107

W. EDWARDS DEMING

UNCERTAINTIES IN STATISTICAL DATA, AND THEIR RELATION TO THE DESIGN AND MANAGEMENT OF STATISTICAL SURVEYS AND EXPERIMENTS

Reprinted from the
Bulletin of the International Statistical Institute
VOL. XXXVIII : PART IV
TOKYO 1961

UNCERTAINTIES IN STATISTICAL DATA, AND THEIR RELATION TO THE DESIGN AND MANAGEMENT OF STATISTICAL SURVEYS AND EXPERIMENTS

by

W. EDWARDS DEMING
Consultant in Statistical Surveys
4924 Butterworth Place, Washington 16, D. C., U.S.A.

I. THE UNIVERSE, THE FRAME, AND THE EQUAL COMPLETE COVERAGE

Purpose of the paper

A statistical survey today involves a complex combination of knowledge and skills of various kinds. It is often an operation of large scale. Faulty organization, with confused responsibilities between subject-matter, statistical theory, and operations, is a frequent cause of poor results. A practicing statistician in industry, where faulty logic or faulty organization in a statistical study may lead to costly decisions in the management of a business, and to criticism of statistical methods and of statisticians, must formalize some principles and apply them in the management of statistical surveys.

The statistician is the logician and the architect of a survey or experiment. He is qualified, as a necessary part of his education in statistical method, to classify the responsibilities in the planning, execution, and interpretation of the results (see Example 2 in part V). The purpose of this paper is to point out to statisticians the necessity to do this, and to point out some principles for guidance.

Operational definition

It is well to note first that a concept of definition, to have communicative meaning, must be operational. I quote Shewhart's *criterion of meaning*[1]:

> Every sentence, in order to have definite scientific meaning, must be practically or at least theoretically verifiable as either true or false upon the basis of experimental measurements either practically or theoretically obtainable by carrying out a definite and previously specified operation. The meaning of such a sentence is the method of its verification.

Operational definitions of the adequacy of the frame, of the mathematical bias and of the standard error of a sampling procedure, and of the bias and of the

1) Walter A. Shewhart, *Statistical Method from the Viewpoint of Quality Control* (The Graduate School, Department of Agriculture, Washington 25, 1939), page 94.

accuracy of a survey-technique, will indicate a logical line of responsibility for effective use of various sorts of knowledge and skill that go into a statistical survey.

The universe

The concept of the equal complete coverage will be the foundation of our definitions. It is first necessary, however, to define the universe, and then the frame. The universe, the term used in English, is all the people, firms, material, conditions, concentrations, units, models, levels, etc., that one wishes to study, whether accessible or not. The universe, for any study, becomes clear from a careful statement of the problem, and of the uses intended for the data. An example is all the firms that make a certain product, or that may buy it. Other examples are all housewives; all school children; all the pigs in a country, both in rural areas and in towns; or all the material or piece-parts covered by a certain contract or specification. The universe may be all the records of transactions dated within a specified period of time, where the aim of the study is to estimate the company's revenue from certain types of business. A further example is all people, or old people, or young people, when we wish to compare 2 medical treatments.

The frame

The frame is a means of access to the universe,[1] or to enough of the universe to be worth studying. What census data, lists, maps, will form suitable frame? In case of accounts, the question is very often at what point may we study the records that show the transactions that we are interested in, after final corrections and missing information have been entered? A frame is composed of units of one kinds or another, called sampling units, which enable us to take hold of portions of the universe, piece by piece. Every piece of material that the frame covers will belong to one definite sampling unit, or will have an ascertainable probability of belonging to any given sampling unit. Without a frame there can be neither a complete coverage nor a probability sample, as there would be no way to lay out the work nor to know the probability of selection of any sampling unit.

Every sampling unit in a frame will bear a serial number, or will have a prescribed way of getting one. A random number will thus select a definite sampling unit, and will lead to the investigation of all or of a designated random sample of whatever material in the sampling unit belongs to the universe.

A question of vital importance in the early stages of a survey is how much of the universe does a proposed frame cover? (90%, 96%, 100%?) What groups, classes, areas, or conditions does it omit, wholly or partially? Fig. 1 is a schematic diagram that portrays the material in the universe, part or all of which lies in the frame. The portion of the universe that the frame fails to include, if any, is the gap between the frame and the universe.

A frame may be useless or nearly so for the purpose intended if it omits too much of certain important classes of the universe. It is substantive judgment and not knowledge of statistical theory that must decide whether a proposed frame is satisfactory. Thus, if we carry out a study of the uses and purchases of (e.g.) typewriters in business establishments with a frame that omits small establish-

1) The concept of the frame was first stated by Frederick F. Stephan in "Practical problems of sampling procedure," American Sociological Review, vol. 1, 1936: pp. 569-580.

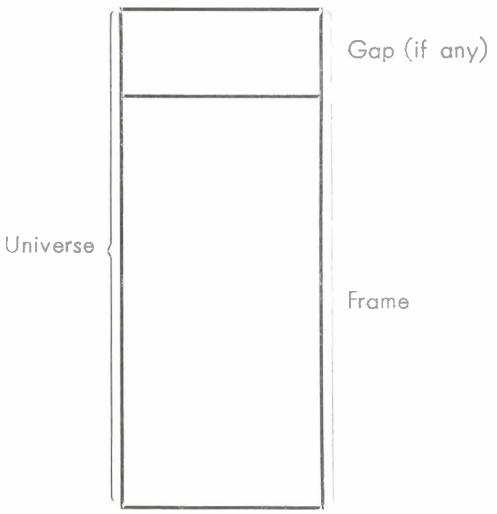

Fig. 1. Schematic diagram to show the relation between the frame and the universe. The gap is the material in the universe that the frame fails to include. There is sometimes no gap.

ments and nonprofit institutions, it is substantive judgment and not statistical theory that must decide whether the results will be useful without information from the small establishments and from the nonprofit institutions.

