SOME THOUGHTS ON THE USE OF
STATISTICAL SAMPLING IN LEGAL
RESEARCH

by Carlos N. BOUZA-HERRERA, Professor at the Faculty of Mathematic of University of Havana, Cuba.

Much of legal research is based on discovering facts through analyzing a lot of papers. Electronically Stored Information (ESI) poses issues on using data stored electronically. With the increase of data volumes, a need of reducing costs, without violating the accepted assumptions poses urgently mid changes in the law firms. The reduction of costs should not be solved by using “low-cost lawyers”.

This paper discusses on the use of Technology Assisted Review and Statistical Sampling for retrieving information and some examples are discussed for illustrating.

A broad definition of legal research is that: it is a process which looks for identifying and retrieving what is needed for supporting legal decision-making. Hence, we may consider that it starts with the analysis of the facts on a particular and ends with the binomial application-communication of the results of the investigation.

Nowadays statistical evidence, sustained by probabilistic reasoning, plays an important role in common life. It is expanding its area of applications to criminal investigations, prosecutions and trials. Particularly, forensic scientific evidence, including DNA, produced by expert witnesses, is one of the emerging areas for statistical applications. That sustains that if you are involved in criminal adjudication, having a comprehension of the basics of probability and statistics is needed. In other legal researches, a similar situation is present: data must be retrieved and analyzed. Misunderstandings of what statistical information at hand are to be processed and interpreted, as well as of the role of the involved probabilities, have contributed towards serious miscarriages of justice. These facts suggest including in the education for lawyers a training on statistical thinking on how it should be used in legal research.

Actually, some processes use statistical sampling for providing evidence at the court yard. The correctness of the statistical procedures used, are being taken into account in the allegation of decisions by the court. Hence, having a good statistic advisor is one of the actual needs of the law firms.
Another problem is related with the need of dealing with Bigdata. They are being used in different legal issues at least in the past 20 years. The presence of Bigdata poses to the investigators to deal with responding to:
– How much data they have?
– Which is the structure of the data (structured, unstructured, text-based, internal and external)?
– Is it possible to analyze the existing data in real time for instantaneous decision-making?
– Are the data reliable?

Much of legal research is based on discovering facts through analyzing a lot of papers. Electronically Stored Information (ESI) poses issues on using data stored electronically. With the increase of data volumes, a need of reducing costs, without violating the accepted assumptions poses urgently mid changes in the law firms. The reduction of costs should not be solved by using “low-cost lawyers”.

To give a modern response when dealing with Bigdata an emerging technology for the retrieval of document information is connected with Technology Assisted Review (TAR) and Statistical Sampling. They are occupying a distinguished place as a tool for the research of law firms, as it reduces risks and improves productivity in eDiscovery processes. TAR is based on statistical models and it is being accepted as some kind of standard statistical tool for analyzing Bigdata problems posed by the existence of ESI.

Actually, many U.S. courts are endorsing the use of predictive-coding technologies. Consequently U.S. law-firms are encouraging structuring task groups for improving Bigdata practice.

§ 1 – SOME USES OF STATISTICAL SAMPLING

The analysis of data always has posed a complicated task to law firms. Nowadays the available data overcomes the capacity of the attorneys if some modern technique is not used for sampling and providing relevant information. Consider the use of applying Statistical Sampling to discovery of relevant and responsive documents. Though it is not a common practice, it is increasing its role in legal research. The reasons are the usual in statistical research. Its use is cost-effective in many tests as its behavior has been reasonably effective in finding relevant and responsive documents. Sampling is currently used in many areas of the Social Sciences. In particular, sociology studies use sampling models for obtaining information. The theoretical frame uses the fact that the study deals with finite population of well identified units (U={u₁, ..., uₙ}).
Selecting a sub set of them generates a sample \((s \subset U)\). Using judgmental sample was initially the approach up to the general acception that probability sampling is the only way of obtaining “representative samples”.

Statistical sampling is of use in many aspects of the administration of justice. Providing facts coming from well supported statistical research is a source of evidence. The court analyzes the results of statistical research but it must be aware of what is scientifically correct and how some models may be used for manipulating the results. Then, statistical experts are to be contracted by law firms for designing their needs of developing statistical inquires. On the other hand, the court must have an adequate counterpart for giving support to the righteous of the conclusions derived by the research.

The use of statistical evidence has proved to be of considerable support in the court. In some areas, they are currently used. Statistical sampling is accepted for estimating Medicare overpayments. Unfortunately, there are not well-established guidelines for sampling methodologies, as in other areas. Hence, there is a basis for considering whether a statistical principle, or method, is to be preferred to another one. There is a need of establishing some standards for considering when a statistical study is valid or not. In USA, the programs of Medicare have established that a statistical sampling evidence should be considered as acceptable, only if it uses a probability sampling design. That is observing any sample \(s\) must have a probability \(P(s) \in [0,1]\).

