Foreword

When a book has many authors—as this book has—it is unusual that one of them should be asked to write a foreword. But I was asked, accepted and found myself in a quandary. I could not evaluate the work of my co-authors without also evaluating my own. The best I can do is to focus on this book's goals rather than on the book itself. In this light, I selected some topics that are central to fractal geometry and to this book. For the beginning of this foreword I have selected a broad topic that goes under several names: Zipf's law, Pareto's law, power-law density, hyperbolic probability distribution, scaling or fractal probablility distribution, and simply fractal statistics. For the end I have selected the lognormal probability distribution.

Most of my working life has been spent in the company of one or another of the above terms that refer to fractal statistics. They may "sound" or "feel" very different, but unless they are made specific by additional qualifying terms I found it safe to view them as synonymous or near-synonymous. The single reality to which they refer is typically a collection of empirical data. One can plot them in different ways, but if doubly logarithmic coordinates are used, one observes consistently a scaling range within which the log-log plot is rectilinear. It is appropriate to use the terms "fractal distribution" and "fractal statistics" to denote this fact, but I find it more discrete to use "scaling" or "hyperbolic" distribution, and my physicist friends favor "power-law distribution." The preceding terms do not seek mystery, and all mystery vanishes if one follows probabilists' notation. Let U denote a quantity whose value is random, for example, the height of a man or the size of an oil reservoir selected at random. Let the corresponding lower case letter, u, denote the sample value, as measured in numbers of inches or in millions of barrels. Then we have a one-parameter relation expressing that "the number Nr of cases where $U \geqslant u$ is proportional to $u^{-\alpha}$." Hence the formula

$$\mathrm{Nr}\{U \geqslant u\} = Fu^{-\alpha} \tag{1}$$

Assimilating the relative number of cases to a probability, this reads

$$\Pr\{U \geqslant u\} = \text{probability that } (U \sim u) = Fu^{-\alpha} \tag{2}$$

So far so good, but one is free to plot (1) on transparent paper and then to turn the sheet around. This yields

$$u = F^{-1/\alpha}(\mathrm{Nr})^{-1/\alpha} \tag{3}$$

Fractal Geometry and its Uses in Petroleum Geology and Earth Sciences.
Edited by Christopher C. Barton & P. R. La Pointe.
New York: Plenum, 1995, vii-xi.

The coordinates Nr and u once caught the fancy of George Kingley Zipf (1902–1950). My book, *The Fractal Geometry of Nature* (FGN, 1982) devotes a full page to that writer (pp. 403–4). The last sentence reads "One sees in him, in the clearest fashion, even in caricature, the extraordinary difficulties that accompany any interdisciplinary approach." The sentences before this observe that "I owe a great deal to Zipf . . . Otherwise, Zipf's influence is likely to remain marginal."

Given Zipf's scant knowledge of statistics, he did not view Eq. 2 as just a way of turning the customary probabilistic coordinates around, but as a powerful way of quantifying complicated reality. Among the specific examples he favored were word frequencies in discourse, and city and firm sizes. Take "items," that may be words, or city or firm sizes, and order them by decreasing frequency, population or income. Then the quantity $\mathrm{Nr}\{U \geqslant u\}$ becomes the rank r of an item in this ordering. For example, the biggest city has rank $r = 1$, the second biggest has rank $r = 2$ and so on. Equation 3 asserts that the size of the city of rank r as function of the rank in this ordering is $u = F^{1/\alpha} r^{1/\alpha}$.

Zipf's book, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949) created quite a stir when I was a young scientist in search of unusual facts to investigate. I focused on the case where the "items" were words in some long document, for example a book, and immediately explained Eq. 2 by an argument that is recorded in *FGN*, Chapter 38 and is so straightforward that a sketch of the simplest case can fit here. Take an alphabet of $M + 1$ letters, L_m, with L_0 denoting the improper letter "space." Have "typing monkeys" use this alphabet to produce a random text in which L_0 is used with the probability p_0, and each of the other letters with the probability $(1 - p_0)/M$. A word made of k proper letters followed by a space will have the probability $p_0[(1 - p_0)/M]^k = e^{-k \log B}$, by definition of B. Such a word's rank is $r \propto M^k$. Therefore $k = \log r / \log M$, and the word's probability k

$$p \propto p_0 \exp(-\log r \log B / \log M) \propto r^{-1/\alpha} \tag{4}$$

with

$$1/\alpha = \frac{\log B}{\log M} = \frac{-\log(1 - p_0) + \log M}{\log M} = 1 + |\log_M (1 - p_0)| > 1 \tag{5}$$

There is nothing more to Zipf's law for words. For example, Markovian discourse and other generalizations yield the same result asymptotically. Every generalization involves a complication: the probability of an "m-gram" formed by m letters is no longer the same for all m-grams of a given m. Zipf's law only holds after all m-grams have been sorted out by decreasing probability.

