*Manuscript submitted to the Organizing Committee of the **Fourth DESI Workshop on Setting Standards for Electronically Stored Information in Discovery Proceedings** on April 20, 2011. Updated May 18, 2011.*

Application of Simple Random Sampling[1] (SRS) in eDiscovery

Doug Stewart
Daegis

Abstract

eDiscovery thought leadership organizations advocate for the use of sampling throughout much of the EDRM process. Additionally, judging from the numerous and frequent references to "sampling" found in the eDiscovery literature and online content, there appears to be wide acceptance of the use of these techniques to validate eDiscovery efforts. At the same time, there are lingering questions and concerns about the appropriateness of applying random sampling techniques to eDiscovery data sets. This paper offers evidence that random sampling of eDiscovery data sets yields results consistent with well established statistical principles. It shows that Simple Random Sampling (SRS) can be used to accurately make predictions about the composition of eDiscovery data sets and thus validate eDiscovery processes.

Introduction

Sampling is often mentioned as the principal method of validating many eDiscovery activities and decisions. Thought leadership organizations such as The Sedona Conference, EDRM and TREC Legal Track have published guides, protocols and reports that explicitly call for the use of sampling techniques in various eDiscovery processes[2]. Also "sampling" is frequently mentioned in the literature, at conferences and in various forms of online content[3] as a key tool for validating results of collection, search, document review and other technology assisted eDiscovery activities. Further, the courts have called for the use of sampling in the eDiscovery process[4].

Despite these strong endorsements, there appears to be some reluctance or inertia toward the adoption and integration of sampling methods into the eDiscovery workflow. To some extent this reluctance may be based on a lack of understanding as most lawyers do not receive training in statistical principles. Lack of understanding may also contribute to the lingering doubts about the suitability of using Simple Random Sampling (SRS) techniques in the eDiscovery process. Additional education and training focused on applying sampling techniques in the eDiscovery process should drive adoption and acceptance of these methods. The Sedona Conference, EDRM and others[5] recognize this need and have provided leadership and advocacy in this area. Additionally, simple demonstrations that these techniques work may prove to be one of the best ways to dispel some of the concerns.

---

[1] A sampling technique where every document in the population has an equal chance of being selected.
[2] See http://www.thesedonaconference.org/content/miscFiles/Achieving_Quality.pdf; http://edrm.net/resources/guides/edrm-search-guide; and http://trec-legal.umiacs.umd.edu/LegalOverview09.pdf
[3] For example, "Using Predictive Coding – What's in the Black Box?" K. Schieneman et al. http://www.esibytes.com/?p=1649
[4] Victor Stanley, Inc. v. Creative Pipe, Inc., 2008 WL 2221841 (D. Md. May 29, 2008).
[5] "Sampling for Dummies: Applying Measurement Techniques in eDiscovery" Webinar by M. Grossman and G. Cormack 01/27/2011

This study sets out to test the efficacy and applicability of SRS techniques to the eDiscovery process. In doing so, it guides the reader through the process of applying sampling methods on eDiscovery data sets. Several sampling methods are described and tested. Additionally, the key parameters including sample size, confidence level and confidence interval are discussed and measured.

Methods and Material

The metadata of six inactive eDiscovery databases was searched and sampled for the purposes of this study. The databases ranged in size from a few thousand to more than a million records. Various fields including author, custodian, date, file type, and responsive were searched and sampled using the following four sampling techniques:

1. **Simple Random Sampling:** Random sample sets created by randomly selecting records from the specified population using the Microsoft .NET 3.5 Random Class to generate random record sets. Required sample size was one of the input parameters.
2. **Systematic Sampling:** Random sample sets created by selecting every n^{th} record from the specified population using a t-SQL script. A calculation was performed to determine the required value of n to produce the appropriate sample size.
3. **MD5 Hash Value Sampling:** Random sample sets created by running a MS SQL Server query to select all records with MD5 hash values beginning with two designated characters (e.g., AF or 4A). This method was used to produce a random sampling of $1/256^{th}$ of the population.
4. **Non-Random Sampling:** Non-random sample sets created by running a search for documents that fell within a certain date range. Not to be confused with a weighted sample.

