We Mean Business:

Turning Validation Metrics into Defensible Decisions

May 5, 2026

Speakers

Stephanie Clerkin

Director of Litigation Support
Korein Tillery

Matt Mahon

Vice President of Client Solutions
Level Legal

level legal®

Agenda

- Foundation
- Key Concepts
- Pre-Process Considerations
- Case Studies
- Q&A

level legal®

Hot Takes

Technology won't slow you down. Human sign-off will.

We need to stop arguing the mechanics of document review and shift the focus to agreeing on a defensible validation protocol.

level legal®

Foundation

Disclaimer

- A singular methodology for validating the results of a technology-assisted review workflow (including GenAI workflows) **doesn't exist.**
- You need to consider:
  - Richness
  - Proportionality
  - Edge Cases
  - Risk Tolerance
  - Custodians
  - Data Sources
  - File Types
  - Foreign Language

level legal®

The Confusion Matrix

		Actual: Positive	Actual: Negative
Predicted	Positive	**True Positive** Relevant documents tagged as relevant	**False Positive** Non-relevant documents tagged as relevant
	Negative	**False Negative** Relevant documents tagged as non-relevant	**True Negative** Non-relevant documents tagged as non-relevant

level legal®

Key Concepts

Richness / Prevalence

Definition: The percent of a population that has a specific characteristic, such as responsiveness.

- Measures the % of relevant documents within a document population.

- Informs review strategy and workflow decisions.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Richness / Prevalence

$$Richness = \frac{True\ Positive + False\ Negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive}$$

level legal®

Recall

Definition: The number of documents retrieved from a search divided by all of the responsive documents in a collection.

- Measure of completeness of your review.

- *Remember:* It's always an estimate unless the full set of documents is reviewed.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Recall

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Precision

Definition: The number of true positives retrieved from a search divided by the total number of results returned.

- Measure of the accuracy of your review.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Precision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

level legal®

F-Measure / F Score / F1 Score

Definition: A measure of a search's accuracy calculated by using precision and recall.

- A weighted average that balances recall and precision.

- In eDiscovery, recall is often more important.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

F-Measure / F Score / F1 Score

$$\text{F-Measure} \\ \text{F Score} \\ \text{F1 Score} = 2x \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

level legal®

Error

Definition: The fraction of documents that are incorrectly coded by a search or review effort.

Keep in mind:

- False positives and false negatives are not equally harmful.

- The error rate itself doesn't tell you the type of error.

- Context matters (e.g., privilege review vs. responsiveness review).

Source: *The Grossman-Cormack Glossary of Technology-Assisted Review*

Error

$$Error = \frac{False\ Positive + False\ Negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive}$$

Random Sampling

Definition: The process of selecting data from a population with no bias or input from the person performing the sampling, in which each item has an equal chance of being selected as any other item.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Stratified Sampling

Definition: A method of data sampling where data is initially divided into subgroups (e.g., by age range or a geographic criteria) or strata, and then each group is sampled in order to ensure that each subgroup is properly represented.

Useful for:

- Different custodians.
- Different document types or sources.
- Uneven document distributions.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Confidence Level

Definition: The percentage of samples for which the results are expected to correctly describe a population parameter within a provided confidence interval.

- Be careful: Percentages can hide large numerical ranges.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Confidence Interval / Margin of Error

Definition: The range of values that is likely to contain the true parameter for a population to the specified confidence level.

- Be careful: The margin of error range grows significantly with large populations.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Point Estimate

Definition: The result of a sample that estimates prevalence in the specific population being sampled.

- Used to estimate collection richness and the total number of relevant documents.

- Be careful: Percentages alone can be misleading.
  - Always convert percentages to document counts.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Elusion

Definition: The percentage of documents of a search's null set that were missed by the search, usually determined with review of a random sample of the null set.

- While widely used, it still has issues:
  - Can be difficult to interpret without total number of relevant documents.
  - The rate can imply different outcomes depending on context.

- Don't rely on elusion alone.

