Accurate Methods for the Statistics of Surprise and Coincidence

Ted Dunning

Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003-0001

ABSTRACT

Much work has been done on the statistical analysis of text. In some cases reported in the literature, inappropriate statistical methods have been used, and statistical significance of results have not been addressed. In particular, asymptotic normality assumptions have often been used unjustifiably, leading to flawed results.

This assumption of normal distribution limits the ability to analyze rare events. Unfortunately rare events *do* make up a large fraction of real text.

However, more applicable methods based on likelihood ratio tests are available which yield good results with relatively small samples. These tests can be implemented efficiently, and have been used for the detection of composite terms, and for the determination of domain-specific terms. In some cases, these measures perform much better than the methods previously used. In cases where traditional contingency table methods work well, the likelihood ratio tests described here are nearly identical.

This paper describes the basis of a measure based on likelihood ratios which can be applied to the analysis of text.

January 7, 1993

Accurate Methods for the Statistics of Surprise and Coincidence

Ted Dunning

Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003-0001

1. Introduction

There has been a recent trend back towards the statistical analysis of text. This trend has resulted in a number of researchers doing good work in information retrieval and natural language processing in general. Unfortunately much of their work has been characterized by a cavalier approach to the statistical issues raised by the results.

The approaches taken by such researchers can be divided into three rough categories:

1) Collect enormous volumes of text in order to make straightforward statistically based measures work well.

2) Do simple-minded statistical analysis on relatively small volumes of text and either 'correct empirically' for the error, or ignore the issue.

3) Perform no statistical analysis whatsoever.

The first approach is the one taken by the IBM group researching statistical approaches to machine translation [Brown, et al., 1989]. They have collected nearly one billion words of English text from such diverse sources as internal memos, technical manuals, and romance novels, and have aligned most of the electronically available portion of the record of debate in the Canadian parliament (Hansards). Their efforts have been Augean and have been well rewarded by interesting results. The statistical significance of most of their work is above reproach, but the required volumes of text are simply impractical in many settings.

The second approach is typified by much of the work of Gale and Church [Gale and Church 1991a, and 1991b, Church, Gale, Hanks and Hindle 1989]. Many of the results from their work are entirely usable, and the measures they use work well for the examples given in their papers. In general, though, their methods lead to problems. For example, mutual information estimates based directly on counts are subject to overestimate when the counts involved are small, and z-scores substantially overestimate the significance of rare events.

The third approach is typified by virtually all of the information retrieval literature. Even recent and very innovative work such as that using Latent Semantic Indexing [Dumais, et al, 1988] and Pathfinder Networks [Schvaneveldt, 1990] has not addressed the statistical reliability of the internal processing. They do, however, use good statistical methods to analyze the overall effectiveness of their approach.

Even such well accepted techniques as inverse document frequency weighting of terms in text retrieval [Salton, 1983] is generally only justified on very sketchy grounds.

The goal of this paper is to present a practical measure which is motivated by statistical considerations and which can be used in a number of settings. This measure works reasonably well with both large and small text samples and allows direct comparison of the significance of rare and common phenomena. This comparison is possible because the measure described in this paper has better asymptotic behavior than more traditional measures.

In the following, some sections are composed largely of background material or mathematical details and can probably be skipped by the reader familiar with statistics, or by the reader in a hurry. The sections that should not be skipped are marked with **, those with substantial background with *, and detailed derivations are unmarked. This 'good parts' convention should make this paper more useful to the implementor or reader only wishing to skim the paper.

2. The assumption of normality *

The assumption that simple functions of the random variables being sampled are distributed normally or approximately normally underlies many common statistical tests. This particularly includes Pearson's χ^2 test and z-score tests. This assumption is absolutely valid in many cases. Due to the simplification of the methods involved, it is entirely justifiable even in marginal cases.

When comparing the rates of occurrence of rare events, the assumptions on which these tests are based break down because texts are composed largely of such rare events. For example, simple word counts made on a moderate sized corpus show that words which have a frequency of less than one in 50,000 words make up about 20-30% of typical English language news-wire reports. This 'rare' quarter of English includes many of the content-bearing words, and nearly all the technical jargon. As an illustration, the following is a random selection of approximately 0.2% of the words found at least once, but fewer than 5 times in a sample of a half million words of Reuters' reports,

abandonment	detailing	landscape	seldom
aerobics	directorship	lobbyists	sheet
alternating	dispatched	malfeasances	simplified
altitude	dogfight	meat	snort
amateur	duds	miners	specify
appearance	eluded	monsoon	staffing
assertion	enigmatic	napalm	substitute
barrack	euphemism	northeast	surreptitious
biased	experiences	oppressive	tall
bookies	fares	overburdened	terraced
broadcaster	finals	parakeets	tipping
cadres	foiling	penetrate	transform
charging	gangsters	poi	turbid
clause	guide	praised	understatement
collating	headache	prised	unprofitable
compile	hobbled	protector	vagaries
confirming	identities	query	villas
contemptuously	inappropriate	redoubtable	watchful
corridors	inflamed	remark	winter
crushed	instilling	resignations	
deadly	intruded	ruin	
demented	junction	scant	

The only word in this list that is in the least obscure is *poi* (a native Hawaiian dish made from taro root). If we were to sample 50,000 words instead of the half million used to create the list above, then the expected number of occurrences of any of the words in this list would be less than one half, well below the point where commonly used tests should be used.