The importance of modeling adequately is exemplified by some trials as the following ones:

- **Transyd Enterprises LLC D/B/A Transpro Medical Transport (Appellant) vs (Beneficiaries) Trailblazer Health Enterprises LLC (Contractor), Claim for Part B Benefits, 2009 WL 5764287 (Sept. 15, 2009).** MAC rejected the appellant’s argument “PSC’s sampling methodology is invalid” because the PSC failed to document that its statisticians possessed at least a master’s degree in statistics or the equivalent.

- **Robert D. Lesser, M.D. & Assocs. (Appellant) vs (Beneficiaries) Pinnacle Business Solutions, Inc. (Contractor), Claim for Part B Benefits, 2011 WL 5263619, Docket No. M-11-358 (Feb. 18, 2011).** The Council noted that AJI relied on the 60-day timeline in the MPIM, which applies to prepayment and post payment review for MR (Medical Review) purposes. The case arose from a statistical sampling review by the Benefit Integrity unit of the ZPIC.

- **The MMPIM General Medicine, P.C. (Appellant) (Beneficiaries) Palmetto, GBA (Contractor), Claim for Part A Benefits, 2010 WL 7232825, Docket No. M-10-1933 (Nov. 24, 2010):** The Council found
Some Thoughts on the Use Statistical Sampling in Legal Research – Carlos N. Bouza-Herrera

The appellant’s case was based on unsupported speculations and conjectures. It addresses claimed that stratification should have been used, stating that the statistical sampling guidelines did not require stratification of every sample in order to make the sampling valid.

§ 2 – Technology Assisted Review (TAR) and Statistical Sampling

A particularly important task in legal work is text classification. Different studies suggest that machine learning techniques outperforms the classic manual document review developed by lawyers. They support that Technology Assisted Review (TAR) and Statistical Sampling increase both productivity and accuracy at a lower cost. Empirical evidence sustains that the use of TAR reduces the review time in a 75% of the time and the cost is only 30% of the classic methods.

Those are the reasons why one of the more accepted sampling procedures is using TAR. There are not many publications on its theoretical properties but the comparison of the cost reduction, due to its use has increased its popularity in legal research. Many law firms are considering how unassisted document review performs in comparison with TAR, which is validated by statistical sampling models.

TAR uses the expertise of attorneys and the methods of machine-learning to automatize the prioritization of documents to be reviewed. The ranking uses a measure of the responsiveness of document to a particular matter. By using it for dealing with Big data the firms reduce costs and key documents are obtained faster.

Some recent documented evidences of the usefulness of using TAR are:

– Da Silva Moore vs Publicis Groupe8. Andrew Peck (US Magistrate Judge) gave his opinions the validity of judgmental and statistical sampling for validating the results of predictive coding. (The Case for Statistical Sampling in e-Discovery7).

– Kleen Products vs Packaging Corporation of America9. Nan Nolan (US Magistrate Judge) heard the testimony, for sustaining the validity of the sampling process, used by defense. The validity of testing the results based on research terms, instead of predictive coding, in finding relevant documents was on trial. the parties had to determine with sampling procedure was acceptable for them. Once an agreement was obtained on the keyword to be looked for using sampling the research and discussion went forward.

– Global Aerospace vs Landow Aviation. The court stated that predictive coding (aka TAR) including a statistical model for validating the
Some Thoughts on the Use Statistical Sampling in Legal Research – Carlos N. Bouza-Herrera

protocol was adequate for locating and retrieving documents for production. The reduction of the costs is important but in addition the consistency of using TAR and sampling is considerably larger than the so called “linear review”. Linear reviews are developed by the reviews, performed by attorneys of the documents. The inconsistency of the reviews due to human error in not measured. Commonly no statistical sampling is used and hence, the reliability of such reviews is not possible. Therefore, the inconsistency of reviewers is unknown. TAR is validated with statistical sampling and it is highly consistent, and hence more reliable, compared with results of unassisted reviews performed by attorneys. Therefore, using it the lawyers assure that the process achieves a large level of success in identifying the relevant and responsive documents.

Well known sampling models as stratification allows improving the quality of the review process. For example, if a ranking of the importance of the documents is made previously, the consistency may be improved by using an unequal probability sampling or ranked set sampling. Such approaches save time as they avoid expecting for “first-level” reviews. For example, documents ranked first receive a preferential treatment in terms of the probability of being selected.

Corporate law departments deal with large amounts of data from invoices, and need to determine the factors influencing rates for negotiating better, deals based on that data. A free mobile application that aggregates data from thousands of law firm invoices is TyMetrix Legal Analytics. TyMetrix RateDriver™ mobile application uses the statistical model from Real Rate Report™. It is a statistical analysis of legal invoices.

