A SAMPLING STANDARD UNDERGOES DEVELOPMENT: REFLECTIONS ON ASTM-E141-91

C. H. Proctor, North Carolina State University
Department of Statistics, NCSU, Raleigh, NC 27695-8203

KEY WORDS: Audit subsample, replicated subsample, skewness rule, sample size

Basically, standards say "Do the thing this way." "The thing" for ASTM (American Society for Testing and Materials) is making measurements and also, in the case of statistical standards, processing and analyzing data. The measurements are for use in engineering, science and commerce. Andrew Carnegie was a founder of the society. His concern was to sell steel without arguing over whose measurements were right.

The standards can, at times, appear complex and arcane but they have been designed with sufficient care to settle disputes. A dispute may be entirely friendly, as between two branches of the same company needing to agree on the quality of some material to be transferred between them, or it may be bitter, as between two parties facing one another in court with a money amount to be settled by a sample survey or by a test result.

Standards are nothing but empty platitudes until the parties agree to follow them, or until some court or commission invokes them. One feature of ASTM and ISO (International Standards Organization) standards that has been advanced as favorable to their widespread acceptance is that they are consensus standards. This means that any published standard has been voted on and approved a number of times with no dissenting votes. If there were negative votes, changes were made and the vote withdrawn or, in a very rare case, a negative vote can be voted to be nonpersuasive.

One might liken this process to legislating. The amount of effort, and even emotion, expended in ASTM committees over the choice of words and, above all, of definitions enhances the resemblance to legislating. The committee members, however, are by and large self-selected, even though they view themselves as representing various constituencies. Perhaps a telling feature is that the intent of the written standards is to avoid having to settle a dispute by confrontation, but rather one follows the guidelines and thereby reaches the correct resolution. In legalistic terms, standards can be invoked during discovery so that matters of fact can be settled before, rather than during, trial.

Standards are also published material. They can be purchased as other written material in the tradebook marketplace. Since their content is seldom fresh and original, nor their style free and bold, they don't usually become popular. On the other hand, each committee usually has some literary talent and the editorial services of ASTM are considerable, so the output is very respectable -- certainly as technical, if not popular, literature.

Perhaps enough has been said about standards in general, but we'll just mention a few special characteristics of statistical standards. These are generally written in the, so called, cookbook style with copious numerical examples. A main aim is that the recommended or required method be workable. This is almost as important as reducing bias and is easily as crucial as reducing variance. The method must also be designed to be widely applicable and to be computationally transparent -- no "black box" software need apply.

The standard under consideration, E141, may have had, at its origin in 1959, a more philosophical and less arithmetic flavor than other statistical standards, but the recent revision we will be considering conforms to the more traditional style -- terse, realistic and specific. As its title ("Acceptance of Evidence Based on the Result of Probability Sampling") expresses, the standard is designed for critically reviewing survey results based on a probability sample. The emphasis on an "equal complete coverage result" as the objective of the survey (as well as on other basic concepts) marks the standard as, at least in great part, the product of W. Edwards Deming, a master expositor of clear and concise statistical methodology.

In addition to the concept of equal complete coverage others of Dr. Demings' principles that form part of E141 include the audit subsample (to check for gross departures from procedures) and the use of replicate subsamples (to allow variance estimation). One can, however, read in this revision in all three instances, slight concessions to recent developments in statistics. There is now explicit mention of a "target parameter or ideal goal" beyond the "equal complete coverage result" and numerical comparison

of the two was invited, but was not specifically illustrated. Suggested sizes of audit subsample are now mentioned with calibration as an additional option when 30 or more sampling units are to be audited. The arithmetic (although not the full Tukey Jackknife) of variance calculations from replicate subsamples has been included and the fpc deemphasized. One can note in these three examples the intrusion of "model-based" considerations, essentially measurement model ones, into the original "design-based" or enumerative approach. If future committees find themselves tempted to introduce the stochastic process underlying the variable of interest, they would do well to proceed cautiously. Although such a process is an essential part of any effort to find assignable causes in process control, it does seem irrelevant to the enumerative setting of E141.

