TITLE: BAYESIAN STRATIFIED SAMPLING TO ASSESS CORPUS UTILITY

AUTHOR(S): J. Hochberg, J. Scovel, T. Thomas, S. Hall

SUBMITTED TO: Sixth Workshop on Very Large Corpora
Montreal, Canada
August 15-16, 1998

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
Bayesian Stratified Sampling
to Assess Corpus Utility

Judith Hochberg, Clint Scovel, Timothy Thomas, and Sam Hall
{judithh, jcs, trt, shall}@lanl.gov
Group CIC-3, Los Alamos National Laboratory
MS B265, Los Alamos, NM 87545

Abstract

This paper describes a method for asking statistical questions about a large text corpus. We exemplify the method by addressing the question, "What percentage of Federal Register documents are real documents, of possible interest to a text researcher or analyst?" We estimate an answer to this question by evaluating 200 documents selected from a corpus of 45,820 Federal Register documents. Bayesian analysis and stratified sampling are used to reduce the sampling uncertainty of the estimate from over 3100 documents to fewer than 1000. A possible application of the method is to establish baseline statistics used to estimate recall rates for information retrieval systems.

1. Introduction

The traditional task in information retrieval is to find documents from a large corpus that are relevant to a query. In this paper we address a related task: answering statistical questions about a corpus. Instead of finding the documents that match a query, we quantify the percentage of documents that match it.

The method is designed to address statistical questions that are:

- **subjective:** that is, informed readers may disagree about which documents match the query, and the same reader may make different judgment at different times. This characteristic describes most queries of real interest to text researchers.

- **difficult:** that is, one cannot define an algorithm to reliably assess individual documents, and thus the corpus as a whole. This characteristic follows naturally from the first. It may be compounded by an insufficient understanding of a corpus, or a shortcoming in one's tools for analyzing it.

Statistical questions asked of small corpora can be answered derived exhaustively, by reading and scoring every document in the corpus. Such answers will be subjective, since judgments about the individual documents are subjective. For a large corpus, it is not feasible to read every document. Instead, one must sample a subset of documents, then extrapolate the results of the sample to the corpus as a whole. The conclusions that one draws from such a sampling will have two components: the estimated answer to the question, and a confidence interval around the estimate.

This paper describes a method for combining traditional statistical sampling techniques [1-2] with Bayesian analysis [3-4] to reduce this sampling uncertainty. The method is well-grounded in statistical theory, but its application to textual queries is novel. One begins by stratifying the data using objective linguistic tests designed to yield relatively homogeneous strata, within which most documents either match or do not match the query. The next step is to sample randomly within each stratum. A reader scores each selected document, and the results of the different strata are combined. If the strata are well constructed, the resulting estimate about the corpus will have a much smaller confidence interval than one based on a random sample of the corpus as a whole.
The method is well suited for subjective queries because it brings a human reader's subjective judgments to bear on individual documents. The Bayesian approach that we apply to this problem allows a second opportunity for the reader to influence the results of the sampling. The reader can construct a probability density that summarizes his or her prior expectations about the results for each stratum. These prior expectations are combined with presampling results to determine the makeup of the final sample. When the final sample is analyzed, the prior expectations are again factored in, and thus affect the estimated mean and the size of the confidence interval. Thus different readers' prior expectations, and their judgments of individual documents, can lead to substantially different results, which is consistent with the subjective probability paradigm.

In earlier work we used this method to analyze textual medical records, asking, "What percentage of the patients are female?" [5] This question was surprisingly difficult: not only did the record format lack a required gender field, but also some records did not specify the patient's gender at all, or gave conflicting clues. Consequently, the question was subjective: determining gender in these difficult cases was up to the individual reader's judgment. We stratified the medical records into mostly male documents and mostly female documents based on linguistic tests such as the number of female versus male pronouns in a medical record, then sampled within each stratum. This technique reduced the sampling uncertainty for the question from fourteen percentage points (based on an overall sample of 200 documents) to five (based on a stratified sample of 200 documents).

In this paper, we describe the method (updated from [5]) and its application to a new corpus, the Federal Register. This corpus is of general interest because it is part of the TIPSTER collection. The question we addressed is likewise of general interest: what percentage of documents are of possible interest to a researcher, or to an analyst querying the corpus? Anyone who has worked with large text corpora will recognize that not all documents are created equal; identifying and filtering uninteresting documents can be a nuisance. Estimating the percentage of uninteresting documents in a corpus therefore helps determine its utility.