If such ordinary words are 'rare', any statistical work with texts must deal with the reality of rare events. It is interesting that while most of the words in running text are common ones, most of the words in the total vocabulary are rare.

Unfortunately, the foundational assumption of most common statistical analyses used in computational linguistics is that the events being analyzed are relatively common. For a sample of 50,000 words from the Reuters' corpus mentioned previously, none of the words in the table above are common enough to expect such analyses to work well.

3. The tradition of Chi-squared tests *

In text analysis, the statistically based measures that have been used have usually been based on test statistics which are useful because, given certain assumptions, they have a known distribution. This distribution is most commonly either the normal or χ^2 distribution. These measures are very useful and can be used to accurately assess significance in a number of different settings. They are based, however, on several assumptions that do not hold for most

textual analyses.

The details of how and why the assumptions behind these measures do not hold is of interest primarily to the statistician, but the result is of interest to the statistical consumer (in our case, somebody interested in counting words). More applicable techniques are important in textual analysis. The next section describes one such technique; implementation of this technique is described in later sections.

4. Binomial distributions for Text Analysis **

Binomial distributions arise commonly in statistical analysis when the data to be analyzed is derived by counting the number of positive outcomes of repeated identical and independent experiments. Flipping a coin is the prototypical experiment of this sort.

The task of counting words can be cast into the form of a repeated sequence of such binary trials comparing each word in a text with the word being counted. These comparisons can be viewed as a sequence of binary experiments similar to coin flipping. In text, each comparison is clearly not independent of all others, but the dependency falls off rapidly with distance. Another assumption that works relatively well in practice is that the probability of seeing a particular word does not vary. Of course, this is not really true, since changes in topic may cause this frequency to vary. Indeed it is the mild failure of this assumption that makes shallow information retrieval techniques possible.

To the extent that these assumptions of independence and stationarity are valid, we can switch to an abstract discourse concerning Bernoulli trials instead of words in text, and a number of standard results can be used. A Bernoulli trial is the statistical idealization of a coin flip in which there is a fixed probability of a successful outcome that does not vary from flip to flip.

In particular, if the actual probability that the next word matches a prototype is p, then the number of matches generated in the next n words is a random variable (K) with binomial distribution $p(K=k) = p^k (1-p)^{n-k} \left[\begin{matrix} n \\ k \end{matrix} \right]$ whose mean is np and whose variance is $np(1-p)$. If $np(1-p) > 5$, then the distribution of this variable will be approximately normal, and as $np(1-p)$ increases beyond that point, the distribution becomes more and more like a normal distribution. This can be seen in Figure 1, below, where the binomial distribution (dashed lines) is plotted along with the approximating normal distributions (solid lines) for np set to 5, 10 and 20, with n fixed at 100. Larger values of n with np held constant give curves which are not visibly different from those shown. For these cases, $np \approx np(1-p)$.

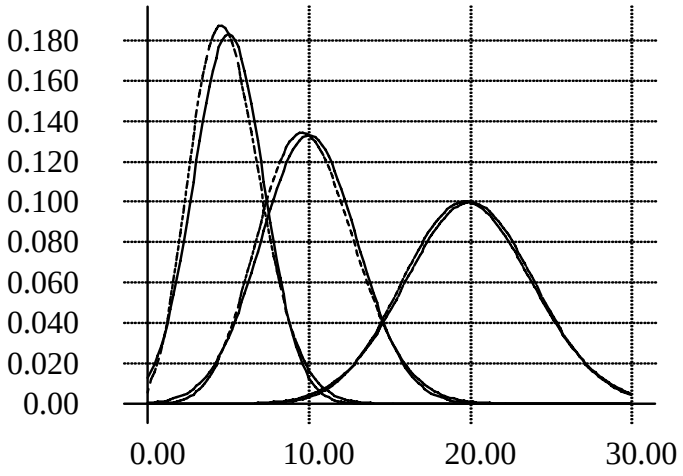

Figure 1: Normal and Binomial Distributions

This agreement between the binomial and normal distributions is exactly what makes test statistics based on assumptions of normality so useful in the analysis of experiments based on counting. In the case of the binomial distribution, normality assumptions are generally considered to hold well enough when $np\ (1-p) > 5$.

The situation is different when $np\ (1-p)$ is less than 5, and is dramatically different when $np\ (1-p)$ is less than 1. First, it makes much less sense to approximate a discrete distribution such as the binomial with a continuous distribution such as the normal. Second, the probabilities computed using the normal approximation are less and less accurate.

Table 1 shows the probability that one or more matches are found in 100 words of text as computed using the binomial and normal distributions for $np = 0.001$, $np = 0.01$, $np = 0.1$ and $np = 1$ where $n = 100$. Most words are sufficiently rare so that even for samples of text where n is as large as several thousand, np will be at the bottom of this range. Short phrases are so numerous that $np \ll 1$ for almost all phrases even when n is as large as several million.