I shall draw upon a summary records of the United Nations Statistical Commission to illustrate confusion between the frame and the universe, and confusion concerning the responsibility for the decision on the frame. The statistical reader may be able to recall, from his own experience, similar illustrations, and consequent misinterpretation of the data. I am not sure whether the confusion existed in the original statement or in the mind of the reporter; hence I omit the name and date, but I quote as follows:

> Mr. ... (of the UN Statistical Commission) congratulated the Sub-Commission on Statistical Sampling on its report, but felt that greater stress should have been laid on the need for common sense in using sampling methods. There are many dangers involved in using random sampling methods. For example, if one took a random sampling of farms for the purpose of finding out the number of pigs (for example) in Norway, the result would be inaccurate, because some pigs live in the towns. Thus, the statistician should always keep in touch with reality ....

It is the task of the statistician, as a logician, to set such matters clear at the outset. First, the definition of the universe for a count of pigs is the responsibility of (e.g.) the Minister of Agriculture, who must decide whether he wishes to count the pigs in the rural areas only, or to include the pigs in the towns. It makes no difference whether the count will be a complete census of pigs, or an estimate made from a sample. If a proposed frame would be unsatisfactory for a complete census, then it would be unsatisfactory also for a sample. Conversely, a frame that would

be satisfactory for a complete count is also satisfactory for a sample (provided the sampling units in the frame are usable and economical for a sample, a question that we do not consider in this paper).

Definition of the equal complete coverage[1)]

The equal complete coverage is by definition the result that would be obtained from investigation of all the sampling units in the frame (segments of area, business establishments, accounts, manufactured articles) by the same field-workers or inspectors, using the same definitions and procedures, and exercising the same care as they exercised on the sample, and at about the same period of time. The concept of the equal complete coverage is fundamental to the use of samples. The adjective *equal* signifies that the same methods must be used for the equal complete coverage as for the sample. Every sample is a selected portion of the sampling units in the frame; hence A SAMPLE IS A SELECTED PORTION OF RESULTS OF THE EQUAL COMPLETE COVERAGE.

A complete coverage may be conceptual or it may be actual. It is easy to point to examples of samples drawn from actual complete coverages. Take, for example, the Census of Population. There is a punched card for every person enumerated in the Census, but many of the volumes of tables published by the Census are made, not from tabulations of all these cards, but from a sample thereof. Sampling thus greatly enlarges the scope of publication in the Census. Many special studies, as of fertility, are made by the examination of a sample of families drawn by a prescribed rule from the original Census records.

The Census is in such examples the equal complete coverage for the sample. Complete census and sample both contain the same proportion of careful responses, of careless responses and of nonresponse, of careful coverage and of careless coverage.

Effective division of responsibility in the planning of a survery

The concept of the equal complete coverage provides a logical basis for effective division of responsibility between (1) the subject-matter (chemistry, demography, sociology, medicine, psychology, engineering, agricultural science); (2) application of statistical theory (sample-design, statistical controls of the operations, interpretation of the sampling and nonsampling errors); **(3) the operation of carrying out** the study.

The expert in subject-matter has the responsibility at the outset, when the question of a possible survey first comes up, to state how he expects to use the results. This statement automatically defines the universe. The statistician, as a logician, has a duty to explain to the expert in the subject-matter that he (the expert in the subject-matter, not the statistician) must decide whether a complete coverage of the proposed frame, by the proposed methods of questioning or testing, would provide useful information. The statistician must make clear that any inference that are to come from the results of the study by use of statistical theory can only cover the frame and the materials, methods, levels, types, and conditions presented for study

1) The concept of the equal complete coverage (without a name) originated with Morris H. Hansen and W. Edwards Deming, "On an important limitation to the use of data from samples", Bulletin de l'Institute International de Statistique, Bern 1950, vol. xxxii, part 2; pp. 214–219.

in the frame; that generalizations to other materials, levels, types, and conditions outside the frame can come only through knowledge of the subject-matter, not from statistical theory, and that the results of a survey or experiment may be a disappointment if the frame proposed and experimental conditions proposed for the study fail to include all the materials, methods, levels, types, and conditions on which the expert in the subject-matter desires information.

Once it is clear to the expert in the subject-matter that a complete coverage of a proposed frame by a proposed procedure of interviewing or testing would provide useful information, and that a more complete frame would not be worth the additional cost, then (and not until then) questions of sampling and of experimental design arise.

The next step is to design a suitable sampling plan or statistical experiment, and to estimate the cost for a few selected levels of precision. The design of a sample or experiment is definitely statistical, being the application of statistical theory. The decision on whether the survey will be worth its cost, however, belongs to the man who will pay the bill for the study, or to the man who will be responsible for using the results.

II. OPERATIONAL DEFINITION OF THE EXPECTED VALUE AND OF THE STANDARD ERROR

Operational definition of the error of sampling

Suppose that we have a frame, and that its sampling units bear the serial numbers 1, 2, 3, and on to N. Now make a complete coverage of this frame (i.e., investigate 100% of the sampling units therein), using prescribed methods of interviewing or of testing.