2. Method

This section is organized as follows. We begin by describing the data corpus we worked with and the query we devised. We then describe two steps in finding a statistical answer to the query: first through an overall sample of 200 documents from the corpus, then through a stratified sample of 200 documents.

2.1 Data

The text corpus used in this study was the Federal Register. The Federal Register is published by the United States Government. It contains the full text of all proposed and final Federal rules and regulations, notices of meetings and programs, and executive proclamations. We used an electronic version of the Register that was part of the 1997 TIPSTER collection distributed by the Linguistic Data Consortium (http://www.ldc.upenn.edu/). It consists of 348 files, each purported to contain one issue of the Federal Register for the years 1988 and 1989. Each separate rule, regulation, etc. within an issue is considered a separate document and is bracketed with HTML markup tags <DOC> and </DOC>. The corpus contains 45,820 such documents.

There are systematic differences between the corpus and the printed version of the Federal Register. The on-line version omits boilerplate seen in the printed version, such as descriptions of the types of documents found in each section. Page numbers from the printed Federal Register are absent in the on-line version; in fact, there is no way to determine what printed page an individual Federal Register document was found on. The order of documents in the on-line and printed versions differ; for example, documents within special Parts following the main body of the printed version are intermixed with the main body of the on-line version.
Other differences are less systematic and appear to be errors in the on-line version. Some Federal Register issues are missing, and some are incomplete. Some files contain two or three Federal Register issues, and some Federal Register documents (in one case, an entire issue) appear in more than one file.

Thus the TIPSTER corpus can in no way be considered a clean, perfect electronic version of the Federal Register. Rather, it should be considered a realistic example of archival records that are not extensively edited for the purposes of information extraction research. All electronic archives will have to some degree the characteristics seen here: missing data, odd additions, truncated files, mislabeled files, etc. The methods to be explored in this paper should be designed to function in spite of these problems. If they depended on an extensive cleaning operation before they could be applied, then they would not be of general interest, since such input rarely exists in the real world -- and producing it is often prohibitively expensive. Therefore we have made no effort to clean the data, but are using the known flaws to help us judge the behavior of our methods in the presence of such phenomena.

2.2 The Query

The query described in this paper grew out of an attempt to establish basic statistics for Federal Register documents. When counting documents and determining their length, we noticed that some purported documents (as judged by <DOC> </DOC> bracketing) were not what we came to define as real Federal Register documents: documents describing the activities of the federal government. Besides real documents, the electronic Federal Register contained pseudo-documents related to the use and publication of the paper version of the Federal Register, such as tables of contents, indices, blank pages, and title pages.

This discovery at first appeared to be a mere nuisance. We assumed that there was an easy way to separate pseudo-documents from real documents, but could not find one. The harder we looked for a way to separate the two document types, the more we realized that this distinction had theoretical interest. Determining the percentage of real documents would serve to evaluate the true size of the corpus, and its usefulness for TIPSTER type applications where documents relevant to topic queries are expected to be returned.

Note that this query matched the two criteria set forth in the Introduction for applicability to our method. As described above, there was no easy way to separate real documents from pseudo-documents. The query was also subjective, since readers might disagree about the classification of particular documents. For example, a document announcing a series of classes about how to use the Federal Register could be considered a real document (since notices of all sorts appear in the Register), or a pseudo-document (since it is promulgated by the Register's office and appears at regular intervals). As another example, readers might disagree about which documents correcting previous documents contain enough topic matter to be real documents, and which merely address meaningless typographical errors and should therefore be considered pseudo-documents.

2.3 Overall estimation and the Bayesian approach

To estimate the overall percentage of real documents in the corpus, we sampled 200 documents at random from the 45,820 total. Sampling was done with replacement (i.e., we did not remove sampled documents from the population); however, no documents were selected twice. One of our researchers then reviewed the documents and judged them as real documents versus pseudo-documents. He did this by reading the first fifty lines of each document.

Of the 200 documents sampled, 187, or 0.935, were judged to be real documents. The binomial likelihood function associated with this result,
is graphed in Figure 1. It shows, given each possible true percentage of real documents, the likelihood that one would find 187 real documents out of 200 sampled. This distribution peaked at 0.935; however, its mean was slightly to the left (at 0.9307). This difference can be understood by keeping in mind that the binomial likelihood is equivalent to a beta distribution with parameters $\alpha = 187$, $\beta = 14$. The mean of this distribution is, by definition, $\frac{\alpha}{\alpha+\beta}$ or $\frac{94}{101}$ (0.9307).