	$p(k > 1)$	
	Using Binomial	Est. using Normal
$np = 0.001$	0.000099	0.34×10^{-217}
$np = 0.01$	0.0099	0.29×10^{-22}
$np = 0.1$	0.095	0.0022
$np = 1$	0.63	0.5

Table 1: Error introduced by normal approximations

Table 1 shows that for rare events, the normal distribution does not even approximate the binomial distribution. In fact, for $np = 0.1$ and $n = 100$, using the normal distribution over-estimates the significance of one or more occurrences by a factor of 40, while for $np = 0.01$, using the normal distribution over-estimates the significance by about 4×10^{20}. When n increases beyond 100, the numbers in the table do not change significantly.

If this over-estimation were constant, then the estimates using normal distributions could be corrected and would still be useful, but the fact that the errors are not constant means that methods dependent on the normal approximation should not be used to analyze Bernoulli trials where the probability of positive outcome is very small. Yet, in many real analyses of text, comparing cases where $np = 0.001$ with cases where $np > 1$ is a common problem.

5. Likelihood ratio tests *

There is another class of tests which do not depend so critically on assumptions of normality. Instead they use the asymptotic distribution of the generalized likelihood ratio. For text analysis and similar problems, the use of likelihood ratios leads to very much improved statistical results. The practical effect of this improvement is that statistical textual analysis can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions, and it allows comparisons to be made between the significance of the occurrences of both rare and common phenomenon.

5.1. Parameter spaces and likelihood functions

Likelihood ratio tests are based on the idea that statistical hypotheses can be said to specify subspaces of the space described by the unknown parameters of the statistical model being used. These tests assume that the model is known, but that the parameters of the model are unknown. Such a test is called parametric. Other tests are available which make no assumptions about the underlying model at all; they are called distribution free. Only one particular parametric test is described here. More information on parametric and distribution free tests is available in [Bradley, 1968] and [Mood, et. al. 1974].

The probability that a given experimental outcome described by $k_1,..., k_n$ will be observed for a given model described by a number of parameters $p_1, p_2,...$ is called the likelihood function for the model and is written as

$$H(p_1, p_2, \cdots ; k_1, \ldots, k_m)$$

where all arguments of H left of the semi-colon are model parameters, and all arguments right

of the semi-colon are observed values. In the continuous case, the probability is replaced by a probability density. With binomial and multinomials, we only deal with the discrete case.

For repeated Bernoulli trials, $m=2$ because we observe both the number of trials and the number of positive outcomes and there is only one p. The explicit form for the likelihood function is

$$H(p\,;\,n\,,\,k) = p^{k}\ (1-p)^{n-k}\ \begin{bmatrix} n \\ k \end{bmatrix}$$

The parameter space is the set of all values for p and the hypothesis that $p = p_0$ is a single point. For notational brevity the model parameters can be collected into a single parameter, as can the observed values. Then the likelihood function is written as

$$H(\omega;\,k)$$

where ω is considered to be a point in the parameter space Ω, and k a point in the space of observations K. Particular hypotheses or observations are represented by subscripting Ω or K respectively.

More information about likelihood ratio tests can be found in texts on theoretical statistics [Mood, et. al., 1974].

5.2. The likelihood ratio

The likelihood ratio for a hypothesis is the ratio of the maximum value of the likelihood function over the subspace represented by the hypothesis to the maximum value of the likelihood function over the entire parameter space. That is,

$$\lambda = \frac{\max\limits_{\omega\,\varepsilon\,\Omega_0} H(\omega\,;\,k)}{\max\limits_{\omega\,\varepsilon\,\Omega} H(\omega\,;\,k)}$$

where Ω is the entire parameter space and Ω_0 is the particular hypothesis being tested.

The particularly important feature of likelihood ratios is that the quantity $-2\log\lambda$ is asymptotically χ^2 distributed with degrees of freedom equal to the difference in dimension between Ω and Ω_0. Importantly, this asymptote is approached very quickly in the case of binomial and multinomial distributions.

5.3. Likelihood ratio for binomial and multinomial distributions

The comparison of two binomial or multinomial processes can be done rather easily using likelihood ratios. In the case of two binomial distributions,

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = p_1^{k_1} (1 - p_1)^{n_1 - k_1} \binom{n_1}{k_1} \; p_2^{k_2} (1 - p_2)^{n_2 - k_2} \binom{n_2}{k_2}$$

The hypothesis that the two distributions have the same underlying parameter is represented by the set $\{(p_1, p_2) \mid p_1 = p_2\}$.