Source: *The Sedona Conference Glossary: eDiscovery and Digital Information Management, Fifth Edition*

Pre-Process Considerations

Before You Start

Questions you should ask to guide your methodology:

- What metrics are most important and why?
- What processes address the impact of those metrics?
- Have opposing counsel or the court set expectations around validation methodology or recall targets?

level legal®

GenAI Considerations

- Will you use GenAI in your workflows?
- How do you ensure that your use of GenAI stands up to mathematical and legal scrutiny?
- How will you validate the GenAI output?
- Does your GenAI workflow require court approval or opposing counsel agreement?

level legal®

Pre-Process Checklist

- Agree on review platform and any GenAI / TAR tools to be used.
- Define scope: custodians, date ranges, data sources, and file types.
- Is there an expected richness of the document population?
- Determine confidence level and margin of error for validation sampling.
- Define what constitutes a "responsive" document for this matter.
- Confirm opposing counsel / court agreement on methodology where required.

level legal®

Case Studies

Case Study #1:
In re Broiler Chicken Antitrust Litigation

> *The court ordered the parties to combine three subsamples of 3,000 documents into a single validation sample. The sample was to be reviewed and coded by a subject matter expert, and the producing party was to prepare a table listing each of the 3,000 documents. The requesting party and the special master were to be provided with the table, a copy of each responsive, non-privileged document in the validation sample, and the statistics and recall estimate. The recall estimate was to be computed to inform the decision-making process.*

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Subcollection C(1)

Documents identified by the review as responsive to at least one Request for Production, including any privileged documents, but not including family members of responsive documents, unless those family members are deemed to be responsive in their own right.

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Subcollection C(2)

Documents coded as non-responsive by a human reviewer, regardless of how the documents were selected for review (e.g., by TAR, manual review, or otherwise).

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Subcollection C(3)

Documents excluded from manual review as the result of a TAR process. If the review process involved only manual review and no TAR, the Collection will not include Subcollection C(3).

