Consilio / ADVANCED LEARNING INSTITUTE

Consilio Institute: White Paper

SAMPLING TECHNIQUES FOR LITIGATION AND INVESTIGATIONS

By Matthew Verga, Esq.
Director of Education

Consilio Institute: White Paper

SAMPLING TECHNIQUES FOR LITIGATION AND INVESTIGATIONS

TABLE OF CONTENTS

Disclaimers

The information provided in this publication does not, and is not intended to, constitute legal advice; instead, all information, content, and materials available in this publication are provided for general informational purposes only. While efforts to provide the most recently available information were made, information in this publication may not constitute the most up-to-date legal or other information.This publication contains links to third-party websites. Such links are only for the convenience of the reader; Consilio does not recommend or endorse the contents of the third-party sites.

Readers of this publication should contact their attorney to obtain advice with respect to any particular legal matter. No reader of this publication should act or refrain from acting on the basis of information in this publication– without first seeking legal advice from counsel in the relevant jurisdiction. Only your individual attorney can provide assurances that the information contained herein – and your interpretation of it – is applicable or appropriate to your particular situation.

Use of this publication, or any of the links or resources contained within, does not create an attorney-client relationship between the reader and the author or Consilio. All liability with respect to actions taken or not taken based on the contents of this publication is expressly disclaimed. The content of this publication is provided "as is." No representations are made that the content is error-free.

FINDING OUT HOW MANY RED HOTS ARE IN THE JELLYBEAN JAR

A candy store is running a contest. In the front window is a comically enormous jar of jelly beans, all different kinds and colors. Mixed in among them are a secret number of red hot cinnamon candies, similar in size and shape, all red. Whoever can guess closest to the true number of red hots mixed into the jar wins the prize. How do you guess? Do you try to count the red candies you can see, hoping they're all red hots, and then guess at how many you can't see? Do you try to count all the candies? Do you try to estimate volumes?

What if you were allowed to take one scoop of candies out of the enormous jar for closer examination, to determine exactly which ones in the scoop were red hots? Could you extrapolate from the scoop to the jar? How much better might your guess be then?

Sampling in eDiscovery

Despite years of discussion in the eDiscovery industry about the power and importance of sampling techniques – particularly in the context of technology-assisted review (TAR), many practitioners remain unfamiliar with what they can accomplish with them, and when, outside of TAR, they might do so. Beyond just being an essential part of TAR, however, there are opportunities across the phases of an eDiscovery project – whether for litigation or an investigation – to replace guesses based on anecdotal evidence with actual estimates based on formal sampling.

Courts have actually been encouraging parties to leverage sampling techniques in eDiscovery since before TAR existed, suggesting its use for the validation of search terms and document review processes:

- "Common sense dictates that **sampling and other quality assurance techniques must be employed** to meet requirements of completeness," *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650 (M.D. Fla. 2007)[1] [emphasis added]
- **"The implementation of the methodology selected should be tested for quality assurance;** and the party selecting the methodology **must be prepared to explain the rationale** for the method chosen to the court, **demonstrate that it is appropriate** for the task, and **show that it was properly implemented,"** *Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251 (D. Md. 2008)[2] [emphasis added]

And they have continued encouraging its use for those purposes, even outside of TAR, to this day:

- "Just as it is used in TAR, **a random sample of the null set provides validation and quality assurance of the document production when performing key word searches.** Magistrate Judge Andrew Peck made this point nearly a decade ago. *See* William A. Gross Constr. Assocs., 256 F.R.D. at 135-6 (citing ***Victor Stanley, Inc. v. Creative Pipe, Inc.***, 250 F.R.D. 251, 262 (D. Md. 2008)); ***In re Seroquel Products Liability Litig.***, 244 F.R.D. 650, 662 (M.D. Fla. 2007) (requiring quality assurance)." *City of Rockford v. Mallinckrodt ARD Inc.*, 326 F.R.D. 489 (N.D. Ill. Aug. 7, 2018)[3] [emphasis added]

[1] *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650 (M.D. Fla. 2007), available at https://casetext.com/case/in-re-seroquel-products-liability-litigation-16.

[2] *Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251 (D. Md. 2008), available at https://casetext.com/case/victor-stanley-inc-v-creative-pipe.

