**Details on the Zipf Law**

Most corpora contain very many low-frequency words (words that occur only once or twice in the entire corpus), but only very few high-frequency words (words that occur hundreds or thousands of times in the corpus). If one characterizes each distinct word in the corpus by its number of occurrences, a histogram of the number of word occurrences *x* shows a pattern of high frequencies for (words with) low number of occurrences *x* and low frequencies for (words with) high number occurrences *x*. For each corpus one can draw the histogram of the number of word occurrences and calculate its characteristics such as its mean (that is, the average number of word occurrences) and its median. Histograms of the word frequencies and histograms of the logarithms of the word frequencies exhibit rapid exponential decays.

What is so striking is that the distribution of the number of word occurrences  is mathematically very simple. It is described by a **power law**:the numbers of word occurrences (exactly one, two, … occurrences in a corpus) have frequencies. The distribution decays exponentially with the increasing logarithm of the number of word occurrences, giving high probabilities to low-frequency and small probabilities to high-frequency words. The exponential decay depends on the corpus. For  close to one, the distribution of the number of word occurrences decreases very slowly and the mean number of word occurrences is large. For large , the exponential decay is steep and the mean number of word occurrences is close to 1, implying that (most) words in the corpus are essentially singletons (unique words) occurring just once. The larger the, the more concentrated the distribution of the number of word occurrences.

It is certainly an interesting property of human language that word frequencies vary according to such a simple law – in particular, a law that does not depend upon a word’s meaning; see Wigner (1960) and Piantadosi (2014) for further discussion. The law states that given some corpus of natural language utterances, the frequency of a word is inversely proportional to its rank in the frequency table. This is known as Zipf’s law, named after the American linguist George Zipf (1936, 1949).

For random variables with continuous outcomes, the distribution with densityfor  is known as the Pareto distribution; the exponent is the parameter of the distribution. The Pareto distribution is commonly used in survival analysis. The constant of proportionality is given by  (as the area under the density curve must be one); hence for . More generally, the Pareto distribution can be defined from any minimum possible valueonwards; then  for .

For  close to one, the density decreases very slowly. The larger the , the steeper the exponential decay. It can be shown that the mean of the Pareto distribution is given by (and in the general case). The mean reflects the average of the distribution of the number of word occurrences. A small  close to 1 implies a large mean. A large implies a mean close to 1, implying that (most) words in the corpus are essentially singletons (unique words) occurring just once. The larger the , the more concentrated the distribution of the number of word occurrences.

The Pareto distribution is a continuous distribution. Note that here we are dealing with a discrete distribution as the numbers of word occurrences *x* in a corpus are discrete and positive (starting from 1 onwards). Since probability mass functions must add to 1, it is straightforward to determine the proportionally constant. It follows that the probability mass function is , for , and parameter . More generally, the distribution can be defined from any minimum possible integer valueonwards; then  for . While this material may appear overly technical and complicated, it is needed to explain the estimation of the exponent in Zipf's power law; see the discussion below. Also note that the normalization constant of the probability mass function can be calculated from the Riemann zeta function].

How can one check from the data whether this law applies to a particular corpus of *N* distinct words? Plotting the number of words of given frequency of occurrence in the corpus against the frequency of occurrence should show an exponential decay with a long right tail. In such plots you may have to omit high-frequency words (frequencies far out on the right tail of the distribution) as otherwise you would not be able to appreciate the exponential decay. A better graphical display, motivated by applying the logarithm to both the left and the right hand side of Zipf’s law, , plots the logarithm of the number of words of given frequency of occurrence in the corpus against the log frequency of occurrence. If Zipf’s law applies, one would observe a linear pattern in the log-log scatter plot and the negative value of the slope in this linear function would provide an estimate of the parameterof the power distribution. However, note that the plotting results in the tail towards the right-hand side of the graph can become very noisy. This happens because the numbers of words that occur with high-frequency become small, implying that there will be a large range of word frequencies for very similar low number of occurrences.

One could use the negative of the least squares estimate from the log-log scatter plot,

  ,

as an estimate of the exponent . See, for example, Chapter 2 in Abraham and Ledolter (2006). Here *n* is the number of distinct word frequencies among the *N* words of the corpus, say *n=*656, with *i* from 1 (words mentioned once, and), 2 (words mentioned twice, and), … , to 656 with if the most-frequent word in the corpus is mentioned 4,123 times. The, for , are the number of words with given word frequency.

However, the least squares line is being "pulled" by the small number of occurrences of high-frequency words. The points on the log-log scatter plot with small numbers on the y-axis (such as log(1) and log(2)) represent a very large range of word frequencies, and the points with the largest word frequencies on the x-axis have a large influence on the estimate of the slope. It turns out that the maximum likelihood estimate (MLE) is a much better estimate.

For the *N* distinct words of the corpus with frequencies, where some of the will be the same, the log-likelihood function of the discrete distribution implied by Zipf’s law is given by

.

