Reprinted from INFORMATION AND CONTROL, Volume 3, No. 1, March 1960
 Printed in U.S.A.

INFORMATION AND CONTROL **3**, 80–88 (1960)

Some Further Notes on a Class of Skew Distribution Functions

HERBERT A. SIMON

Carnegie Institute of Technology, Pittsburgh, Pennsylvania

This note takes issue with a recent criticism by Dr. B. Mandelbrot of a certain stochastic model to explain word-frequency data. Dr. Mandelbrot's principal empirical and mathematical objections to the model are shown to be unfounded. A central question is whether the basic parameter of the distributions is larger or smaller than unity. The empirical data show it is almost always very close to unity, sometimes slightly larger, sometimes smaller. Simple stochastic models can be constructed for either case, and give a special status, as a limiting case, to instances where the parameter is unity. More generally, the empirical data can be explained by two types of stochastic models as well as by models assuming efficient information coding. The three types of models are briefly characterized and compared.

1. INTRODUCTION

In a recent note in this journal, Dr. Benoit Mandelbrot has raised some objections to a stochastic explanation of certain well-known data on word frequencies. A number of fundamental points in Dr. Mandelbrot's note appear incorrect, others are debatable. Some of these relate to the empirical properties of the distributions, some to the mathematical analysis. Since the words frequency data have attracted a great deal of attention, it is perhaps worth while to try to clarify the points at issue.

Let $f(i,k)$ be the number of different words, each of which occurs exactly i times, in a sample of k words of text. In a wide range of cases, the observed data can be fitted quite well by a function of the form:

$$f(i,k) = G(k)i^{-(\rho+1)} \tag{1}$$

and even more satisfactorily, particularly for low values of i, by the function:

$$f(i,k) = AB(i,\rho + 1) \tag{2}$$

where A and ρ are constants, and $B(i,\rho + 1)$ is the beta function of i, $\rho + 1$. As i increases, (2) approaches (1) asymptotically.[1] For both (1) and (2), the expected value of i is finite if and only if $\rho > 1$.

Function (1) has a long history in statistics; in economics, it is usually associated with the name of Pareto, in linguistics, with the names of Estoup and Zipf. Zipf was particularly interested in the case where $\rho = 1$. Function (2) was first introduced by Yule (1924) to explain certain taxonomic data of Willis, and hence I have proposed calling it the Yule Distribution.

2. THE EMPIRICAL DISTRIBUTIONS

A great deal of Dr. Mandelbrot's critical discussion depends on his claim that for the empirical word-frequency distributions, $\rho < 1$. He states categorically (1959, p. 92):

"One finds, in general, that $\rho < 1$ for word frequencies . . . The few cases where $\rho > 1$ are also quite exceptional in other respects (e.g., Modern Hebrew about 1935)."

He makes an almost identical statement on page 498 of (1953). Unfortunately, he does not in either case present his evidence, and the source, Zipf, on which he chiefly relies, contradicts him. The data that Zipf report show ρ to be greater than 1 more often than not, and *almost always to be very close to 1*—a point to which I shall return. I find in Zipf the following least-squares estimates of ρ: for Joyce's *Ulysses*, between 0.99 and 1.01 (p. 34); Plautus, .98 (p. 34); the *Iliad*, 1.15 (p. 34); Nootka and Plains Cree holophrases, 1.36 and 1.14, respectively (p. 84); Nootka morphenes, 0.67 (p. 85); Nootka varimorphs, two values, 0.67 or 1.12, depending on the curve-fitting method (p. 85); Dakota words, 1.29 (p. 86); Gothic words, 1.025 (p. 94); old high German words, 0.98 (p. 116). In addition, a large number of values, all close to 1, are reported for children's speech.

In addition to the calculated values, Zipf presents a large number of graphs of distributions, on a double-log scale, in virtually all of which ρ is very close to 1—sometimes a little greater, sometimes a little less. The figure on page 25 of Zipf, for example, strikingly conforms to the hypothesis, with ρ indistinguishable from unity. In most of the distributions (see, e. g., Zipf, pp. 123, 125), there is a little curvature, usually a convexity upward. Under these circumstances, neither function (1)

[1] For additional detail see page 426 of Simon (1955), which I shall refer to by the short title, "Yule Distribution."

nor (2) fits exactly, and it is difficult to know how best to estimate ρ. An unweighted least-squares fit to the distribution on a logarithmic scale is perhaps not the most plausible method.

Several estimators are proposed in "Yule Distribution." If k is the size of sample, n_k the number of different words in the sample, and $f(1)$ the number of different words each of which occurs exactly once, then, by Eq. (2.19) and (2.12) of "Yule Distribution," we have $\alpha = n_k/k$ and $\rho = 1/(1 - \alpha)$. Using these relations, we find the following values for ρ: *Ulysses*, 1.13; Eldridge's word count, 1.16; Yule's count of nouns in Macaulay, 1.33; Plautus, 1.34. (In fairness, it should be pointed out that when this method of estimating is used, ρ is necessarily greater than 1.) Alternatively, we can estimate α by Eq. (2.21): $(2 - \alpha) = n_k/f(1)$. We then find the following values: *Ulysses*, 1.24; Eldridge, .983; Macaulay, .935; Plautus, 1.81. (See the discussion of this estimator on page 431 of "Yule Distribution.")

Finally, it should be observed that if $\rho < 1$, neither (1) nor (2) can hold through the entire range, for in this case the mean of the distribution would be infinite. No model (and this applies to Dr. Mandelbrot's as well as to mine) that requires $\rho < 1$ can hold for indefinitely large values of i.[2] Empirically, this shows up in the curvature of the observed distributions for large i.

