A Better Model for Term Frequencies

Jason D. M. Rennie
jrennie@gmail.com

April 7, 2005

1 The Poisson

The most traditional term frequency model is the Poisson,

$$p_P(x) = \frac{1}{e^\mu}\frac{\mu^x}{x!}, \qquad x = 0, 1, 2, \ldots, \tag{1}$$

where x is the frequency. It is a discrete distribution; $e^\mu = \sum_{x=0}^\infty \frac{\mu^x}{x!}$ is the normalization constant. The mean is μ:

$$E[x] = \sum_{x=0}^\infty \frac{x\mu^x}{x!e^\mu} = \mu \sum_{x=1}^\infty \frac{\mu^{x-1}}{(x-1)!e^\mu} = \mu \sum_{x=0}^\infty \frac{\mu^x}{x!e^\mu} = \mu. \tag{2}$$

The second moment is $\mu^2 + \mu$:

$$E[x^2] = \sum_{x=0}^\infty x^2 p(x) = \mu \sum_{x=0}^\infty \frac{(x+1)\mu^x}{e^\mu x!} = \mu(\mu+1). \tag{3}$$

Thus, the variance, $E[x^2] - \mu^2$, is also μ.

It is well-known that the Poisson serves as a poor model of term occurrence. Figure 1 shows the empirical term distribution from a collection of threads on a restaurant discussion board, and the Poisson distribution with mean parameter set to the empirical mean. The Poisson clearly underestimates the probability of a word occurring repeatedly in a document. In the set of threads, there are 17 cases of a word occurring exactly 20 times; this corresponds to a empirical rate of 2.2×10^{-5}. However, the Poisson predicts an empirical rate of less than 10^{-40}—the chance of even seeing a single word occur 20 times is extremely small. Church [1] discusses the flaws of the Poisson and suggests alternatives such as the negative binomial [4] and Katz' mixture model [3].

2 The Binomial

The Binomial might be thought of as a document-length-aware version of the Poisson. While the Poisson puts non-zero probability on all non-negative integers, the Binomial stops at the length of the documents. However, since

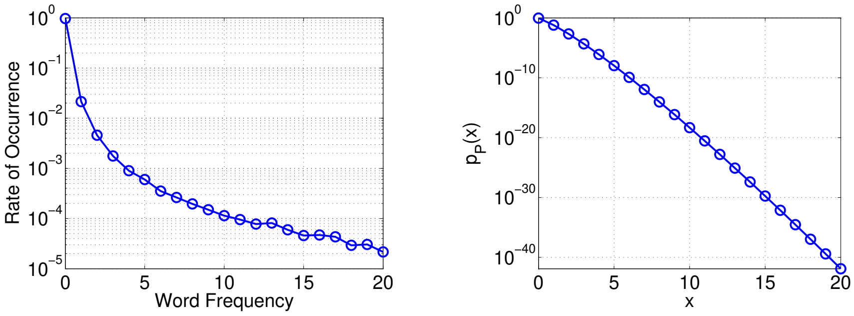

Figure 1: The left plot shows the empirical distribution of frequencies from a collection of postings to a restaurant discussion board. The right plot shows the Poisson distribution with mean parameter set to the mean empirical word frequency. Note the different ranges for the log-scale y-axes.

for most words, the expected number of occurrences is a fraction of document length, the amount of density the Poisson applies to infeasiable frequency values is miniscule. In practice, the Binomial and Poisson share much in common.