The key to the preceding argument is simple: it relies solely on compensation between two exponentials. That is, it involves an auxiliary quantity k and two constants $\mathrm{A} > 0$ and $\mathrm{B} > 0$ such that rank $\propto \mathrm{A}^k$, and probability $\propto \mathrm{B}^{-k}$, hence,

$$u \propto r^{-1/\alpha}, \text{ where } \alpha = -\log \mathrm{A} / \log \mathrm{B} \tag{6}$$

To identify a suitable k was an easy task in the case of words, as we already know. There are other cases where Zipf's law reduces to the compensation between two exponentials, and the interpretation of k varies from case to case. For example, the Gutenberg-Richter law is an example of Zipf's law in which $k = \log$ (energy of an earthquake) is the seismologist's *magnitude*. The Pareto law of income can be written in Zipf's format, and arises if log (income), the economists' "moral wealth," is exponentially distributed. The exponent has been "explained" in many different ways, both "static" and "dynamic." For

example, log (income) becomes exponentially distributed in the limit if log (income) performs a random walk with downward drift and reflecting lower barrier, or some equivalent assumption. (One serious problem with this is that alternative assumptions that seem before the fact to be equally compelling yield very different results after the fact; for example, eliminating the reflecting barrier would yield the lognormal distribution as will be discussed below.)

For word frequencies and for city or firm sizes, the compensation between the two exponentials can be phrased in several additional ways. There is a "thermodynamical" or "information-theoretical" restatement that brings nothing new, but appears learned. There are also cases where Zipf's law remains altogether unexplained.

In 1953, shortly after the publication of my explanation of Zipf's law for words, I joined the Massachusetts Institute of Technology as a postdoctoral fellow and found myself to be very popular among the linguists. It turned out that many of them had been taken by Zipf's claim that Eq. 3 has a deep meaning for syntax, or perhaps for semantics, and I gained durable praise for showing that this claim is nonsense, that there is nothing in Zipf's law for linguistics. However, it proved very interesting in probabilistic terms and (as told in the last chapter of *FGN*) it somehow started me on a path that led to fractals.

It is a good thing that professional statisticians had no influence on me at the time. Listening to them, I might have reached the conventional conclusion that is recorded in a once-influential book by J. Aitchinson and J. A. C. Brown, *The Lognormal Distribution* (Cambridge, 1957). On pp. 101–2 of that book, we read that "A number of distributions are given by Zipf, who uses a mathematical description of his own manufacture on which he erects some extensive sociological theory; in fact, however, it is likely that many of these distributions can be regarded as lognormal, or truncated lognormal, with more prosaic foundations in normal probability theory." Incidentally, this proves that Aitchinson and Brown did not know what they were talking about. Few other persons knew. Nevertheless, Zipf became a repulsive magnet to professional students of randomness, but also an attractive magnet for nonprofessional dabblers of all kinds.

Once again, the Zipf episode eventually proved to be useful to me, yet I soon concluded that its usefulness had been exhausted, and felt it was buried forever. But I was mistaken. In recent years, Zipf's law has enjoyed a spurt of renewed fascination and overselling. The "bad vibes" that blind overselling had created in the 1950s having been forgotten, it is now a fresh (and incredibly mysterious) key to every form of complexity or to a "linguistic" analysis of DNA structure.

This replay of old dreams as if they were new confirms the magic power of certain words. Past probabilists used to speak of a "law" where today's probabilists speak of a "distribution;" the former is far more impressive. For example, calling a distribution "hyperbolic," "scaling," or "fractal" does not in any way promise that its occurrences have much in common. By contrast, repeated experience shows that "Zipf's law" suggests a mysterious commonalty. Of the same ilk is "$1/f$ noise," a term that necessity often forces me to both use and fight, because experience proves that it suggests to many readers a single underlying phenomenon . . . a suggestion that happens to be very far off the mark. I really hope that "Zipf's law" will crawl back into its grave.

While the occurrence of "Zipf's law" in "linguistics" has been fully accounted for, other examples of scaling distributions range from fully explained to largely mysterious. This brings to light some interesting observations on the history of the scientific method and my own contribution to this book. This also brings up an important historical quirk

associated with the fact that theory is frequently overemphasized in places where it does not belong. This overemphasis explains the orphan status that often befalls empirical discoveries that are not explained and not yet embedded in an over-reaching theory. Allow me to dwell on these issues.

My contribution to this book was published by IBM and copyrighted and it was reasonably widely read, though it gained little immediate influence. On the other hand, no one was willing to publish this work properly, and I knew why: it negated a very powerful dogma. Some petroleum data had come to my attention and I found that the best fit was given by the scaling distribution, which in my circle of economists was ordinarily called a Pareto distribution. In the early 1960s, as already stated, statisticians took it for granted that nothing of interest was fitted by a scaling distribution. But everyone believed that powerful arguments had definitely established that everything of interest was fitted by the Gaussian or by the lognormal distributions. This was not a casual belief, but a consequence of a dearly held cliché and of a dearly held prejudice. That cliché goes back to Auguste Comte (1798–1857) and holds that there is a unique pecking order among the fields of study, moving down from the most to the least scientifically perfect. It followed that the more perfect fields were expected to instruct and guide the less perfect ones, but the converse was inconceivable. In particular, I was often told that empirical evidence in favor of scaling distributions was considered suspect because they were never encountered in physics. In due time, a few odd examples did surface, but in 1950 they were not known. At best, a statistician knew only the examples collected by Zipf, all of which concerned social sciences.

