" as used below means the total number of calls including first call and call-backs.)

Scheme A: If no call-backs or substitutions are made and there are no refusals, a sample of 143 farms must be specified to get 100 schedules, assuming one call to each yields schedules from 70 percent of the farms. Thus 143 calls are required to get 100 schedules.

Scheme B: On the other hand, suppose a sample of 100 farms is specified and an indefinite number of call-backs is required so that close to 100 sched-

[3] Hansen, Morris H., and Hurwitz, William N. THE PROBLEM OF NON-RESPONSE IN SAMPLE SURVEYS. Amer. Statis. Assoc. Jour. 41 (236): 517–529 1946.

816634—49——2

ules are obtained. If the proportion of schedules obtained on successive calls remains about 70 percent, 143 calls will yield very close to 100 schedules, if there are no refusals. The average number of calls per schedule (1.43 in the above illustration) is approximately the same, regardless of whether call-backs are specified or not, if the proportion of interviews obtained in each successive call remains about constant. A sample of 100 schedules under Scheme B is worth more than a sample of 100 schedules under Scheme A. An alternative would be to select a sample of 100 farms and make substitutions for those who are not at home on first call, but making substitution does not appear to be a satisfactory solution to the non-response problem.

Actually, it is conceivable that in some situations a sample with specifications for call-backs might have lower costs of collection per schedule than a scheme where no call-backs are made. Much depends upon how the field work is organized and the method of sampling involved. As area sampling has been applied when sampling farms, an appreciable amount of time is required to identify the farms in a segment and establish their headquarters, before determining which farms should be included in the sample. Under Scheme A, about 40 percent more segments are required than under Scheme B to yield the same number of schedules. The cost of selecting 40 percent more segments and identifying the farms in them is a factor leading to higher costs under Scheme A.

The objective here is not to discuss all of the factors making one scheme more or less costly than the other, but to indicate that making call-backs may be less expensive than supposed on first thought.

5. *Expansion of the sample*

Methods of expansion or estimation should be considered during the planning stages of a survey. Sampling specifications may be affected by them; for example, if a highly efficient method of expansion is available, perhaps a smaller sample can be used. Or the field instructions may be affected, or the control or identification information on the questionnaire. Specifications of a sampling design are not complete until consideration has been given to the question of expansion.

When deciding upon a method of estimation, statistical efficiency is only one of the factors involved. In many surveys there are deficiencies which may affect some methods of estimation more than others. An example is the use of an area sample for the purpose of getting a sample of farms according to the Census definition. Appreciable incompleteness might occur among small borderline farms.

Three methods of estimation may be considered: (1) Reciprocal of the sampling rate, (2) mean per farm, and (3) ratio to farmland. For the second and third the expansion factors are assumed to be Census number of farms and farmland, respectively. The first method would be satisfactory for items that are found in very small quantities or nonexistent on the small borderline farms, as hired labor. Estimates of total hired workers would be unaffected by omission of farms having no hired workers; but this would not be true with the mean-per-farm method of estimation. On the other hand, by the first method estimates of total farm population could be seriously affected by the omission of some of the small farms; the effect on estimates made by the second method would be less serious. For many crop acreages an estimate made by the mean-per-farm method would be affected more than the estimates made by either of the other two methods.