Confidence intervals on likelihood estimates for estimating association strengths

Mark Johnson
Brown University

16th June, 1999

This is a paper I posted on my Web site several months ago. While the basic idea is valid, I realize now that the discussion here is incomplete. There are many good statistics books on this topic: Agresti (1996), Cox and Snell (1989), and Lloyd (1998) are good places to start. I intend to post an updated version of this paper soon.

Yes, I know that the code at the back of this paper does not work! When I wrote this paper I used routines from the wonderful "Numerical Recipies in C" book (if you don't have this book, get it!), which are unfortunately copyright. In earlier versions of this paper I included code based on approximations from Abramowitz and Stegun—which I threw together after the fact so I could provide something to the Web public legally—and which didn't work (my bugs, not theirs!). Several helpful people corrected that bug, but I really wanted to provide the more accurate code based on the Numerical Recipies routines. It looked as if I would have to write my own versions of the Numerical Recipies routines, which I naturally put off doing. Since then, I have discovered the publically-available library CEPHES on

`http://www.netlib.org/cephes/`

which is claimed to calculate the inverse incomplete Beta function very accurately (I have not checked this!). I have recoded my code to use the CEPHES routines, which I can distribute. So download the code from my Web site

`http://www.cog.brown.edu/~mj`

and tell me if it works!

Thanks, and happy computing!

Mark Johnson

Abstract

Sparse data causes errors in the maximum-likelihood estimates of event probabilities that are often large enough to render measures of association such as pointwise mutual information useless for small sample sizes. This squib describes a procedure for estimating event probabilities that produces a confidence interval estimate rather than a point estimate. Using these confidence intervals in calculations of measures of association results in reasonable association strength rankings even if the data is drawn from very small sample sizes.

1 Introduction

In computational linguistics one often wants to identify strongly associated pairs of items from a corpus and compare the strength of their association with that of other pairs. Finding word collocations is an obvious example of such a problem, where one seeks strongly associated pairs of adjacent words, but formally similiar problems abound. Measures of association based on estimates of the relative likelihoood of the items appearing together, particularly information-theoretic measures such as (pointwise) mutual information, would seem to be natural ways of quantifying such associations. (There is no generally agreed mathematical definition of association, and different measures may be appropriate in different applications.) However, the sample sizes of many linguistically interesting events can be very small, and consequently there is extremely high variance in the estimated likelihoods and hence also in the association measures. This high variance means that the estimated association measures for such events may be much higher than their "true" value, leading to a large number of "false positive" association pairs.

Pointwise mutual information is a measure of association that is affected in this way. As explained in section 3, the pointwise mutual information of words w_1 and w_2 is

$$\mathrm{MI}(w_1, w_2) \quad = \quad \log_2(\Pr(W_1 = w_1 | W_2 = w_2)/\Pr(W_1 = w_1)) \tag{1}$$

where W_1 and W_2 are random variables ranging over the first and second words of word pairs respectively. Now in sections 2–21 of the Penn WSJ corpus the words "accidentally" and "smother" once in the pair "accidentally smother", and that is the only time either appears. Suppose we take the maximum likelihood estimates for the probabilities in (1). The maximum likelihood estimate of $\Pr(W_1 =$ "accidentally"$|W_2 =$ "smother"$)$ is 1.0, and since there are $950,028$ words in this corpus, the maximum likelihood estimate of $\Pr(W_1 =$ "accidentally"$)$ is approximately 10^{-6}, so the pointwise mutual information for this pair is approximately 20. Further, this is the maximum pointwise mutual information for any word pair from this corpus: in (1) it is impossible to find a numerator greater than 1, and no smaller non-zero denominator can be found for this corpus. Yet this seems quite wrong: "accidentally" and "smother" do not appear particulary strongly associated.

A number of remedies have been suggested in the literature, to which Manning and Schütze (1999) provides a good introduction. For example, one can simply ignore all samples of less than a certain size, say, 5 events. Alternatively, one can compute other kinds of statistics which are accurate for smaller sample sizes. For example, organizing the data as a contingency table, one can apply statistical tests of independence and take the significance level at which the independence assumption is rejected as a measure of association. Chi-squared is probably the most well-known of such tests, but Dunning (1993) argues that a likelihood ratio test yields better performance with small sample sizes.

However, while these tests of independence avoid the problem of low sample sizes, they are arguably biased towards pairs with large sample sizes. A test for independence measures how surprising it would be to observe the data if the two variables are in fact independent. Informally, one could obtain equally surprising data either by drawing a small sample from strongly dependent variables, or else by drawing a large sample from weakly dependent variables. Thus even though in both circumstances one might be equally certain that the samples were not drawn from a distribution in which the variables are independent, it seems reasonable to say that in the first case the variables are more strongly associated. Intuitively, strength of association should not depend on sample size.

For example, while Dunning's likelihood ratio statistic assigns low frequency pairs such as "accidentally smother" a reasonably low score, the pairs it assigns a high significance score to typically consist of high frequncy words. Table 3 lists the highest ranked word pairs using this statistic. Note that the second highest ranked pair is "of the"; this is a very high frequency pair, but these words do not seem particularly closely associated.

