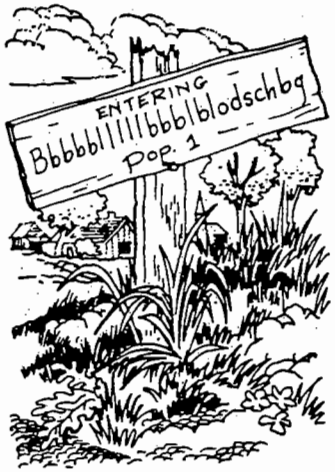

The Sizes of Things

Herbert A. Simon *Carnegie-Mellon University*

On Figure 1 are drawn four lines. The lowest one, a simple straight line inclined at a 45° angle, serves merely for purposes of comparison in describing the three slightly wavy lines. The three wavy lines—and particularly the two just above the straight line—depict some curious facts about the world. Whether they are significant facts as well as curious facts is a question we examine.

The lower broken line relates, on a logarithmic scale,[1] the 1980 populations of the 20 largest cities in the United States to the ranks by size of the cities,

[1]The common logarithm is probably familiar as a tricky device for multiplying numbers through a process of addition. Another way of looking at the logarithm is that taking the logarithm compresses the scale of numbers so as to create a new scale, one that makes multiplying the old number by 10 equivalent to adding one unit to the new number. For example, the logarithm of 10 is 1, of 100 is 2, of 1,000 is 3, and so on. The logarithm of 2,000 is about 3.300 and that of 20,000 is about 4.300. If a city has a population of 5,000,000, then the logarithm of its population is about 6.70. The compression achieved by a logarithm scale increases as the numbers do.

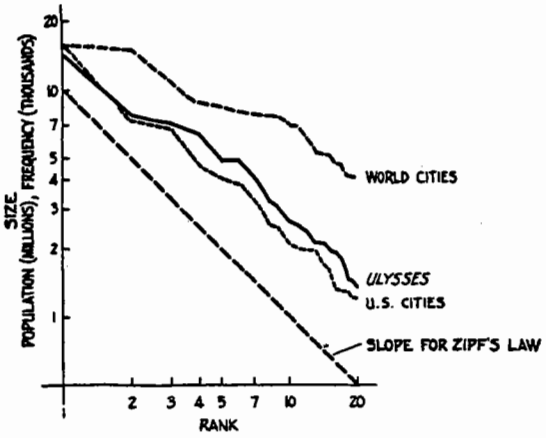

Figure 1 *Logarithm of size plotted against logarithm of rank for frequencies of words and for populations of cities.*

arranged with New York, ranked 1, down to Cincinnati, ranked 20. The population of each metropolitan statistical area was used, not just that within the city limits. The horizontal axis shows, also on a logarithmic scale, the city ranks, from 1 through 20; on the vertical scale are shown the corresponding logarithms of populations in millions of persons. Ignoring the two largest cities (New York and Los Angeles), we can see that the rest of the line is nearly straight and inclined nearly at a 45° angle, parallel to the straight line below. Straightening out the left end of the curve would involve raising New York from about 20 million people to about 30 million and Los Angeles to 15 million (a heavy price to pay for a straight line), but the remaining 18 cities would require very little adjustment—generally less than 10% up or down.

The solid line, just above and very close to the line for cities, shows (again on logarithmic scales) the number of occurrences of each of the 20 words most frequently used in James Joyce's *Ulysses,* when the words are arranged in descending order of frequency of occurrence. For this line, the ordinates show the frequencies of occurrences in thousands. The most frequent word in *Ulysses, the,* occurred 14,887 times; the twentieth most frequent, *all,* occurred 1,311 times. As with the city sizes, the word frequencies lie almost on a straight line, although straightening the line would again require adjustment of the first few words; *the* would have to be increased to about 26,000 occurrences, *of* to 13,000, and *and* to about 8,700. The remaining 17 frequencies are extremely close to a straight line inclined at 45°.

Observe that in these distributions the product of the rank of each item by its size remains constant over the whole scale. If the first item (rank 1) has size 1,000,000, the tenth item will have size about 100,000 (10 × 100,000 =

1,000,000), and the twentieth item will have size 50,000 (20 × 50,000 = 1,000,000). The task before us is to explain why these regularities hold, why the product of number and rank in these distributions is almost constant, and—even more mysterious—why the size distribution of U.S. cities should obey the same law as the frequency distribution of words in a stylistically unusual book such as *Ulysses* (or in any book for that matter). Let's begin with the words.