Figure 1. Binomial likelihood function given 187 real documents out of 200 sampled

We evaluated the likelihood function at a high degree of granularity -- at x intervals corresponding to five significant digits -- so that we would later be able to map percentages of documents onto exact numbers of documents. For example, the difference between 0.65 and 0.66 of our 45,820 document set maps onto a difference of 458 documents (29,783 versus 30,241). The difference between 0.6543 and 0.65432 maps onto a difference of a single document (29,980 versus 29,981).

It remained to quantify the uncertainty associated with using 0.9307 as an estimate of the percentage of real documents in the population as a whole. We did this using a Bayesian approach in which these sampling results were combined with prior expectations for the population. We chose a non-parametric prior by inputting a personal likelihood: one researcher's subjective opinion about the population based on a first look at the corpus. The likelihood was evaluated at multiples of 0.1, then splined to obtain a likelihood function and normalized to obtain a probability density. The resulting density was discretized at five significant digits to match the granularity of the binomial likelihood function. The original and splined likelihoods are graphed in Figure 2.

Figure 2. Prior personal likelihood for proportion of real documents

An alternative to the above procedure is to choose from a parametric family such as beta distributions. This approach simplifies later calculations [5]. However, the non-parametric prior allows the researcher more freedom in choosing a probability density that expresses his or her best understanding of a population.

Once the prior distribution was established, we applied Bayes' theorem to calculate a posterior probability distribution for the population. We did this by multiplying the prior distribution (Fig. 2) by the binomial likelihood function (Fig. 1), then normalizing. The non-zero portion of the resulting posterior is graphed in Figure 3.

Figure 3. Posterior probability distribution for proportion of real documents

For comparison, Figure 3 contrasts this posterior density with the binomial likelihood function from Figure 1, also normalized. From a Bayesian perspective, the latter density implicitly factors in the standard non-informative prior in which each possible percentage of real documents has an equal probability. Comparing the two densities, one can see that the main effect of the non-parametric prior was to shift the density slightly to the left.

The next step was to find the 95% confidence interval surrounding the estimated mean of the posterior density based on the non-parametric prior. In other words, we wanted to find the range on the x axis that we were 95% confident contained the actual percentage of real documents on the posterior density. The traditional way to do this, assuming a normal
distribution, is to compute the mean \( \mu \) and variance \( \sigma^2 \), defined as

\[
\sum_{k=1}^{l} x_k f(x_k) \quad \text{and} \quad \sum_{k=1}^{l} f(x_k)(x_k-\mu)^2,
\]

respectively, where \( l \) is the number of points evaluated for the function. Then one can set the confidence interval at \( \mu \pm 1.96 \sigma \).

Instead we calculated the confidence interval exactly, in a numerical fashion that yielded the tightest possible interval and thus reduced the final uncertainty of the estimate. To do so we moved outward from the peak of the distribution, summing under the curve until we reached a total probability of 0.95. At each step outwards from the peak we considered probability values to the right and left and chose the larger of the two. This method finds a tighter interval than the numerical method used in [5], which was based on finding the left and right tails that each contained 0.025 of the distribution.

The confidence interval found for the posterior probability distribution is summarized in Table 1, both in percentages of real documents and in numbers of real documents. The document range was calculated by multiplying the percentage range by the number of documents in the dataset (45,820). Again, for comparison's sake the table includes the parallel results obtained using a non-informative prior. The two intervals were almost identical, with a slight advantage to the non-informative prior, implying that the non-parametric prior was poorly chosen. But regardless of the prior used, the size of the confidence interval, expressed in numbers of documents, was over 3100 documents. This was a lot of uncertainty -- enough to taint any decision about the usage of documents in the on-line Federal Register.