[3] *City of Rockford v. Mallinckrodt ARD Inc.*, 326 F.R.D. 489 (N.D. Ill. Aug. 7, 2018), available at https://casetext.com/case/city-of-rockford-v-mallinckrodt-ard-inc-1.

And, of course, the importance of sampling comes up again and again in discovery decisions and orders related to TAR use.

Industry publications, too, have taken repeated notice of the power and importance of sampling in eDiscovery. For example, sampling features prominently in The Sedona Conference's Commentary on Achieving Quality in the E-Discovery Process,[4] and the EDRM organization has released a commentary specifically on leveraging sampling in eDiscovery.[5]

Informal Approaches to Sampling

Many practitioners do engage in informal types of sampling already. As practitioners have done since the early days of discovery, it is common for a knowledgeable team member to test potential search terms and phrases by informally "poking around" in some of the results returned by them. The same thing goes for poking around in the materials collected from different sources or different custodians to determine the relative importance of different tranches of materials. The same also goes for quality control checks of document review efforts, with more senior attorneys poking around in the batches of documents reviewed by less-experienced attorneys to double-check their relevance or privilege determinations.

These informal approaches to sampling are inarguably valuable for gathering anecdotal evidence, making instinctual assessments, and learning about your materials or your efforts. Some information is always better than no information. But there are limits to what can be learned through these informal approaches and to how reliable such insights are.

Formal Approaches to Sampling

Formal approaches to sampling, on the other hand, facilitate more precise estimates with known reliability. It is these approaches that make sampling so valuable in TAR specifically and in eDiscovery generally. For example, formal sampling approaches can be used to generate:

- Reliable estimates of how many relevant documents are in a given tranche
- Reliable projections of the amount of redaction or privilege logging to do
- Reliable measurements of relevant materials missed by a given process
- Reliable reporting on the efficacy of a given search or other classifier

These measurements and many more can be taken using the same basic sampling techniques at various points in the discovery project lifecycle.

About this White Paper

In this white paper, we are going to review key sampling concepts and processes that are relevant to litigations and investigations, from winning the jellybean jar contest described above, to planning at the beginning of a project and checking completeness at the end, to testing your classifiers, both human and machine.

[4] The Sedona Conference, *Commentary on Achieving Quality in the E-Discovery Process,* 15 SEDONA CONF. J. 265 (2014), available at https://thesedonaconference.org/publication/Commentary_on_Achieving_Quality_in_the_E-Discovery_Process.

[5] EDRM, *Statistical Sampling Applied to Electronic Discovery,* https://www.edrm.net/resources/project-guides/edrm-statistical-sampling-applied-to-electronic-discovery/ (Feb. 18, 2015).

KEY SAMPLING CONCEPTS FOR WINNING THE CANDY CONTEST

In order to use sampling to estimate how many red hots are mixed into the jellybean jar, we need to understand some basic sampling concepts, including: sampling frame, prevalence, confidence level, and confidence interval, as well as how each affects required sample size. We also need to understand that whenever we refer to sampling here, we are referring to simple random sampling in which **any item within the sampling frame has an equal chance of being randomly selected** for inclusion in the sample.

Sampling Frame

Sampling frame refers to the set of materials from which a sample will be taken. In the context of our jellybean example, the sampling frame would be the full contents of the enormous jar of jellybeans and red hots. In the context of eDiscovery, your sampling frame will typically be the pool of materials available after any initial, objective culling has taken place (*i.e.*, what's left for assessment and review after initial de-NISTing, deduplication, and date restriction during processing).

In addition to being your sample source, your sampling frame also affects the size of the samples you will need to take. As we will discuss below, sample size is primarily determined by how reliable and precise you want your results to be, but the size of your sampling frame also affects your needed sample size to some extent. As your sampling frame gets bigger, your sample size will also need to get bigger – but only up to a point. Beyond that point, the effect levels off, so the sample size needed for a frame of 100,000 items (e.g., jellybeans, documents, etc.) is roughly the same as the sample size needed for 1,000,000 of them, which is roughly the same as the sample size needed for 10,000,000 of them. Sampling frame size has the weakest effect on sample size.