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

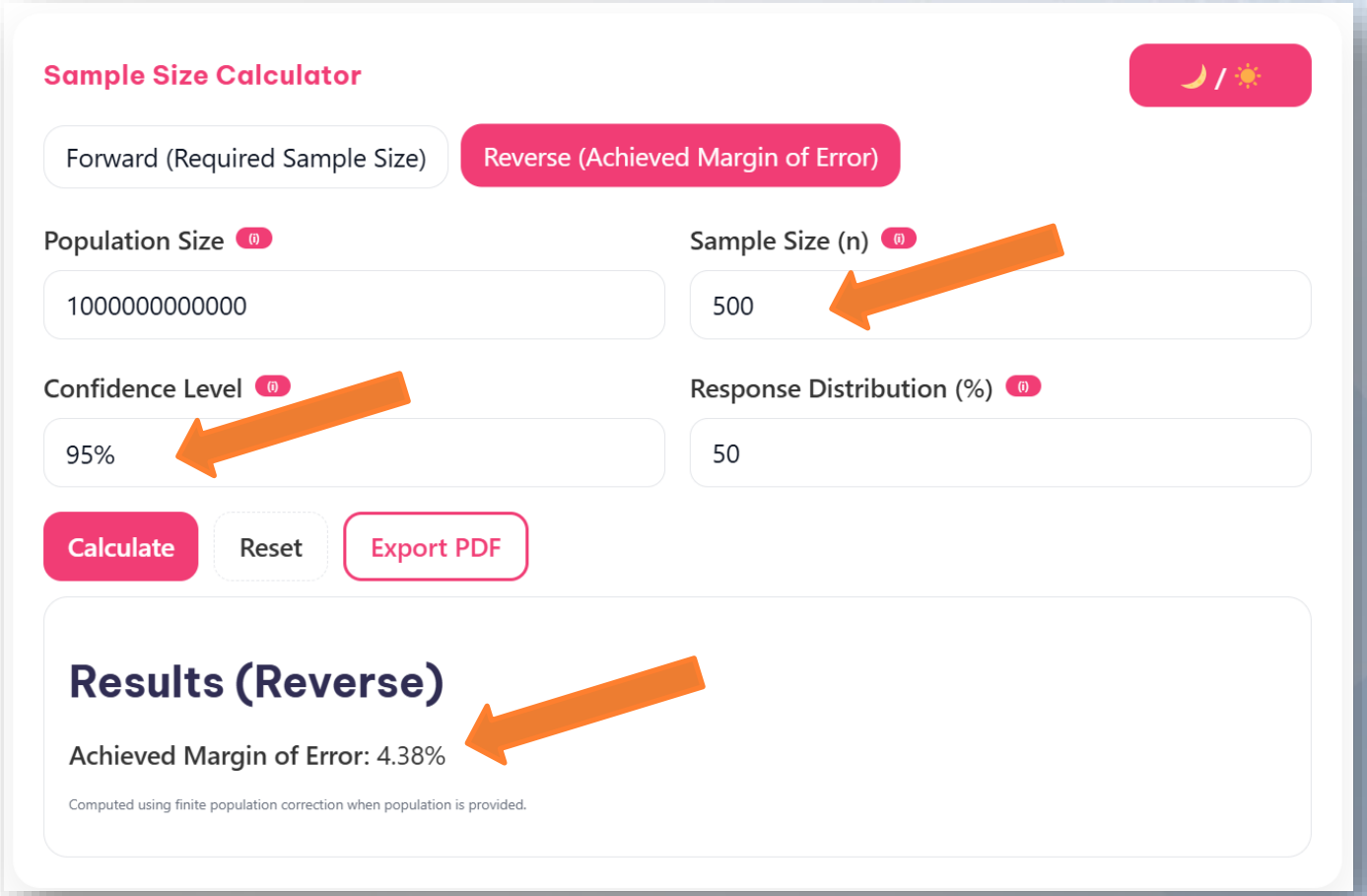
Stratified Random Samples

- **Subsample D(1):** 500 documents selected at random from Subcollection C(1).
- **Subsample D(2):** 500 documents selected at random from Subcollection C(2), if TAR was used, otherwise 2,500 documents selected at random from Subcollection C(2), if manual review was used.
- **Subsample D(3):** 2,000 documents selected at random from Subcollection 1(c) if TAR was used.

