LAW360

Portfolio Media. Inc. | 860 Broadway, 6th Floor | New York, NY 10003 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

How To Get The Most Out Of Your Statistical Sampler

Law360, New York (October 23, 2012, 3:56 PM ET) -- Statistical sampling is a powerful tool and a recognized statistical specialty. While many people are capable of drawing simple samples, it often requires different and more advanced expertise to design the types of complex samples needed for litigation. This article is intended to serve as a guide to assist attorneys who are working with their statistical sampling experts.

The sampling process includes three basic steps: (1) designing the sampling plan, (2) selecting the sample and (3) calculating standard errors and confidence intervals for the resulting estimates. The confidence interval is the range of possible values above and below the sample estimate in which the true but unobserved population value is likely to fall.

A scientifically acceptable sample has two key features. The first is that every population element has a known, non-zero probability of selection. The probabilities need not be equal, but they must be known. The second is that the method of estimating the standard errors is consistent with the sample design. Different sampling plans require different methods for estimating the standard errors. Achieving the consistency between selection and estimation method is the core element of the expertise that a statistician brings to a sampling project.

Statisticians are trained to make technical decisions. They can stratify a sample, decide whether or not to use “dollar-based sampling,” choose to sample with or without replacement and calculate the optimal sampling rates in the various strata. The expert statistician can resolve these technical issues in a noncontroversial manner that will maximize the precision of the sample estimate.

However, there are related issues that go beyond the technical expertise of the sampler. These include deciding just how accurate the estimates need to be, working around practical data problems and getting the information needed to create an optimal sampling plan. The statistician and client must work together to resolve these issues. The goal of this article is to describe the statistician’s technical expertise and to suggest ways that attorneys can work with their experts to solve the types of “real world” problems that are likely to arise in any sampling case.

Deciding How Accurate the Estimates Should Be

The issue of sampling is that the sample estimate will almost surely be unequal to the population value even if the sample design is excellent. Taking cost constraints into account, the expert statistician can increase the chances that the sample estimate and population value will be “close,” but will need the client’s guidance to decide how close is “close enough.”

Time and budget constraints limit the sample size, and a smaller sample size reduces the precision of the sample estimate. To illustrate, let us assume that the litigation concerns a set of insurance claims whose collective value is $100 million.

The attorney's client believes that a large number of claims, guessed to be worth $30 million, were processed improperly. To keep the study costs reasonable, the statistician has designed the sample such that the anticipated margin of error is $5 million. This means that we expect the 95 percent confidence interval for the value of the improperly processed claims to extend from $25 million to $35 million. Is the accuracy good enough? Statistical principles will not provide an answer and the attorney and client must help to define the desired level of accuracy.

In addition, the only way to narrow the confidence interval is to increase the sample size and this can be expensive. To reduce the margin of error to $2.5 million, the sample would need to be four times as large. Is it worth multiplying the costs by four to narrow the anticipated confidence interval to between $27.5 and $32.5 million? The statistician cannot decide this alone, but (s)he can explain the risks and benefits inherent in different sample size choices. There is no "statistically correct" level of error, and the statistician, client and attorney must work together to determine the best compromise between the goals of an accurate estimate and lower costs.

Perhaps the key element in making this decision is the client's confidence that the share of "bad claims" will be high. If the sample estimate is $30 million and the margin of error $5 million, then the entire confidence interval will indicate a large damages estimate. If, however, the sample estimate is only $10 million and the margin of error is $5 million, the confidence interval will extend from $5 to $15 million, and it will be unclear whether the overall damages are small or large.

Missing and Incomplete Files

Returning to the example, let us assume that the sample size is 500, but for 100 selected claims, 20 percent of the sample, the claim files are either missing or are so incomplete to be unusable. How do we generalize from the 400 available and complete files to the 100 unusable files? The two sets are likely to differ in many ways. Factors making it more likely that the files are missing or incomplete may also be factors making it more likely that the procedures were improper. The correlation creates a bias.

To illustrate, among the 400 usable files, assume that 100 of the claims are complex, and 90 of these (90 percent) were improperly processed. Among the 300 remaining applicants with simpler claims, 150 (50 percent) were improperly processed. Overall, 60 percent of the usable sample of claims was improperly processed.

Turning to the 100 unusable files, we know that 75 are complex and just 25 are simpler. We can use this information to adjust the estimate by assuming that 90 percent of the more complex unobserved files and 50 percent of the simpler unobserved files were improperly processed. We would then calculate an assumed estimate for the unobserved files by multiplying, i.e., (.75 * .90 + .25 *.50 =) 80 percent. This is considerably higher than the 60 percent rate for the observed files.

This estimate is only as good as the assumptions it is based on. The statistician, without substantive expertise regarding the complexity of claims and the use of improper procedures, is not in a position to justify the assumptions. The client or a subject matter expert needs to provide the substantive expertise to justify the assumptions.

Deciding on the Sample Size

You might think that the decision about the sample size would be purely statistical, but even this requires nontechnical assumptions. Let us return to the problem of the insurance claims. Before retaining the statistician, the attorney and the client together believed that as much as 30 percent of the claim value was processed incorrectly, but there was also a good chance that the share was as low as 10 percent. The total value of all claims is $100 million. The statistician and the attorney decide to start with a sample size for which the likely confidence interval is plus or minus $5 million.

Assuming first that the share of “bad claims” is 30 percent, the statistician would expect the sample estimate of the “bad claim amount” to be $30 million and the 95 percent confidence interval to extend from $25 to $35 million. This confidence interval indicates that the true but unobserved total is likely to be within $5 million, or 17 percent of the sample estimate, $30 million. If we were to double the sample size, the anticipated confidence interval would narrow to $26.5 to $33.5 million, and the true but unobserved total would probably be within 12 percent of the sample estimate. The likely gain from doubling the sample size is possibly not worth the added expense.

On the other hand, if the share of bad loans was only 10 percent, we expect the best estimate to be $10 million and the 95 percent confidence interval to extend from $5 to $15 million. The true but unobserved total is likely to be within 50 percent of the best estimate, $10 million. The upper limit of the confidence interval is three times as large as the lower limit, and the precision of the sample estimate is less likely to be satisfactory than it was under the “30 percent” assumption.

If we doubled the sample size under the “10 percent” assumption, the 95 percent confidence interval would narrow to $6.5 to $13.5 million, and we would be 95 percent certain that the true but unobserved value is within 35 percent of the sample estimate. Expanding the sample seems to create more improvement if the rate of bad claims is 10 percent than it does when the rate is 30 percent.

The statistician, as a technical but not substantive expert, cannot choose between the 10 and 30 percent assumptions. It is a job for either the client or a substantive expert to help the statistician decide which assumption is more realistic.

Conclusion

The expertise of the sampling statistician is deep, but may be narrow with regard to the substantive areas for which expertise is also needed. Good sampling statisticians know how to work in fields where they have little or no previous experience by asking the right questions to set up their samples. They need to rely on the substantive knowledge that someone else, usually a subject matter expert or a combination of the attorney and the client brings to the table. The answers to these questions will allow the sampling statistician to make assumptions that reflect the subject matter reality.

With these assumptions in hand (s)he can design a valid sample and properly estimate the standard errors and confidence intervals. A good sample design therefore results from a cooperative enterprise combining the technical knowledge of the statistician with the subject matter expertise of others.

--By Dr. Eugene P. Ericksen, NERA Economic Consulting

Dr. Eugene Ericksen is a special consultant with NERA Economic Consulting in the firm's Philadelphia office.