Westlaw Today

# Statistical sampling in 'Big Data' litigation

By Carlos D. Brain, Joseph B. "J.B." Doyle, and Robert Doles, Cornerstone Research

OCTOBER 24, 2025

Statistical sampling, already a common tool in legal disputes, is seeing increasing applications as the collection and analysis of big data become ubiquitous. Sampling is often used where reviewing or analyzing the entire population of interest would be prohibitively expensive or time-consuming. Sampling methodologies can circumvent this issue by selecting a sample of the data to analyze and then utilizing statistical methods to extrapolate the results to the broader population.

This approach has been accepted in a variety of legal contexts as a practical and cost-efficient way to answer case-specific questions, typically related to establishing liability and estimating damages. In this article, we discuss recent trends in the applications of statistical sampling in legal disputes, particularly in the context of the pervasiveness of big data.

## Applications of statistical sampling in 'Big Data' litigation

The advent of big data has introduced new complexities in litigation, particularly in cases involving massive datasets. Many companies, especially those in the technology sector, are increasingly generating large amounts of data daily and incorporating these data into their core business practices.

The largest social media networks, for example, have billions of users who spend over two hours per day on average on social media platforms, generating data.[1] Many other large technology companies similarly collect and store data in large quantities, but such practices can also be found in other sectors, such as healthcare, which is estimated to generate nearly a third of all data.[2]

In cases where allegations relate to broadly defined universes of the data that companies collect, it may be impractical to analyze *all* relevant data. In particular, the data may be too large to analyze using desired analytical methods while keeping computation time (i.e., the amount of time a program takes to run) reasonably low.

While the cost of computational processing power has continued to decline substantially, such advances have not obviated the need for sampling where sufficiently large datasets are involved. As a result, sampling is increasingly being used to mitigate computational and storage constraints by analyzing only a sample of the data at issue.

Additionally, sampling may be used in cases where producing the entirety of a company's proprietary data may expose it to unnecessary costs and risks. There are often good reasons to prevent counterparties from engaging in exploratory analysis to identify new potential allegations (i.e., "fishing expeditions") after receiving more data than may be necessary in a given matter.

*The advent of big data has introduced new complexities in litigation, particularly in cases involving massive datasets.*

Potential data mining exercises along these lines are prone to producing spurious correlations, which, in litigation contexts, could lead to costly and unnecessary disputes. Further, for many companies, their valuation can be tied in large part to the confidential and proprietary nature of their data. Thus, producing all, or a large portion, of a company's data to counterparties may expose it to unnecessary business risks. This issue is particularly salient given the rise of data breach litigation in recent years.

For these reasons, when the amount of at-issue data is immense, sampling can be a valuable tool to reduce the amount of data required to reach meaningful conclusions. However, there are important considerations to be aware of in such matters, including some issues that uniquely arise in "big data" litigation.

## Key statistical sampling concepts in 'Big Data' litigation

One novel statistical issue that has arisen in "big data" litigation is the ability to analyze additional samples to assess the reliability (or lack thereof) of an initial sample. In many cases where statistical sampling is applied, measuring the variable of interest in the sampled observations is expensive — often, mitigating these costs is why statistical sampling is employed in the first place.



Thomson Reuters™

Thus, in these cases, it is typically not practical to draw and analyze additional samples. Alternatively, in cases where the purpose of sampling is instead to limit computation time, measurement costs may not necessarily be high. As a result, selecting, measuring, and analyzing a new sample might be relatively inexpensive. In such cases, rebuttal experts may draw and analyze additional samples to assess the reliability of a sampling analysis.

*When there are meaningful differences across strata, stratified sampling can be more efficient than simple random sampling, and it can allow the researcher to achieve the same level of precision with a smaller sample size.*

If materially different results are obtained by analyzing alternative samples (e.g., due to issues in how the sample was drawn or due to the quality of the sampling methodology), then it is possible the original sample was not sufficient to reach reliable conclusions.

While "big data" litigation typically involves massive amounts of data, one important — but potentially counterintuitive — concept to understand is that extremely large datasets do not necessitate extremely large samples to perform statistical inference (e.g., estimating a proportion of observations with a certain feature or defect). That is, required sample sizes do not scale linearly with the size of the relevant dataset.