The key parameters used to create the random samples for this study included:

1. **Confidence Interval:** Also called the "margin of error", the Confidence Interval indicates the precision of the sample's estimate by providing upper and lower limits on the estimate (e.g., plus or minus 2%).
2. **Confidence Level:** An indication of how certain one can be about the results. A 95% confidence level means that 95 times out of 100 the estimate will reflect the population's composition within the margin of error provided by the Confidence Interval.
3. **Sample Size:** Determined by using a sample size calculator. Required inputs include the desired Confidence Level and the desired Confidence Interval. The Sample Size is related to the Population Size but does not scale linearly. For example, the required Sample Size needed to achieve a 95% confidence level with a +/-2 % confidence interval is shown below for a variety of Population Sizes:

Population	Sample Size
1,000	706
10,000	1,936
100,000	2,345
1,000,000	2,395
10,000,000	2,400

4. **Population or Population Size:** The total number of documents in the source data set.

5. **Percentage or Prevalence:** The percentage of documents in the population that have the property being measured (e.g., percentage of the documents that are responsive). If the value is known it can be used to fine tune the Confidence Interval. If not known then 50% must be used to provide the most accurate estimates.

Sample sizes, confidence levels and confidence intervals were calculated using the sample size calculator found at:

http://www.surveysystem.com/sscalc.htm

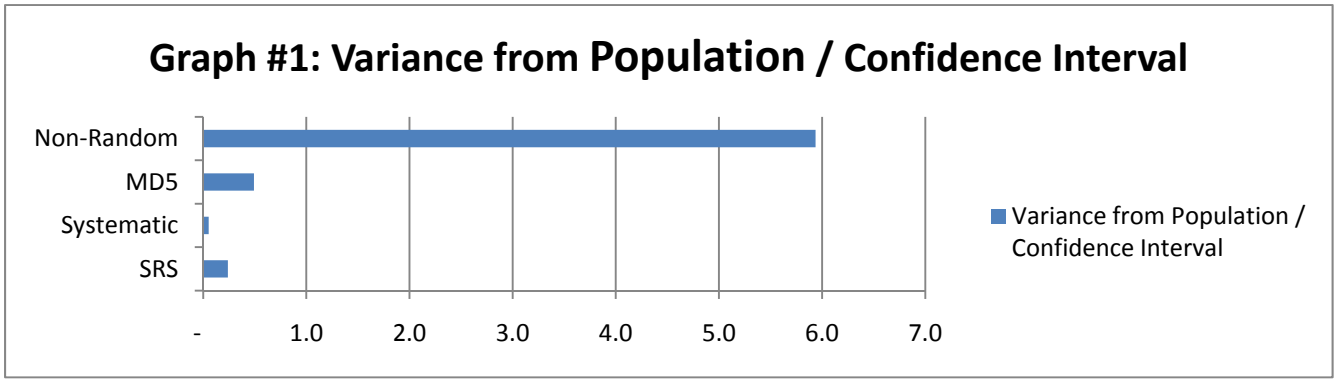
All analysis work was done using Microsoft Excel 2007.

Results

Graph #1: This graph shows the relative precision of each sampling method based on a single iteration of each. It shows how well the sampling techniques performed relative to each other. The precision is represented by the ratio of the absolute value of the sample's variance from the overall population for the property under investigation divided by the sample's confidence interval (or margin of error) as determined by using the sample size calculator. For instance, if the property under investigation were "ABC = Yes" the precision ratio would be calculated as follows:

Precision = abs((% of ABC = Yes in sample) – (% of ABC = Yes in population))/Sample's confidence interval