WORDS: COMMON AND RARE

In the late nineteenth century, several linguists (among them de Saussurre in France) discovered the surprising rank-frequency regularity in the relative contributions of different words to any body of text. Obviously, certain words, such as *of,* will occur rather frequently in almost any English text, while other words, such as *conundrum,* will occur infrequently or not at all. The frequency of any specific word may vary widely from one text to another.

Whenever you arrange the various words occurring in a particular text in the order of their frequency of occurrence—first the word that occurs most often in that text, then the word that occurs next most often, and so on—the regularity depicted in Figure 1 will reappear. The twentieth word on your list will occur about half as often as the tenth word.[2] If you enjoy this kind of numerology, you will find equally startling regularities at the other end of the distribution among the rare words.

About one-half of the total number of *different* words in the text occur exactly once each, about one-sixth occur exactly twice each, and about one-twelfth occur three times each (see Table 1). The ratio $1/[n(n + 1)]$ gives the fraction of all the distinct words in the text that occur exactly n times each. This regularity in frequency of occurrence of the rare words is, of course, the same rank-size law we have been observing at the other end of the distribution, for the rank of a word is simply the cumulated number of different words that have occurred as frequently as it has, or more frequently. Suppose then, as the rank-size rule requires, that $K/(n + 1)$ words occur $n + 1$ or more times each, and K/n words occur n or more times each. Then the number of words occurring exactly n times will be $K/n - K/(n + 1) = K/[n(n + 1)]$.

The rank-size law, often called *Zipf's law* in honor of a U.S. linguist who wrote a great deal about it, holds for just about all of the texts whose vocabularies have been counted, in a great range of languages, not excluding native American languages. But while it holds for *Ulysses,* it fails for Joyce's *Finnegans Wake* (possibly because of the freedom Joyce exercises in creating all sorts of word fragments and variants of dictionary words). A count of ideograms in Chinese texts seemed to show that the law failed; but a recent frequency count of Chinese *words* (each word may consist of one, two, or more ideograms) shows that it fits the Chinese language just as well as it fits others.

[2]In most cases, the first two or three frequencies are substantially lower than the rule predicts, as in Figure 1.

Table 1 The number of rarely occurring words in James Joyce's *Ulysses*

Number of Occurrences (n)	*Number of Words*	
	Actual	*Predicted**
1	16,432	14,949
2	4,776	4,983
3	2,194	2,491
4	1,285	1,495
5	906	997
6	637	712
7	483	534
8	371	415
9	298	332
10	222	272

*Predicted number = $K/[n(n + 1)]$; K = 29,899, the total number of *different* words in *Ulysses.*

Why does this regularity hold? Why should the balance between frequent and rare words be exactly the same in a daily newspaper as in Joyce's *Ulysses,* the same in German books as in English books, or the same in most (not all) schizophrenic speech as in normal speech?

Several answers have been proposed, one of which is typical of the explanations that are provided by probability theory. Probability theory often explains the way things are arranged on average by conceding its inability to explain them in exact detail. To explain the laws of gasses it avoids tracing the path of each molecule.

To explain the word distribution, we make some assumptions that might be thought outrageous if applied in detail, but that might be plausible if only applied in the aggregate. We assume that a writer generates a text by drawing from the whole vast store of his or her memory, and by drawing from the even vaster store of the literature of the language. The former of these processes we might call *association,* the latter *imitation.* Specifically, we assume that the chance of any given word being chosen *next* is proportional to the number of times the word has previously been stored away—in memory or in the literature. Remember, these assumptions are intended to apply only in the large. To accept them, we need not believe that Shakespeare wrote sonnets by spinning a roulette wheel any more than we believe the individual molecules of a gas chart their courses by shaking dice.

If we accept the assumptions, then it becomes a straightforward mathematical matter but one beyond the scope of this essay to derive the probability distribution they imply. The derivation yields what is known as the *Yule distribution.* In the upper range, among frequently used words, the Yule distribution agrees with the rank-size law of Zipf; in the lower range, among rarely used words, it gives precisely the observed fractions $1/[n(n + 1)]$.

Now we see why the *same* distribution can fit texts of diverse kinds drawn from the literatures of many languages. The same distribution can fit because it does not depend on any very specific properties of the process that generated

the text. It depends only on the generator being, in a probabilistic sense, an associative and imitative process. We might even suspect that substantial departures from exact proportionality in association and imitation would not greatly change the character of the distribution. To the extent that the consequences of changing the assumptions have been explored, mathematically and by computer simulation, the distribution has indeed proved robust. We can give Shakespeare and Joyce a great deal of latitude in the way they write without altering visibly the gross size-rank relation of their vocabularies, but as *Finnegans Wake* shows, we can't give them infinite latitude.