Prevalence

Prevalence is how much of something there is within your sampling frame. For example, it could be how many red hots there are in your jellybean jar, or it could be how many relevant documents there are in your collected materials. It could also be how many documents are privileged, how many require redaction, or any other binary property you want to measure.

In the math underlying sampling, the prevalence of what you are seeking is also a factor that can have an effect on the required sample size for some purposes. When what you are doing sampling for is to estimate prevalence, however, you need to plug in an assumption for this value, and to be safe, you plug in the most conservative value (*i.e.*, the one that results in the largest sample size). For prevalence, this is 50%, meaning that half the sampling frame is what you're looking for and half is not. Most sampling features in eDiscovery tools and online calculators will default to this value and may not even given you the option to change it.

Confidence Level

Confidence level is a measurement of how reliable your results are. It is expressed as a percentage out of 100, and most commonly, you will see discussion of 90%, 95%, or 99% confidence levels. What these numbers technically mean is that, if you reran the same sampling process 100 times in a row, you would expect to get similar results 90 times out of 100, or 95 times out of 100, or 99 times out of 100.

The higher you want your confidence level to be, the larger the sample size you will need to use to achieve it, and confidence level has a stronger effect on sample size than sampling frame size or prevalence does. For example, if you were taking a sample from a sampling frame of 100,000 items, and you wanted a margin of error of +/-2% (which we will discuss further below), here is how your required sample size would vary with your desired confidence level:

- For a confidence level of **90%**, a sample size of **1663**
- For a confidence level of **95%**, a sample size of **2345**
- For a confidence level of **99%**, a sample size of **3982**

Confidence Interval

Confidence interval is a measurement of how precise your results are. It is expressed as a percentage out of 100, and most commonly, you will see discussion of confidence intervals of 2%, 4%, and 10%. Even more commonly, you will see discussions refer to margin of error with references to +/-1%, +/-2%, and +/-5%. These margins of error are actually the equivalents of those confidence intervals. The latter is just framed in terms of plus or minus half the range, and the former is framed in terms of the full range.

The narrower you want your range of uncertainty to be, the larger the sample size you will need to use achieve it, and confidence interval (or margin of error) has the strongest effect on needed sample size. For example, if you were taking a sample from a sampling frame of 100,000 items, and you wanted a confidence level of 95%, here is how your required sample size would vary with your desired range of uncertainty:

- For a confidence interval of **10%**, a.k.a. a margin of error of **+/-5%**, a sample size of **383**
- For a confidence interval of **4%**, a.k.a. a margin of error of **+/-2%**, a sample size of **2345**
- For a confidence interval of **2%**, a.k.a. a margin of error of **+/-1%**, a sample size of **8763**

RED HOTS, HOT DOCS, AND THE ONES THAT GOT AWAY

Now that we understand the necessary sampling concepts, let's apply those concepts to our candy contest and figure out how many red hots we think are in the jellybean jar. In order to do so, we will need to identify our sampling frame, select our desired confidence level, and select our desired confidence interval.

For this example, our sampling frame is all the candies in the enormous jellybean jar, which a sign indicates holds approximately 100,000 candies. For our confidence level, let's use 95%, which has been referenced in a variety of cases and articles as a potentially acceptable level of confidence, and for our confidence interval, let's use 4% – also known as a margin of error of +/-2%, which has also been widely discussed and used. (For example, 95% and +/-2% were the proposed values used in the plan in the *da Silva Moore*[6] case and in many other TAR cases.)

[6] *Moore v. Publicis Groupe,* 287 F.R.D. 182 (S.D.N.Y. 2012), available at https://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2011cv01279/375665/96.

So, How Many Red Hots?

Now that we have our required values (**100,000, 95%, +/-2%**, and an assumed **50%** prevalence), we are ready to plug them into our sampling tool or calculator to find out how large our simple random sample will need to be. Most modern document review tools have some form of sampling tools built into them, but sampling calculators[7] are also readily available online and random document selections can be made in manual ways if needed (e.g., by using the RAND function in Microsoft Excel). Plugging the values we've chosen into a sampling calculator reveals that we need a simple random sample of **2,345** pieces of candy to make our desired estimate, which is just **2.345%** of the total sampling frame.