<table>
<thead>
<tr>
<th>Posterior based on which prior</th>
<th>Interval</th>
<th>Size of confidence interval (in documents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-parametric</td>
<td>( 0.89042 &lt; p &lt; 0.95892 )</td>
<td>40799-43937</td>
</tr>
<tr>
<td>Non-informative</td>
<td>( 0.89540 &lt; p &lt; 0.96376 )</td>
<td>41027-44159</td>
</tr>
</tbody>
</table>

2.4 Reducing uncertainty using stratified sampling

To reduce the uncertainty in the estimate of the percentage of real documents in our corpus, we performed a stratified sampling. The goal of this step was to divide the data into two relatively homogeneous strata, one containing mostly real documents, the other mostly pseudo-documents. The variance of a binomial distribution, \( \frac{p(1-p)}{n} \) (where \( n \) is the number sampled, and \( p \) the percentage of "yes" answers), shrinks dramatically for extreme values of \( p \). Therefore, one can generally reduce sampling uncertainty by combining results from several homogeneous strata, rather than doing an overall sample from a heterogeneous population. As with our overall sample, we performed the stratified sampling within the Bayesian framework.

We observed that most pseudo-documents were of the following types:

1. Part dividers (title pages for subparts of an issue)
2. Title pages
3. Tables of contents
4. Reader Aids sections
5. Instructions to insert illustrations not present in the electronic version
6. Null documents (no text material between <TEXT> and </TEXT> markers)
7. Other defective documents, such as titles of presidential proclamations that were separated from the proclamation itself.

We wrote an expert system, implemented in a short Perl script, that recognized the pseudo-document types 1-4 using key phrases (e.g., />Part [IVXM]/ for Part dividers), and recognized types 5-7 by their length -- always under twenty lines (including blanks). This test stratified the data into 3444 apparent pseudo-documents and 42,376 apparent real documents.

Exploration of the strata showed that this stratification was not perfect -- indeed, if it were, we could no longer call this query difficult! Some real documents were misclassified as pseudo-documents, usually Part dividers or title pages, because they accidentally triggered the key phrase detectors. A document correcting the incomprehensible "Federal Register-ese" error

"<ITAG tagnum=68>BILLING CODE 1505-01-D </ITAG>"

was misclassified as a real document. However, as we will see, the stratification was good enough to sharply reduce the confidence interval.

Before doing a stratified sampling of the corpus, we had to decide how many documents to sample from each stratum. Rather than sampling each stratum equally, or in proportion to its occurrence in the population, we used a Bayesian modification of Neyman allocation; this is a departure from [5]. Traditional Neyman allocation requires a presampling of each stratum to determine its heterogeneity; heterogeneous strata are then sampled more intensively. In the Bayesian version developed by Newbold [7], prior expectations for each stratum are combined with presample results to create a posterior density for each stratum. These posteriors are then used to determine the allocation.

This technique therefore required creating posterior densities for each stratum that blended a prior density and a presample. Accordingly, we devised non-parametric priors for the two strata -- apparent pseudo-documents, and apparent real documents -- based on our exploratory analysis of the strata. As in the overall analysis, we splined the priors to five significant digits. These priors (before the spline) are graphed in Figure 4.

Figure 4. Prior likelihoods for proportion of real documents in the strata

For the presample, we randomly chose ten documents from each stratum (with replacement) and read and scored them. The presample results were perfect -- all apparent pseudo-documents were pseudo-documents, and all apparent real documents were real. We applied Bayes' theorem to calculate the posterior distribution for each stratum, multiplying the binomial likelihood function associated with the stratum's presample by the relevant prior density, and normalizing.

With these posteriors in hand, we were ready to determine the optimum allocation among the strata. Newbold [7] gives the fraction $q_i$ allocated to each stratum $i$ by

$$q_i = \frac{C_i^{1/2}A_i^{1/2}(n_i+1)^{1/2}}{\sum_{j=1}^{k} C_j^{1/2}A_j^{1/2}(n_j+1)^{1/2}}$$  \hspace{1cm} (2)$$

where $k$ is the number of strata, $C_i$ is the cost of sampling a stratum (assumed to be a constant 1), $n_j$ is the number of documents in the presample for the stratum, and $A_j$ is
where \( \Pi_i \) is the fraction of the overall population that comes from the stratum, and \( P_i \) is the population mean for the posterior distribution in the \( i^{th} \) stratum.

The outcome of this procedure was an allocation of 15 apparent pseudo-documents and 185 apparent real documents. Since we had already sampled ten documents from each stratum, we now sampled an additional 5 apparent pseudo-documents and 175 apparent real documents, choosing documents randomly with replacement, and judged each document subjectively as above. To our surprise (knowing that the stratification was error-prone), this sampling again gave perfect results: all apparent pseudo-documents were pseudo-documents, and all apparent real documents were real.

