Relativity®

White Paper

Common Statistical Concepts and Their Influence on Computer-Assisted Review

Dr. Gideon Frieder

Contents

Foreword

One of the primary challenges of introducing new technology to a group of users—or introducing an established technology to a new group of users—is the learning curve that must be addressed regarding the vocabulary that accompanies the technology

Computer-assisted review is no exception, and while the technology has established a foothold in the legal community, the industry still has some work to do regarding its understanding of the assisted review vocabulary and solidifying standard definitions.

Perhaps the most daunting topics in the entire computer-assisted review glossary are statistics and sampling. There are, after all, very good reasons why some of us elected to attend law school as opposed to a pursuit of the sciences, and the avoidance of numbers might be high on the list.

The necessary re-introduction of this branch of mathematics to the legal world has resulted in many attorneys scrambling to make proper use of terms like population, confidence level, and margin of error, often before a true understanding of these terms has had a chance to sink in. Yet the ability to participate in conversations regarding statistical sampling is essential for both internal and external reasons.

> **“ …the ability to participate in conversations regarding statistical sampling is essential. ”**

Internally, a team needs to effectively discuss sampling and validation criteria to finalize a project plan. The tech team needs the legal team to understand what the process is, and what to expect. The legal team also needs to convey their requirements to the tech team in order to satisfy the case's specific defensibility needs. In addition, the legal team must convey all of this information to the end client in a clear and accessible manner.

Externally, this understanding is important so that the parties of a dispute are able to negotiate production expectations effectively. We have already witnessed, via *Da Silva Moore*[1], what happens when parties agree to use computer-assisted review technology, but do not agree on how it will be applied.

Dr. Gideon Frieder—the author of the following paper—is the A. James Clark Professor of Engineering and Applied Science and a professor of statistics at the George Washington University. An expert in computer architecture and computational methods, among other specialties, Dr. Frieder previously served as the dean of the School of Engineering and Applied Science at the university.

Dr. Frieder's paper provides an accessible resource for those who wish to have a better understanding of the statistics and sampling methods which are utilized in a computer-assisted review workflow. Take a look at the callout text before each section for more information about how each statistical component relates to computer-assisted review.

The goal of this paper is to be platform-agnostic and address each term with relatable real-world examples. We hope it can help you better participate in the conversation.

Constantine Pappas, Product Specialist
kCura

[1] Da Silva Moore v. Publicis Groupe, et al., No. 11 Civ. 1279 (ALC) (AJP), 2012 U.S. Dist. LEXIS 23350 (S.D.N.Y. Feb. 24, 2012). Magistrate Judge Peck's decision was affirmed by District Judge Andrew Carter of the Southern District of New York on April 25, 2012. In subsequent proceedings, the plaintiffs were denied a motion to require that Judge Peck recuse himself on June 15, 2012, and that decision was upheld on appeal to the Second Circuit Court of Appeals on April 10, 2013.

The Basic Concepts

This first section discusses basic terms such as population which, in an Assisted Review project, defines the scope of a project—typically the number of documents. Each document will have a discrete value based on a binary choice, such as responsive or not responsive. In addition, understanding discrete distributions is important when interpreting rank scores for categorized documents in Relativity, which typically range from 70 to 100.

- Constantine

In computer-assisted review, a team's experts review a small subset of documents and—by training the computer based on their decisions—amplify their expertise in the case across the entire data set. Sampling is a key component of this, as it can assist a case team in reviewing millions of diverse, scattered documents for e-discovery.

Sampling is the process by which a part of a population is measured for a specific property, and that measurement is used to estimate the value of that property for the entire population. We use sampling when the whole population is too large, too difficult to access, or unavailable for actual contact.

To ensure that the precise meaning of sampling—the process involved, and its advantages and shortcomings—is more clearly understood and therefore better applied during a computer-assisted review project, this paper will explore sampling as a common statistical process. How does it work, and why should we believe it?