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

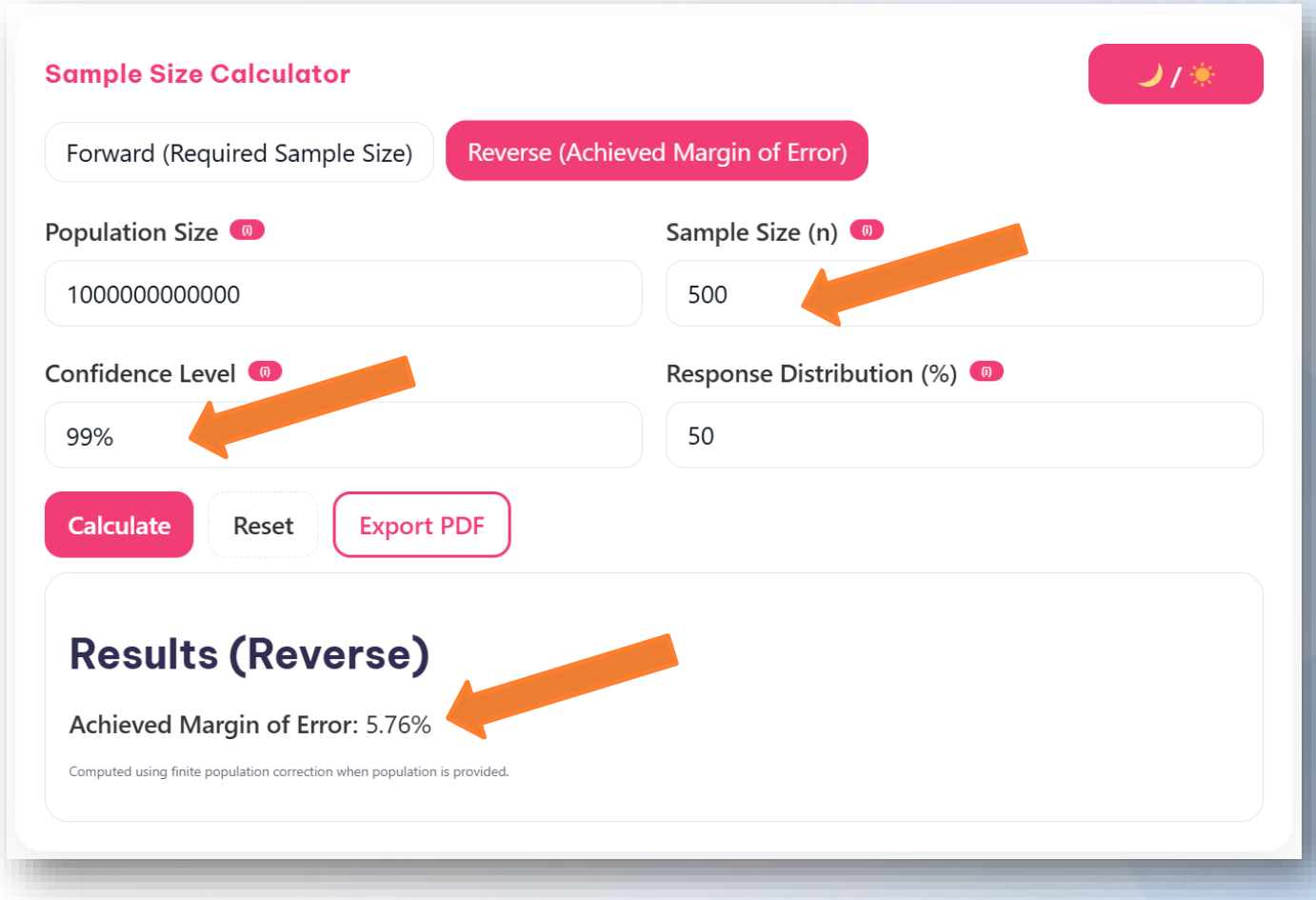
Were the Subsample sizes selected at random?



Source: [Raosoft Sample Size Calculator 2026]