MEGALOPOLIS AND METROPOLIS

Having stripped away some of the mystery of the vocabularies of literary texts, we are perhaps prepared to tackle the corresponding regularity in U.S. city sizes. We have seen (Figure 1) that the city populations obey the same rank-size law, to a quite good approximation.[3] If two cities have ranks j and k, respectively, in the list, their populations' ratio will approximate k/j.

The regularity is not just a happenstance of the 1980 Census. It holds quite well for all the Censuses back to 1780. It does *not* hold, however, for cities in arbitrarily defined geographical regions of the world, which are not relatively self-contained economic units. It does not hold, for example, for Austria, or for individual Central American countries, or for Australia. Nor does it hold if we put the cities of the whole world together (see the uppermost curve in Figure 1). In that case, the distribution is still relatively smooth and regular, but population does not drop off with rank as fast as Zipf's law demands. The distribution is flatter, and the largest metropolises are "too small," though, I hasten to add, this phrase should not be interpreted normatively.

(The definition of size for the world's cities differs from that of the U.S. cities and so the actual magnitudes are not quite comparable. The lists from which they are compiled differ in their year, 1975 vs. 1980, and the notion of metropolitan area may not have been used in deciding the size of a member of the list of world cities.)

In the case of city sizes, then, we must be prepared to explain *two* things: why Zipf's law has held for more than two centuries for the cities of the United States, and why it doesn't hold for many other aggregates of cities. Let's start with the former question and ask what the analogues might be to the association and imitation processes that explained the word distributions. More precisely, let's ask what processes would lead cities to grow at rates proportional, on average, to the sizes already achieved (sometimes called *Gibrat's principle*); for that is the main assumption the mathematical derivation requires.

Cities grow by the net balance of births over deaths, and they grow by the net balance of inward over outward migration. With respect to births and deaths,

[3]We can take either the populations of cities as defined by their corporate boundaries, or populations within metropolitan areas as defined by the U.S. Census. The regularity shows up about as well in either case—perhaps it is a little more satisfactory if we use metropolitan statistical areas.

we need assume only that, on average, birth and death rates are uncorrelated with city size. With respect to migration, we assume that migration outward is proportional, on average, to city size (that is, that per capita *rates* are independent of size), and migration inward (from rural areas, from other cities, or from abroad) is also proportional to city size. The last assumption means that the cities in a given size group form a "target" for migration, which is larger, in total, as the total population already living in the cities of that group is larger. (I leave it to the reader to consider the reasons why this might be a plausible assumption, at least as an approximation.)

If we make these assumptions, we are again led by the mathematics of the matter to Zipf's rank-size law. But now it is instructive to ask: Under what circumstances would we expect a collection of cities to fit the assumptions? The answer is that the cities should form a "natural" region within which there is high and free mobility of population and industry, and which is not an arbitrary slice of a still larger region. The United States fits these requirements quite well, while an area playing a specialized role in a larger economic entity might not fit at all (for example, Austria after the dissolution of the Empire, or a country specializing in agricultural exports and having a single large seaport).

If we put together a large number of distributions, each separately obeying the rank-size law, we get a new distribution of the same shape, simply displaced upward on the graph, but with the top few omitted. We would expect the totality of the world's cities to fit the rank-size distribution, except for a deficiency of extremely large metropolises at the very top, and so it does. If we take the published figures at their face value (the definitional problems are severe, and the census counts of varying accuracy), there are somewhat more than 50 urban aggregations in the world having more than 2 million people each. Zipf's law would then call for a New York or a Tokyo of 100 million people, instead of the mere 20 million who now inhabit each of those cities. But the deficiency of cities at the very top (mostly the top 10) is soon largely made up by the numerous cities of over 5 million population each. Already, the tenth city on the list, Paris, has a population of 9.2 million, only 10% fewer than the number demanded by Zipf's law.

The sizes of cities are of obvious importance to the people who live in them, but it is not obvious what practical conclusions we are to draw from the actual size distribution. One *possible* conclusion is that the distribution isn't going to be easy to change without strong governmental or economic controls over places of residence and work. Or, to put the matter more palatably—because we generally wish to avoid such controls—the mathematical analysis that discloses the forces governing the phenomena teaches us that any attempt to alter the phenomena requires us to deal with those forces with sophistication and intelligence.

BIG AND LITTLE BUSINESS

Economists have generally been more interested in the sizes of business firms than they have been in the sizes of cities. Concentration of industry in the hands

of a few large firms is generally thought to be inimical to competition and is generally also supposed to have proceeded at a rapid rate in the United States during the present century.