The likelihood ratio for this test is

$$\lambda = \frac{\max\limits_{p} H(p, p; k_1, n_1, k_2, n_2)}{\max\limits_{p_1, p_2} H(p_1, p_2; k_1, n_1, k_2, n_2)}$$

These maxima are achieved with $p_1 = \frac{k_1}{n_1}$ and $p_2 = \frac{k_2}{n_2}$ for the denominator, and $p = \frac{k_1 + k_2}{n_1 + n_2}$ for the numerator. This reduces the ratio to

$$= \frac{\max\limits_{p} L(p, k_1, n_1)\, L(p, k_2, n_2)}{\max\limits_{p_1, p_2} L(p_1, k_1, n_1)\, L(p_2, k_2, n_2)}$$

where

$$L(p, k, n) = p^k \; (1 - p)^{n - k}$$

Taking the logarithm of the likelihood ratio gives

$$-2 \log \lambda = 2 \left[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2) \right]$$

For the multinomial case, it is convenient to use the double subscripts and the abbreviations

$$P_i = p_{1i}, p_{2i}, \ldots, p_{ji}, \ldots,$$

$$K_i = k_{1i}, k_{2i}, \ldots, k_{ji}, \ldots,$$

$$Q = q_1, q_2, \ldots, q_j, \ldots,$$

so that we can write

$$H(P_1, P_2; K_1, n_1, K_2, n_2) = \prod_{i = 1,2} n_i ! \prod_j \frac{p_{ji}^{k_{ji}}}{k_{ji} !}$$

The likelihood ratio is

$$\lambda = \frac{\max\limits_{Q} H(Q, Q; K_1, n_1, K_2, n_2)}{\max\limits_{P_1, P_2} H(P_1, P_2; K_1, n_1, K_2, n_2)}$$

This can be separated in a similar fashion as the binomial case by using the function

$$L(P, K) = \prod_j p_j^{k_j}$$

$$\lambda = \frac{\max_Q L(Q, K_1)\, L(Q, K_2)}{\max_{P_1, P_2} L(P_1, K_1)\, L(P_2, K_2)}$$

This expression implicitly involves n because $\sum_j k_j = n$.

Maximizing and taking the logarithm,

$$-2 \log \lambda = 2 \left[\log L(P_1, K_1) + \log L(P_2, K_2) - \log L(Q, K_1) - \log L(Q, K_2) \right]$$

where

$$p_{ji} = \frac{k_{ji}}{\sum_i k_{ji}}$$

and

$$q_j = \frac{\sum_i k_{ji}}{\sum_{ij} k_{ji}}$$

If the null hypothesis holds, then the log-likelihood ratio is asymptotically χ^2 distributed with $k/2 - 1$ degrees of freedom. When j is 2 (the binomial), $-2 \log \lambda$ will be χ^2 distributed with one degree of freedom.

If we had initially approximated the binomial distribution with a normal distribution with mean np and variance $np\,(1-p)$ then we would have arrived at another form which is a good approximation of $-2 \log \lambda$ when $np\,(1-p)$ is more than roughly 5. This form is

$$-2 \log \lambda \approx \sum_{ij} \frac{(k_{ji} - n_i q_j)^2}{n_i q_j\, (1-q_j)}$$

where

$$q_j = \frac{\sum_i k_{ji}}{\sum_{ij} k_{ji}}$$

as in the multinomial case above and

$$n_i = \sum_j k_{ji}$$

Interestingly, this expression is exactly the test statistic for Pearson's χ^2 test, although the form shown is not quite the customary one. Figure 2 shows the reasonably good agreement between this expression and the exact binomial log-likelihood ratio derived earlier where $p = 0.1$ and $n_1 = n_2 = 1000$ for various values of k_1 and k_2.

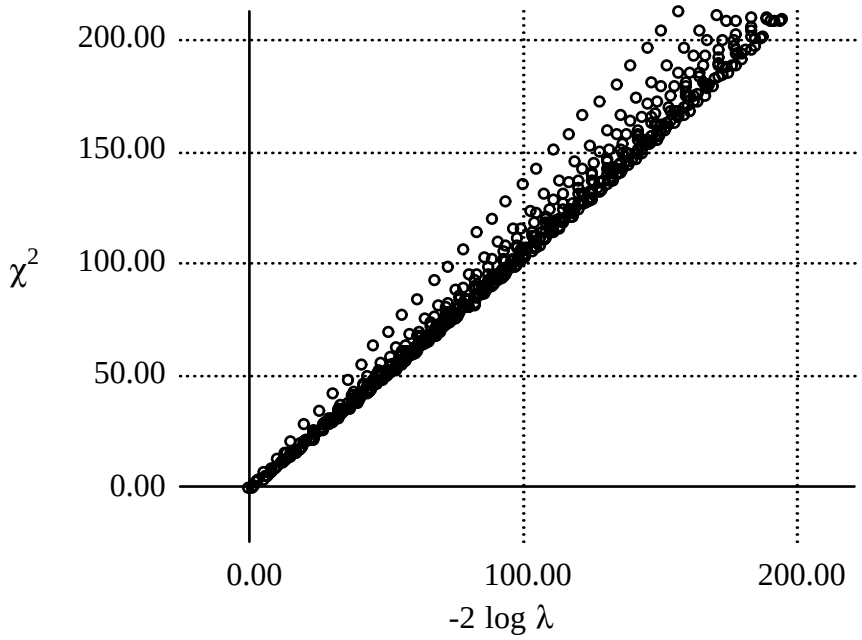

Figure 2: Log-likelihood versus Pearson χ^2

Figure 3, on the other hand, shows the divergence between Pearson's statistic and the log-likelihood ratio when $p = 0.01$, $n_1 = 100$ and $n_2 = 10000$. Note the large change of scale on the vertical axis. The pronounced disparity occurs when k_1 is larger than the value expected based on the observed value of k_2. The case where $n_1 < n_2$ and $\frac{k_1}{n_1} > \frac{k_2}{n_2}$ is exactly the case of most interest in many text analyses.