The investigators have had a certain course of training, or maybe none at all. They may follow a certain ritual. Some of them may be careful; some may be careless. They may fail to find all the dwelling units, all the people, all the material, whatever it be. They may report on nonexistent sampling units. They may make mistakes. Some respondents may misunderstand some questions. Some people may not be at home when the interviewer calls; some will refuse to be interviewed. There may be errors in the original records that constitute the frame. The complete count, of whatever quality, is the equal complete coverage for all the samples that may be drawn and processed in the manner prescribed.

However carried out, and whatever be the rules for coding and for adjustment for nonresponse, the comlete coverage of the N sampling units will yield the N numerical values

$a_1, a_2, a_3, \cdots, a_N$ for the x-characteristic

$b_1, b_2, b_3, \cdots, b_N$ for the y-characteristic

Denote the sum of these N individual populations by

(1) $$A=a_1+a_2+a_3+\cdots+a_N=Na$$

(2) $$B=b_1+b_2+b_3+\cdots+b_N=Nb$$

We may have interest in A, B, and various other characteristics of the frame such as

(3) $$\varphi = A : B$$

and should like to estimate them by use of a sample.

The numbers a_1, a_2, etc., are not "true" values of the populations in the N sampling units; they are instead only the results of the complete coverage. They contain all the errors mentioned above. There are no true values.

An operational definition of the error of sampling is contained in the following experiment.

1. Write each of the N observed values a_i on a card, and number the cards serially 1, 2, 3, etc., to N (Fig. 2).

2. Draw in the manner specified in the sampling plan a sample of n cards. Let

x_1 be the observed value on the sampling unit drawn by the 1st random number
x_2 be the observed value on the sampling unit drawn by the 2d random number
x_3 be the observed value on the sampling unit drawn by the 3d random number

x_n be the observed value on the sampling unit drawn by the nth random number and likewise for y_1, y_2, ···, y_n.

The frame			*The sample*	
1	a_1		1	x_1
2	a_2		2	x_2
3	a_3		3	x_3
⋮	⋮		⋮	⋮
N	a_N		n	x_n
Total	A			Σx_i
Average per sampling unit		a		$\bar{x}$
Standard deviation		σ		s

Fig. 2. Pictorical representation of the complete coverage of a frame, and of a sample drawn therefrom. x_1 in the sample is some one of the a_i in the complete coverage, drawn at random: likewise x_2, x_3, ···, x_n. Statistical theory enables us to make predictions about the sampling variation of the results from future samples, all drawn from the same complete coverage and processed according to the same rules.

3. Form estimators by the formula specified in the sampling plan. To be specific, we may focus attention on possible functions like

(4) $$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n)$$

(5) $$\bar{y} = \frac{1}{n}(y_1 + y_2 + y_3 + \cdots + y_n)$$

(6) $$X = N\bar{x}$$

(7) $$Y = N\bar{y}$$

(8) $$f = \bar{x} : \bar{y}$$

If we use these functions as estimators of a, b, A, B, φ, respectively, we could compute the errors of sampling

(9) $$\Delta\bar{x} = \bar{x} - a$$

(10) $$\Delta\bar{y} = \bar{y} - b$$

(11) $$\Delta f = f - \varphi$$

Etc.

In practice, we do not usually have the complete coverage, and can not compute the sampling errors for our sample.

It is an exciting fact, however, that a single sample, provided it is big enough and provided it is laid out properly, will provide an estimate of the margin of sampling variation of all the estimates that one can form by repeatedly drawing samples from a given complete coverage and processing them by the prescribed sampling procedure. The same theory enables the statistician to design in advance a sample that will deliver about the precision required. This is the great contribution of modern statistical theory.

Operational definitions of the expected value, standard error, and bias of a sampling procedure

We continue our experiment.[1)]

4. Return the sample of n cards to the frame, and repeat Steps 2 and 3 to form a new estimate by the same sampling procedure. Repeat these steps again and again, 10,000 or more times.

5. Plot the distribution of $\bar{x}$, using any suitable class-interval. Compute the mean and the standard deviation of this distribution. Any one of the above samples is a random selection from all the $N!/(N-n)!n!$ or $\binom{N}{n}$ possible samples of size n, all of which samples have the same probability. $\bar{x}$, $\bar{y}$, X, Y, f, are therefore random variables, whereas A, B, φ, and other results of the complete coverage are constants (not random variables) in this experiment. The $N!/(N-n)!n!$ possible values of any estimator x form the theoretical sampling distribution of x. The mean $E\bar{x}$ and the standard deviation σ_x of this theoretical distribution are of special interest, and have names. Ex is by definition the "expected" values of x. Let ξ be the characteristic of the complete coverage that x estimates: then if

(12) $$Ex = \xi$$

the sampling procedure is said to be unbiased. But if

(13) $$Ex = \xi + C \quad (C \neq 0)$$

the sampling procedure has the mathematical bias C. In any case, the variance of the distribution of x is

(14) $$\sigma_x^2 = E(x - Ex)^2$$

By definition, this is the variance of the sampling procedure for the estimator x, and its square root (σ_x) is the standard error of the sampling procedure for the estimator x. Thus, a sampling procedure has, for any estimator, an expected value,

1) Taken largely from the paper by Hansen and Deming (1950) cited earlier.

a standard error, and possibly a mathematical bias C. The bias C, if it exists at all, disappears rapidly as the number of sampling units in the sample increases. It is not related to the bias of poor performance, nor to bias built into the questionnaire (*vide infra*).

Some other sampling procedure will have a different bias and a different variance.