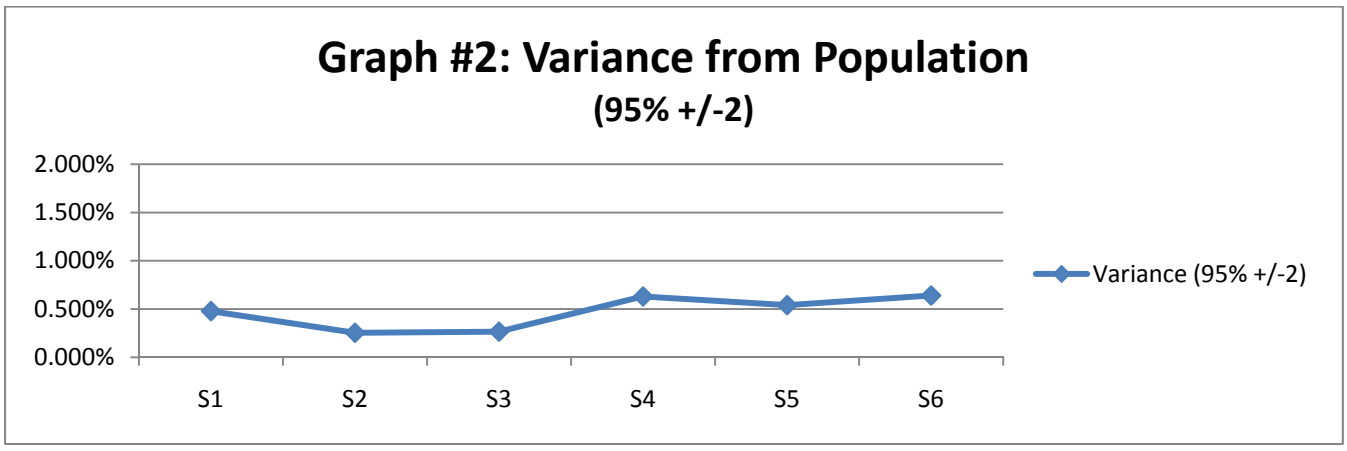
A result of 1 or less indicates the results fell within the confidence interval and thus indicates a sample that conforms to the principles of SRS and accurately characterizes the entire population. A result greater than 1 indicates a sample that does not accurately estimate the population. For example, precision score of 0.50 indicates the sample estimate varied from the actual population by half of the margin of error or confidence interval. A score of 5.0 indicates the sample estimate exceeded the margin of error by a factor of 5.



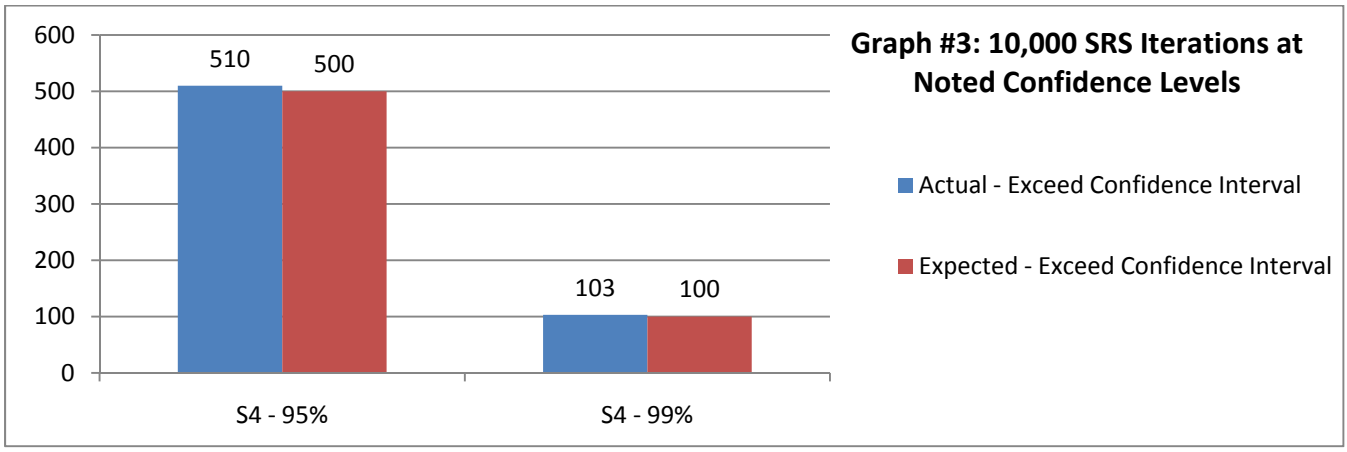
Graph #2: This graph shows the variance of the SRS derived sample from the population for six different eDiscovery databases. The sample size calculator was used to determine sample sizes based on a 95% confidence level and +/-2% confidence interval. The property analyzed was responsive (yes/no) that had been assigned in the review phase of each project's lifecycle. The variance was calculated as follows:

Variance = abs((% of Responsive = Yes in sample) – (% of Responsive = Yes in population))

The data sets (S1 to S6) ranged in size from approximately 4,000 to 1,400,000 records. The property under investigation ranged from an approximate 2% prevalence in the population to over 85% prevalence. The experimental data easily fit within the allowable margin of error.