After illustrating the arithmetic for getting $se(\hat{\theta})$, the standard error of the estimate $\hat{\theta}$, the standard calls for the report about θ to be in the form "$\hat{\theta}$ with a standard error of $se(\hat{\theta})$ on ν degrees of freedom." There then follows definitions of "bounds" and of "confidence limits" that employ $se(\hat{\theta})$ along with t-values. I should confess that I would prefer the discussion of confidence limits be omitted entirely. In most cases where an estimator of θ is needed to settle a dispute, the availability of a confidence interval allows one party to grab the lower limit and the other the upper limit, and they continue to battle. At least in this presentation there is no advocacy of interval estimates.

There is considerable space devoted to setting an upper bound on a population proportion when zero cases are observed in a sample. This is one instance where an upper bound is of practical importance. The calculations are illustrated both for a large population, as well as for a moderate-sized finite population where hypergeometric probabilities need to be calculated. The standard illustrates the use of a hand-held calculator for this task and likely overdoes the arithmetic. The method of setting the bound uses half-integer numbers of elements having the attribute in the population and thus avoids quibbles caused by allowing "or equal to" in the definition of the bound.

Another arithmetic method is illustrated for resolving the question of skewness in the estimate. Rather than repeating the old admonition about "check for skewness," the standard now prescribes that the estimated sampling skewness coefficient not be greater in absolute value than 0.3. This corresponds roughly to replacing the 25 in Cochran's "$n > > 25G_1^2$" rule by 10. Reducing this action limit from 25 to 10 is, in part, due to sampling uncertainty but I would hope it does not raise too many false alarms over skewness.

Another case where vague admonition has been replaced by data-based arithmetic is in re-negotiating sample size. After noting some provisos (one is that more observations can be obtained and another is that there has been no-peeking-at-$\hat{\theta}$-but-only-at-$se(\hat{\theta})$), the standard suggests comparing the loss averaged over going one $se(\hat{\theta})$ above, and one $se(\hat{\theta})$ below, θ to the cost of quadrupling sample size. Roughly speaking, if the average loss exceeds the cost of quadrupling sample size then one should seriously consider increasing sample size. Nonlinearities in the loss function would seem to be offset by difficult-to-judge (overhead?) survey costs and the rule thus has a fair chance of working.

In describing its scope the E141 standard rather piously states: "One purpose ... is to describe straightforward sample selection and data collection procedures so that courts, commissions, etc. will be able to verify whether such procedures have been applied." Although the standard has been around for 30 years it appears not to have penetrated very far into legal circles. I do not find it mentioned in recent books on the use of statistics in the courts. I can't say this is surprising. When I worked as statistician on North Carolina's case for "In re Antibiotic Antitrust Litigation, 410 F. Supp. 669 (D. Minn. 1974)" (see Fienberg, 1989), I wasn't aware of its existence but I wish I had been. The revision now includes data from that case as examples.

The Fienberg (1989) report does cite the ASTM standard E678-80 on "Evaluation of Technical Data." Such a citation provides recognition that ASTM standards can play a role. The Fienberg (1989) report also provides (in an Appendix) "Recommended Standards on Disclosure of Procedures Used for Statistical Studies to Collect Data Submitted in Evidence in Legal Cases," which predictably has some overlap with E141. The differences between the two further accentuate E141's specificity and use of arithmetic examples.

To enable your survey results to stand up in court it may be helpful to consult the standard, or if you wish to critique someone else's survey procedures, here is the guide to use.

References

ASTM (1991). **Annual Book of ASTM Standards, Vol. 14.02,** "Standard Practice for Acceptance of Evidence Based on the Results of Probability Sampling, E141-91," Philadelphia, PA.

Fienberg, S. E., editor (1989). **The Evolving Role of Statistical Assessment as Evidence in the Courts,** Springer-Verlag, NY.