The most fundamental requirement for the use of sampling is the existence of a **population**—a set of items with at least one common property. By this definition, people are a population, but so are cars, shirts manufactured in the U.S., or a collection of documents.

As mentioned, sampling helps estimate the value of a common property in a given population. For example, if the population is a set of documents, we may be interested in their size, the presence of some keywords, and other indicators of relevance.

Let's say that, for a given population, we have the value of the property—sometimes called the attribute—for every member. We are interested in the probability mass function: a table that contains all possible values of the measured attribute and the number of members in the population that have that value. Figure 1 provides an example for a mass function describing the income of a group of U.S. residents.

Figure 1: Distribution of U.S. salaries (rounded in increments of $1,000)

Salary	30	35	40	45	50	55	60
Count	50	60	80	80	70	30	10

Such a distribution is considered **discrete**, because the data has a finite number of values that can be counted, like the number of documents in a collection. There are also **continuous** distributions, which include data that can have any value in a range. Examples include a runner's time in a race, which can be measured down to fractions of a second. In other words, discrete distributions describe a finite number of quantities, while continuous distributions measure values in which additional values can be found between any two of those values. You can't have half a person, but there are many lengths of time that sit between one second and one and a half seconds. We'll explore continuous distributions later in this paper.

So, in our example of measuring salaries, we may want to know the average salary in the population. Before we make the calculation, however, we want to know if it's relevant to the overall data set. Is the average a good estimate of what to expect in a larger population? If we're not sure, how can we find out?

The **average**, or **mean**, of a measurement in a distribution is found by calculating the weighted sum of the values in the distribution and dividing by the size of the population. The mean is denoted by the Greek letter µ (mu). Figure 2 calculates this value for the distribution measured in Figure 1.

Figure 2: Computing the average (mean) for distribution of U.S. salaries

Salary	30	35	40	45	50	55	60
Count	50	60	80	80	70	30	10

Total population: 380

Product	1,500	2,100	3,200	3,600	3,500	1,650	600

Total: 16,150

Mean: 16,150/380 = 42.5

In this population, the mean salary is $42,500, but how do we know if this average is reflective of the larger population?

Assume that the characteristics of the population are changed, and the 10 people with a salary of $60,000 now have an income of $1 million. As a result of that change, the mean will jump to $67,237. We can reasonably say that number is not reflective of the actual salaries in the population, as 370 out of 380 residents earn less than that salary. For this reason, there can be problems in using the mean as a predictor of a population's attribute.

To address this possibility, two functions help assess the value of the mean as a measure reflecting the population. They are the **variance** and the **standard deviation**, customarily denoted by the Greek letter σ (sigma).

Variance measures the dispersion of values in a distribution. A larger variance indicates a wider dispersion, suggesting the mean won't be a very helpful measure of the values. The standard deviation—calculated as the square root of the variance—quantifies the mean's value as a wider measure.

In evaluating measurements, the expression "within one sigma" means that the value measured deviates from the mean—above or below—by less than the value of the standard deviation. The standard deviation thus serves as a measurement of dispersion: the lower the deviation, the tighter the measurement. Therefore, a lower deviation means the mean is more representative of the population.

Additionally, the standard deviation plays a large role in assessing the potential margin of error—which we'll explain shortly—in estimating characteristics of a population.

Normal Distribution

> ***The next section concerns distributions, which can be helpful in assessing the conceptual make-up of an Assisted Review project. The more similar the population is conceptually (evidenced by a steeper curve), the fewer concepts the system will need to identify and learn, which in turn could result in the project stabilizing more quickly.***
>
> ***- Constantine***

In our salaries example, the distribution was depicted by a table whose entries were salaries in units of $1,000, rounded to the nearest $5,000. Now imagine that we're measuring these salaries in one-cent increments. That table would be enormous and quite difficult to use. Clearly, when such distributions are required, a table is not the right tool.

