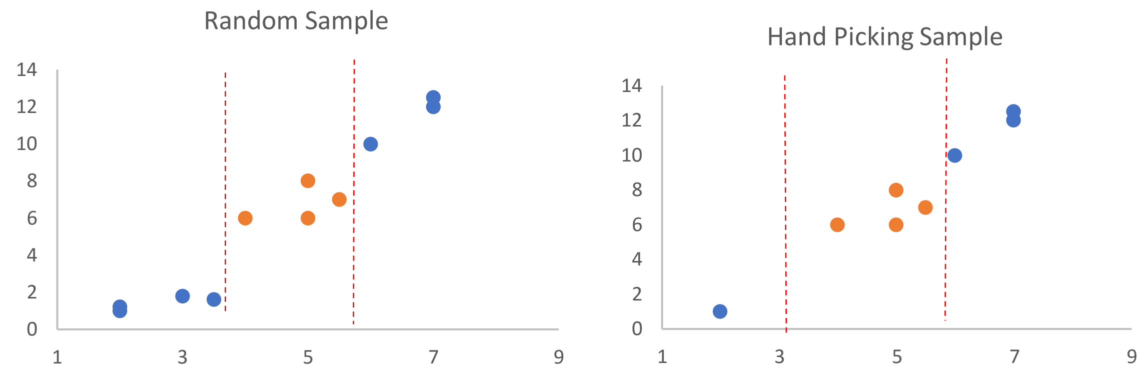
Setting up Datasets - Sampling

A crucial step of setting up datasets is to sample subsets of the population to accurately and truthfully represent the huge corpuses of documents in training, validation, and testing.

The representativeness and accuracy of the sample hinge on two factors: 1) the number of documents sampled, 2) sampling technique used. By the Law of Large Numbers, larger sample sizes lead to more accurate estimations on the population parameters, such as the prevalence of responsive documents or the error rate of the classifier used. For example, a sample reveals that 5 percent of the documents are pertinent. However, without knowing whether the sample size is adequate, one cannot extrapolate the 5 percent prevalence to the entire population with certainty. Likewise, one cannot extrapolate the validation classification error rate to the test dataset without knowing whether sample size is sufficient. In practice, we balance the cost of reviewing the sampled documents and the degree of uncertainty tradeoff.

Sampling Techniques

The most common sampling technique is the simple random sampling, where each document has equal probability being sampled. Though simple random sampling is fast and easy to implement, there are certain drawbacks. If the prevalence is low, reviewers need to go through a lot more documents to obtain enough responsive documents for training the model. Non-probability sampling methods that based on human judgements and domain knowledge may help to capture more responsive documents. Yet, hand-picking documents is likely to reinforce human bias in model training. The graph below illustrates the bias problem introduced by hand-picking sampling. In this example, the hand-picked sample misses informative data points, therefore leads to incorrect decision boundaries in training.



Stratified random sampling can be used to improve the representation of each subgroup and to achieve a higher level of precision in estimating population parameters. The steps to perform stratified random sampling are:

- Divide the pool of documents into subgroups (strata) based on chosen characteristics, such as custodian or department
- Do simple random sampling within each subgroup

The smaller the sample size, the more likely that simple random sampling may underrepresent small yet important subgroups. Stratified random sampling can improve the balance of the sample. The simplified example below illustrates the benefit of using stratified random sampling. Each department represents a distinct subpopulation, where the percentage of relevant documents varies dramatically. Suppose we need to draw a random sample of 400 documents. Under simple random sampling, because of the relative small size of finance documents and the natural sampling variance, we can end up with a very small amount of finance documents and a smaller number of pertinent documents. However, under stratified sampling, we first divide the population based on department, then draw a random sample within each subgroup based on their proportion. For example, with finance counts for 5 percent of total population, we draw 20 (5%*400) documents from the finance subpopulation. Therefore, all subgroups are represented in a balanced way, leading towards more accurate estimations on population mean.

Stratification Variable - Department	Subpopulation Size	Subpopulation mean (unknown)	Stratified Sample Size	Simple Random Sample Size
Finance	1,000 (5% of total)	10%	20	5
Operation	10,000 (50%)	1%	200	210
Legal	9,000 (45%)	5%	180	185

In this example, the expected population mean estimation under the random sampling is: (5*10%+205*1%+185*5%)/400=2.75%. And the expected population mean estimation under the stratified random sampling is: (20*10%+200*1%+180*5%)/400=3.25%, which is the true population mean. Though the difference in this example is only 0.5%, it can be amplified by sampling variance.

