Estimating the Zipf law

Richard D. Gill

Bachelor seminar project 2015

The Zipf law is a discrete probability distribution on the positive integers: in fact, one can say that it is "just" a generalised power law. The probability that X equals x, call it $p(x)$, is proportional to $x^{-\beta}$, $x = 1, 2, \ldots$. Obviously, we need to have $\beta > 1$, otherwise the series is not summable and we cannot find a normalisation constant. However, often one does allow $\beta \leq 1$ by assuming that the random variable X is actually bounded by some (large) constant, say M. Also, a further shift parameter $C > 0$ is often introduced: the probability that X equals x is proportional to $(x + C)^{-\beta}$, $x = 1, 2, \ldots, M$.

Note that the probabilities $p(x)$ are decreasing.

This model is frequently used in linguistics and other fields. The index x labels words in some language, $p(x)$ stands for the probability of the xth most frequent word. Now if we take a finite sample of words from spoken or written texts in a hitherto unknown language, we do not know in advance which word is the most frequent word, which word is the second most frequent word, and so on. If we observe N words we can reduce the data to the ordered observed frequencies $R_1 \geq R_2 \geq ...$; non-negative integers, adding up to N (in fact, the data is a number-theoretic *partition* of the integer N). The point here is that the most common word in our sample, which was observed R_1 times, is not necessarily the most common word in the population – the word whose probability, by definition, is $p(1)$. And so on: the second most common word in our sample, which we actually observed R_2 times, is not necessarily the word with the second largest probability $p(2)$.

I have noticed in some applications in genomics that a certain simple plot appears to give us a rather simple way of estimating the power β. This was confirmed by some simulation experiments. I have not found much literature on the problem yet, but there must be a lot. The aim of the project is to find out if what I saw in these plots is for real.

https://en.wikipedia.org/wiki/Zipf's_law

http://rpubs.com/gill1109/simulation