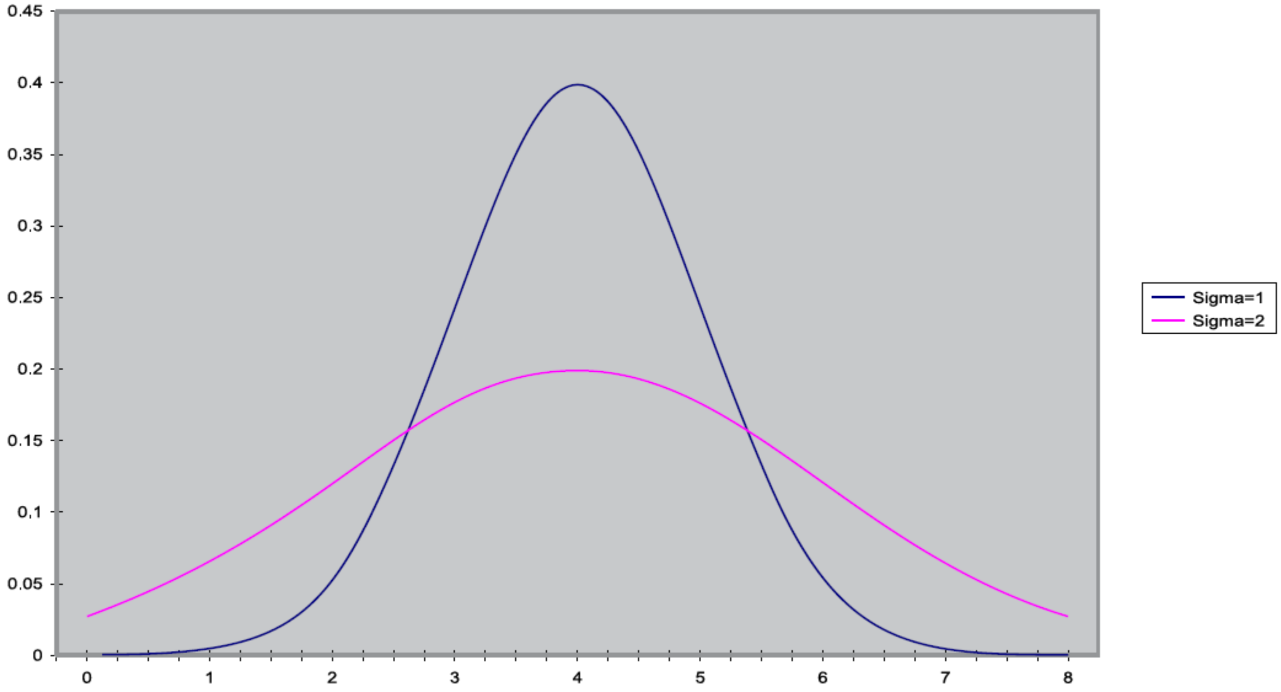
Instead, that scenario would be better described by a continuous distribution. The distribution graph is derived from a mathematical formula, which enables us to compute the **probability** of a measurement, as well as the mean of the distribution and the standard deviation.

In a discrete distribution described by a mass function, the probability of a given measurement of a value is equal to the relative frequency of that value. This is computed by the count divided by the size of the total population. In a continuous distribution, the probability is computed by the area below the distribution graph

The **normal distribution** is an example of a continuous distribution, and describes many natural phenomena. In particular, it describes the behavior of repeated measurements of the mean of any population characteristics.

The normal distribution is defined by two parameters: the mean μ and the standard deviation σ. Its graph is a bell curve, centered around the value μ. The width is shaped by the value of σ. The smaller the value of σ, the more concentrated and narrow the shape of the graph will be. This is what we expect, as we already noted that small values of σ denote less dispersion, while the large values signal large deviation from the mean. Figure 3 shows two normal distributions, with a common mean but with a different standard deviation.

Figure 3: Normal distributions



Sigma-1: μ = 4, σ = 1. The small standard deviation makes this graph narrow, indicating that the mean is more representative of the entire population.
Sigma-2: μ = 4, σ = 2. A larger standard deviation means a broader distribution, which indicates a more diverse set of data and a less representative mean.