In terms of choosing the stratification variables, the estimation will be more precise if the population is partitioned into strata in such a way that within each stratum, the units are as similar as possible. For example, different departments may have different key words in text documents, yet within a department the documents and key words are much more similar. Stratifying by department helps to ensure that different key words are well represented in the sample.

Other methods such as using clustering algorithm to partition the documents into heterogeneous groups can also be used to generate stratification variables.

Sample Size on the Validation Set

The most common method to determine the sample size required for measuring a statistic is the frequentist's approach, which assumes a normal distribution for the unknown parameter based on the Central Limit Theorem (CLT). The approach calculates the margin of error of the unknown mean of a binomial variable, for example the average prevalence of relevant documents or classification error rate. There are three inputs: 1) confidence level, 2) population prevalence (which is assumed to be 50% for a larger and more conservative sample size estimation), and 3) population size (to adjust sample size when population is small[1]). The exact equation used to determine the sample is listed below:

$$Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}*\frac{N-n}{N-1}} \leq \varepsilon$$

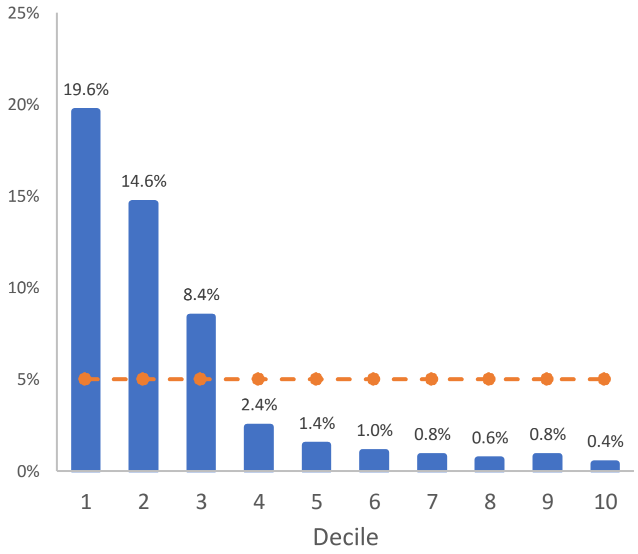
where p is the unknown population mean (assumed to be 50 percent), Z is the critical value of a normal distribution at the given confidence level, N is the population size, and n is the sample size. The right handset of the equation, ε, is the maximum margin of error that can be tolerated. When population size is very large, the finite population adjustment ($\frac{N-n}{N-1}$) can be ignored. For example, at 95 percent confidence level and with 60,000 documents in the population, the minimum sample size required is 2,309 for the margin of error to be under 2 percent.

Measuring sample mean with high certainty and small error band is essential in quality checking validation and scoring. The example below explains why this matters using *recall*, an important metric to test predictive coding model performance. Suppose we built a model on a training and a validation sample of 5,000 documents each, and the prevalence of relevant documents in the sample is 5 percent. This model is then scored on the remaining 50,000 documents in the testing set. Since the sample size is larger than the minimum requirement (see the example above), one can conclude with high certainty that the population prevalence is also 5 percent (the margin of error in this case is 0.38% percent), i.e. there are approximately 2,500 relevant documents (plus or minus 190 documents) in the testing set. After reviewing the top 30 percent documents with the highest probability of being relevant, 2,200 additional relevant documents are discovered. The estimated recall in this example is 88 percent (2,200 over 2,500), i.e. we identified 88 percent of the relevant documents with the cost of only reviewing 30 percent of the documents. The lower bound on recall in this example is 82 percent (2,200 over 2,690).

Similar to recall, lift curve is a tool to determine the optimal review methodology. All validation documents are scored and ranked based on their estimated probability, where the first decile represents the top 10 percent most relevant documents. The lift chart on the right can assist reviewers to allocate resources in the most efficient way. For example, the top 10 percent documents can be sent to more senior lawyers and the top 10 to 20 documents can be sent to more junior lawyers. The optimal allocation depends the cost of reviews by different types of reviewers.

[1] For finite population corrector: https://onlinecourses.science.psu.edu/stat414/node/264

Decile	Validation Size	# Positive	Percent Relevant
1	500	98	19.6%
2	500	73	14.6%
3	500	42	8.4%
4	500	12	2.4%
5	500	7	1.4%
6	500	5	1.0%
7	500	4	0.8%
8	500	3	0.6%
9	500	4	0.8%
10	500	2	0.4%



There are other less common approaches to calculate the sample size, such as using the Hoeffding Inequality and the Bayesian confidence interval calculation. Based on the Hoeffding Inequality, the minimum sample size required for measuring the mean of a random variable is:

$$n \geq \frac{1}{2\varepsilon^2}\log(\frac{2}{\delta})$$

where ε measures 'error band', i.e. the distance between the empirical mean from the true mean, and δ measures 'uncertainty', i.e. probability that we allow the error to exceed ε. It is easy to see that sample size required grows exponentially with decreasing error band. Thus, it is very expensive to trade for high precision. If we set our error band to be 2 percent, at 95 percent certainty (δ is 5%), the minimum sample size required is 4,611 documents.

