96

ON STATISTICAL STANDARDS FOR LEGAL EVIDENCE

By

W. Edwards Deming
Graduate School of Business Administration
New York University

ECONOMIC BENEFITS OF STANDARDS. Commerce as we know it today in industrial countries like our own rests on the use of commercial and statistical standards. Thousands of dollars' worth of product change hands every hour in all parts of the world by the use of statistical acceptance plans.

The basis for a statistical standard is statistical theory, good statistical practice, an understanding of the meaning and limitations of statistical calculations, agreement on their meaning, and usage. Statistical theory and practice have arrived at a stage where there is universal agreement on certain practices and interpretations. For example, any two competent statisticians will agree that certain sample designs will give results that possess valid standard errors. They will agree that other procedures will not do so. Moreover, they will agree on the usage and on the amount of knowledge contained in the standard error.

The purpose of this paper is to explain a few points about the meaning and limitations of statistical theory and calculations, in order that statistical standards may render better service to commerce and industry through wider usage in auditing, accounting, in the appraisal of inventory, in commercial standards, in contracts, and in legal evidence.

ACCEPTED STATISTICAL STANDARDS. There are a number of statistical standards in common use, and more are in the process of creation. Each standard exists for a specific purpose. Some of the existing standards are in the list below.

1. Dodge and Romig SAMPLING INSPECTION TABLES, papers in the Bell System Technical Journal, 1926; published in book form by John Wiley, 1944.

2. Military Standard Sampling Procedures and Tables--Mil-Std 105A, 11 Sept. 1950 (Munitions Board, Washington).

3. Recommendations concerning the preparation of reports on sampling surveys, produced by the United Nations' Sub-Commission on Statistical Sampling, 1948.

4. Tentative recommended practice on the sampling of materials (American Society for Testing Materials, 1956).

5. Choice of sample-size (American Society for Testing Materials, 1956).

Another standard, being prepared under the auspices of a society of lawyers and accountants, will state the conditions under which the results of a sample will be acceptable in legal evidence (next section).

I might add that there is really nothing special about presenting results for legal evidence. One should only obey the rules of good science: say what you know, and stop with that. Don't take sides. Stay within your field of competence. This is a good rule for teaching as well as for statistical practice.

WHEN IS THE RESULT OF A SAMPLE ACCEPTABLE? If a sample of 100% of all the units in a frame would be acceptable evidence, then a sample from the same frame is also acceptable evidence, provided the precision of the sample is sufficient for the purpose.

This answer does not tell when a complete coverage (100% sample) would be acceptable, but it does answer a very important question about sampling. The question of whether a complete coverage would be acceptable is not a question that knowledge of the theory of sampling can answer. It is a question in subject-matter (engineering, chemistry, physics, agriculture, medicine, psychology, marketing) and is the responsibility of the man who is paying for the job and expects to use the results.

Confusion between (a) the universe and (b) the frame; (c) the precision of a sample and (d) the proper method of test or of interview, and the usefulness or relevance of the information, have been responsible, I believe, for a great deal of misguided counsel in respect to the uses and misuses of data in legal cases, and in other places also, I fear.

Actually, as is well known, the result of a sample, especially for a study that is difficult, is for many purposes demonstrably more reliable than the result of a complete count would be. The reason for the difference, when it exists, lies in the superior workmanship that is possible in a small sample. However, superior workmanship in either a sample or a complete count does not just happen: it requires knowledge of the material, knowledge of how people perform their duties, knowledge of theory, skilful planning, and directed effort in the training and in the supervision throughout the job.

The margin of sampling error of a prescribed sampling procedure refers to the margin of difference, for a stated probability, between a result of applying this sampling procedure and the result of using an equal complete coverage OF THE SAME FRAME. (Cf. a later section on the equal complete coverage.) High precision of a sample only means that the equal complete coverage of the same frame would give very nearly the same result as the sample. It does not guarantee that the method of test was appropriate, nor that the work was done well, nor that the results are relevant to the problem.

The universe is the material that one wishes to study. The defini-tion of the universe follows directly from a careful statement of the problem, and even more directly from the proposed tabulation plans and uses of the data. In contrast, the totality of the sampling units in the frame constitutes only the material offered for sampling.

There is often a gap between a proposed frame and the universe. The gap consists of material named in the universe but not included in the frame. Sampling can only cover the frame that is presented for sampling. A complete coverage can do no more. The objective inference in regard to the precision of a sample refer only to the frame, and not to the universe, unless the frame covers the whole of the universe.