The normal distribution has some important properties. Irrespective of the values of μ and σ, for a large number of measurements, about 68 percent will fall within one sigma, about 95 percent will fall within two sigma, and 99.73 percent will fall within three sigma of the mean. We'll use this prevalent distribution to explain and illustrate the concepts of inferential statistics.

Inferential Statistics

> ***The following section concerns inferential statistics—the ability to draw conclusions about a larger population (census) by analyzing a smaller sample. Learning about the whole by analyzing a portion is how time and cost savings are possible with computer-assisted review.***
>
> ***This section also addresses the differences between random and stratified sampling, both of which are employed in Assisted Review projects. Random sampling is utilized when seeking good coverage of an unknown population, whereas stratified sampling can be used when a certain subset of documents is known to be especially rich in content that will be useful to train the system.***
>
> ***- Constantine***

Using probability computed from a distribution, we can compute the likelihood that a certain value will be measured in a large population. But how can we approach it from the opposite direction? That is, given measurements for some part of the population, can we infer values for the whole population? This inference is integral to computer-assisted review, where we use sample sets to make decisions on much larger data collections. Let's dive in and see how it works.

The complete set of values for all the members of a population is called the **census**. It is easy to collect the census if you can perform all measurements, but that may be costly or impossible. In that case, can we get an estimate of the value without having a census? If so, what is the accuracy of that estimate?

We'll begin by performing a partial census on a small, accessible part of the population. We call this subset the **sample**. A sample can be selected in two principal ways:

1. If all members of the population possess the same property without any subdivision, we can create a **random sample**. In such a sample, every member of the population has the same likelihood to be included.
2. In some cases, members of the population are divided into groups characterized by a second property. This is a stratified population, and each stratum is characterized by different values of the distinguishing property. In that case, we would approach each stratum separately, performing random sampling in each of the strata in which we are interested.

To illustrate this concept, let's consider assessing the average height of all students of a university. We can—but we do not need to—stratify the students by gender, major, or year. If we do so and then sample each stratum, our results would be applicable to each stratum separately, calculating average height by gender, major, and class. If we do not separate the strata, we would be measuring the average of the entire population.

Confidence Intervals

> ***Confidence level and confidence interval directly affect how many documents are reviewed in an Assisted Review project. The larger the confidence level, the more documents need to be sampled. The inverse is true for confidence interval: the higher the interval, the fewer documents need to be sampled.***
>
> ***Settings for these values will vary from project to project based on resources available and other situational variables.***
>
> ***- Constantine***

One of the basic theorems of statistics—the Central Limit Theorem—states that if multiple samples are randomly chosen and the mean of each sample is computed, the result will have a normal distribution. Additionally, the distribution's mean will be equal to the mean of the whole population.

This means that, based on the means of samples, we can infer the mean of the whole population—with a margin of error that we can adjust by changing the number of samples. This is the role that statistics play in computer-assisted review.

In a computer-assisted review project, a case team wants to know what percentage of documents in their collection is relevant to the matter. Let's assume that the documents are not stratified. For the mathematical calculations, we'll assign a value of 1 to each responsive document, and 0 to non-responsive documents. That way, the mean of this variable is the fraction of the documents that are relevant, and that mean will reflect the percentage of the relevant documents.

According to the Central Limit Theorem, after computing the mean of each sample, we'll find they fit into a normal distribution. That distribution's mean will also indicate the percentage of relevant documents, while the standard deviation will be related to the deviation of the entire distribution of all documents.

But how can we tell if our computation of the mean is accurate? Two concepts that are particularly relevant to computer-assisted review are **confidence interval** and the associated **confidence level**.

A confidence level provides a measure of assurance that the estimated value of a measurement will fall within a given interval of values. That assurance is given as a percentage, and the interval itself measures the error spread. The length of the interval—or the range of the potential error—depends on the actual value that we wish to estimate, the number of samples, the probability distribution of the data, and our level of confidence in the margin of error.