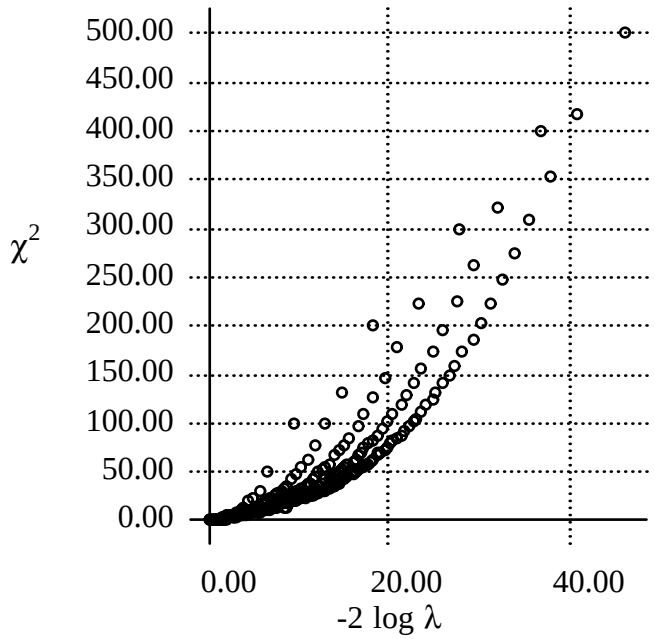

Figure 3: Log-likelihood versus Pearson χ^2

The convergence of the log of the likelihood ratio to the asymptotic distribution is demonstrated dramatically in Figure 4. In this figure, the straighter line was computed using a symbolic algebra package and represents the idealized one degree of freedom cumulative χ^2 distribution. The rougher curve was computed by a numerical experiment in which $p = 0.01$, $n_1 = 100$ and $n_2 = 10000$ which corresponds to the situation in Figure 3. The close agreement shows that the likelihood ratio measure produces accurate results over 6 decades of significance even in the range where the normal χ^2 measure diverges radically from the ideal.

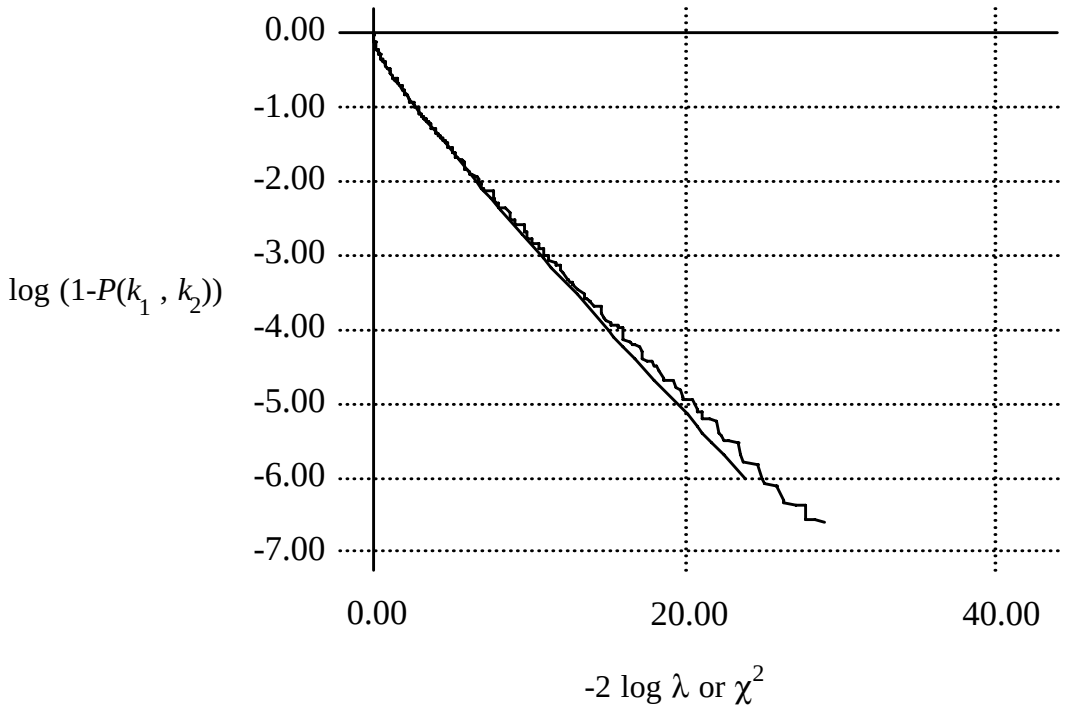

Figure 4: Ideal versus simulated Log-likelihood

6. Practical results

6.1. Bigram analysis of a small text

To test the efficacy of the likelihood methods, an analysis was made of a 30,000 word sample of text obtained from the Union Bank of Switzerland, with the intention of finding pairs of words which occurred next to each other with a significantly higher frequency than would be expected based on the word frequencies alone. The text was 31,777 words of financial text largely describing market conditions for 1986 and 1987.