Once a candy store employee has retrieved for us a randomly selected assortment of **2,345** pieces of candy from the jar, we can then review those sample candies up close to determine exactly which ones are cinnamon red hots. Let's say our review reveals there are **142 red hots** among the **2,345** sample candies, or **6.1%**. We can now say – **with 95% confidence** – that the overall prevalence of red hots in the jar is **between 4.1% and 8.1%**, or **between 4,100 and 8,100 total red hots**.

If we were willing to review a larger sample of **8,763** candies, we could even narrow that range to **between 5,100 and 7,100 total red hots**.

So, How Many Hot Documents?

This same process can be employed in an eDiscovery project to make any number of useful estimations about a new collection of materials, including: **the prevalence of relevant materials, the relative prevalence of relevant materials in different sources, and the prevalence of materials requiring special review efforts (e.g., privilege logging, redactions, technical knowledge, etc.)**. These estimates can in turn be used to more accurately estimate your needed project resources, optimal project workflows, and likely project costs and durations. They can also be valuable in assessing the viability of a TAR solution or the need for additional objective culling. As projects progress, they can also provide a yardstick against which to measure progress and completeness.

It should also be noted that, when applying these techniques to eDiscovery, it is important to use the highest quality document review possible. While identifying cinnamon red hots is very straightforward, making legal or process determinations about documents can be quite nuanced, and the nature of sampling (extrapolating from a little to a lot) means that mistakes in classification during sampling will have amplified effects on the reliability of your estimates.

And, How Many Did We Miss?

One of the most common applications of prevalence estimation is in testing for completeness at the end of a TAR process or after the application of keyword searches. This is sometimes referred to as measuring **elusion**, *i.e.* the quantity of materials that eluded identification by the filtering and review process employed. For such estimations, the sampling frame is the pool of unreviewed materials eliminated before human review, by either the TAR software used, or by the keyword searches applied. The process is otherwise identical to the one described above.

There is no way to perfectly identify and produce **all** relevant electronic materials – and no legal requirement that you achieve such perfection, but there can be great value in being able to say with some certainty how little (or how much) has been missed. A reliable estimate can provide concrete evidence of the adequacy or inadequacy of a completed process, or a basis for arguing the proportionality or disproportionality of any additional discovery efforts.

[7] *See e.g.* Raosoft, *Sample Size Calculator*, http://www.raosoft.com/samplesize.html (2004).

TESTING CLASSIFIERS

What Is a Classifier?

Classifiers are mechanisms used to classify documents or other materials into discrete categories, such as those requiring review and those not requiring review, or relevant and non-relevant, or privileged and non-privileged. That mechanism might be a search using key words or phrases. It might be the decisions of an individual human reviewer or the aggregated decisions of an entire human review process. It might be the software-generated results of a technology-assisted review process. The binary classification decisions of any of these classifiers are testable in the same basic way. To start, we will focus on searches as the classifiers to be tested.

What Properties of a Search Classifier Do We Test?

When testing search classifiers, we are actually measuring two things about them: their **recall** and their **precision**, which correlate to their **efficacy** and their **efficiency**:

- **Recall is how much of the total stuff available to find the classifier actually found**, so higher recall (*i.e.*, finding more) means greater efficacy, and lower recall (*i.e.*, finding less) means lower efficacy
- **Precision is how much other, unwanted stuff the classifier included along with the stuff you actually wanted**, so higher precision (*i.e.*, less junk) means higher efficiency, and lower precision (*i.e.*, more junk) means lower efficiency

Both recall and precision are expressed as **percentages out of 100**:

- For example, if there are **500** relevant documents somewhere in a dataset, and a search finds **250** of those documents, then that search has a recall of **50%** (*i.e.*, 250/500)
- If the search returned **750** non-relevant documents along with the **250** relevant ones, that search would have a precision of **25%** (*i.e.*, 250/1000)

There is also generally a tension between the two criteria. Optimizing a search to maximize recall beyond a certain point is likely to require lowering precision and accepting more junk, and optimizing a search to maximize precision beyond a certain point is likely to require accepting lower recall and more missed relevant materials. Deciding what balance between the two is reasonable and proportional is a fact-based determination specific to the needs and circumstances of each matter.

What Sample Is Needed to Test a Search Classifier?