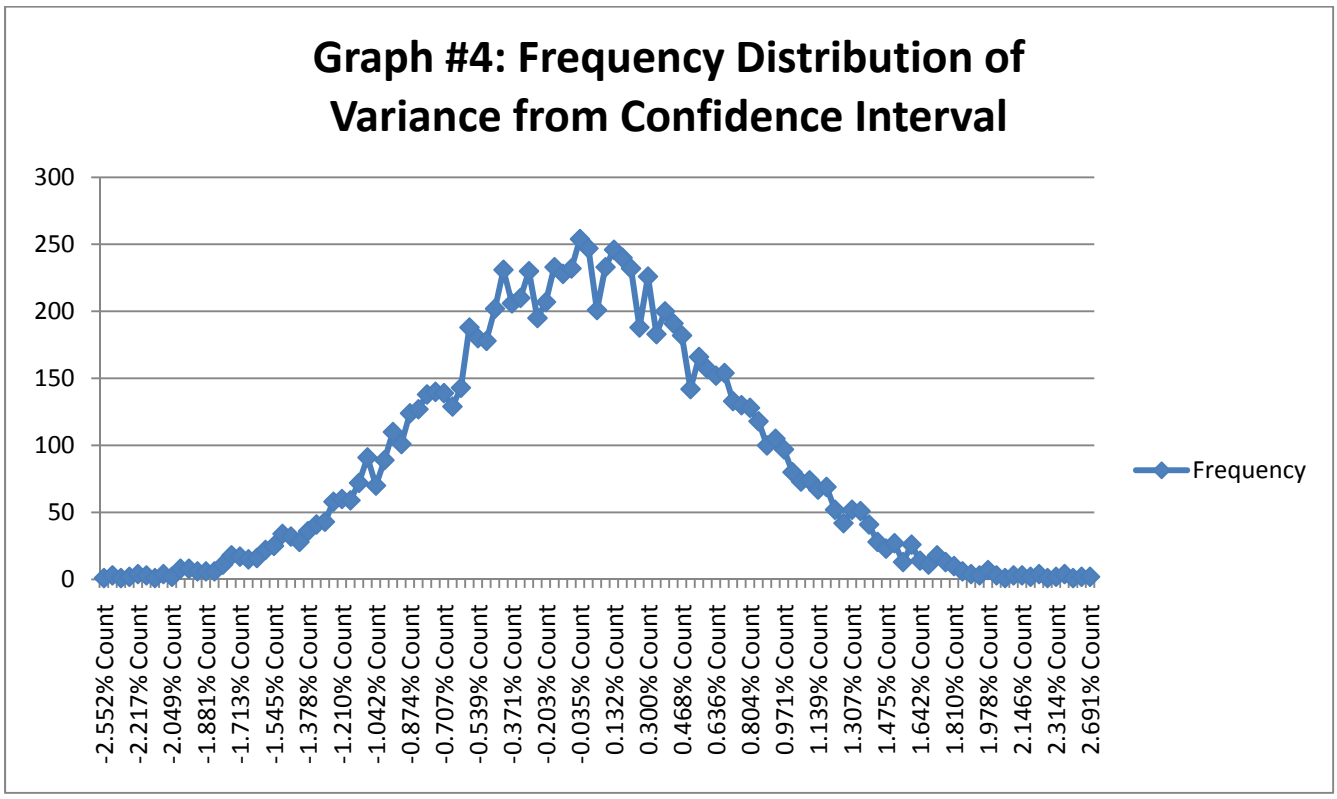


Graph #3: This graph shows the results of running 10,000 iterations of SRS on a single database two times and counting the number of samples that exceeded the confidence interval. The sample size calculator was used to calculate the confidence interval based on a specified sample size, confidence level and the known prevalence (percentage) of the record property under investigation. A confidence level of 95% predicts that 9,500 samples out of the 10,000 analyzed would produce an estimated prevalence that matched that of the population within the confidence interval range —500 (5%) samples would estimate a prevalence that fell outside the calculated confidence interval. A confidence level of 99% predicts that 9,900 samples out of the 10,000 analyzed would produce an estimated prevalence that matched that of the population within the confidence interval range—100 (1%) samples would estimate a prevalence that fell outside the calculated confidence interval. The experimental data match the SRS predictions with extraordinary accuracy.