Case Study #1:
In re Broiler Chicken Antitrust Litigation

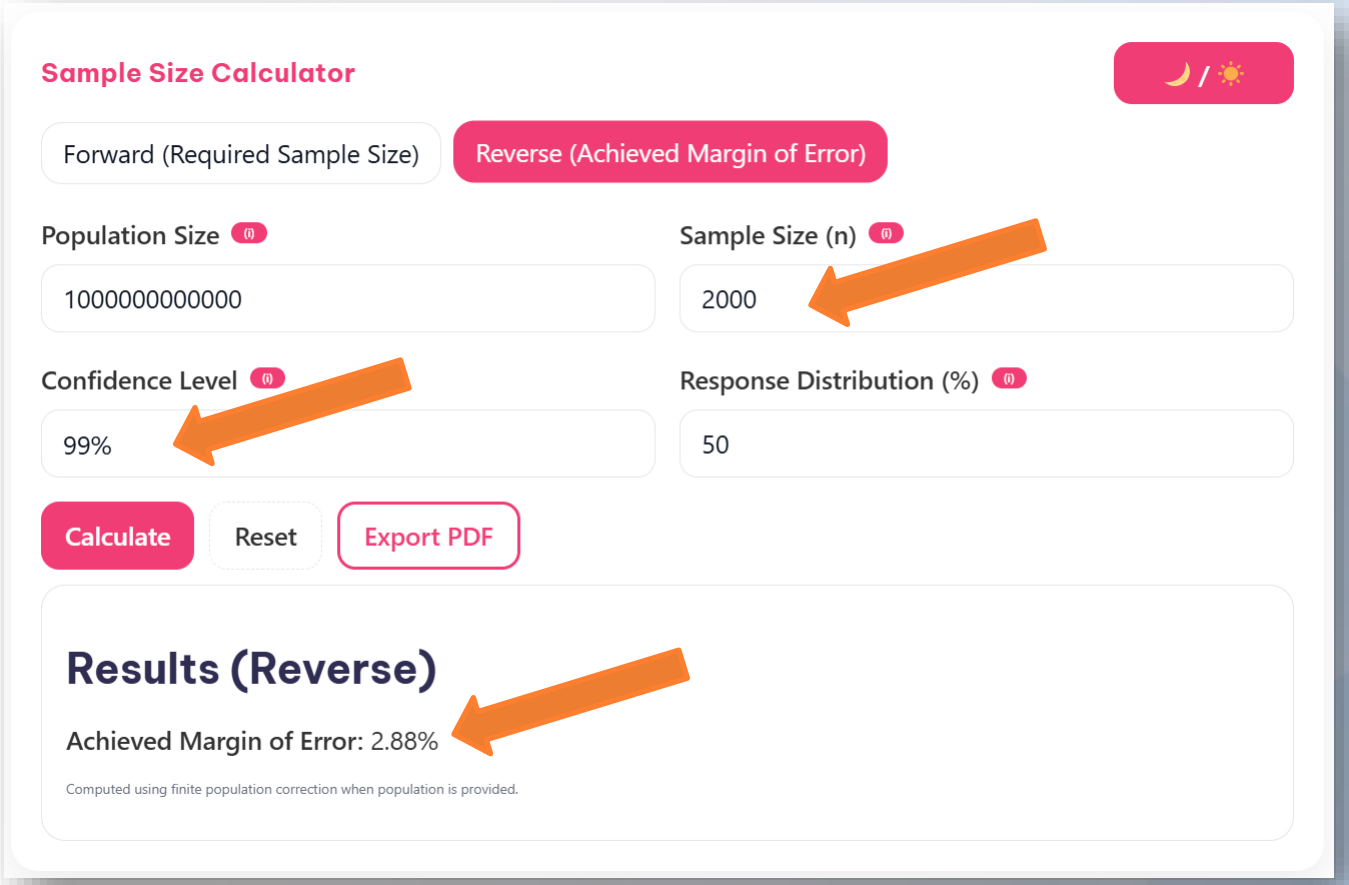
Were the Subsample sizes selected at random?



Source: [Raosoft Sample Size Calculator 2026]

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Were the Subsample sizes selected at random?



Source: [Raosoft Sample Size Calculator 2026]

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Validation Sample: Ensure the Review Is Blind

- The sample of 3,000 documents comprised of the documents from Subsamples D(1), D(2), and, if TAR was used, D(3), shall be combined into a single Validation Sample, with no indication of the Subcollection from which the documents were derived or how they were previously coded. The Validation Sample shall be reviewed and coded by a subject matter expert ("SME") who is knowledgeable about the subject matter of the litigation.

level legal®

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Recall Estimation Method

- The number of responsive documents found ≈ the size of Subcollection C(1) × the number of responsive docs in Subsample D(1) ÷ 500.
- The number of responsive documents coded incorrectly ≈ the size of Subcollection C(2) × the number of responsive documents in Subsample D(2) ÷ 500.
- The number of responsive documents not reviewed ≈ size of Subcollection C(3) × the number of responsive documents in Subsample D(3) ÷ 2,000.
- **Estimated recall** ≈ the number of responsive documents found ÷ (the number of responsive documents found + the number of responsive documents coded incorrectly + the number of responsive documents not reviewed).