The Binomial distribution is

$$p_B(x) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \tag{4}$$

where θ is the mean rate of occurrence and n is the document length. Note that for n large and x small,

$$\frac{\mu^x}{x!} \approx \frac{n!}{x!(n-x)!}\theta^x \tag{5}$$

(since $\frac{n!}{(n-x)!} \approx n^x$). The mean of the Binomial is $\mu = n\theta$,

$$\begin{aligned} E[x] &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} && (6)\\ &= n\theta \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}\theta^{x-1}(1-\theta)^{n-x} && (7)\\ &= n\theta. && (8) \end{aligned}$$

Similarly, the second moment is $n\theta - n\theta^2 + n^2\theta^2$,

$$\begin{aligned} E[x] &= \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} && (9)\\ &= n\theta \sum_{x=0}^{n-1} (x+1) \frac{(n-1)!}{x!(n-1-x)!}\theta^x(1-\theta)^{n-1-x} && (10)\\ &= n\theta(n-1)\theta + n\theta. && (11) \end{aligned}$$

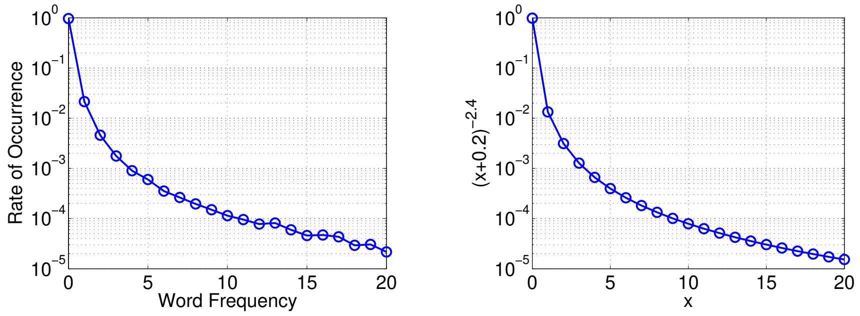

Figure 2: The left plot shows the data empirical frequency distribution (repeat). The right plot shows $(x + 0.2)^{-2.4} = e^{-2.4 \log(x+0.2)}$. Note the log-scale y-axis.

Thus, the variance is $n\theta(1-\theta)$. Note that when θ is small, the variance is nearly the same as the Poisson variance.

As one might expect, if we let $n \to \infty$, and $\theta \to 0$ in such a way that $n\theta = \mu$ remains constant, then the Poisson Distribution tends to the Binomial Distribution. See [2] (§4.3) for the proof.

In practice, the difference between the Poisson and Binomial is small. The Binomial models word frequency distributions as poorly as does the Poisson.

3 A Log-Log Model of Word Frequency

Note that the empirical word frequency plot has a shape similar to that of of an inverted log function, $-\log(x)$. A bit of fiddling with constants gives us a close visual match to the empirical distribution. In figure 2, we plot $e^{-2.4 \log(x+0.2)}$. Like the empirical distribution, this distribution gives significant weight to large values of word frequency. It appears to be an excellent fit. However, there are two major issues we have yet to address: (1) will this work as a good model for individual words with varying rates of frequency? and (2) how do we account for changes in document length?

4 Individual Word Frequency Distributions

We have seen that our log-log model of word frequency can model the frequency distribution for all words over an entire collection of text. But, can it effectively model the word frequency distribution of a single word? In particular, what about a word that is extremely frequent. One could imagine that an extremely common word, such as "the" might have a higher rate of single occurrences than zero occurrences. Figure 3 gives the empirical frequency distribution of "the" in a collection of restaurant discussion board postings. In fact, there appears to be a gradual decrease in rate of occurrence for increasing word frequency; there is no clear "peak," but it is clear that lower word frequencies have a higher rate

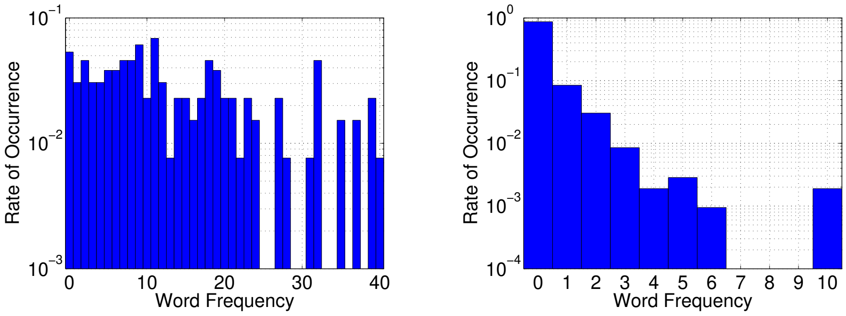

Figure 3: The left plot shows the empirical frequency distribution of "the" in our collection of restaurant discussion board postings. The right plot shows the empirical frequency distribution of eight words that each occur a total of 30 times in the posting collection. Both distributions can be modeled effectively by the log-log word frequency model.