The results of such a bigram analysis should highlight collocations common in English as well as collocations peculiar to the financial nature of the analyzed text. As will be seen, the ranking based on likelihood ratio tests does exactly this. Similar comparisons made between a large corpus of general text and a domain-specific text can be used to produce lists consisting only of words and bigrams characteristic of the domain-specific texts.

This comparison was done by creating a contingency table which contained the following counts of each bigram that appeared in the text,

$k(A\ B)$	$k(\sim A\ B)$
$k(A\ \sim B)$	$k(\sim A\ \sim B)$

where the $\sim A\ B$ represents the bigram in which the first word is *not* word *A* and the second is word *B*.

If the words *A* and *B* occur independently, then we would expect $p(AB) = p(A)\,p(B)$ where $p(AB)$ is the probability of *A* and *B* occurring in sequence, $p(A)$ is the probability of *A* appearing in the first position and $p(B)$ is the probability of *B* appearing in the second position. We can cast this into the mold of our earlier binomial analysis by phrasing the null hypothesis that *A* and *B* are independent as $p(A\,|\,B) = p(A\,|\,\sim B) = p(A)$. This means that testing for the independence of *A* and *B* can be done by testing to see if the distribution of *A* given that *B* is present (the first row of the table) is the same as the distribution of *A* given that *B* is not present (the second row of the table). In fact, of course, we are not really doing a statistical test to see if *A* and *B* are independent; we know that they are generally not independent in text. Instead we just want to use the test statistic as a measure which will help highlight particular *A*'s and *B*'s which are highly associated in text.

These counts were analyzed using the test for binomials described earlier, and the 50 most significant are tabulated in Table 2. This table contains the most significant 200 bigrams and is reverse sorted by the first column which contains the quantity $-2 \log \lambda$. Other columns contain the four counts from the contingency table described above, and the bigram itself.

-2 log λ	k(AB)	k(A~B)	k(~AB)	k(~A~B)	A	B
270.72	110	2442	111	29114	the	swiss
263.90	29	13	123	31612	can	be
256.84	31	23	139	31584	previous	year
167.23	10	0	3	31764	mineral	water
157.21	76	104	2476	29121	at	the
157.03	16	16	51	31694	real	terms
146.80	9	0	5	31763	natural	gas
115.02	16	0	865	30896	owing	to
104.53	10	9	41	31717	health	insurance
100.96	8	2	27	31740	stiff	competition
98.72	12	111	14	31640	is	likely
95.29	8	5	24	31740	qualified	personnel
94.50	10	93	6	31668	an	estimated
91.40	12	111	21	31633	is	expected
81.55	10	45	35	31687	1	2
76.30	5	13	0	31759	balance	sheet
73.35	16	2536	1	29224	the	united
68.96	6	2	45	31724	accident	insurance
68.61	24	43	1316	30394	terms	of
61.61	3	0	0	31774	natel	c
60.77	6	92	2	31677	will	probably
57.44	4	11	1	31761	great	deal
57.44	4	11	1	31761	government	bonds
57.14	13	7	1327	30430	part	of
53.98	4	1	18	31754	waste	paper
53.65	4	13	2	31758	machine	exhibition

52.33	**7**	**61**	**27**	**31682**	**rose**	**slightly**
52.30	**5**	**9**	**25**	**31738**	**passenger**	**service**
49.79	**4**	**61**	**0**	**31712**	**not**	**yet**
48.94	**9**	**12**	**429**	**31327**	**affected**	**by**
48.85	**13**	**1327**	**12**	**30425**	**of**	**september**
48.80	**9**	**4**	**872**	**30892**	**continue**	**to**
47.84	**4**	**41**	**1**	**31731**	**2**	**nd**
47.20	**8**	**27**	**157**	**31585**	**competition**	**from**
46.38	**10**	**472**	**20**	**31275**	**a**	**positive**
45.53	**4**	**18**	**6**	**31749**	**per**	**100**
44.36	**7**	**0**	**1333**	**30437**	**course**	**of**
43.93	**5**	**18**	**33**	**31721**	**generally**	**good**
43.61	**19**	**50**	**1321**	**30387**	**level**	**of**
43.35	**20**	**2532**	**25**	**29200**	**the**	**stock**
43.07	**6**	**875**	**0**	**30896**	**to**	**register**
43.06	**3**	**1**	**10**	**31763**	**french**	**speaking**
41.69	**3**	**29**	**0**	**31745**	**3**	**rd**
41.67	**3**	**1**	**13**	**31760**	**knitting**	**machines**
40.68	**4**	**5**	**40**	**31728**	**25**	**000**
39.23	**9**	**5**	**1331**	**30432**	**because**	**of**
39.20	**5**	**40**	**25**	**31707**	**stock**	**markets**
38.87	**2**	**0**	**1**	**31774**	**scanner**	**cash**
38.79	**3**	**0**	**48**	**31726**	**pent**	**up**
38.51	**3**	**23**	**1**	**31750**	**firms**	**surveyed**
38.46	**4**	**2**	**98**	**31673**	**restaurant**	**business**
38.28	**3**	**12**	**3**	**31759**	**fell**	**back**
38.14	**6**	**4**	**432**	**31335**	**climbed**	**by**
37.20	**6**	**41**	**70**	**31660**	**total**	**production**
37.15	**2**	**0**	**2**	**31773**	**hay**	**crop**
36.98	**3**	**10**	**5**	**31759**	**current**	**transactions**