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation

- Assume the number of documents in the collection (defined as including all documents identified for review for responsiveness and/or privilege following the application of keywords or other culling criteria) = 100,000 documents

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation

- Collection = 100,000 documents
    - Subcollection C(1) = 30,000 documents
    - Subcollection C(2) = 20,000 documents
    - Subcollection C(3) = 50,000 documents

level legal®

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation

- True Positive = 30,000 x 450
  - Subcollection C(1) = 30,000 documents
    - Subsample D(1) = 500 documents
      - **# of responsive documents in D(1) = 450 = True Positive**
  - Subcollection C(2) = 20,000 documents
    - Subsample D(2) = 500 documents
      - **# of responsive documents in D(2) = 25 = False Negative$_{D(2)}$**
  - Subcollection C(3) = 50,000 documents
    - Subsample D(3) = 2,000 documents
      - **# of responsive documents in D(3) = 200 = False Negative$_{D(3)}$**

level legal®

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation

- The # of responsive documents found ≈
  # of documents in C(1) x # of responsive documents in D(1) ÷ 500
  => 30,000 x 450 ÷ 500 = 27,000 documents = True Positive
- The # of responsive documents coded incorrectly ≈
  # of documents in C(2) x # of responsive documents in D(2) ÷ 500
  => 20,000 x 25 ÷ 500 = 1,000 documents = False Negative$_{D(2)}$
- The # of responsive documents not reviewed ≈
  # of documents in C(3) x # of responsive documents in D(3) ÷ 500
  => 50,000 x 200 ÷ 2,000 = 5,000 documents = False Negative$_{D(3)}$

level legal®

Case Study #1:
In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation: Precision and Recall

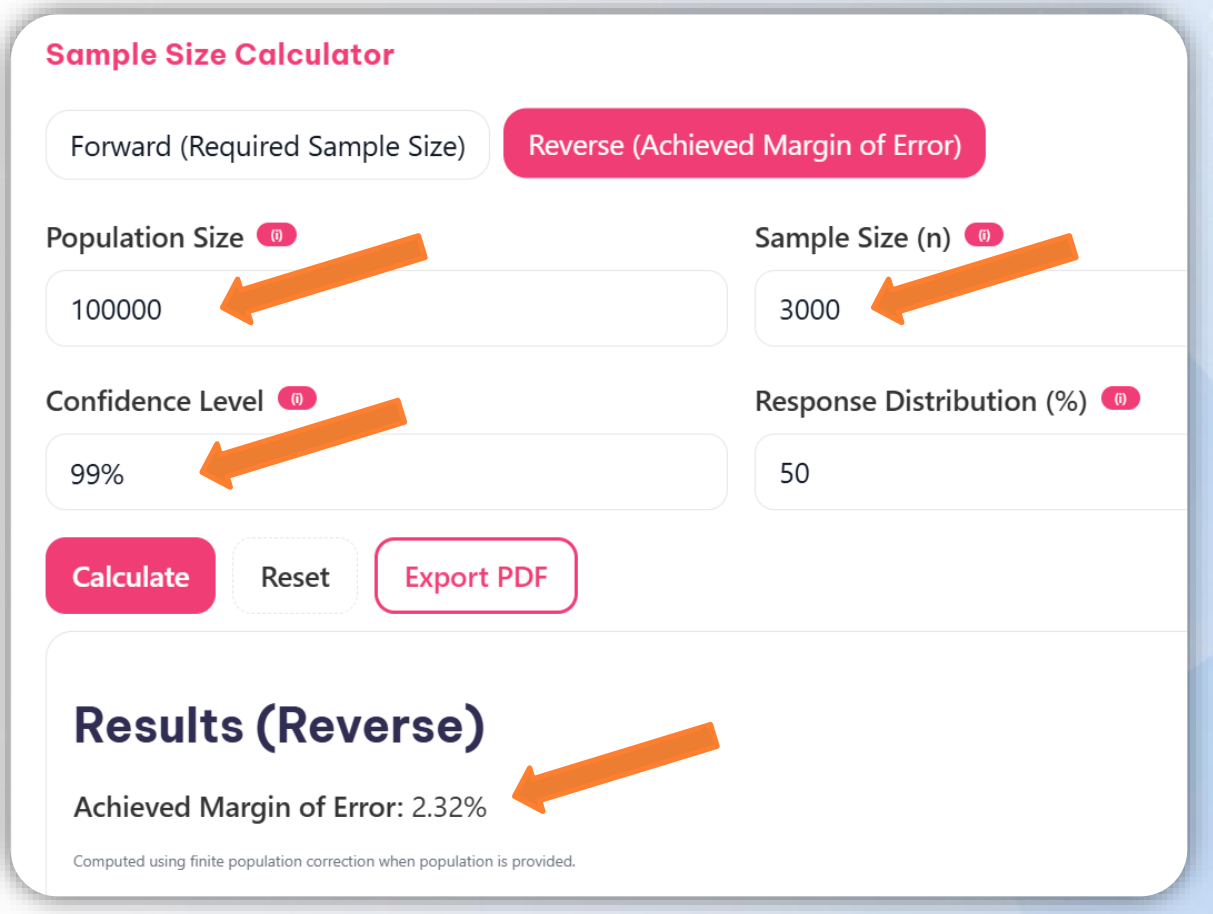
- Recall = True Positive ÷ (True Positive + False Negative)
  => Recall = 27,000 ÷ (27,000 + 1,000 + 5,000) = 81.82%
- Precision = True Positive ÷ (True Positive + False Positive)
  => Precision = 27,000 ÷ (27,000 + 3,000) = 90%