of occurrence. Ignoring the clear randomness in the data, it should be clear to the reader that appropriate selection of constants would lead to a good fit of the log-log model to this data.

What about less frequent words? Figure 3 also gives the empirical distribution for the eight words[1]that each occur exactly 30 times in the data. This distribution, though less "smooth," shares much similarity with the overall empirical distribution. Again, appropriate setting of constants would lead to a log-log distribution that would fit the data well.

5 Document Length

A lingering issue is: how does the model account for different document lengths? Figure 4 shows empirical word frequency distributions for a collection of short (less than 200 words) threads, and for a collection of long (500 words or more) threads. Possibly most striking is the fact that the plots are not so different. A close inspection reveals that the the "long" plot has higher rates of occurrence for all word frequencies of 1 or more. But, the overall shape of the two distributions is very similar. Whereas one might expect that a different parameterization would it appears that a slight change in parameter values would be sufficient. Here we suggest that the necessary modification might be to change the constant in the exponent in order to account for different document length. I.e. the -2.4 from the plot in figure 2 would become smaller (more negative) for shorter documents, larger (more positive) for larger documents. We explore this in more detail in a later writeup.

[1]For the curious, those eight words are "things," "buffet," "indian," "stuff," "meal," "bad," "grilled," and "thought."

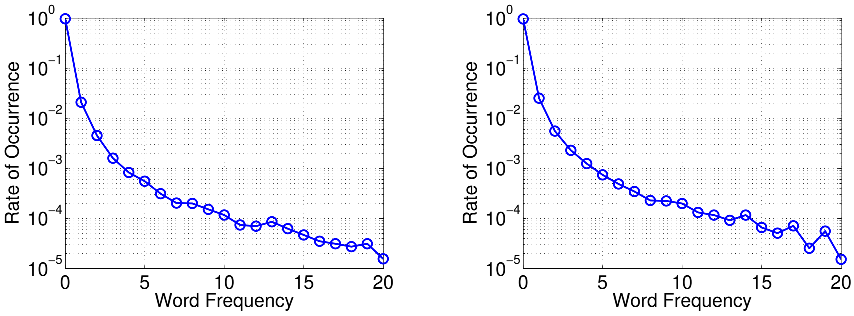

Figure 4: Both plots show empirical word frequency distributions from the collection of restaurant discussion board postings. The left plot represents statistics from (short) threads with less than 200 words. The right plot is compiled from (long) threads with 500 or more words. As one would expect, for frequency values greater than 1, word frequency empirical rate of occurrence is higher for the long threads.

6 Appendix

A useful identity, taken from [2] (§2.1).

$$\binom{n}{r-1} + \binom{n}{r} = n!\left(\frac{1}{(r-1)!(n+1-r)!} + \frac{1}{r!(n-r)!}\right) \tag{12}$$

$$= n!\left(\frac{r}{r!(n+1-r)!} + \frac{n+1-r}{r!(n+1-r)!}\right) \tag{13}$$

$$= \binom{n+1}{r} \tag{14}$$

References

[1] K. W. Church and W. A. Gale. Poisson mixtures. *Journal of Natural Language Engineering*, 1995.

[2] J. G. Kalbfleisch. *Probability and Statistical Inference, Volume 1: Probability.* Springer-Verlag, 1979.

[3] S. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–60, 1996.

[4] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts, 1964.