The above definitions are operational because they describe a procedural basis by which to measure the expected value, the mathematical bias, the variance, and the standard error of an estimator.

Operational definitions will not get us into trouble with impossible words like true value, perfect questionnaire, perfect complete coverage, none of which, so far as I know, has meaning. The definitions given here are based on a complete coverage as actually carried out, imperfections and all. They make a clear separation between what is sampling and what is not.

Limitations of the standard error

The standard error as defined above includes automatically not only the variability that arises from new selections in repetitions of the sampling procedure, but also the original variability in the complete coverage that arose from fluctuations in the investigators' judgment and performance, which may be different before and after lunch; also the variable effect of the order of interviewing and the variable effect of the weather and of other conditions that change the material or change the investigators' judgment over the period of the survey. It includes the variance between interviewers, unless the design of the sample allotted the interviewers orthogonally over the sample. This is so because the effects of these variations are all built into the numbers $a_1, a_2, \cdots$ which constitute the complete coverage. We never know anything about these variations in a complete coverage unless we design the complete coverage as a composite of interpenetrating sub-samples.[1)]

The standard error of an estimator only gives us a measure of the variation between the results of repeated samples from the equal complete coverage. It does not detect nor measure the constant component of any persistent nonsampling errors that were built into the equal complete coverage. These one measures by the statistical audit or control, and by outside comparisons. A small standard error of an estimate means: (1) that the variation between repeated samples must be small; hence also (2) that the accidental blemishes and variations from all sources must be small; and (3) that the result of the sample agrees well with the result of the equal complete coverage *of the same frame*. It does not mean that the persistent nonsampling errors are small, nor that the frame was satisfactory.

The measure of sampling variation

The results (X) of repeated samples from the same complete coverage will distribute themselves as a random variable about EX. The maximum variation between the results of repeated samples all drawn from the same complete coverage, and following a prescribed sampling procedure, is usefully placed at 3 standard deviation ($3\sigma_X$) in either direction from EX. This rule is a statistical standard long

1) I am indebted to my colleague Professor P. C. Mahalanobis, F. R. S., for pointing this out to me.

used in industry.[1] It gives the client or the user of an estimate what he needs to know about the sampling variation. One may wish to widen his estimate of 3 standard errors, to be conservative, when there are only a few degrees of freedom to work with. One may also in rare instances wish to calculate the effect of extreme skewness or other departures from normality.

In my own practice, I steadfastly refuse to estimate or to discuss the interpretation of the standard error when large operational nonsampling errors are obvious. The standard error, under such circumstances, is sure to mislead the user of the data.

The correction of the operational nonsampling errors is the responsibility of the supervisor of operations (interviewing, testing, pricing, computing). It is not the responsibility of the statistician, although statistical methods may be very helpful in detecting the existence of blemishes and blunders in procedure. Once these mistakes are apprehended and corrected then one may usefully discuss the standard error.

III. CLASSIFICATION OF UNCERTAINTIES, AND RESPONSIBILITIES FOR REDUCING THEM

Reasons for studing all the sources of uncertainty

To the user of the data of a survey, it is only the total error that counts: he does not care whether it is a standard error or some other kind of error.[2] The more we know about the limitations of a figure, or of a procedure, the more useful it becomes. Once we learn something about the nature and cause of any uncertainty, we may find some way to reduce it. We accordingly turn attention now to the various uncertainties in surveys, other than those that arise from the random selection of sampling units. Reduction in any type of the nonsampling errors is a step toward improvement of the quality of data from complete counts and samples alike.

Economic balance of errors

When the nonsampling errors are large, it is uneconomical and ineffective to waste funds on a big sample, as a big sample, though it decreases the sampling error. will reduce the total error only very little. One must face the fact in the management of statistical surveys that he may enhance the overall usefulness and reliability of a survey by cutting down on the size of the sample and using the money so saved to reduce the nonsampling errors. In the sampling of records, this might mean tracing and correcting wrong and missing information. In a survey of human populations, this might mean more time and money on the preparation of the questionnaire, hiring fewer and better interviewers, providing better training and better supervision in the field, and making more recalls on people not at home on a pre-

1) See, for example: (1) Report of Committee on Standards of Probability Sampling for Legal Evidence. Current Business Studies (Society of Business Advisory Professions, New York University, March 1957). (2) Tentative recommended practice, "Acceptance of evidence based on the results of probability samples", E 141-59T (American Society for Testing Materials, 1916 Race Street, Philadelphia 3).

2) As stated by Alfred N. Watson at a meeting of the American Statistical Association in Chicago in December 1942.

vious call. In fact, one of the early papers[1] on the various sources of error arose from the standpoint of management: how to achieve, in sample-design, an economic balance between the sampling error and the various nonsampling errors.

Classification of uncertainties and deficiencies common to complete coverages and to samples

The classification that follows has been helpful to the author in statistical practice, as it shows where to lay responsibility and emphasis in the planning of the questionnaire and in the operations. Responsibility for decreasing uncertainties of Type I rests definitely with the expert in the subject-matter, and with experts in questioning, interviewing, or testing. The uncertainties of Type II are completely different in nature, as they arise from operational blemishes. Responsibility for holding the uncertainties of Type II to a minimum rests with the supervision of the job. The important point is that the 2 types of uncertainty exist, and that their correction requires knowledge and action of entirely different sorts, neither being knowledge of statistical theory.

Type I. Built-in deficiencies; missing the point; measuring properties of the material not well suited to the problem. The distinguishing characteristic of this type of uncertainty is that it is built into the questionnaire, or into the method of test, or into the rules for coding. It does not arise from flaws in carrying out the specified survey-procedure: a recanvass (audit or control; *vide infra*) will not discover it. It is independent of the size of the sample.

