The following article was published in 1985 in the UK Journal of Naval Science Volume 10, No. 3 pp 180 - 185

For a recent review of Zipf's Law see http://en.wikipedia.org/wiki/Zipf%27s_law

ZIPF'S LAW

Geoff Kirby

Abstract

In 1949 G. K. Zipf published a 'law' which relates the size of organisations and their ranking. Although apparently instantly forgotten, it was revived in 1975 because it was believed by some to throw valuable insight into the way that organisations grow and die. The validity of this 'law' is examined with examples. A tactical model of organisational behaviour has been simulated on a home computer which gives excellent agreement with Zipfs Law.

Introduction

Make a list of the sizes of organisations forming a well defined group, such as student populations at English universities, and arrange them into rank order. Plot the sizes against rank on double-logarithmic graph paper and you will often find that the relationship is close to a straight line. If we denote the size of an organisation of rank R by S_R then Zipf noted [1] in 1949 that for a very large number of organisations (and that term is used in a very wide sense as we shall see below) we have

$$S_R = S_1/R^k$$

where k is a constant. What has, more recently, become known as Zipf's Law is the case where k is equal to one.

Is this 'Law' a real effect or an artefact of the selection of data? If the 'Law' has some general validity what does it tell us about the manner in which organisations evolve?

Is the 'Law' a real effect?

I said that the term 'organisation' would be used in a very wide sense. Consider the frequency with which words are used in the English language. From rudimentary origins the vocabulary has grown and evolved. Bright modern words, such as 'Punk` spring up to oust old tired words such as 'Flapper' and words proliferate as we struggle to invent ever more superlative superlatives.

The development of vocabulary has much in common with other human competitive endeavours.

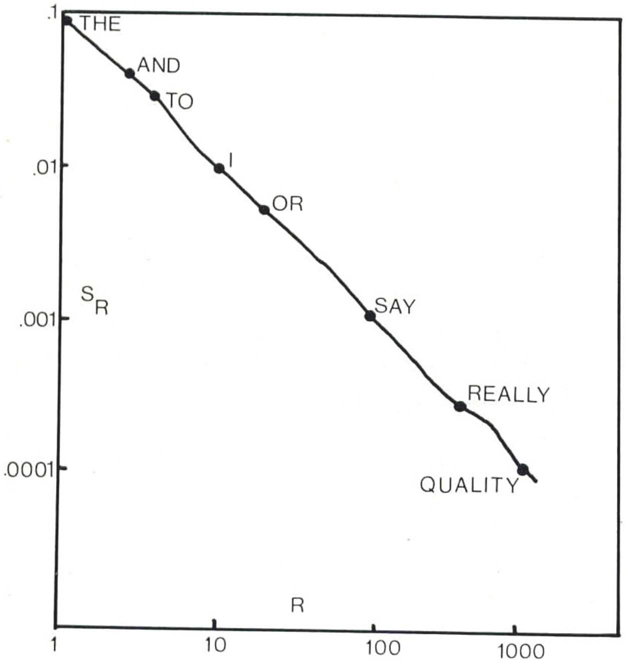

Figure 1 Frequency of word usage in English

Figure 1 shows a plot of frequency of word usage in English plotted against rank. The most popular word, at least in polite conversation, is **THE** which is used about once in every twelve words. This has a rank of one. As we move to higher ranks we encounter less well known words. **QUALITY** occurs about once in every thousand words. The curve is remarkable. For over three orders of magnitude it follows very closely Zipf's Law in its currently used form with k equal to one.

This, and some of the following examples, are taken from a paper by Scarrott [2]. More modern examples have recently been generated by the author.

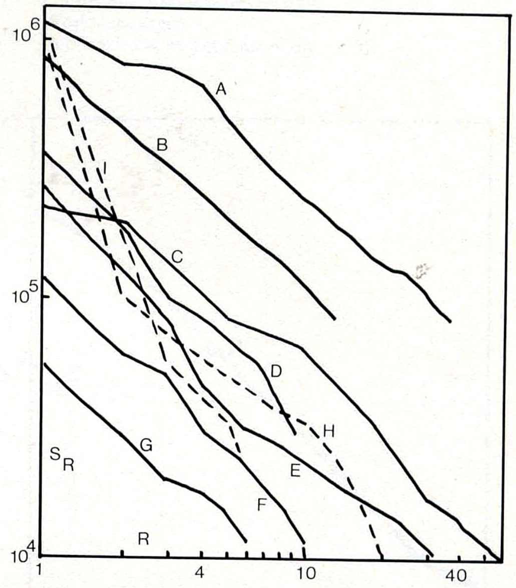

Figure 2 Ranking of world cities by population, see table 1 for key.