§ 3 – A study

Less documented is its use in providing evidence on reclamations on the contamination due to enterprises. A question is: are the levels of contamination acceptable? The enterprise produces reports to the governmental agencies. On doubts on the accuracy of the reports environmentalists supported claims of farmers that the water used for agriculture was being contaminated. Their claim is based on the observed behavior of the production of the land.

The case was Farmers, F. (Appellant) vs (Beneficiaries) Chemical enterprise, CE (Contractor), Claim for Part A contamination of the water is affecting the fertility of lands: The appellant considered that the reported data which supported that the contamination levels were within the accepted interval were not correct. The arguments of the enterprise were unsupported speculations and the conjectures
cannot be proved without a statistical study. The statisticians supporting the appellants claimed that the measurements of the sensors at the factory output were providing not accurate information. They selected some points in the course of the water source and obtained their own measurements. A sample of them were compared with the ones made at the output of the factory by the sensor of the enterprise owners.

The set of measures of the outputs were considered as binary (0, 1) indicating whether they coincided with the ones of the other sensors (correctly classified=1, incorrectly classified=0). The results of N measurements are summarized in the Table 1.

Considering that the classification is equivalent to a double-blind method that is they are made independently. Each measurement generates a value

\[ Y_t = \begin{cases} 
1 & \text{if } t \text{ is correctly classified by both sensors} \\
0 & \text{other wise} 
\end{cases} \]

Summarizing is obtained the next table

**Table 1.** Classification of N measurements of 2 sensors

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>(n_{11} = A)</td>
<td>(n_{12} = B)</td>
<td>(n_{1+} = A + B)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>(n_{21} = C)</td>
<td>(n_{22} = D)</td>
<td>(n_{2+} = C + D)</td>
</tr>
<tr>
<td>Total</td>
<td>(n_{+1} = A + C)</td>
<td>(n_{+2} = B + D)</td>
<td>(N = A + B + C + D)</td>
</tr>
</tbody>
</table>

Different agreement indexes were considered. They are function of

\[ p_i = \frac{n_{ij}}{n}, \quad q_i = \frac{1}{2}(p_{2i} + p_{i2}), \quad q_i = \frac{1}{2}(p_{i+1} + p_{1i}), \]

where \( p_{i+} = \frac{n_{i.}}{2}, p_{+i} = \frac{n_{.i}}{n} \)

Were evaluated the following indexes

**Dice**

\[ I_D = \frac{2p_{11}}{p_{1+} + p_{+1}} = \frac{p_{11}}{p} \]

A value close to zero means that the sensors have a small “agreement”.

**Correlation coefficient**

Note that we are dealing with attributes (categorical variables). In this case, the correlation coefficient of Pearson may be rewritten, in terms of Table 1 as

\[ \rho = \frac{A \times D - B \times C}{\sqrt{(A + B) \times (C + D) \times (A + C) \times (B + D)}} \]
The values of $\rho$ will be in the interval [-1, 1]. If $\rho=1$, the sensors behave similarly, $\rho<0$ means that they highly disagree and they are “independent” if $\rho=0$.

**Measure of Differences**

$$D = \frac{B + C}{N}$$

An increase of $D$ means that they largely disagree

**Kappa**

$$\kappa = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{+i} p_{+i}}{1 - \sum_{i=1}^{k} p_{+i} p_{+i}}$$

A large value of it means the existence of a high level of agreement.

3 sensors were placed and data were collected during a month. The values of the indexes were computed for each one and compared with the reports of the enterprise. Each one was evaluated considering the belonging to the accepted levels of contamination fixed by the law. They are reported in the next table

**Table 2. Values of the indexes of 3 sensors**

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Dice</th>
<th>Correlation Coefficient</th>
<th>Measure of Differences</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0,801</td>
<td>0,128</td>
<td>0,333</td>
<td>0,302</td>
</tr>
<tr>
<td>2</td>
<td>0,823</td>
<td>-0,001</td>
<td>0,301</td>
<td>-0,010</td>
</tr>
<tr>
<td>3</td>
<td>0,774</td>
<td>0,300</td>
<td>0,352</td>
<td>0,364</td>
</tr>
</tbody>
</table>

Then, it was documented that the lectures of the enterprise had a low agreement when classifying the violating of the accepted level of contaminator with the other sensors. The court fixed a fine to the enterprise for avoiding their responsibilities with the environment and a calculation of the damage to the farmers is in progress. The statisticians of the enterprise alleged that they assumed that the measurements were normally distributed but the appellant’s proved that this probability assumption was incorrect and that categorical data analysis must have been used by then for controlling.
Some Thoughts on the Use of Statistical Sampling in Legal Research – Carlos N. Bouza-Herrera

References


BALDING D.J., Weight-of-Evidence for Forensic DNA Profiles, Chichester, Wiley, 2005


