LJN LAW JOURNAL NEWSLETTERS

LJN'S LEGAL TECH Newsletter®

An ALM Publication

Volume 30, Number 11 • March 2014

Statistical Sampling

Saves Time and Money and Enables Proportionality

By Howard Edson and Dean Gaida

The focus on proportionality in high-profile cases such as *Apple v. Samsung* (N.D. Cal. Aug. 14, 2013), coupled with the recent proposed amendments to the Federal Rules of Civil Procedure (FRCP) are driving attorneys to reevaluate the methods by which they uncover crucial electronic evidence for a case.

The proposed changes to the FRCP focus on proportionality, putting the burden on counsel to both maintain proportionality and ensure the thoroughness of an e-discovery effort. Statistical sampling, while not necessarily new, is emerging as a reliable method for addressing the new e-discovery standards the courts may soon be enforcing.

Sampling allows for statistically-valid inferences about a large set of documents, while requiring only a small subset of those documents to be measured, or reviewed. A sample must be an "unbiased" representative of the population. One widely accepted technique is simple random sampling, which as the term suggests, takes a document sample at random from the population.

Sampling in Practice

For example, suppose in a matter involving one million documents, the team wants to know what percentage will be relevant so they can determine the time and cost required to find those relevant documents. They would like a confidence level of 90%, and a confidence interval of +/- 2%. They review a random sample of 1,700 documents, and determine that 12% of the sample is relevant. Although the exact percentage of relevant documents in the population cannot be known precisely, the team can, through the power of statistical sampling, say with 90% confidence that it lays between 10% and 14% — that is, they can expect to find between 100,000 and 140,000 relevant documents in the population.

Any other observation made about the sample could be extrapolated to the population with the same statistical confidence, such as the number of foreign language documents or number of privilege documents. For example, if they observed that 4% of the sample documents contained no text and thus needed manual review, they could say with 90% confidence that the exact percentage of documents in the population containing no text lies between 2% and 6%. Having this knowledge makes it much easier to develop case strategy and move forward with the next steps of e-discovery.

There are a few scenarios in which sampling can streamline e-discovery efforts, thus aiding in managing proportionality. These include early case assessment, quality control for human review and validation for predictive coding workflows.

The Benefits

Some of the benefits of sampling follow.

Early Case Assessment

Statistical sampling at the outset of a case can help attorneys quickly and cost-effectively measure key characteristics of a large dataset, such as the number or percentage of relevant documents ("richness" or "prevalence"), and the amount of work, including the estimated time and cost, likely required to find those documents. These are important attributes for any discussion around proportionality. By evaluating the sample, attorneys can gain insight on

Howard Edson is a managing director in FTI Technology Research and Development, based in Seattle. He specializes in Web software product management and business intelligence. **Dean Gaida** is a managing director with the FTI Technology practice based in New York. Gaida's areas of expertise include computer forensic acquisition and analysis methodologies, data mining/database analysis, and review workflow strategy development.

the preferred culling strategy and whether or not the use of predictive coding would be reasonable for the matter at hand.

Quality Control (QC) During Review

During the first legal review on a case, contract attorneys often work through the documents and end the project with a deeper understanding of the case than at the beginning. If samples are taken from the responsive and non-responsive data populations and re-reviewed by experts, attorneys can confirm that things are moving forward according to the parameters of the case. Sampling can be used periodically throughout the review to determine how many documents were coded incorrectly to extrapolate the overall coding accuracy of the review team. This helps guard against "review drift," and enables the evaluation of human coding decisions to identify training opportunities where mistakes were made.

This technique can provide insight into how effective coding decisions were made on a particular section of documents, and reviewers' abilities early on in a case before they get too deep into the review. It can also determine which reviewers on the team are the strongest or weakest. All of these steps can be done quickly and with minimal cost to add to the overall defensibility of the e-discovery process.

For example, a QC sample of 400 documents could be taken from a review population of 30,000 documents initially reviewed. If the secondary QC review of that sample found 92% of the documents were coded correctly, then it could be inferred with 95% confidence that between 87% and 97% of the 30,000 documents have been assigned the most appropriate coding. This sample sizes provides a 95% confidence +/-5% confidence interval relating to how the review team is doing.

Foundation for Predictive Coding

Just as with human review, statistical sampling provides a reliable and cost-effective means to measure the accuracy of a predictive coding review. Models must be validated against a random sample to appropriately measure the recall and precision rates delivered by the predictive coding model (recall measures how well a process retrieves relevant documents; precision measures how well a process retrieves *only* relevant documents).

The entire process can hinge on what is found by applying the model to the validation set, giving attorneys and the courts peace of mind that the results provided by predictive coding are accurate and thorough. For recall in particular, it is important to understand that as a measurement of the percentage of *positive* documents located, the confidence level is based on the validation set's quantity of available positive documents vs. simply the overall size of the sample taken. For example, with a validation sample taken of 4,000 documents, when measuring prevalence one may expect to easily achieve a confidence of 99% (+/-2%). However, if the validation set only contains 660 *positive* documents, the confidence interval *for recall* will need to be adjusted to +/-5.

A Proper Algorithm

With the above points in mind for how to leverage sampling tools, it is important to understand the most critical functionality to expect of sampling software. First and foremost, a simple random sampling algorithm must assure that each document in the population has an equal chance of being selected for the sample. When evaluating options for sampling software, be sure to test that the algorithm is both effective in providing a random sample of the parent population and computationally efficient. This can help ensure that it can scale to sample large datasets quickly.

Further, users should have the option to lock samples so that their integrity is not compromised during the course of the discovery. If the set of documents in either a sample or its population are changed after the sample is taken, then sampling bias can be introduced, undermining statistical validity. When working with samples it is important to have the ability to both create and delete samples, and also to resize them upward (or downward) when greater statistical confidence is required. When training a predictive coding model, this also allows a small training set to begin with, while enabling the user to iteratively increase the size until the model achieves acceptable recall and precision.

Conclusion

As outlined above, sampling is an important e-discovery tactic that addresses key issues with proportionality, defensibility and predictive coding. It is one of the primary ways attorneys can quickly and reliably understand the data they are dealing with on a case. When applied properly, it is a defensible process and has generally wide acceptance from the courts. Just as importantly, it can help give validation that the e-discovery process is thorough, accurate and sound.