For example, Figure 2 shows the population of cities in several countries in rank order. Table 1 gives the key.

A	United States	B	China
C	West Germany	D	Spain
E	France	F	East Germany
G	Switzerland	H	United Kingdom
I	Mexico		

Table 1. Key to figure 2

We see that Zipf's Law is broadly obeyed by all the countries plotted except for the United Kingdom and Mexico. However, in both these exceptions the distortion is due to a disproportionately large principal city. If the ranking of the first city is ignored, the resulting curve has a slope close to -1.

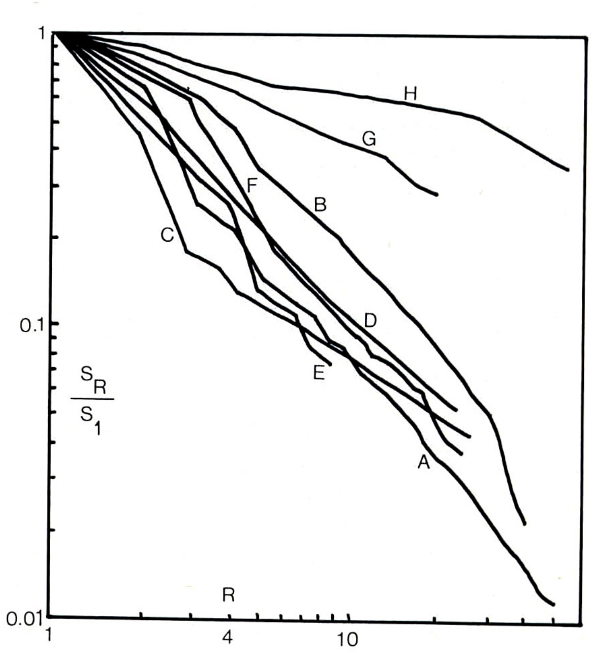

Figure 3 Ranking of various groups, see table 2 for key

A	Populations of all countries
B	Number of ships built by all countries
C	Students at English universities
D	Building Societies by assets
E	Populations of World's religions
F	US insurance companies by staff
G	World languages
H	English public schools by students

Table 2. Key to figure 3

Figure 3 shows a variety of organisations. Here again the curves have a slope close to -1 except for two cases, G and H.

In this figure the size of groups are normalised to the size of the largest group.

The evidence appears to be strong that Zipf's Law is trying to tell us something about the way the organisations evolve.

Are we being too gullible?

For example, the nature of the plot means that we will always have a curve starting from the top left hand corner which slopes down towards the right. This is already a biasing factor. Could a random distribution of sizes, when ranked, show a similar curve to Zipf's curve? Only with some difficulty. A set of random numbers drawn from a uniform distribution has a linear variation of S_R with R which is totally different from Zipf's Law. The distribution of random numbers needed to reproduce Zipf's Law would be so unusual as to demand a rational mechanism for its introduction.

Consider curves C and H on Figure 3. The curve for student population in English universities obeys Zipf's Law very accurately. The curve for student population in public schools is far from Zipf's Law.

Why are these two populations fundamentally different?

A competition for students is present in both cases. Both types of educational establishments have similar time spans over which the present population has evolved. This type of comparison of closely related organisations suggests that Zipf's Law may be an artefact of data selection.

Two examples of this will now be shown.