Examples:

1. Failure to perceive what information would be useful; eliciting (perhaps accurately) information that is of little help on the problem. In the sampling of accounts, errors in source-documents will carry through into the final estimates, whether one covers the documents by a complete coverage or by a sample. Failure to know about these errors, or to correct them (best done in the sample) is an error of Type I.
2. Too big a gap between the frame and the universe. An example occurs when one applies an interpretation or forecast to domains and universes not covered by the frame used in this survey. (A gap is not mere mistakes of omission in the preparation of the frame.)
3. Ineffective rules for coding.
4. Ineffective tabulations.
5. Failure to recognize secular changes that take place in the universe before the results are written up and recommendations made.
6. Bias arising from bad curve-fitting; wrong weighting; incorrect adjustment.
7. Unwarranted deductions from the results, with a report that may lead to misunderstanding and to misuse of the survey. The report concerning the findings of the survey should make clear the limitations of the data. It should take into account the fact that the users of the figures may lack survey-experience, and be unable to comprehend uncertainty in a figure. The report should evaluate and interpret the margin of sampling error, and the possible effect of blemishes and

1) W. Edwards Deming, "On errors in surveys," *American Sociological Review*, vol. 9, 1943: pp. 359–369.

blunders made in carrying out the survey-procedure. It should call attention especially to the possible misinterpretation that could arise from nonresponse, or from any gap between the frame and the universe.

Although a recanvass will not discover the existence of an uncertainty of Type I, an outside comparison may do so.

Type II. Blemishes and blunders made in carrying out the field-work, the testing, the interviewing, the coding, the computations, and other work. These errors have their origin in imperfect workmanship. They are discoverable and measurable by repetition or recanvass (called the audit or control) of a sample of the main sample. All of them can occur in complete coverages as well as in samples.

8. Failure to find or to visit all the sampling units that were drawn into the sample.

9. Failure to provide definite boundaries or clear definition of a sampling unit. As a result, or possibly through carelessness or by accident, the investigators may fail to test or to interview some part of a sampling unit, or may go out of bounds and test or interview units not intented for the sample, or not even in the frame.

10. Failure to cover a sampling unit completely, such as failure to find all the dwelling units or all the people therein.

11. Covering some material twice.

One can avoid this error in either a complete coverage or a sample of physical material if the inspectors will mark any unit that they test, so that anybody can see that it had already been tested. In a destructive test, a 2d test is impossible. Human populations usually report and halt a 2d coverage.

12. Failure to ask some of the questions, or to make all the tests prescribed. Getting wrong answers.[1] Asking questions not on the questionnaire.

13. Using the wrong test-instrument. Errors in counting and in weighting. Looking up the wrong price, or computing it incorrectly.

14. Nonresponse and refusal.

15. Mistakes in calculation and in transcription.

Persistent omission or inclusion of material above or below average value, or persistent mistakes in one direction, will cause biases. The only way to evaluate them is by the audit or statistical control, or with the help of outside sources of information.

Will 2 samples agree? Will 2 complete counts agree?

The precision of a sample is not established by comparison against a complete census UNLESS the complete census is THE equal complete coverage for this sample. Simultaneous trials of complete count and sample, just to see whether sampling will give the same results, is in my opinion a woeful waste of funds. In my own practice, I have steadfastly refused to engage in such tests. A simultaneous test, for the sake of comparison, to see if sampling will work, is almost sure, I believe, to impair the results of both the complete coverage and the sample. The 2 results might still agree, of course. However, we know by theory, in advance, better than

1) Morris H. Hansen, William N. Hurwitz, Harold Nisselson, and Joseph Steinberg, "The re-design of the Census Current Population Survey," J. Amer. Statist. Assn., vol. 50, 1955: pp. 801–819.

any number of comparisons could possibly establish, what the performance of a sampling procedure will be, provided we really carry it out according to plan.

IV. OPERATIONAL DEFINITIONS OF THE BIAS AND OF THE ACCURACY OF A TECHNIQUE

The preferred technique and the working technique

We may speak of the definitions, the method of test, the questions, the methods of interviewing, the method of supervision, the treatment of nonresponse, etc., as the survey-technique. What ever be the survey-technique, a complete coverage of all the sampling units in the frame will produce some result, like the numbers a_1, a_2, a_3, etc. in Eq. 1. Another complete coverage carried out with the same survey-technique, before changes have taken place, would give results slightly different from the results of the first coverage. There is an element of randomness in a complete coverage, even when the questionnaire and procedure of interviewing are fixed. This is so because people do not always give the same answer when you ask them a 2d time. Or, some other member of the household may answer on the 2d complete coverage. Some other survey-technique (different definitions, different questions, different procedures) would give still another figure. Neither figure is right or wrong. and neither of them is a true value. Physical measurements, and data transcribed from records, will also show differences from one complete coverage to another.

Any result, whatever it be, is the result of applying some set of operations. Althought there is no true value, we do have the liberty to define and to accept a specified set of operations as preferred, and the results thereof as a *master standard* (so-called by Harold F. Dodge). Thus, there may be, by agreement of the experts in the subject-matter, for any desired property of the material, a *preferred* survey-technique.