It is chastening to recall that two sets of statistical regularities had been discovered in the 19th century. One set is well known because it inspired the old textbook examples of the fitting of Army conscripts' heights by the normal distribution and of Army conscripts' horsefalls by the Poisson distribution. Actually, a far greater and more motivated effort had been expended by Vilfredo Pareto (1848–1923) in fitting large personal incomes by a power-law distribution. This fit's excellence, however, was mostly disregarded, challenged and even ridiculed by the specialists, and its study was mostly left to amateurs. It may have been forgotten, were it not for the fact that Pareto's name became famous for an altogether different reason, namely, for an attempt to define and study economic equilibrium in direct imitation of physical equilibrium. Equilibrium soon led to beautiful theoretical work; so did the normal and lognormal distributions, while the Pareto law of personal incomes remained suspended in an intellectual vacuum.

Zipf may have felt the low esteem given to empirical results not accompanied by a theory. Indeed, the findings he collected and expanded upon, including Pareto's law, were "explained" by him as the outcome of a grandiose "principle of least effort," but this principle promptly evaporated upon examination. Lacking a theory, scientists did not expect to encounter scaling distributions; therefore, they either did not face them or failed to see them.

Most relevant for this book is the fact that geologists knew of scattered examples of scaling, but largely disregarded them. The situation is altogether different today, and this book shows that scaling is recognized throughout the earth sciences, and even that some examples can—after the fact—be traced far back to astonishingly early authors. But, once again, I am told that those examples were never faced in those early authors' time. In this context, there was no wonder that the 1962 IBM report of mine, which this book reproduces, was not publishable material. However, a paper that J. M. Berger and I published in 1963, (on errors in data transmitting telephones) combined with my 1962 report on oil fields made

me seek fresh examples of scaling in the hard sciences. By encouraging me to do so, it eventually led to fractal geometry—whose existence is, one main reason why new instances of scaling can now be faced unflinchingly. They might have been (but were not) faced later in the 1960s when statistical physicists developed the theory of critical phenomena and the renormalization group. But those physicists were in their Ivory Tower, while I was in the trenches with communications engineers, and the students of oil fields, hydrology, and turbulence.

Acting as a hybrid between a mathematician and an experimental scientist and engineer authorizes me to give practical advice: Don't indulge casually in mere data fitting. However, it is bound to happen again that careful study of large data sets will suggest that the best fit is provided by formulas that theoreticians will call strange. If so, argue hard with the theoreticians; don't expect them to defer to your authority without a whisper, but don't meekly defer to theirs.

Referees advised me thirty years ago to give up the scaling distributions and acknowledge the authority of the lognormal; they kept reminding me that "everyone knew" that scaling had no theoretical basis, while the lognormal had plenty. I read carefully the motivations for the lognormal and remained unconvinced. For example, rightly or wrongly, everyone expected the Gaussian in additive phenomena and the lognormal in multiplicative phenomena. But I found that not all multiplicative phenomena are lognormal and besides no one had advanced a full explanation of why an oil field's capacity should involve a multiplicative process. Therefore, the acceptance of the lognormal largely hinged on some authors' good names. But those good names were aging poorly. Who in the U.S. in 1994 knows of Robert Gibrat, a Frenchman whose fame as an economist largely rested on his Ph.D. thesis on the lognormal, and led him to become a high official and manager of technology, briefly a Minister of Transportation? Gibrat was a prime booster for the lognormal, but his work left me thoroughly unconvinced. Nevertheless, I was obliged to face his authority in the 1960s, when I authored, not only the report reproduced in this book, but also several IBM reports that were not otherwise published. Allow me to paraphrase two points from these reports. The first point is, once again, that neither the lognormal nor the scaling distribution has a fully-explored justification for many of the examples known to me. However, each distribution benefits, before any actual testing, from a kind of circumstantial evidence: the lognormal because of its links with the central limit theorem, and the scaling because it is indissolubly linked with fractals, and fractals are now recognized as being very widespread in nature.

Which distribution should be adopted by someone faced with a new batch of data? Curve-fitting is an unavoidable activity but it is difficult, not respected, and never perfect. A formula that best fits one part of an experimental curve will ordinarily fit the other parts poorly. Since the costs of different errors are assessed differently by different authors, curve-fitting is unavoidably subjective. For example, given the same oil fields, the lognormal will not, and scaling will, predict additional very large fields. In other words, contradictory claims must be judged by criteria other than fitting of one sample. History is of great help. Look up old oil field data and extrapolate both the lognormal and scaling distributions. Favor the distribution that has been treated better by history. There is every evidence that you will pick the scaling distribution.

Benoit B. Mandlebrot

New Haven