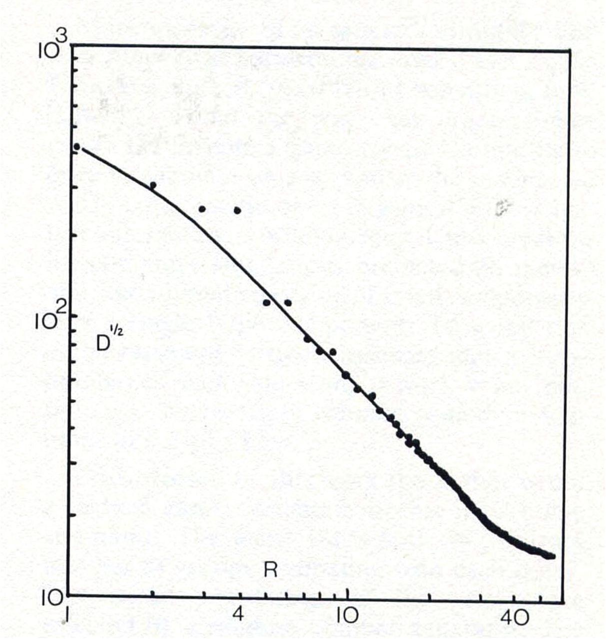

Figure 4 Ranking by size of Solar System bodies

Figure 4 shows a very good demonstration of Zipf's Law with the slope of the data lying very close to -1.

So, what is being plotted? It is the square root of the diameters of the solar system bodies - planets, satellites and asteroids. Because Zipf's Law is so remarkably well obeyed should we speculate on the origin of the solar system with the square root of body diameter as an ‘organisationally significant’ parameter? It would seem foolish to do so because the square root of diameter was chosen to fit Zipf’s Law and not vice versa.

As a second spurious example, run your eye down a long list of physical properties of nature concentrating only on those with dimensional units. A suitable source [3] yields a large number of values. Now note the first digit in the values and rank them.

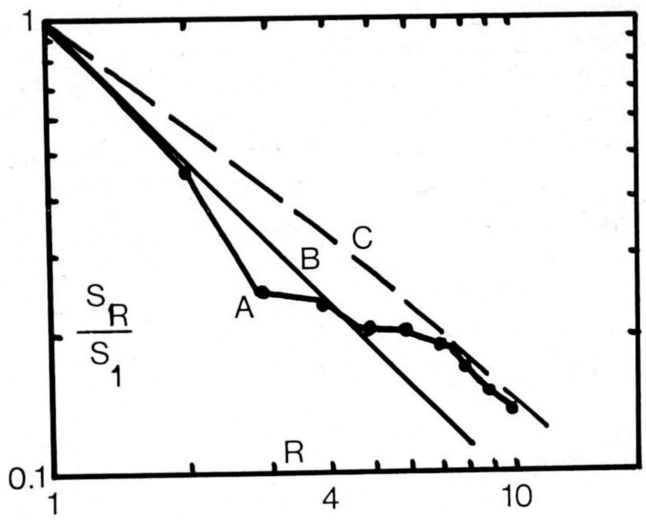

Figure 5 Frequency of occurrence of first digits in tables of physical properties

Figure 5 shows such a plot. The curve A shows the ranking obtained from 175 first digits in tables of thermal conductivity, surface tension and densities of the elements. Curve B is the Zipf’s curve. Is this broad agreement fortuitous or have the numerical values of the physical properties of natural materials arranged themselves to suit Zipf?

In fact, the distribution of initial digits theoretically follows the law shown by curve C which is given by

$$S_R = \frac{\log_{10}(R+1)-\log_{10}(R)}{\log_{10}(2)}$$

and, given a bigger sample of initial digits, the data would fall upon curve C.

This relationship occurs because the distribution of initial digits must be independent of the choice of units used to describe the physical value of being recorded. If we change the units of measurements, say from cm-grm-second to foot-ounce-fortnight then the changes in the numerical values that occur must give an invariant initial digit distribution. The logarithmic variation is the only one to maintain the same distribution of initial digits under all changes of measurements units. (This is not a precisely valid argument and the logarithmic distribution only works for physical properties spanning several decades.)

The point to note here, as in the solar system example, is that some distributions can be found that mimic Zipf’s Law without the slightest hint that the same mechanisms underlie their production as underlie the production of, say, a word frequency count.

A mechanism for Zipf’s Law

We have failed to demonstrate unambiguously that Zipf’s Law is on a par with Newton’s Law of gravity or any of the other Laws of Nature. However, many workers believe that Zipf’s Law is a manifestation of a fundamental but poorly understood ‘Law of Nature’ whose investigation can lead us to great understandings of natural organisational mechanisms.