We applied Bayes' theorem a final time to derive a new posterior probability density for each stratum based on the results of the full sample. For each stratum, we multiplied the binomial likelihood function corresponding to the full sampling results (0/15 and 185/185) by the prior probability density for each stratum (i.e., the posterior density from the presample), then normalized.

The final step was to combine the two posteriors to obtain an overall estimate for the mean and confidence interval of the population as a whole. The traditional approach would be to find the mean and variance for each stratum's posterior and combine these according to each stratum's weight in the population. Newbold [7] gives the weighted mean as

\[
\sum_{i=1}^{k} \Pi_i \frac{b_i}{n_i}
\]

where \( b_i \) is the number of real documents found in stratum \( i \) out of \( n_i \) sampled. As an alternative technique, we used a Monte Carlo simulation [8] to compute the distribution of

\[
\sum_{i=1}^{k} \Pi_i \frac{b_i}{n_i}
\]

this distribution to provide a best estimate and a corresponding confidence interval.

To perform the simulation, we randomly sampled the two posterior densities a million times. For each pair of points picked, we determined the weighted average of the two points, found the corresponding point on the overall density, and incremented its value by \( 10^{-6} \). For example, if we picked 0.1 from the posterior for the first stratum (apparent pseudo-documents) and 0.9 from the second stratum (apparent real documents), then, given the respective weights 0.075 (3444/45,820) and 0.925 (42,376/45,820) of the two strata, we would increment the value of 0.84, or 0.1*0.075 + 0.9*0.925, by \( 10^{-6} \).

The resulting density is graphed in Figure 5 along with the posteriors. Since there were more apparent real documents in the corpus than apparent pseudo-documents, the combined density was closer to the former than to the latter.

Figure 5. Posteriors after full sample, and Monte Carlo combination of posteriors

Using the same method as in section 2.3, we then found the exact 95% confidence interval for the combined density. The results, summarized in Table 2, show a better than 3:1 reduction from the overall sample, from 3138 to 919 documents. Table 2 also shows the results obtained using a non-informative prior -- that is, based on the sampled results alone, without any specific prior expectations. Here we clearly see the benefit of vigorously applying the Bayesian
approach, as the prior knowledge helps reduce the confidence interval by three percentage points, or 235 documents.

Table 2. Results from stratified sampling (200 documents)

<table>
<thead>
<tr>
<th>Posterior based on which prior</th>
<th>Interval in percent real documents</th>
<th>Interval in number of documents</th>
<th>Size of document interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-parametric</td>
<td>0.91157 &lt; p &lt; 0.93163</td>
<td>41,768-42,687</td>
<td>919</td>
</tr>
<tr>
<td>Non-informative</td>
<td>0.91172 &lt; p &lt; 0.96916</td>
<td>41,775-42,929</td>
<td>1154</td>
</tr>
</tbody>
</table>

3. Discussion and Conclusion

Using the Bayesian method and a stratified sampling of 200 documents, we have addressed the question of how many Federal Register documents are useful documents that reflect the activities of the Federal government. The answer was a 95% confidence interval between 91% and 93%, or between 41,768 and 42,687 documents. This was a substantially tighter estimate than could be obtained using either an overall sample, or a stratified sample without prior expectations.

The question remains whether this estimate is precise enough. For many applications the answer is yes. For example, a two percent range of error would probably not pose a problem in comparing the utility of different corpora. But other applications might call for greater precision. For example, one might want to find the estimated number of documents that match a query in order to calculate recall rates for an information retrieval system. In that case a difference of 919 documents could make a big difference.

When higher precision is called for, the simplest way to further narrow the confidence interval is to do more sampling. In a follow-on experiment, it took less than a half hour to read an additional 200 documents (this turned up two mis-stratified documents, confirming our expectations from exploratory analysis). The new data sharpened the posteriors, reducing the combined confidence interval to 624 documents, or 1.3 percentage points. Further reductions could be obtained as desired.

References


Acknowledgements

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. We thank the TIPSTER program and the Linguistic Data Consortium for making the *Federal Register* corpus available. We also thank Mike Cannon and Tony Warnock for helpful discussions on statistical issues.
Likelihood of observing exactly 187/200 real documents

Possible true proportion of real documents in entire population
Figure 5

Posterior for apparent real documents

Monte Carlo combination

Posterior for apparent pseudo-documents

Possible true proportion of real documents

Probability