It has long been known that business firms in the United States, England, and other countries have size distributions that resemble Zipf's rank-size law, except that size decreases less rapidly with rank than in the situations described previously (that is, the ratio of the largest firm to the tenth largest is generally less than ten to one[4]). The slower the decrease in size with increase in rank, the less concentrated is business in the largest firms.

Economists have been puzzled by the fact that the rate of decrease in size with rank, which is one way of measuring industrial concentration, appears to be about the same for large U.S. manufacturing firms at the present time as it was 35 years ago or even at the turn of the century. Even during periods of frequent mergers, the degree of industrial concentration, as measured by the rank-size relation, has changed only slowly.

From our previous analyses, we should be ready to solve the puzzle. Indeed, it can be shown mathematically that under appropriate assumptions about the firms that disappear by merger, and those that grow by merger, mergers will have no effect on concentration. Moreover, the assumptions required for this mathematical derivation fit the United States statistical data on mergers fairly well. In analogy to the processes for words and cities, we can guess what those assumptions—and the data that support them—are like:

- The probability of a firm "dying" by merger should be approximately independent of its size.
- The average assets acquired by surviving firms through mergers should be roughly proportional to the size they have already attained.

And these are indeed not very far from the truth.

Thus a line of scientific inquiry that began with a linguistic puzzle over word frequencies leads to an explanation of a paradox about industrial concentration in the United States. That explanation opens new lines of research for understanding business growth and arriving at public policy for the maintenance of business competition.

Our fascination with rank-size distributions need not stop with the three examples examined here. We may expect the Zipf distribution to show up in other places as well, and each new occurrence challenges us to formulate plausible (and testable) assumptions from which the rank-size law can be derived and

[4]Let m and n be the ranks of two members of a rank-size distribution, and let S_m and S_n be their respective sizes. Then the rank-size law, in this generalized form, requires $S_m/S_n = (n/m)^k$, where the exponent k is a proper fraction. When k approaches unity as a limit, we get the special case of the Zipf distribution. The general distribution is usually called by economists the *Pareto distribution*. If we graph the logarithmic distribution, taking logarithms of both ranks and sizes, we again obtain a straight line with a slope equal to the fraction k. The steeper this straight line (the larger k), the larger are the first-ranking firms compared with the firms further down the list (that is, the larger is k, the greater the concentration).

the occurence explained. I will leave a final example as an exercise for the reader. List the authors who have contributed to a scientific journal over a span of 20 years, or whose names have appeared in a comprehensive bibliography, such as *Chemical Abstracts*. Note the number of appearances for each author, and rank the authors by that number. Then about one-half of all the authors will have appeared exactly once, one-sixth will have appeared twice, and so on; the data will not stray far from the Yule distribution. What are the ways of authors that can provide a naturalistic explanation for that fact?

PROBLEMS

1. From Figure 1, what roughly is the population of Philadelphia, the fifth largest city in the United States?
2. Consider Table 1.
   a. 483 distinct words appear 7 times in the text of *Ulysses*. How was the predicted value of 534 computed?
   b. Suppose the predicted number of words occurring n times is 164. Approximately what is n? (Hint: You will have to use the quadratic formula from high school algebra.)
3. What does the author mean by *association*? By *imitation*?
4. How can the same distribution fit the population of U.S. cities and the frequency of words in a text?
5. What is *Gibrat's principle*? How does it relate to Zipf's law? To the sizes of U.S. cities?
6. State in words the mathematical assumptions that lead to the Yule distribution, first in the case of literary texts, then in the case of city sizes.
7. Can you think of a reason why the few largest cities in the United States might not satisfy the rank-size law?
8. a. How are the Zipf and Pareto distributions related?
   b. In a logarithmic graph of the Pareto distribution, what is k?
9. Suppose IBM's and Xerox's sales rank first and fourth, respectively, Xerox's sales are $1 billion, and business-firm sales obey a Pareto distribution with $k = 1/2$. What would IBM's sales be?
10. If R_n is the rank of a thing of size n, state Zipf's law.
11. Assume the fourth largest city has a population of 10 million. What ranking would you expect a city of 2½ million to hold, according to Zipf's law?
12. The author states that the number of different words occuring exactly n times in a given text equals $K/[n(n + 1)]$. What is K? (Hint: See Table 1.)

13. Refer to the exercise outlined in the final paragraph. What two assumptions would you postulate for this distribution? (Hint: Use the assumptions for mergers or city sizes as a close guide.)

REFERENCE

Yuji Ijiri and Herbert A. Simon. 1977. *Skew Distributions and the Sizes of Business Firms.* Amsterdam: North-Holland Publishing Company.