Mandelbrot wrote in 1977 [4]: *‘The failure of applied statisticians and social scientists to heed Zipf helps to account for the backwardness of their fields.’*

More recently, in an article entitled ‘Will Zipf Join Gauss’, Scarrott has argued [2] that Zipf’s Law does indeed try to tell us something fundamental about the way that organisations grow. He invents a game called Cosmic Patience to simulate on a computer the manner in which components of an organisational structure interact to grow and decay. Unfortunately his computer simulations predicted an inevitable and ultimate merging of small components into a single large component. To avoid this embarrassment Scarrott assumes that a large number of such games are at work at any one time and, by averaging, obtains something that looks like Zipf’s Law.

Unconvinced by this work the author wrote a tactical game simulation on his BBC home computer.

The game starts with an arbitrary number of groups competing with each other. The initial populations of the groups are selected by a random number generator. The initial distribution of members in the groups has no effect because after many iterations the distribution of members in size reaches an equilibrium. The rules are:

1. At each iteration a group chosen at random absorbs any one smaller group also chosen at random.

2. A new group springs into the sample and steals a fraction from each of the other groups.

The justification of these rules comes from observing the growth of towns and cities. A large city expands and absorbs surrounding villages which lose their identity in a counting exercise, just as London has absorbed ancient and picturesque villages such as Sudbury. The members of more distant small villages also migrate to the cities to work but there is a flow out of the cities and towns to the villages.

The latter is particularly well observed in UK south coast resorts which have a growing elderly population. Bournemouth is now a very large town having been fed by elderly people retiring from larger towns and cities. In Victorian times it was a small village over-shadowed by Poole.

Thus, some small villages grow at the expense of their neighbours and become large enough to be counted in the game whilst others lose their identities.

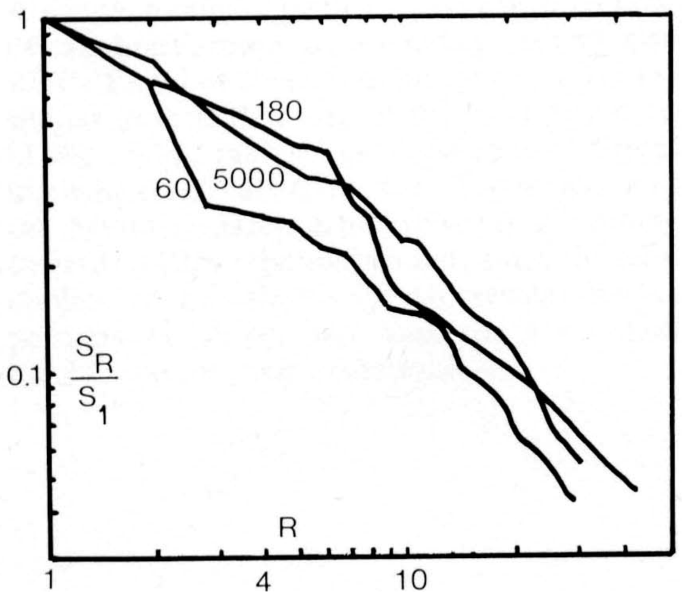

Figure 6 Results of gaming simulation

Figure 6 shows the result of a Monte Carlo simulation. The three curves show the sizes of the groups plotted against ranking after 60, 180 and 5000 iterations.

We see from the curves that the sizes of the groups above about the fourth ranking have stabilized about a Zipf slope of -1. This is an improvement over Scarrott`s model [2] since only one game is assumed active within an ensemble of groups, whereas Scarrott has to assume a large number of simultaneous games in progress.

Acknowledgement

I wish to thank my children for allowing me occasional access to my computer during short breaks in their (seemingly) never ending battle to defend Earth against the alien hordes.

References

1 Zipf G. K. *‘Human Behaviour and the Principles of Least Effort`*, Addison-Wesley, 1949.

2 Scarrott G. S. *‘Will Zipf Join Gauss?’*, New Scientist, 16 May 1975, pp. 402-404.

3 Kaye G., and Laby T. *"Tables of Physical and Chemical Constants'*, Longmans, 1974.

4 Mandelbrot B. B. *‘The Fractal Geometry of Nature`*, W. H. Freeman, 1977.