Table 2: Bigrams Ranked by Log-Likelihood Test

Examination of the table shows that there is good correlation with intuitive feelings about how natural the bigrams in the table actually are. This is in distinct contrast with Table 3 which contains the same data except that the first column is computed using Pearson's χ^2 test statistic. The overestimate of the significance of items that occur only a few times is dramatic. In fact the entire first portion of the table is dominated by bigrams rare enough to occur only once in the current sample of text. The misspelling in the bigram 'sees posibilities' is in the original text.

χ^2	k(AB)	k(A~B)	k(~AB)	k(~A~B)	A	B
31777.00	**3**	**0**	**0**	**31774**	**natel**	**c**
31777.00	**1**	**0**	**0**	**31776**	**write**	**offs**
31777.00	**1**	**0**	**0**	**31776**	**wood**	**pulp**
31777.00	**1**	**0**	**0**	**31776**	**window**	**frames**
31777.00	**1**	**0**	**0**	**31776**	**upholstery**	**leathers**
31777.00	**1**	**0**	**0**	**31776**	**surveys**	**expert**
31777.00	**1**	**0**	**0**	**31776**	**sees**	**posibilities**
31777.00	**1**	**0**	**0**	**31776**	**practically**	**drawn**
31777.00	**1**	**0**	**0**	**31776**	**poultry**	**farms**
31777.00	**1**	**0**	**0**	**31776**	**physicians'**	**fees**
31777.00	**1**	**0**	**0**	**31776**	**paints**	**varnishes**
31777.00	**1**	**0**	**0**	**31776**	**maturity**	**hovered**
31777.00	**1**	**0**	**0**	**31776**	**listeriosis**	**bacteria**
31777.00	**1**	**0**	**0**	**31776**	**la**	**presse**
31777.00	**1**	**0**	**0**	**31776**	**instance**	**280**
31777.00	**1**	**0**	**0**	**31776**	**cans**	**casing**
31777.00	**1**	**0**	**0**	**31776**	**bluche**	**crans**
31777.00	**1**	**0**	**0**	**31776**	**a313**	**intercontinental**
24441.54	**10**	**0**	**3**	**31764**	**mineral**	**water**
21184.00	**2**	**0**	**1**	**31774**	**scanner**	**cash**
20424.86	**9**	**0**	**5**	**31763**	**natural**	**gas**
15888.00	**1**	**1**	**0**	**31775**	**suva's**	**responsibilties**

15888.00	**1**	**1**	**0**	**31775**	**suva's**	**questionable**
15888.00	**1**	**1**	**0**	**31775**	**responsible**	**clients**
15888.00	**1**	**1**	**0**	**31775**	**red**	**ink**
15888.00	**1**	**1**	**0**	**31775**	**joined**	**forces**
15888.00	**1**	**1**	**0**	**31775**	**highest**	**density**
15888.00	**1**	**1**	**0**	**31775**	**generating**	**modest**
15888.00	**1**	**1**	**0**	**31775**	**enables**	**conversations**
15888.00	**1**	**1**	**0**	**31775**	**dessert**	**cherry**
15888.00	**1**	**1**	**0**	**31775**	**consolidated**	**lagging**
15888.00	**1**	**1**	**0**	**31775**	**catalytic**	**converter**
15888.00	**1**	**1**	**0**	**31775**	**bread**	**grains**
15888.00	**1**	**1**	**0**	**31775**	**bottlenecks**	**booking**
15888.00	**1**	**1**	**0**	**31775**	**bankers'**	**association's**
15888.00	**1**	**1**	**0**	**31775**	**appenzell**	**abrupt**
15888.00	**1**	**1**	**0**	**31775**	**56**	**513**
15888.00	**1**	**1**	**0**	**31775**	**56**	**082**
15888.00	**1**	**1**	**0**	**31775**	**46**	**520**
15888.00	**1**	**1**	**0**	**31775**	**43**	**classified**
15888.00	**1**	**1**	**0**	**31775**	**43**	**502**
15888.00	**1**	**0**	**1**	**31775**	**wheel**	**drive**
15888.00	**1**	**0**	**1**	**31775**	**shops**	**joined**
15888.00	**1**	**0**	**1**	**31775**	**selected**	**collections**
15888.00	**1**	**0**	**1**	**31775**	**propelled**	**railcars**
15888.00	**1**	**0**	**1**	**31775**	**overcapacities**	**arising**
15888.00	**1**	**0**	**1**	**31775**	**listed**	**job**
15888.00	**1**	**0**	**1**	**31775**	**liquid**	**fuels**
15888.00	**1**	**0**	**1**	**31775**	**incl.**	**cellulose**
15888.00	**1**	**0**	**1**	**31775**	**fats**	**oils**
15888.00	**1**	**0**	**1**	**31775**	**drastically**	**deteriorate**
15888.00	**1**	**0**	**1**	**31775**	**completing**	**constructions**
15888.00	**1**	**0**	**1**	**31775**	**cider**	**apples**
15888.00	**1**	**0**	**1**	**31775**	**bicycle**	**tags**
15888.00	**1**	**0**	**1**	**31775**	**auctioning**	**collections**
15887.50	**2**	**0**	**2**	**31773**	**hay**	**crop**