To be acceptable, the frame whence a sample will be drawn must yield, when the sample is increased to 100% of the sampling units of th frame, a satisfactory portion of the universe that one desires to study Statistical sampling is often very helpful in studies aimed at measurir the size and importance of the gap between the frame and the universe. However, the final decision on whether a proposed frame includes enough of the universe to justify the cost of a sample or of a complete count, either one, belongs to the man who is paying for the job and who expect to use the data. As a statistician, I never make this decision.

There should be no necessity, with the adoption of statistical standards, to present in evidence the theory of sampling, nor even the details of the sample-design that was used. The layman will not under-stand them anyhow. The standard error conveys all the information that there is to convey in respect to the possible difference between the sample and an equal complete coverage. Knowledge about the sampling plan adds no new information. The only relevant evidence is a statemen by a competent statistician (a) that the sampling plan was in fact (or was not) a probability sample of a certain frame, and (b) that probes of the execution of the sampling plan revealed certain departures or no departures from the plan prescribed; (c) that in his opinion these departures will likely cause certain inaccuracies in the results. A statistician's certificate could thus well be a paragraph similar to an accountant's certificate, which refers to standard and accepted procedures, followed by notations of exceptions and their possible effects. The sampling plan and the frames, and all pertinent details, should of course be available for probe if any question arises about the execution of the plan, or about the existence and treatment of rare populations and of extra large sampling units, as these might, if ignored, affect the interpretation of a standard error.

THE FORM OF THE RESULT OF A SAMPLE. The result of a sample will be in the form of a numerical upper (lower) limit that one may accept, with a stated probability, as above (below) the result that would have come from an equal complete coverage of the same frame. The result of a sample is sometimes in the form of both an upper limit and a lower limit that will include, with a stated probability, the result that would have come from an equal complete coverage.

The upper and lower limits are computed from the estimate (typified by $\bar{x}$) and from its standard error. They take forms such as $\bar{x} + ts/\sqrt{n-1}$, or $\bar{x} - ts/\sqrt{n-1}$, or $\bar{x} \pm ts/\sqrt{n-1}$.

The interpretation of the stated probability is the usual one of relative frequency. Any one result, such as $\bar{x} + ts/\sqrt{n-1}$, is one of many thousands that repetitions of the same sampling procedure could produce. Statistical theory enables us to compute the proportion of these results that will fall above or below the result of the equal complete coverage, which I now proceed to define.

THE EQUAL COMPLETE COVERAGE. The result of an equal complete coverage is the result of a 100% sample of the same frame that the sample was drawn from, under the stipulation that the complete coverage be carried out with the same testing or interviewing procedures for eliciting the information, and by the same workers, using the same definitions, exercising the same care, and occurring at about the same time.

The equal complete coverage is a very important concept. You will note that I do not speak of the true value of a result, as I do not know how to define it operationally. It is possible, however, to define operationally in principle and often in actuality the equal complete coverage.

VALIDITY OF A SAMPLE. The result of a sample is valid if it has a standard error. The most important property of modern sampling methods is that the margin of sampling error is calculable from the sample itself. It is the standard error that enables one to compute the margin of sampling error for a stated probability.

Only a probability sample can have a standard error. A probability sample is one that is designed and carried out so that one may use the theory of probability to compute the margin of sampling error in the result.

Knowledge of the nonsampling errors is of course necessary in the uses of the data from either a sample or from a complete coverage. Measurement of the nonsampling errors and the detection of blemishes in the procedure are accomplished through systematic probes of the main sample.

It is a fairly simple matter to design a perfectly valid statistical sample, and to understand the margin of error. It is only necessary to define identifiable sampling units, whether they be large or small, equal or unequal in size, economical or uneconomical, and to allot to them one by one the serial numbers 1, 2, 3, and up; then to draw a sample of them by the use of random numbers. Random No. 17 designates sampling unit No. 17 for the sample. The next step is to process the data to form estimates by following the prescribed procedure or formula whose bias and variance are known. Then, compute the variances and the standard errors of the estimates. Lastly, interpret the statistical significance of the standard errors, such as by the 3-sigma rule, or by a test of significance.

In more detail, a sampling procedure contains 5 parts:

1. The frame, which must be capable of identifying any sampling unit.

2. The procedure for selecting the sample of sampling units from the frame.

3. The formula or procedure for the formation of an estimate.

4. The formula or the procedure for calculating the standard error of an estimate.

5. A definite statistical plan by which to probe the testing or the interviewing, the coverage, and the supervision, to be able to assess quantitatively the nonsampling errors. The standard error, and hence also the margin of sampling error, does not include the errors of incomplete coverage, nor of other persistent failures to do the work properly. These are the nonsampling errors, which would afflict an equal complete count.