The bias and the accuracy of the working technique

Unfortunately, it often happens that the preferred technique, usable on a laboratory-scale, is too expensive to apply in a full-scale survey, or it may be objectionable otherwise. Experts in the subject-matter must then supply also a working technique. Thus, the preferred technique by which to define a person's age might be to compute the difference in time between today and the date shown on his birth-certificate. But some people don't have birth-certificates at all, and few people have them handy. Moreover, some people would not be happy with an interviewer who asked for birth-certificates. The Passport Division can ask for birth-certificates, but interviewers may only ask the person how old he is, and record the result. This would be the working technique by which to measure age.

The preferred technique and the working technique will give different results. A working technique is acceptable to the experts if it gives results not too far, in their judgment, from the results of the preferred technique.

The difference in the 2 techniques, applied to a complete coverage of the frame, is the *bias* of the working technique. A working technique is *accurate* (in the judgment of the experts in the subject-matter) if its bias is small.

It is important to remember that the bias of a working technique is not an error of sampling. The result of a sample will possess the bias of whatever technique is

built into the equal complete count; it will also possess sampling error. The sampling error will disappear as the size of the sample increases, but the bias of the working technique will remain fixed, independent of the size of the sample. The sampling error is calculable from the results of the sample. The bias of a working technique is measurable only by a properly designed experiment, which will compare by the use of interpenetrating samples, with proper randomization of the interviewers, the results from the 2 techniques, the preferred technique, and the working technique.

V. EXAMPLES OF STATISTICAL REPORTS

Purpose of showing these examples

I show here 4 examples of statistical reports. The reader will note (1) that the reports stay within the bounds of statistical inference, or cite authority for any substantive judgment used in the inferences; (2) that the defense of the frame, the questions, methods of investigation, the field-work and the processing, were entirely the responsibility of the client or of his designated experts in the subject-matter; (3) that the statistician, as the logician, prescribed the responsibilities for each phase of preparation and execution; (4) that the statistical report gives no advice on what action to take as a result of the information derived from the study. Such advice, however important and necessary for the user of the data, is not properly, in my judgment, part of statistical practice.

In fact, these reports do not even say that the results are precise or accurate enough for the purpose. Instead, they stamp the results with a label of the precision and accuracy actually found. The user has the privilege of accepting the results, or of discarding them.

Example 1. The 1st example is a brief one, as there was no formal statistical control of the field-work. It refers to a survey conducted by the firm O'Brien-Sherwood of New York in 2 counties, to estimate certain financial, economic, and social characteristics of the readers of a certain newspaper. The statistician's statement was included in a bulletin that the newspaper published to describe the purpose of the survey and to exhibit the results, along with the standard errors of the most important results. (The purpose was to show that advertisements in this newspaper reach purchasing power considerably above the average.)

Statistician's Statement in Regard to a Survey of Limestone and Cherokee Counties

The specifications of the sample for this survey followed generally accepted theory and practice of probability sampling. The specifications if followed would yield results for the responses whose standard errors have the usual interpretation.

The standard error of a result does not measure the effect of nonresponse nor of persistent omissions, inclusions, or departures from procedure. It does include, however, in this survey, the effects of variable performance of an interviewer; also the differences between interviewers.

I had no responsibility for the questionnaire. The sampling plan in this survey did not call for any formal statistical audit of the field-work, nor of the tabulations

or of the computations, nor do I take any responsibility therefor. I did satisfy myself that the firm O'Brien-Sherwood understood the sampling procedure, including the formation of the estimates and of the standard errors. I was on hand at strategic times (my substantive judgment) to ask and to answer questions, and could always be reached by telephone.

I may say, however, in respect to coverage, that the sample gave an estimate of 470,000 dwelling units in the 2 counties combined, with a standard error of about 1.5%. The Census count, taken about a year previously, was 454,400. The difference is 2 standard errors, which could arise from sampling error, or from growth, or from some of both. The direction and magnitude of the difference appear to indicate successful coverage of the selected segments by the field-workers : incomplete coverage would have produced a deficit.[1]

The figures on which I base this estimate of the total number of dwelling units in the whole area, and the standard error thereof, came from O'Brien-Sherwood at my request for the results of the sample.

The firm also informed me that the interviewers obtained responses in 87.3% of the households visited, and that the nonresponses were distributed amongst all interviewers, and in all areas, not being confined nor concentrated in any one class. My instructions asked the firm to make no adjustment for nonresponse, but to show in the tables the figures that came from the households that actually responded : also to show the proportion of nonresponses. I offer no adjustment for the nonresponses.

Example 2. The 2nd example is an excerpt from legal testimony[2], in which a telephone company had carried out an inspection of the various classes of telephone plant through the aid of sampling, to arrive at a figure for the overall per cent physical condition of the entire plant that was subject to sampling. Phrases in parenthesis are explanatory, and were not part of the testimony.

Direct examination

Q. Would you please explain the nature of your engagement with the Illinois Bell Telephone Company?

A. Mr. B., General Staff Engineer of the Company, informed me that he wished to make a survey to determine the overall physical condition of the Company's plant, and he asked me to draw up the proper sampling procedures.

Q. What was the scope of your engagement?

A. To furnish sampling plans for the plant that Mr. B asked me to sample. These plans included instructions on how to serialize the sampling units, exactly how to construct by use of a table of random numbers, the sampling tables for the selection of the sample, though for some classes of plant, I furnished the sampling table myself, procedures for forming the estimate desired, and for estimating its

1) Even a simple statement like this requires substantive knowledge, specifically, some crude knowledge about field-work (pointed out to me by my colleague Leon Pritzker).