Further, tests for independence do not distinguish cases where the pair occur significantly *less* frequently than would be expected if the variables were independent, yet in this case one would probably want to say that the variables are disassociated rather than associated. For example, the 21st ranked pair according to the likelihood ratio statistic is "the the"; in this case, the pair (presumably a typo) is much less frequent than would be expected if the words were independently distributed. Again, this seems counter-intuitive.

We can avoid the "false positive" problem of the likelihood based measures of association by replacing the point likelihood estimates with an interval estimate in which we are reasonably confident that the true likelihood actually lies. We use the bounds on this interval to obtain a conservative estimate of the strength of association, reducing the number of false positives by in effect trading recall for precision. The next section describes how this confidence interval can be estimated, and the following section compares the results of this method with Dunning's likelihood ratio for finding pair collocations in the Wall Street Journal corpus.

2 Confidence Intervals on Likelihood Estimates

Suppose one observes an event occuring k times in n Bernoulli trials from a binomial distribution, and we wish to estimate a confidence interval $[\theta_-, \theta_+]$ in which the unknown true event probability θ lies. For any θ besides 0 and 1 there is a non-zero likelihood of observing any number k of events in n trials, so it is not possible to find a non-trivial hard bound on θ. The best one can hope for is an interval such that if θ lies outside of this interval it is unlikely (but not impossible) for events such as the one observed to occur. Now the likelihood of observing i events in n trials is $\Pr_{n,\theta}(i) = \binom{n}{i}\theta^i(1-\theta)^{n-i}$. Given a confidence

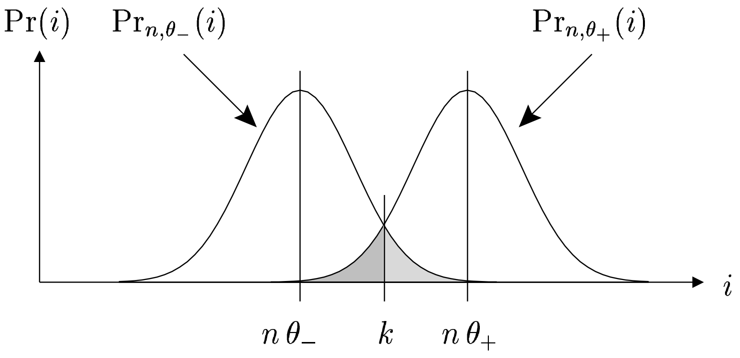

Figure 1: The confidence interval $[\theta_-, \theta_+]$ on the binomial parameter θ. k is the number of times the event was actually observed in n trials. The area contained in each of the shaded areas above and below k is $(1-s)/2$.

level s, we seek a lower bound θ_- and an upper bound θ_+ such that:

$$\sum_{i=k}^{n} \mathrm{Pr}_{n,\theta_-}(i) = (1-s)/2 \tag{2}$$

$$\sum_{i=0}^{k} \mathrm{Pr}_{n,\theta_+}(i) = (1-s)/2 \tag{3}$$

Thus, if $\theta \leq \theta_-$ then the likelihood of observing k or more events in n trials is less than or equal to $(1-s)/2$; similarly if $\theta \geq \theta_+$ then the likelihood of observing k or fewer events in n trials is greater than or equal to $(1-s)/2$. Figure 1 depicts the relationship between k, s, θ_- and θ_+.

The confidence interval parameters $\theta_- = \theta_-(n,k)$ and $\theta_+ = \theta_+(n,k)$ can be calculated as follows. The cumulative binomial distribution with event probability θ is related to the incomplete beta function I_θ in the following way (Abramowitz and Stegun, 1965):

$$\sum_{i=k}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} = I_\theta(k, n-k+1)$$

Thus the required bounds are the solutions of the following equations:

$$I_{\theta_-}(k, n-k+1) = (1-s)/2 \tag{4}$$

$$I_{1-\theta_+}(n-k, k+1) = (1-s)/2 \tag{5}$$

The results presented below were obtained by numerically solving these equations using Brent's method for zero finding: the program is essentially "glue" code that calls the routines for the incomplete beta function and Brent's method given in Press et al. (1992). It is also possible to directly compute the confidence interval using the approximation to the inverse beta function given in Abramowitz and Stegun (1965); this method is faster but less accurate than the numerical method just described.