If any of the first 4 parts is missing, the sample is not a probability sample. If the 5th part is missing, we are unable to say how the nonsampling errors might affect intended uses of the data. The 5th part is as necessary for a complete count as for a sample, and is included today in carefully designed censuses.

Every item of material that is covered by any portion of the frame must definitely be in one sampling unit or another, before the random numbers select the sample from that portion of the frame. The boundaries of any sampling unit must be definite and unmistakable. No piece of material may have the possibility of being included in more than one sampling unit.

The selection of the sampling units for the sample will invariably be made with random numbers. Every sampling unit shall have a serial number before the random numbers select the sample. These serial numbers need not be stencilled on the units, nor even on a map, but there must be a definite rule that will establish unequivocally the boundary and the serial number of any sampling unit.

The estimate shall be formed by a definite formula or procedure whose bias and standard error are known. It may suffice in some cases, in the calculation of the margin of error, to know only upper limits for the bias and for the standard error. The standard error will depend on the method of selection and on the formula for the estimate.

If these requirements are met, the result and its upper or lower margin of error for any designated probability will then have the standard interpretation, viz., that the outside margin of sampling error is 3 standard errors. However, if the distribution of an estimate is skewed, or if the number of degrees of freedom in the sample is small, the interpretation of the margin of error must be made with the aid of the proper statistical theory.

THE CONFIDENCE THAT ONE MAY PLACE IN A VALID SAMPLE. The confidence that one may place in the interpretation of the standard error was described in masterful words by Professor Leslie Kish of the Survey Research Center in the University of Michigan. He had been employed by the Illinois Commerce Commission to investigate a certain sampling plan which had been introduced in evidence. His testimony read in part as follows: *

> The soundness of the statistical theory which underlies the computations of this average, and of its standard error, as well as the meaning of the confidence intervals formed from them, are universally accepted and incontrovertible. By this I mean that there is no controversy; no personal judgment enters into their meaning; they are well-established mathematical consequences. Every trained statistician will make the same interpretation of the results of those computations. ...
>
> I have made a detailed, independent, and critical examination of the sampling plan, both in its theoretical and practical aspects. I took it to be my duty to find any fault that might make the sampling plan invalid. It is my professional opinion that the sampling plan represents an excellent example of combining theoretical knowledge with sound understanding of the practical requirements to produce a valid, and reliable statistical sampling plan. ...
>
> The examination of the theoretical aspects of a sampling plan and of the formulas connected with it, is a relatively brief assignment for a trained sampler. However, the practical procedures necessary to fit the plan to the intricacies of an actual situation require more detailed, lengthy, and painstaking examination. I have examined the procedures as outlined in company testimony. It is my firm judgment that the sampling procedures, as given, constitute a sound sampling plan which, if followed through, will yield the statistic mentioned above. Secondly, it is my firm opinion that the procedures outlined constitute an eminently practical vehicle for putting the plan into practice. They are clear; they are complete and comprehensive; and they are tailored to the situation.

---

* Illinois Commerce Commission: Illinois Bell Telephone Company, in the matter of the proposed advance in rates ... , Docket 41066, 17 August 1954.

VALIDITY CONTRASTED WITH EFFICIENCY. Although it is fairly simple to achieve validity, it is an entirely different story to construct an economical sampling plan; that is, one that will deliver a prescribed precision with economy, and to fit theory to practice so that the actual work moves smoothly. To do this requires skill in statistical theory, experience, and maturity.

A sampling plan may be valid, yet inefficient. A sampling plan may be valid, but useless. It may be efficient but useless. It may be precise but useless. It may be a perfect 100% sample, but useless.

Statistical theory does not state the problem nor tell us what kinds of information might be relevant or admissible; nor does it tell us the methods, definitions, and techniques by which information is to be elicited from people, business establishments, or records, nor does it prescribe methods for testing industrial product. Such questions are substantive. They are the same whether the survey is to be conducted by a sample, large or small, or by a complete coverage. Judgment on them must come from experts in the subject-matter and from experts in the techniques of interviewing and testing; not from statistical theory. Statistical theory is nevertheless essential if one would estimate reliably and economically by trial the numerical differences between the results to be expected from different methods, definitions, and techniques, or between two sets of inspectors or interviewers. It is statistical theory and not knowledge of the subject-matter that discovered the importance of interactions, and other astounding facts that have extended substantive knowledge in many lines.