For example, to achieve a five percentage point margin of error for estimating a proportion at the 95% confidence level for a binary (i.e., "yes/no") variable, one only needs a maximum sample size of 385, regardless of the size of the underlying *population*.[3] Indeed, the required sample size to achieve those parameters quickly approaches 385 as the population size increases:

| Population Size | Required Sample Size |
|---|---|
| 100 | 080 |
| 1,000 | 278 |
| 10,000 | 370 |
| 100,000 | 383 |
| 1,000,000 | 384 |
| 10,000,000 (and higher) | 385 |

Hence, while it is sometimes argued, for example in discovery discussions, that a sample should reflect a certain percentage of the underlying population, these arguments are often misplaced.

As is the case in virtually all statistical sampling contexts, sample representativeness is often a key issue when sampling from large datasets, and arguments related to representativeness have already come up in "big data" matters. If a sample is not representative of the underlying population, then it may not be appropriate to extrapolate findings from the sample.

For example, if a sample is drawn by sampling all data within a list of selected dates, then it would be important to establish that the selected dates could produce a sample representative of the broader population. Additionally, in "big data" matters in which all at-issue data are not produced, it may not be possible to conduct standard representativeness tests; in such cases, it may be particularly relevant to understand the method by which the sample was drawn.

One statistical tool that can be valuable in "big data" matters is stratified sampling, which can help ensure a sample is representative along key dimensions. To employ stratified sampling, the population of interest is divided into subgroups (or "strata"), and separate samples are drawn from each stratum.

Typically, the strata are chosen based on factors that are either expected to be related to the variable of interest or are related to the goals of the analysis. For example, if the goal of an analysis is to compare two different segments of the population, the researcher can use stratified sampling to ensure sufficiently large samples of each segment are present to facilitate the comparison.

Similarly, if representativeness of the sample across certain categories is deemed critical to the study, the researcher can use stratified sampling to ensure the proportions of subjects in the sample across those categories match those in the population.

When used properly, stratified sampling can also increase the precision of an analysis. In particular, if there are certain key variables that are expected to drive the outcome of interest, stratifying on those variables can be more efficient than simply drawing randomly from the full population. This is because stratification divides the sample into subgroups (some of which may be sparsely populated) within which variability in the outcome of interest is reduced compared to the variability in the overall population.

For this reason, when there are meaningful differences across strata, stratified sampling can be more efficient than simple random sampling, and it can allow the researcher to achieve the same level of precision with a smaller sample size. Big data cases are particularly well suited to take advantage of these efficiency gains where it may be possible to inexpensively analyze an initial sample to get preliminary estimates about

variation within strata that can be used to inform the ultimate research design.

## Conclusion

In sum, the continuing rise of big data has yielded an increasing range of new applications for sampling in a variety of litigation contexts. Understanding the potential benefits of sampling, as well as the issues that often arise when sampling methods are employed, can be helpful for litigators working on matters involving very large datasets.

For instance, in many cases, disputes often relate to whether the sample is adequately representative of the target population. As a result, more complex sampling techniques, such as stratification, which are designed to ameliorate these concerns as well as reduce the cost of sampling, can be potentially worthwhile approaches.

*The views expressed herein do not necessarily represent the views of Cornerstone Research.*

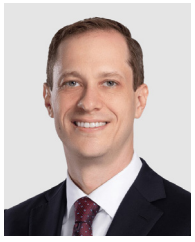
## Notes:

[1] "Most popular social networks worldwide as of February 2025, by number of monthly active users," *Statista*, available at https://bit.ly/3KWCqEH; "Daily time spent on social networking by internet users worldwide from 2012 to 2025," *Statista*, February 2025, available at https://bit.ly/476NJTo.

[2] "The healthcare data explosion," *RBC Capital Markets*, available at https://bit.ly/49gd4LN.

[3] Calculations are based on certain assumptions and are for illustrative purposes only. For ease of exposition, calculations ignore other factors that may be relevant to any given case.

## About the authors


**Carlos D. Brain** (L) is a vice president and the head of the Silicon Valley office of **Cornerstone Research**, where he applies economic and statistical analytical methods. He can be reached at cbrain@cornerstone.com. **Joseph B. "J.B." Doyle** (C) is a principal specialist in the firm's Boston office. As a member of the firm's applied research center, he consults on matters involving sophisticated statistical analysis and modeling and can be reached at jdoyle@cornerstone.com. **Robert Doles** (R) is a manager in the Boston office. He provides economic and financial analysis and expert support in commercial litigation, with expertise in statistical sampling, private equity and venture capital, and mutual funds. He can be reached at rdoles@cornerstone.com.

**This article was first published on Westlaw Today on October 24, 2025.**