Mathematical programming
The Joy of Computers 1983
- Peter Laurie
001.6404 LAU

ZIPF'S LAW

Most of the problem – and the challenge – and the fun of computing is that it lets you deal with very much greater volumes of information much faster than you can using paper. You can handle more things in more different ways; and, not surprisingly, new kinds of behaviour emerge.

One of the most interesting insights into the behaviour of masses of information was discovered by an American sociologist named George Zipf back in the 1940s.* Although the work has great relevance to a computerate view of the world, it was done – with immense labour – using paper and pencil. Zipf started out looking at the frequency of words in English text. He went through large chunks of prose, counting how often each word was repeated. He then arranged the results in order of rank, so that the most frequently used word came first, then the next and so on. Then he drew a graph of the results.

This looks much as you would expect: as words get rarer, they are used less often. However, the graph does not dive straight into the bottom axis because there are always rare new words coming in to trickle the curve out to the right.

The next thing he did was to plot rank against the *logarithm* of the frequency. Now this looked altogether more intriguing. He got a straight line. To the nonmathematical, this may not be very interesting, but it meant that frequency and rank were connected by an equation like this:

$$\log F = -k(\log R) + 1$$

where k and 1 are constants. Now, we can look at 1 as the log of some other constant, say m, so this equation becomes:

$$\log F = -k(\log R) + \log m$$

or:

$$\log F = \log ((m/R)+(k))$$

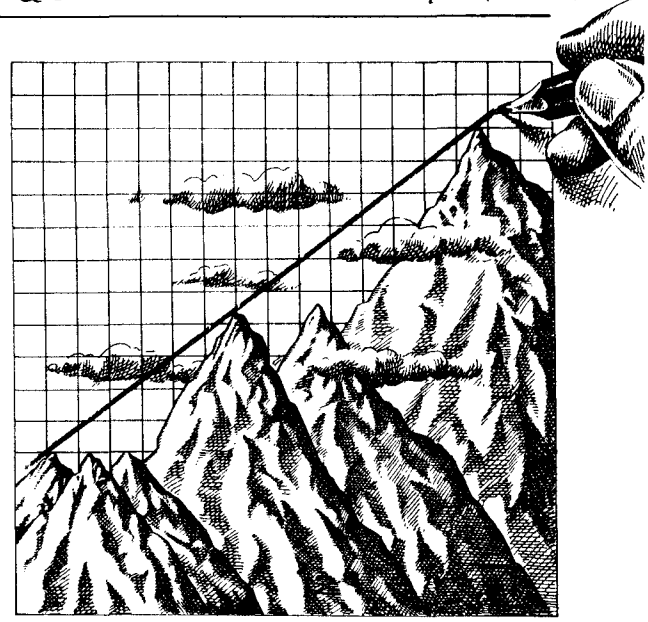
It turned out that k was pretty well 1, so, taking antilogarithms:

$$F = m/R$$

which, in plain English, meant that the frequency of a word was proportional to 1 over its rank. The third most common word, for instance, was found just one third as often as the most common word. The hundredth most common word, one hundredth as often as the most common.

All we need to know now is: how common is the most common word? And we do not have to spend weeks counting words in Shakespeare to find out. If we add up: 1/1 + 1/2 + 1/3 + ... 1/n, we find that although the sum gets big quite rapidly to begin with, it soon starts to level out. You can do it with this program:

```
10 K=0:N=0
20 N= N+1/K
30 ? K; N;" ";
40 K=K+1
50 GOTO 20
```



While I was writing this I had the program running on another machine, and got these results:

K	N
10	2.929
100	5.187
1000	7.485
10000	9.788
100000	12.091

As k gets bigger, n stops growing so fast, and when it gets very big, n tends to about 12. You can, of course, carry on running the program until k is 1 million, 10 million and so on, but it will take an awfully long time and will not tell you much more than you knew already.

This is interesting, odd and useful. Zipf found that the same rule, or something very like it, applied all over the place. It applied to the sizes of cities and towns: in any country the second largest city was roughly half the size of the largest; the third one third the size, right down to the tiny villages. So, if you knew the number of people in a country, you could calculate, say, how many towns there ought to be with between 100 and 200,000 inhabitants.

Records of people's addresses in grid-patterned American cities showed that half as many people had walked two blocks to find their spouses as those who discovered love on the block in which they themselves lived; one third as many had walked three blocks, and so on.

Zipf's law also applied to the sizes of businesses – if you know the total value of microcomputers sold in a year, you can use Zipf's Law to work out roughly the turnovers of the different companies in the industry. The biggest company should sell twice as many as the second, and so on down the line. If there are already a hundred companies, and you want to start another, you can work out how many computers you ought to sell each year.

*G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, 1949.