The confidence level measures our confidence that in a large number of measurements, the actual value measured will be within the margin of error.

Once we select the confidence level—which is defaulted at 95 percent for many real-world cases—we can compute the margin of error. That will be our measure of the potential error, so the estimates will be described as the computed value plus or minus the margin of error. In computer-assisted review, a 95 percent confidence level would mean that, in 95 percent of our estimates, the true value of overall accuracy will be within the reported error.

A single formula relates the confidence level, sample size, confidence interval, and standard deviation for most probability distributions. Once three of these numbers are known, the fourth can be computed accordingly. In a computer-assisted review project, the confidence level and the allowed error are chosen by the case team, and the standard deviation may be known or estimated based on the mathematics previously described. Therefore, with three data points determined, we can compute the sample size that will assure the specified confidence level and the error bound.

Concluding Remarks: Estimation and People Confidence

The basis of the statistical processes we've described is anchored in two assumptions:

1. We have an adequate number of observations on properly selected samples.
2. We cast all of our measurements into the measurement of a mean, enabling us to use the Central Limit Theorem and the resulting normal distribution, as well as compute the sample size to suit the selected confidence level and confidence interval.

The nature of our sampling and the number of samples affects our accuracy. In statistics, the measurement will always have an inherent error. That can only be avoided if every member of a population is measured—which, in the case of computer-assisted review, could mean a manual review of millions of documents. An expert review of sufficient sample documents, however, is a defensible and statistically valid way of tackling a large data set more quickly and cost-effectively.

This brings us to the question of people confidence, or how to interpret statistical estimates. Statistics do not lie, as the resulting predictions are mathematically reliable. That said, human error does exist. In most cases, it occurs as sampling or data errors that can affect these calculations.

Regarding the second type of error, because data is gathered by subjective methods—like surveys people may not convey the truth. In an election poll, for example, a voter may say that they will vote for one candidate but, in the end, submit their ballot for another. These types of errors are not of concern in the mathematics behind our statistics, but they do affect the belief of the populous in statistically inferred results. This can occur in computer-assisted review when the definition of relevance changes: the calculated predictions may be mathematically sound, but if the matter has changed course, that information may no longer be important to the case.

Additionally, when it comes to computer-assisted review, sampling errors can be avoided if a case team has closely adhered to the parameters of relevance when coding documents and training the computer. If a small number of subject matter experts are inputting those original examples for the system, fewer miscalculations are likely to occur during categorization.

Finally, case teams who approach a computer-assisted review project with these statistics in mind can make thoughtful decisions on how much expert time should be spent on reviewing example documents, how many samples will be reviewed, and what kind of results they need to consider the project complete. Approaching a project with these items in mind—as well as the previously mentioned defenses against error—will yield more accurate statistics, fewer disagreements, and more reliable results.

kCura®

231 South LaSalle Street, 8th Floor, Chicago, IL 60604
T: 312.263.1177 • F: 312.263.4351
sales@kcura.com • www.relativity.com

© kCura LLC. All rights reserved.

Glossary

Attribute — a property measured for members of a given population

Census — the complete set of values of all attributes for the entire population

Confidence interval or margin of error — a measure of the potential error in an estimate

Confidence level — a measurement, expressed as a percent, that quantifies confidence in an estimate's actual value to be within the margin of error

Continuous distribution — a distribution whose data can have any value in a range

Discrete distribution — a distribution whose data has a finite number of values that can be counted

Mass function — a table that contains all possible values of the measured attribute and the number of members in the population that have that value

Mean — the average value of measurements in a population

Normal distribution — a continuous distribution which describes the behavior of many natural phenomena, and also the behavior of repeated measurements of the mean in a population

Population — a set of items with at least one common property

Probability — the frequency or possibility that a given measurement will occur in a population

Sample — a subset of a population

Standard deviation — a quantification of how much dispersion from the average exists in a distribution

Stratification —t he division of members of a population into groups characterized by a second property

Variance — a measurement of the dispersion of values in a distribution