In order to test a search classifier's recall and precision, you must already know the numbers of documents in the classifications you are testing. For example, to determine what percentage of the relevant material is found, you must know how much relevant material there is. Since it is not possible to know this about the full dataset without reviewing it all (which would defeat the purpose of developing good searches), classifiers must be tested against a control set drawn from the full dataset.

Much as we did for estimating prevalence, control sets are created by taking a simple random sample from the full dataset (after initial, objective culling) and manually reviewing and classifying the materials in that sample. Just as with estimating prevalence, it is important that the review performed on the control set be done carefully and by knowledgeable team members. In fact, in many cases you may be able to use the same set of documents you reviewed to estimate prevalence as a control set for testing classifiers.

Unlike estimating prevalence, however, figuring out the size of the sample needed for your control set is not so cut and dry. As we will discuss below, the reliability of the results you get when testing classifiers is related to how many potential things there were for the classifiers to find in the control set. For example, if you are testing searches designed to find relevant documents, the more relevant documents there are in your control set the more reliable your results will be.

This means that **datasets with low prevalence may require larger control sets to test classifiers than datasets with high prevalence**, depending on how reliable you need your results to be. The results of a prevalence estimation exercise can help you figure out how large of a control set you need (and whether your prevalence estimation set can just be repurposed for this exercise).

SHOW YOUR WORK: CONTINGENCY TABLES AND ERROR MARGINS

Once you have run a search classifier you are testing against your control set, you can calculate recall and precision for it by using contingency tables. Contingency tables (also sometimes referred to as cross-tabulations or cross-tabs) are simple tables used to break down the results of such a test into four categories: true positives, false positives, false negatives, and true negatives. These four categories are comparisons of the results of the search classifier to the prior results of your manual review of the control set:

1. **True positives** are documents that your search classifier returns as relevant results that your prior review of the control set also marked as relevant, *i.e.* **the right stuff**
2. **False positives** are documents that your search classifier returns as relevant results that your prior review of the control set had determined were not relevant, *i.e.* **the wrong stuff**
3. **False negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set marked as relevant, *i.e.* **missed stuff**
4. **True negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set had determined were not relevant, *i.e.* **actual junk stuff**

An Example Application

As we discussed above, sample sizes of a few thousand documents are common for taking prevalence measurements about large document collections. So, let's assume a hypothetical in which you have **a randomly selected set of 3,982 documents that you previously reviewed** to take a strong measurement of prevalence (99% confidence level, with a margin of error of +/-2%) within your collection of 100,000 documents. Let's also assume that **your review of that random sample revealed 1,991 relevant documents**.

In addition to knowing prevalence within the overall collection (48-52% prevalence, with 99% confidence), you now have **a 3,982 document control set for testing search classifiers, containing 1,991 relevant documents for them to try to find**. The next step is running your search classifier against it and seeing how its classifications compare to those of your prior review. Let's assume your hypothetical search returns 1,810 total documents which break down into the four categories as follows:

	Deemed Relevant by Prior Review	Deemed Not Relevant by Prior Review
Returned by Search Classifier (*i.e.*, Deemed Relevant)	1,267 (True Positives)	543 (False Positives)
Not Returned by Search Classifier (*i.e.*, Deemed Not Relevant)	724 (False Negatives)	1448 (True Negatives)

As we can see on this contingency table, the 1,810 results from your hypothetical search included 1,267 documents that were also deemed relevant in your prior review, which are your true positives. It also included 543 documents that were deemed not relevant in your prior review, which are your false positives. And, finally, we can see it missed 724 documents that were deemed relevant in your prior review, which are your false negatives.

You can use the results shown in this contingency table to easily estimate the recall and precision of the hypothetical search classifier you tested. As we discussed above, recall is the percentage of all the available relevant documents that were successfully identified by the search classifier being tested. So, in this example, your search identified 1,267 out of 1,991 relevant documents, **which gives you a recall of about 63.6%**. Also as discussed above, precision is the percentage of what the search classifier identified that was actually relevant. So, in this example, the search returned a total of 1,810 documents including 1,267 relevant documents, **which gives you a precision of 70%**.

Your hypothetical search, then, has high precision and good recall. The search could probably be revised to trade off some of that precision for higher recall or, possibly, to improve both numbers. Subsequent iterations of the search can be easily tested in the same way to measure the effect of your iterative changes.