level legal®

Case Study #1:

In re Broiler Chicken Antitrust Litigation

Hypothetical Calculation: Confidence Interval / Margin of Error



Case Study #1:
In re Broiler Chicken Antitrust Litigation

Did We Get The Validation We Need?

- Evaluate the Relevance of the False Negatives
- Calculate the Recall Range
  - Point Estimate is 33,000 documents
  - Margin of Error (Confidence Interval) = +/- 2.32%
  - Estimated # of Relevant Documents = 33,000 +/- 2.32% x 100,000
    => 33,000 +/- 2,320
    => Estimate # of Relevant Documents = [30,680, 35,320]
  - Recall Range = [27,000 ÷ 35,320, 27,000 ÷ 30,680]
    => Recall Range = [76.44%, 88.01%]

level legal®

Case Study #2:
Applying Broiler Chicken To GenAI Reviews

Validation Framework: Subcollections

- **Subcollection C(1):** Documents coded as responsive by a human reviewer.
- **Subcollection C(2):** Documents coded as responsive by GenAI.
- **Subcollection C(3):** Documents coded as non-responsive by a human reviewer.
- **Subcollection C(4):** Documents coded as non-responsive by GenAI.
- **Subcollection C(5):** Documents in the review population excluded from manual and GenAI review.

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #2:
Applying Broiler Chicken To GenAI Reviews

Validation Framework: Stratified Random Samples

- **Subsample D(1):** 500 documents selected at random from Subcollection C(1).
- **Subsample D(2):** 500 documents selected at random from Subcollection C(2).
- **Subsample D(3):** 500 documents selected at random from Subcollection C(3).
- **Subsample D(4):** 500 documents selected at random from Subcollection C(4).
- **Subsample D(5):** 2,000 documents selected at random from Subcollection C(5).

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #2:
Applying Broiler Chicken To GenAI Reviews

Validation Framework: Other Subsample Considerations

- Key Custodian(s)
- Data Source(s)
- File Type(s)
- Date Range(s)
- Foreign Language
- Search Term Hits

Source: *In re Broiler Chicken Antitrust Litigation* (2018)

Case Study #3:
Inbound Production

Sampling and Validation For Documents Produced by Third Parties

- **800,000 documents received from a defendant's production:** TAR was applied instead of linear review for deposition prep and expert work.
- **Random + Stratified Sampling:** random for richness baseline; stratified by custodian.
- **3% Elusion:** withing threshold; model accepted.
- **Worthless form Templates:** pulled into the model, caught by precision sampling, excluded without hurting recall.
- **Validation:** makes TAR an easier sell to case team and more defensible.

level legal®

Q&A

Level Legal is an eDiscovery, managed review, and consulting company that delights law firms, corporations, and government agencies through industry-best customer service. This concierge approach to outsourced legal services delivers peace of mind.