Random Sampler: Technical Description

The random sampler application is intended to select files randomly from a given data population and the sample size is determined by given *confidence intervals (CI)* and *confidence levels (CL)*.

In a typical random sampling, we consider reviewing a random sample of documents from the whole document population, classifying them as valid or invalid, and projecting *the percentage of relevant documents* found in the sample onto the whole population. If the sample is completely random and the sample size is sufficiently large, we can determine a *Confidence Level* within a certain error margin (i.e., *Confidence Interval*). The Confidence Level is expressed as a percentage (e.g., CL 95% means you can be 95% certain). A commonly used CL is 95%, which is derived from the normality assumption of random samples from a population. The Confidence Interval is expressed in percentage points (e.g., $\pm 2\%$) of the confidence (i.e., CL), which indicates the reliability of an estimate. For more details about random sampling techniques used in the legal community, the readers are recommended to review the article by Ralph Losey [3]. For a more in-depth technical description of random sampling, please refer to the book by Cochran [1].

One question we commonly ask in statistical sampling is that "How do we determine the sampling size for a population?" We use the most standard method to compute the sample size by Cochran [1], which is described as follows.

$$n_0 = \frac{z^2\, p\,(1-p)}{e^2} \qquad (1)$$

The variable n_0 represents the required sample size and z represents the z-statistic that is computed based on the desired Confidence Level, e.g., the CL 90% has z-value 1.645, CL 95% has z-value 1.96, etc. The variable p represents the estimated proportion of an attribute that is presents in the population (i.e., the estimated probability that a sampled document in the population is *relevant*). For Random Sampler, we assume maximum variability in the target data, i.e., $p = 0.5$. The variable e represents the desired Confidence Interval in terms of precision, e.g., the CI $\pm 2\%$ represents the precision $e = 0.02$. Suppose, we assume a CL of 95% and CI of 2%, we get n_0 as

$$n_0 = \frac{1.96^2\ 0.5\ (1-0.5)}{0.02^2} = 2401$$

Equation 1 yields a representative sample under the assumption of a large population. For a finite population, we need a correction [2] for $n_0$as follows:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}} \qquad (2)$$

The variable n represents the corrected sample size and N represents the population size. For example, if we assume $N = 2{,}000$ for the previous example, we get:

$$n = \frac{2401}{1 + \frac{(2401 - 1)}{2000}} = 1091.3636$$

Finally, we round off the sample size n to an integer value and randomly select n samples from the population for review.

Software Implementation

We implement the Random Sampler algorithms and user interface using the Python [5, 6] programming language and the wx-Python [4] GUI toolkit. To select random samples of documents from the whole population, we need randomly generated document indexes.

We generate the document indexes using the Python random number generator module [5], which generates pseudo-random numbers based on the industry-standard Mersenne Twister [7] algorithm. The sample size based on a user specified Confidence Level and Confidence Interval is computed using the method described in the introduction. For the PST formatted email documents, we unpack all the emails and their attachments contained in the PST documents and employ the random sampling algorithm on individual documents. We also provide a user interface to evaluate the generated sample and generate HTML reports, which can be viewed using the default system HTML viewer (e.g., Internet Explorer, Mozilla Firefox, Google Chrome, etc.).

References

1. Cochran, W. G. 1963. *Sampling Techniques*, 2nd Ed., New York: John Wiley and Sons, Inc.
2. Glenn D. Israel. 2009. Determining Sample Size, Technical Report, URL: http://edis.ifas.ufl.edu/pd006
3. Ralph Losey. Random Sample Calculations And My Prediction That 300,000 Lawyers Will Be Using Random Sampling By 2022, URL: http://e-discoveryteam.com/2012/05/06/random-sample-calculations-and-my-prediction-that-300000-lawyers-will-be-using-random-sampling-by-2022/
4. http://www.wxpython.org/
5. http://docs.python.org/2/library/random.html
6. http://www.python.org/
7. Makoto Matsumoto and Takuji Nishimura. 1998. Mersenne twister: a 623-dimensionally equi-distributed uniform pseudo-random number generator