How Reliable Are These Estimates?

We noted at the beginning of this hypothetical that your control set was a random sample of 3,982 documents that had been taken and reviewed previously to estimate prevalence in the full document collection with a confidence level of 99% and a margin of error of +/-2%. That same confidence level and margin of error, however, **do not carry over** to the estimates of recall and precision that you have made using the same documents. Because of how the math in question works, **your sample sizes for recall and precision are effectively smaller, which in turn makes your margins of error a little wider**.

The effective sample size for a recall estimation performed in this way is not the total number of documents in the control set, but rather **the number of relevant documents within it that are available to be found**. In this example, the search classifier is looking for the 1,991 relevant documents contained in the control set, which are effectively a random sample of 1,991 relevant documents from among all the relevant documents in the full document collection (a sampling frame you've already estimated to be about 50,000 documents).

The effective sample size for a precision estimation is also not the total number of documents in the control set, but rather **the number of documents identified by the search classifier**. In this example, the search classifier identified 1,810 documents, which are effectively a random sample of 1,810 documents from among all the documents the search would return from the full document collection (a sampling frame you can estimate to be about 45,500 documents).

Some tools will provide you with these calculations automatically, but you can also plug these numbers into a sampling calculator[8] yourself to work backwards and see what margin of error would apply to your recall and precision measurements. In this example, your recall estimate would carry a margin of error of about +/-2.83% (at a confidence level of 99%), and your precision estimate would carry a margin of error of about +/-2.97% (also at a confidence level of 99%). **Thus, you could be very confident that your tested search had a recall between 60.77% and 66.43% and a precision between 67.03% and 72.97%.**

GRADING PAPERS: MEASURING HUMAN REVIEW

As we discussed above, a classifier can be a search, a TAR process, or other things – **including a human reviewer or a team of human reviewers**. Just as a search or a TAR tool is making a series of binary classification decisions, so too are your human reviewers, and the quality of those reviewers' decisions can be assessed in a similar manner to how you assessed the quality of a search classifier above. Depending on the scale of your review project, employing these assessment methods can be more efficient than a traditional multi-pass review approach, and in general, they are more precise and informative.

Human Classifiers and Control Sets

In this context, the reviewers doing the initial review work are the classifier being tested. **The control set is effectively generated on the fly by the more senior attorney performing quality control review.** Their classification decisions are the standard against which the initial reviewer's classification decisions can be judged. If an appropriate document tagging palette is employed (or if a sufficiently sophisticated review tool is being used), it is not hard to track and compare both sets of decisions to assess your human classifiers the same way we assessed searches.

In this context, however, we are not typically measuring the recall and precision of the human reviewers, although that could be done as well. **For human reviewers, it is more common to measure accuracy and error rate.** Accuracy is expressed as a percentage out of 100, and it represents **the total number of correct classification decisions made by the initial reviewers.** Error rate is also expressed as a percentage out of 100, and it represents **the total number of incorrect classification decisions made.** Together, accuracy and error rate should add up to 100%.

In terms of a contingency table, accuracy is derived from **the combination of all true positives and true negatives**, and error rate is derived from **the combination of all false positives and false negatives**.

[8] *See e.g.* Raosoft, *Sample Size Calculator*, http://www.raosoft.com/samplesize.html (2004).

An Example Application to Human Review

Let's look at an example of how this works. **Let's assume that you perform quality control review of a random sample of 350 of the 2,000 documents reviewed this week by a particular member of your initial review team.** After completing your classifications and comparing them to those of the initial reviewer, you get the following breakdown of results:

	Deemed Relevant by the QC Reviewer	Deemed Not Relevant by the QC Reviewer
Deemed Relevant by the Initial Reviewer	70 (True Positives)	40 (False Positives)
Deemed Not Relevant by the Initial Reviewer	65 (False Negatives)	175 (True Negatives)

As with your search classifier, it is now straightforward to calculate an estimated accuracy and error rate for this reviewer's work this week. As noted above, accuracy is a combination of all the correct classification decisions, *i.e.* true positives + plus true negatives. So, in this example, your reviewer made 245 correct decisions out of 350 total decisions. **That gives you an accuracy of 70%.** As also noted above, error rate is combination of all the incorrect classification decisions, *i.e.* false positives + false negatives. So, in this example, your reviewer made 105 incorrect decisions out of 350 total decisions. **That gives you an error rate of 30%.**