This function can be maximized with respect to the exponent . The normalization constant can be calculated from the Riemann zeta function. The attached R program evaluates the log-likelihood function over a grid of and finds its maximum empirically. Alternatively, one can use the R library **poweRlaw** to obtain the MLE estimate directly. This is simpler as one doesn’t have to write a code for these calculations.

Sometimes distributions follow a power law only after some minimum value, and in this case one talks about variables having a power tail. A judgement is required to determine the value. One strategy is to perform a scan over all values of , and choose the for which the log-likelihood function is maximized. Once is determined, the usual MLE estimate for can be used. This can all be done with functions in the R library **poweRlaw**.

**Take-away points from Zipf's law**: Finding that there are only few words that occur often and many words that occur rarely has implications for the text analysis. The few words that occur frequently probably occur in most of the documents, and such words will not distinguish one document from the other. The diminished importance of such words is reflected in the tf-idf measure which down-weights the frequency of a word in a document by the document frequency of the word. On the other hand, many of the rare words are often the results of misspelled words, and one would not want such rare words to have an undue influence on the results either. Hence, our requirement that words used for text classification must occur in at least a certain minimal proportion of documents.

**A further comment:** Software packages such as Voyant talk about the word (or vocabulary) density of a document. **Vocabulary density** is a measurement of vocabulary usage in comparison to the length of the document. It relates the number of unique words in the document to the total number of words in the document. For example, assume that the document contains a total of 5,904 words and 1,489 unique word forms. Vocabulary density, the ratio of the number of words in the document to the number of unique words in the document, is 0.252. Vocabulary density is low if the same words get used over and over again. Vocabulary density is high if many different words are used in a document.

**Simulated Example**

The R code shown on our website (in file **ZipfRCode**) generates *N* = 100,000 word frequencies from the discrete Pareto distribution with exponent . The resulting word frequencies are stored in the vector *x*. Figure 1 plots the number of words with a given word frequency against the word frequency and, as expected, it shows an exponential decay with a very long right tail. Occasionally, very large word frequencies are generated; the very large word frequencies are not shown in Figure 1 as word frequencies larger than 50 are truncated. In Figure 2 we plot the logarithm of the number of words with given word frequency against the logarithm of the word frequency, show the superimposed least squares line, and calculate the least squares estimate . While the log-log graph appears approximately linear, the least squares line gets "pulled" down by the few words with large word frequencies, making the least squares estimate too small; here it actually violates the constraint that .

The log-likelihood function is shown in Figure 3. We calculate the maximum likelihood estimate  and confirm that it is much closer to the true value . Finally, we use the package poweRlaw to carry out the maximum likelihood estimation which confirms the estimate we obtained.



**Figure 1:** Plot of the number of words with given word frequency against the word frequency. Simulated data. 100,000 words with word frequencies generated from the discrete Pareto distribution with exponent .



LS estimate of alpha = 0.9038

**Figure 2:** Plot of the logarithm of the number of words with given word frequency against the logarithm of the word frequency. Simulated data.



MLE of alpha = 1.33

**Figure 3:** Plot of the log-likelihood function implied by the discrete Pareto distribution

(Zipf’s law) against the exponent . Simulated data.

**Applying Zipf's Law to the speeches of the 39th Congress**

Among the roughly 258,000 different words in our corpus of about 97,000 speeches (speeches of any word length), about 188,000 words occurred just a single time, while 22,000 words occurred exactly twice. The most frequent word ("state") occurred about 65,000 times, with the second most-frequent word ("will") occurring 63,000 times. The histogram of the logarithms of the word frequencies is shown in Figure 4, with a very large number of singleton words at log(1)=0. The logarithm of the largest word frequency log(65,000)=11.08 and its frequency 1 are shown at the right tail of the distribution. The exponential decay supports Zipf's power law. The scatter plot of the logarithm of the number of words of given frequency of occurrence against the log frequency of occurrence is shown in Figure 5. The maximum likelihood estimate of the exponential decay is estimated as . For details, see the R program file **Chapter4RCode2**. However, the power law may start from a larger minimum value. Thus we use the R-library poweRLaw to estimate the best minimum valuefor the power law to take effect; we find  and exponential decay .



**Figure 4:** Histogram of the logarithm of the word frequencies for: Speeches of the 39th Congress



**Figure 5:** Plot of the logarithm of the number of words with given word frequency against the logarithm of the word frequency: Speeches of the 39th Congress

**References**

Abraham B, Ledolter J (2006). Introduction to Regression Modeling. Duxbury Press.

Piantadosi ST (2014). Zipf’s word frequency law in natural language: A critical review and future directions. Psychon Bull Rev. 21(5): 1112–1130.

Wigner EP (1960). The unreasonable effectiveness of mathematics in the natural sciences. Communications on Pure and Applied Mathematics.13(1):1–14.

Zipf G (1936). The Psychobiology of Language. London: Routledge.

Zipf G (1949). Human Behavior and the Principle of Least Effort. New York: Addison-Wesley.