2) The Illinois Commerce Commission, Docket No.39126, 1951, and Docket No. 41606, 1954 : the Illinois Bell Telephone Company, Chicago, in the matter of the proposed advance in rates. The passage printed here is testimony prepared in advance, and is not necessarily the same word for word in the record. Moreover, I have supplied some lines from several other subsequent dockets.

standard error. My engagement also covered the statistical interpretation of the results, by which I would explain to Mr. B., on the basis of figures that he would furnish to me as the result of applying the sampling procedures that I would supply, and as the result of an audit (statistical control) that I would prescribe to examine the inspectors' performance, what was the reliability of the overall per cent condition derived from the sample, using as a norm a 100% inspection (equal complete coverage) of every one of the millions of items on the lists that he presented to me for sampling, carried out by the same definitions and methods of inspection as were used on the samples, and calculated in the same way. I satisfied myself that he (and the men directly responsible to him) understood the sampling procedure. I was on hand at strategic times to ask and to answer questions, and could always be reached by telephone.

Q. Were there any special terms about your engagement?

A. No, there was nothing unusual about it. I accepted the engagement subject to my code of professional conduct,[1] which binds me to complete technical responsibility with respect to the sampling procedures, and which binds the company to follow them in every detail; to make no departures without authorization from me.

Q. Did you explain to Mr. B. what his responsibilities would be?

A. I explained to him that he must take full responsibility for the completeness and the accuracy of the engineering records and other lists (the frame) that he would present to me for sampling: that he would be responsible for the methods of inspection, and for the supervision of the inspectors; for the weights of the various classes of property, and for the accuracy of the computations that I would prescribe. I told him that I would assist him to introduce statistical controls on the supervision and on the summaries and on the computations, but that he alone would be responsible for the final product.

(I omit the rest of this example, having covered some essential points.)

Example 3. This example is the report on the results of a sample whose purpose was to estimate certain components of the inventory of parts on hand of a large manufacturing concern, and the LIFO adjustment (change in value over the year) on the inventory. This statement is a legal document, as it forms the basis for the corporation's income tax, as well as for information for the management.

Statement to the Comptroller of the Corporation

> Statement in respect to the reliability of the estimates of the dollar-value of 1957 year-end corporate material, of prior-plant conversion-costs, and of unrealized earnings, for the portion of plant in the paint-and-glass-products pool subject to sampling.

This statement refers to the reliability of the results that you derived from a sample that I prescribed. I understood from you that the lists that you presented to me for sampling were prepared from records maintained by the Corporation for purposes of production programing and inventory control. They showed serial numbers and descriptions of items, and they met a fundamental requirement, namely, your assurance that processing all these part-numbers would constitute a 100%

1) Available on request; or see Chapter 1 of my book *Sample Designs in Business Research* (Wiley, 1960).

evaluation of the problem.

My responsibility is limited to the statistical methodology—the procedure of selection, the procedure for forming the estimates that you required, along with the standard errors thereof, and their interpretation; statistical tests of compliance with the sampling procedure specified, and an audit to test the performance of your people; and finally, the statistical evaluation of the reliability of the results. Your responsibility covers those aspects of the study that would be the same whether you used sampling or not.

I designed a sampling plan to apply to the lists (the frame) that you provided. I worked from time to time with your people on the selection of the sample and on the sample for the audit. I worked with them on the forms, controls, and verifications to apply to the selection of the sample and to the arithmetic processing. I have confidence in their ability and desire to follow accurately the whole procedure. I have reason to believe, by my own subjective judgment based on experience, that the numerical results of the sample are an accurate summary of the figures fed into the routine of processing.

According to figures that you furnished to me at my request, the results of the sampling are in the table herewith. The book-inventory came from the financial statement; the other figures came from ratios estimated from the sample.

Book-inventory, 1957 year-end (from financial statement)	$202,850,010
Corporate material	170,243,916
Prior-plant conversion costs, active	19,995,647
Prior-plant conversion costs, inactive	1,210,756
Unrealized earnings	11,399,691

The design of the sample made it possible to calculate objectively by standard methods, from the results of the sample itself, the tolerance to allow for the outside margin of difference between any of these results and the result that would have come from a complete processing of all the items on the lists that you provided. The outside margin of difference (3-sigma limits) for material falls within 1/2 of 1% of the figures in the table. The outside margin of difference for the corporate conversion-costs, active, falls within 3% of the figure in the table. The outside margin of difference for the unrealized earnings falls within 5% of the figure in the table. Theory and experience show that limits so calculated include the results that you would have gotten from a complete processing of all the items on the lists that you provided, were you to carry out the complete processing under the same rules and with the same care that you exercised on the samples.

The above tolerances include the possible effects of any accidental errors of a canceling nature that might have occurred in the pricing and in the processing, as well as the uncertainty that arises from sampling, but they do not detect nor evaluate the effect of any possible persistent error that there might have been in the pricing or in the processing. The sampling plan therefore called for an audit by which to detect persistence, if any, and to evaluate what effect it could have on the results of a complete pricing and processing of all the items on the lists that you presented to me for sampling, were you to carry out the complete pricing and processing with the same care that you exercised on the sample.