There is also no reason that the same measurements could not be performed for more than one classification criteria based on the same quality control review (e.g., relevant and not relevant, privileged and not privileged, requiring redaction and not requiring redaction, etc.). **Any binary classifications for which you and your reviewers are making classification decisions can all be measured the same way.** The specific criteria measured and the specific results you get can then guide you in your ongoing reviewer training efforts, review oversight steps, and project staffing decisions.

How Reliable Are These Estimates?

As with testing search classifiers, you can work backwards from these results to determine how reliable these estimates of accuracy and error rate are, based on the size of the sampling frame (*i.e.*, the total number of reviewed documents from which you pulled the sample) and the size of the sample you took. In this example, the sampling frame would be 2,000 and the sample size would be 350. Using those numbers, we find that your estimates of accuracy and error rate have a margin of error of +/- 4.76% at a confidence level of 95%. **Thus, you could be 95% confidant that the rest of the work from the reviewer in question was between 65.24% and 74.76% accurate.**

A Note about Lot Acceptance Sampling

Lot acceptance sampling is an approach to quality control that is employed in high-volume, quality-focused processes such as pharmaceutical production or military contract fulfillment. **In this approach, a maximum acceptable error rate is established, and each batch of completed materials is randomly sampled to check that batch's error rate at a predetermined level of reliability.** If the batch's error rate is below the acceptable maximum, the batch is accepted, and if the error rate is above the acceptable maximum, the batch is rejected.

Large-scale document review efforts have a lot in common with those other high-volume, quality-focused processes, and some particularly-large review projects have employed lot acceptance sampling in a similar way. Individual batches of documents reviewed by individual reviewers are the batches being accepted or rejected, and random samples are checked from each completed one. Those with a sufficiently low error rate move on to the next phase of the review and production effort, those with too high of an error rate are rejected and re-reviewed (typically by someone other than the original reviewer). Error rates and batch rejections can be tracked by reviewer, by team, by classification type, or by other useful properties to identify problem areas for process improvement or problem reviewers for retraining or replacement.

Many practitioners become uncomfortable at the idea of deliberately identifying an acceptable error rate, or even of actively measuring the error rate at all, but **avoiding knowledge of your errors does not prevent their existence**. It just prevents you from being able to address them or being prepared to defend them. After all, **the standards for discovery efforts are reasonableness and proportionality – not perfection.**[9]

KEY TAKEAWAYS

There are five key takeaways from this white paper to remember:

1. Formal sampling can replace intuitive assessments and assumptions with precise, reliable estimates, and judges have often expressed a preference for argument and negotiation based on actual data and specific estimations rather than guesswork
2. Formal sampling has a variety of applications in litigation and investigations beyond just validation of technology-assisted review processes, including planning at the beginning of a project, checking completeness at the end, and testing your classifiers, both human and machine
3. Prevalence estimation can be accomplished by reviewing only a small percentage of large document collection, and it can be used to reliably estimate how many relevant documents are in a given tranche, the amount of redaction or privilege logging to do, the quantity of relevant materials missed by a given process, and more
4. Testing classifiers can also be accomplished by reviewing only a small percentage of a large document collection, and it can be used to iteratively improve your own searches, to evaluate those proposed by others, and to QC human document review
5. When using these sampling techniques, it is important to make sure you know how strong (confidence level) and how accurate (confidence interval/margin of error) your estimates need to be and will be, and consultation with an experienced expert is recommended

[9] "The second myth is the myth of a perfect response. The [respondent] is seeking a perfect response to his discovery request, but **our Rules do not require a perfect response. . . . Likewise, 'the Federal Rules of Civil Procedure do not require perfection.'** Like the Tax Court Rules, the Federal Rule of Civil Procedure 26(g) **only requires a party to make a 'reasonable inquiry' when making discovery responses,**" *Dynamo Holdings Ltd. P'ship v. Comm'r of Internal Revenue,* 2016 WL 4204067 (USTC 2016) [internal citation omitted; emphasis added].