Rank	**Likelihood Ratio**		**Mutual Information**	
1	New	York	TREASURY	BILLS
2	of	the	MERRILL	LYNCH
3	,	which	LATE	EURODOLLARS
4	,	”	INTERBANK	OFFERED
5	in	the	BANKERS	ACCEPTANCES
6	a	share	Zoete	Wedd
7	.	”	Pitney	Bowes
8	more	than	H.F.	Ahmanson
9	will	be	Kuala	Lumpur
10	,	but	IG	Metall
11	has	been	Dalkon	Shield
12	said	it	Bare-Faced	Messiah
13	”	says	Aga	Khan
14	the	company	Polly	Peck
15	Stock	Exchange	READY	ASSETS
16	San	Francisco	Fulton	Prebon
17	vice	president	Guzman	Cabrera
18	do	n’t	U	S
19	Wall	Street	negotiable	C.D.s
20	to	be	gon	na
21	the	the	Wastewater	Treatment
22	have	been	Peat	Marwick
23	“	We	Chez	Panisse
24	chief	executive	Della	Femina
25	.	.	LYNCH	READY
26	this	year	HOME	LOAN
27	,	or	CALL	MONEY
28	did	n’t	Violin	Concerto
29	on	the	COMMERCIAL	PAPER
30	year	earlier	Khmer	Rouge
31	net	income	Wedding	Song
32	York	Stock	PRECIOUS	METALS
33	real	estate	Nissho	Iwai
34	at	least	Nihon	Keizai
35	does	n’t	Exporting	Countries
36	Dow	Jones	Cobb	Parkway
37	would	be	Castle	Harlan
38	the	U.S.	Bonwit	Teller
39	third	quarter	Palo	Alto
40	“	I	Rupert	Murdoch

Table 1: Rank ordered lists of most significant word pairs using the likelihood ratio statistic (Dunning 1993) and the point-wise mutual information statistic defined in equation (8) at a $s = 0.99$ confidence level.

3 Finding strongly associated word pairs

The previous section described how to estimate a confidence interval $[\theta(n,k)_-, \theta(n,k)_+]$ for a binomial parameter θ given a confidence level s and data consisting of k observations of an event in n trials. This section examines how this confidence interval can be used to provide a conservative estimate of pointwise mutual information, which seems to provide a more intuitive measure of the strength of association of word pairs than do significance tests.

The pointwise mutual information of a pair of words $w_1\, w_2$ is defined as:

$$\begin{aligned} \mathrm{MI}(w_1, w_2) &= \log_2 \frac{\Pr(W_1 = w_1, W_2 = w_2)}{\Pr(W_1 = w_1)\Pr(W_2 = w_2)} \\ &= \log_2 \frac{\Pr(W_1 = w_1 | W_2 = w_2)}{\Pr(W_1 = w_1)} \qquad (6) \\ &= \log_2 \frac{\Pr(W_2 = w_2 | W_1 = w_1)}{\Pr(W_2 = w_2)} \qquad (7) \end{aligned}$$

Suppose that in our corpus the pair of words $w_1\, w_2$ occurs n_{12} times, that w_1 and w_2 each appear n_1 and n_2 times respectively, and that the total number of words in the corpus is n. Based on equation (6) we conservatively estimate the pointwise mutual information of $w_1\, w_2$ as follows:

$$\widehat{\mathrm{MI}}(w_1, w_2) = \log_2 \frac{\theta_-(n_{12}, n_2)}{\theta_+(n_1, n)} \qquad (8)$$

This statistic conservatively estimates the pointwise mutual information by using a lower bound in the numerator and an upper bound in the denominator. This reduces the number of false positive association pairs, effectively trading recall for precision.

Table 3 lists the word pairs ranked highest by this statistic at the $s = 0.99$ confidence level and by the likelihood ratio test described by Dunning (1993). The corpus used was sections 2-21 of the Penn WSJ treebank; no preprocessing was used to remove punctuation or normalize capitalization.

A striking feature of the conservative pointwise mutual information statistic is that it ranks multiword names and titles extremely highly. This is not just a property of the first entries in the rank ordered list: approximately 70% of the 1,000 most highly ranked pairs are names of one kind or another. Presumably this reflects the fact that the words making up such names are truly very strongly associated.

The estimator becomes more conservative as the confidence level approaches unity, which in effect reduces the value of the statistic for lower frequency items. For example, at the $s = 0.999$ confidence level the three most highly ranked pairs are "Guzman Cabrera", "Polly Peck" and "MERRILL LYNCH", each of which occurs more than 10 times in the corpus.

Finally, it is worth noting that while pointwise mutual information is a symmetric measure, the conservative estimate is not. That is, the statistic based on equation (7), namely:

$$\widehat{\mathrm{MI}}'(w_1, w_2) = \log_2 \frac{\theta_-(n_{12}, n_1)}{\theta_+(n_2, n)} \qquad (9)$$

is in general not the same as (8).[1] In practice, however, the $\widehat{\mathrm{MI}}$ and $\widehat{\mathrm{MI}}'$ statistics produce

[1] Note that the likelihood ratio statistic proposed by Dunning (1993) is asymmetric in this sense also.

very similiar ranks: the first 15 pairs on the rank ordered lists are identical, for example.

4 Conclusion

This squib has shown how confidence intervals can provide a way of conservatively estimating the strength of association between pairs of elements that is robust in the presence of extremely small sample sizes, effectively trading recall for precision. The pointwise mutual information statistic based on these confidence intervals does not show an obvious bias toward either low or high frequency pairs, and the confidence level s provides an effective means of adjusting the precision/recall tradeoff. Finally, it is straightforward to use confidence interval estimators to obtain conservative versions of other statistics also.

References

Abramowitz, Milton and Irene A. Stegun. 1965. *Handbook of Mathematical Functions.* Dover, New York.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Manning, Chris and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, England, 2nd edition.