The audit consisted of a probe of a subsample of items drawn from the main

sample. It called for repetition of the entire procedure for the items in the audit, by use of the original instructions, including recalculation, with other investigations that seemed warranted. Analysis of the differences found in the audit indicates the possibility of a small amount of persistence and that it could act in either direction to affect any of the figures in the above table. If there should be any persistence, it would affect the complete pricing in exactly the same way that it would affect the sample. With respect to the total inventory, the maximum overestimate that could arise from persistence, if there be an overestimate from this source, does not exceed 5 parts in 10,000. The maximum underestimate, if there be an underestimate, does not exceed 13 parts in 10,000. With respect to the material in the inventory, the maximum overestimate that could arise from persistence, if there be an overestimate from this source, does not exceed 2.4%. The maximum underestimate of the material, if there be an underestimate, does not exceed 9 parts in 1,000. With respect to the conversion costs, active plus inactive, the maximum overestimate that could arise from persistence, if there be an overestimate from this source, does not exceed 2.6%. The maximum underestimate of the conversion costs, active plus inactive, if there be an underestimate, does not exceed 15%. With respect to the unrealized earnings, the maximum overestimate that could arise from persistence, if there be an overestimate from this source, does not exceed 11.7%. The maximum underestimate of the unrealized earnings, if there be an underestimate, does not exceed 6.6%.

I recommend that you accept the results of the sample as figures whose reliability is objectively evaluated in the statements contained above.

Example 4. This example is a report on the results of a sample whose purpose was to estimate the dollar-value of the inventory of material-in-process of a large manufacturing company. This statement is a legal document, as it will go into the company's financial statement, and is subject to review by the auditors and by any stockholder.

Statement from the cosulting statistician
to the Comptroller of the Company

This statement is predicated on figures and other information furnished to me by your Company, on the assumption that your people followed correctly my sampling procedures. I may point out that the method of counting, the pricing, the extensions, and the verification of the existence of the inventory, including the existence of the materials in process, are outside my province, and I undertake no responsibility on these aspects of the inventory nor for anything other than for the statistical methodology and for the interpretation of the results that you have furnished to me.

The sampling plan that I designed for your inventory provided procedures for (1) the selection of lots for the sample; (2) the formation of an estimate of the aggregate inventory of the materials in process; (3) the calculation of the margin of sampling error in this estimate; (4) a probe of a subsample of the main sample to evaluate some of the nonsampling errors.

I shall deal first with the margin of error of the sampling itself. In my opinion, the results that your Company obtained for the inventory of the materials in process in June 1957 falls within a maximum sampling tolerance of $224,000 in either

direction from what your Company would have obtained had you counted and physically processed every lot of the designated inventory of the materials in process with the same care and with the same degree of skill that you exercised in applying the sampling procedures. The maximum sampling tolerance, $224,000, is 1.9% of $12,098,069, this being the figure that your Company furnished to me for the estimated total regular inventory, including the materials in process and other and additional items.

I turn my attention now to the nonsampling errors, which are dependent on human observation and have not the objectivity of the calculation of a sampling tolerance. The sampling plan contained within itself a systematic probe for the evaluation of certain nonsampling errors, viz: lots missed; wrong count of parts; wrong part number, wrong name for the part; wrong operation-number; missing operation number; mixed parts on one ticket.

The error of sampling, mentioned above, includes the effect of the variable part of the nonsampling errors, such as wrong counts, wrong part number, wrong operation number, mixed parts. It does not include the constant or systematic part of the nonsampling errors, such as a persistent tendency to over-count or to under-count.

The probe for lots missed and for lots counted twice was total. It detected no lot counted twice, and only 2 lots missed, out of the 47,370 or so lots in the regular inventory. This flaw was corrected, so it should lead to no error whatever, and I shall make this assumption.

I have evaluated the other nonsampling errors with the aid of a probability model, with figures furnished by you. The results indicate a possible overestimate. The maximum overestimate, if there be an overestimate, can hardly exceed $58,000. It is possible that there is no overestimate at all, as the probability model gives $1,650 as the limit of any underestimate attributable to the nonsampling errors.

The limits of error from the combination of the sampling and the nonsampling errors are in my opinion a maximum overestimate of $255,300, and a maximum underestimate of $199,000, those figures being respectively 2.1% and 1.7% of $12,098,069. The actual magnitude of the overestimate or of the possible underestimate lies, in my opinion, well inside these two extremes.

RÉSUMÉ

La divisicn entre les contributions de la matière et du statistique est aussi importante pour le dessein du projet échantillonage que la théorie de la probabilité. Au statisticien, comme logicien, tombe la responsibilité pour la specification, aux étages initials, sur la contribution (1) d'expert du matière pour la formulation initial du problem, pour decidir si un frame proposé est satisfactaire, pour le plan pour usage des resultats d'experiment ou d'échantillonage; (2) d'expert de statistique pour plus exacte formulation du problem à la langue statistique; pour la dessein d'experiment ou d'échantillonage, qui compri le choix du frame ou substrate propre à l'investigation; pour les règles pour tirer les unités du frame; le règle pour la formule mathematique pour faire le sommaire des résultats, peut-être à la forme de la moyenne, ou à la forme de la somme, ou d'autre description statistique du frame; aussi la formule pour calculer les écarts-types des sommaires variés, ou pour faire les probes de significance, appropries à la methode de la selection des unites du frame et à la formation des sommaires; pour les controles statistiques avec quel es on découvriria

les fautes non-statistiques et leurs effet vers les résultats; en fin, l'interpretation de les résultats, spécialment au regard de leur qualité et leur fauts statistiques.

Les résultats d'investigation referent au frame, ne point au universe, que est la totalité des unités et des conditions que on desire étudier. Il faut que l'expert de la matière de jeter un point, par jugement, ne point par la théorie de statistique, sur l'écart entre le frame et l'universe.

La decision sur un frame proposé est la responsibilité d'expert du matière. La decision reste à la question si le frame sera satisfactoire pour en échantillonage complet, 100 %.