Table 3: Bigrams Ranked by χ^2 Test

Out of 2693 bigrams analyzed, 2682 of them fall outside the scope of applicability of the normal χ^2 test. The 11 bigrams which were suitable for analysis with the χ^2 test are listed in Table 4. It is notable that all of these bigrams contain the word *the* which is the most common word in English.

χ^2	k(AB)	k(A~B)	k(~AB)	k(~A~B)	A	B
525.02	**110**	**2442**	**111**	**29114**	**the**	**swiss**
286.52	**76**	**104**	**2476**	**29121**	**at**	**the**
51.12	**26**	**2526**	**66**	**29159**	**the**	**volume**
6.03	**4**	**148**	**2548**	**29077**	**be**	**the**
4.48	**1**	**73**	**2551**	**29152**	**months**	**the**
4.31	**1**	**71**	**2551**	**29154**	**increased**	**the**
0.69	**4**	**70**	**2548**	**29155**	**1986**	**the**
0.42	**7**	**62**	**2545**	**29163**	**level**	**the**
0.28	**4**	**60**	**2548**	**29165**	**again**	**the**
0.12	**5**	**2547**	**67**	**29158**	**the**	**increased**
0.03	**18**	**198**	**2534**	**29027**	**as**	**the**

Table 4: Bigrams where χ^2 Analysis is Applicable

7. Conclusions

Statistics based on the assumption of normal distribution are invalid in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only the very most common words (that is, the ones least likely to be of interest). This fact is typically ignored in much of the work in this field. Using such invalid methods may seriously overestimate the significance of relatively rare events. Parametric statistical analysis based on the binomial or multinomial distribution extends the applicability of statistical methods to much smaller texts than models using normal distributions and shows good promise in early applications of the method.

Further work is needed to develop software tools to allow the straightforward analysis of texts using these methods. Some of these tools have been developed and will be distributed by the Consortium for Lexical Research. For further information on this software, contact the author or the Consortium via email at `ted@nmsu.edu` or `lexical@nmsu.edu`.

In addition, there are a wide variety of distribution free methods which may avoid even the assumption that text can be modeled by multinomial distributions. Measures based on Fischer's exact method may prove even more satisfactory than the likelihood ratio measures described in this paper. Also, using the Poisson distribution instead of the multinomial as the limiting distribution for the distribution of counts may provide some benefits. All of these possibilities should be tested.

8. Summary of formulae **

For the binomial case, the log likelihood statistic is given by

$$-2\log\lambda = 2\left[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)\right]$$

where

$$\log L(p, n, k) = k\log p + (n-k)\log(1-p)$$

also, $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$, and $p = \frac{k_1 + k_2}{n_1 + n_2}$.

For the multinomial case, this statistic becomes

$$-2\log\lambda = 2\left[\log L(P_1, K_1) + \log L(P_2, K_2) - \log L(Q, K_1) - \log L(Q, K_2)\right]$$

where

$$p_{ji} = \frac{k_{ji}}{\sum_j k_{ji}}$$

$$q_j = \frac{\sum_i k_{ji}}{\sum_{ij} k_{ji}}$$

$$\log L(P, K) = \sum_j k_j \log p_j$$

9. References

Bradley, James V., Distribution-Free Statistical Tests. Prentice Hall.

Brown, Peter F.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Frederick; Lafferty, John D.; Mercer, Robert L. and Roossin, Paul S. 1989. A statistical Approach to Machine Translation. *Technical Report, RC 14773 (#66226) 7/17/89, IBM Research Division.*

Church, Ken W.; Gale, William A.; Hanks, Patrick and Hindle, Donald. 1989. Parsing, Word Associations and Typical Predicate-Argument Relations. *International Workshop on Parsing Technologies, CMU, 1989.*

S. Dumais, G. Furnas, T. Landauer, S. Deerwester, and R. Harshman, 1988 Using latent semantic analysis to improve access to textual information. *Proceedings of CHI '88, 281-285.*

Gale, William A. and Church, Ken W. 1991a A program for Aligning Sentences in Bilingual Corpora, *In press.*

Gale, William A. and Church, Ken W. 1991b Identifying Word Correspondences in Parallel Texts *To appear.*

McDonald, James E.; Plate, Tony; and Schvaneveldt, Roger. 1990. Using Pathfinder to extract semantic information from text. In: Schvaneveldt, Roger (Ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization.* Ablex. (pp. 149-164)

Mood, A.M.; Graybill, F.A. and Boes, D.C. 1974. *Introduction to the Theory of Statistics* McGraw Hill.

Slaton, Gerald and McGill, M.J. 1983. *Introduction to Modern Information Retrieval* McGraw Hill, New York