Graph #4: This graph shows the results of running 10,000 iterations of SRS on a single database and then plotting the frequency distribution of each sample's percentage variance from the population. The sample size calculator was used to calculate the sample size based on the desired confidence level and confidence interval.

The data reveal that the distribution of the all the sample estimates centers on the actual prevalence percentage found in the population and then trails off as one moves out from the center as is predicted by SRS. As a result, this graph conforms to a normal distribution.



Discussion

The data represented in Graph #1 agree with established statistical principles and support the common assumption that random sampling techniques create samples that make more precise estimates or predictions about populations as a whole than non-random sampling techniques. In this study the non-random sample varied from the population by nearly six times the expected confidence interval or margin of error. The randomly generated samples all fell within the expected confidence interval.

Graph #2 demonstrates that SRS methods can be used across a variety of eDiscovery data sets to make predictions about the full population that fall within the calculated confidence intervals. The results shown indicate that regardless of the population size the SRS techniques were able to accurately estimate the population to within roughly 0.5 percent. The consistency in the accuracy of the estimates is even more astonishing when one considers that the prevalence of the property in question ranged from just over 2% to over 85% prevalence in the six data sets and the data sets themselves ranged in size from approximately 4,000 to 1,400,000 documents.

Graph #3 indicates that SRS of eDiscovery databases will produce results that fall within the calculated confidence levels and confidence intervals. The confidence levels are supported by the iteration data with remarkable accuracy—out of 10,000 iterations the results varied by only 10 samples and three samples from what was predicted by SRS.

The normal distribution seen in Graph #4 strongly suggests that SRS of eDiscovery data sets produces results that adhere to the well established statistical principles and body of knowledge. Specifically, the variance from the population for the 10,000 samples follows the distribution predicted by the Central Limit Theorem[6].

Conclusions

The prevailing assumption that SRS, when applied to eDiscovery data sets, produces results in line with accepted statistical principles is supported. This study provides compelling empirical evidence that supports the widely held belief that SRS is one of the best means of validating search and other eDiscovery activities.

The fact that a sample of fewer than 2,400 records from a population of one million can be used to accurately estimate the population as a whole may defy intuition. The best way to get comfortable with SRS is to employ the techniques and test them. Firsthand experience seems to be the best teacher.

Future work should include the creation of protocols and standards for further incorporating SRS methods into the eDiscovery workflow. This effort should also include standardized protocols for reporting on the sampling methods employed and the results obtained to ensure transparency in the process. Standardized protocols for the use of sampling techniques may also serve to educate and familiarize those that may have gaps in their understanding of these established techniques.

Sampling will play an increasingly important role in the eDiscovery process as the industry continues to mature, as data volumes continue to rise and as technology continues to advance. As such, the eDiscovery industry and thought leadership should continue their educational and training efforts to ensure that the relevant segment of the legal community is comfortable with the application of these techniques. Transparency in process, standardization, further training and practical demonstrations of how well sampling techniques work will go a long way toward achieving this goal.

Doug Stewart has over 25 years of IT, security and management expertise in the field of electronic discovery and litigation support. As Daegis' Director of Technology, Doug has been instrumental in the development and deployment of Daegis' eDiscovery Platform, which includes functionality for hosted review, on-site deployment, iterative search and much more. In 2009, Doug oversaw Daegis' ISO 27001 Certification for information security management, which includes a rigorous annual audit process. In addition, Doug manages several departments at Daegis including IT, data collection, and information security.

[6] The Central Limit Theorem states that as the sample size increases, the sample means tend to follow a normal distribution.