A similar approach is to use the Chernoff Bound, which is used to approximate the generalization of a binomial distribution. Sample size required to approximate the mean of a is:

$$n \geq \frac{2+\varepsilon}{\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$$

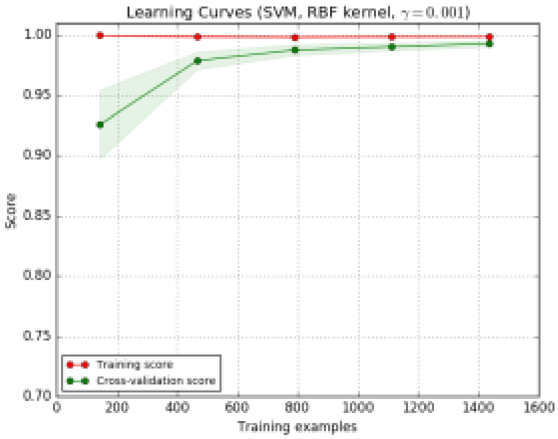
where ε measures 'error band', i.e. the distance between the empirical mean from the true mean, and δ measures 'uncertainty' (the opposite of confidence). Suppose we want to measure the sample mean with less than 2 percent error and 95 percent confidence, the sample size required would be approximately 18,600. Notice that both Hoeffding Inequality and Chernoff Bound tend to give a high number than needed (can be prove by binomial simulations).

In practice, sample size depends on not only statistical calculation but also on real-world constraint such as cost and timeline.

Sample Size on the Training Set

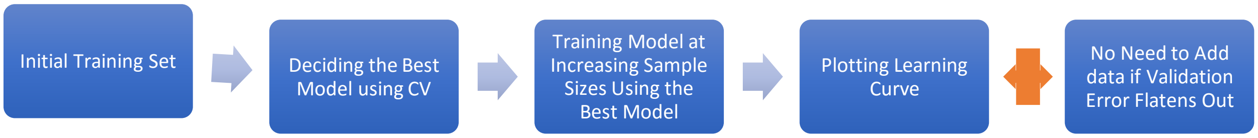
In contrast with the straightforward sample size calculation used to measure the sample mean, it is more of an art than science to determine the size of the training set. Training size required is influenced by 1) model/classifier used and its complexity 2) features space and its dimensionality. In general, more complex classifiers and higher dimensionality require more training sample. However, because both factors are not static during model training, it is difficult to know in advance how many training documents are needed.

In practice, scholars have been using validation and *learning curve* to determine if more training data are needed. Learning curve measures the performance of a classifier (training and validation error rates) in respect to the sample size. Validation error rate follows an inverse power law as training sample increases. Adding training sample data is likely to help complex classifiers to perform better. However, if the classifier suffers from high bias, increasing training data will not by itself help the performance. The graph below give an example on how learning curves can help to determine if additional sample data are needed in the training set.



Learning curves in this example is plotted using the Support Vector Machine model with fixed kernel and parameter. Y-axis is a performance metric, AUC, and the x-axis is the sample size in the training set. Validation performance improves at a decreasing rate as training set increases. The marginal benefit of adding additional training examples is small after 600 training examples. Therefore, the modeler decide that 1,400 training examples are sufficient for the model building.

A workflow to determine training sample size is shown below:


Initial Training Set
Deciding the Best Model using CV
Training Model at Increasing Sample Sizes Using the Best Model
Plotting Learning Curve
No Need to Add data if Validation Error Flatens Out

When the data are highly imbalanced, e.g. proportion of relevant documents is less than 10 percent, adding positive training examples to the training set can improve the model performance. Stratified random sampling result can assist reviewers to target subgroups with the most relevant documents, thus expedite the review and avoid going through a large corpus of irrelevant documents. Based on the sample size and sample mean within each stratum, we can estimate the margin of error by each stratum. For example, a 5,000 samples are stratified on department. In the operation department, the estimated percentage of relevant documents has an upper bound of 1.4 percent with 95 percent confidence interval, which may be too low to call for additional reviews.

Department	Population Size	Sample Size	Sample Mean	Upper Bound (95 C.I.)	Lower Bound (95 C.I.)
Finance	1000	250	10%	13.7%	6.3%
Operation	10000	2500	1%	1.4%	0.6%
Legal	9000	2250	5%	5.9%	4.1%