We must conclude that Dr. Mandelbrot has not established his case that, in general, $\rho < 1$. On the contrary, the data suggest that generally $\rho \sim 1$. But what is the significance of this? Several derivations of (1) in Mandelbrot (1953 and 1954) require that $\rho < 1$ (page 495), and therefore fail to handle any of the empirical distributions for which the parameter exceeds unity. On the other hand, the *first* derivation (pp. 427–429) of (2) in "Yule Distribution" requires that $\rho > 1$, and therefore fails when the parameter falls short of unity. However, a number of variant models are discussed in "Yule Distribution" which lead, approximately, to (2), and which admit $\rho < 1$. I shall discuss below whether these variants involve "analytic circularity" (Dr. Mandelbrot's term for "lack of parsimony").

In trying to decide whether the parameter is greater than or less than unity, we must not lose sight of the striking fact, already mentioned,

[2] That Dr. Mandelbrot is aware of this is revealed by his comment (1959, p. 91): "These will always be 'weak' laws, in the sense that they break up either for small i or for large i depending upon the specific example." Again, see page 431 of "Yule Distribution."

that it is almost always very close to unity. It is hard to specify *how* close for there are no satisfactory tests of closeness of fit in these matters, and hence it is not surprising that different statisticians, equally "skilled in the art," may experience different degrees of satisfaction with the results. I. J. Good (1953, pp. 258–259), for example, after fitting (2), in the special case where $\rho = 1$, to the Eldridge word-frequency count, concludes that the fit "is remarkably good" for $i \leq 15$, and can be improved by introducing a convergence factor. He fits the same function to Yule's sample of nouns in Macaulay's essay on Bacon, and says (p. 261): "It is curious that this should again give such a good fit for values of i that are not too large ($i \leq 30$). The sample is of nouns only and, moreover, Yule took different inflexions of the same word as the same."

Yule, himself, was much more critical, rejecting the fit of (1) to both Zipf's data and his own (p. 55):

"I spent some time on a re-examination of his data and cannot agree with the claim that the formula holds to any satisfactory degree of precision even for his distributions: it certainly does not hold for any of my own that I have tested."[3]

If we accept Mr. Good's more optimistic conclusion that some of the fits are "remarkably good" for the limiting case, where we take $\rho = 1$, then we would like our theory of the phenomena to explain the special significance of this limiting case. The derivation of (2) in "Yule Distribution" does this, for it shows that as long as the ratio of number of *different* words in the text to total word occurrences is small (say, not more than 0.2), the parameter will be close to 1 (say, not over 1.25).

Before leaving the subject of the empirical distributions, I should like to state my agreement with Dr. Mandelbrot that for the taxonomic examples of Willis, $\rho < 1$, for income distributions, $\rho > 1$. But the data on pages 377–382 of Zipf clearly contradict his assertion that "for non-biological taxonomies such as names of professions, business catalogues, etc., . . . ρ is always less than one, and usually it is close to $\frac{1}{2}$."

[3] I would conjecture that Yule used the chi-square test to reject the hypothesis. We are confronted here with the usual difficulties of testing an extreme hypothesis. Incidentally, Dr. Mandelbrot (1959, p. 93) seriously misinterprets Yule when he uses the passage just quoted to conclude that (1) holds only for inflected words and not for lexical units or nouns alone. Yule's stricture applies to all cases, and almost equally good, or bad, fits are obtained under a wide range of alternative definitions of the unit.

3. THE STOCHASTIC MODELS

In Section II of "Yule Distribution," I formulated a stochastic model that yields (2) as its steady state distribution.[4] As I pointed out there, the definition of "steady state" poses some difficulties. Hence, I reinterpreted the same model in Section III, by means of an alternative urn scheme, in a way that allowed a rigorous definition of "steady state."[5]

Dr. Mandelbrot's principal objections, however, are levelled against the derivations in the case where $\rho < 1$. I have already given the reasons from empirical observation for thinking this is not generally the significant case for word frequencies. Nevertheless, this case certainly does arise in some instances, (e.g., the Thorndike count), and in applications of these kinds of stochastic models to other data (e.g., the taxonomic data of Willis). Hence, I should like to discuss this case a little more fully.

On pages 430–431 of "Yule Distribution" I show heuristically how the case $\rho < 1$ for small i might arise. Dr. Mandelbrot (1959, p. 96), after introducing several approximations, which he does not justify in detail, shows that my approximation can be "exact" only in a very special case. I will go further, and say (as I did already on page 431 of "Yule Distribution") that it cannot be exact even in that special case because of nonconvergence as i increases.

On page 439 of "Yule Distribution" I gave a short sketch of an alternative derivation of (2) for $\rho < 1$, corresponding to Yule's (1924)

[4] Since Dr. Mandelbrot mentions several times that this model is a special case of Champernowne's, I should like to put the record straight. Champernowne never derives (2), but only the approximation, (1). Yule derives (2) for the case $\rho < 1$, but not for $\rho > 1$. Neither Champernowne's derivation nor Yule's discloses the special significance of the limiting case, $\rho = 1$, or the reasons why the word distributions should lie close to this limiting case. Moreover, the assumptions required for my derivation of (2) are much weaker than Yule's. Finally, since Rapoport (1957, p. 157) has suggested that my derivation was a "counter-analysis" to Mandelbrot's, I might mention that at the time I derived (2) I was not familiar with the papers of Mandelbrot, Champernowne, or Yule. I came across these in the course of the search for prior work that one normally makes before publishing.

[5] The alternative derivation of Section III disposes, I think, of Dr. Mandelbrot's assertion (1959, p. 95) that "actually, $f^* \sim k$ cannot be considered as being a steady-state requirement." Since he says he plans to raise this point on another occasion, perhaps we can postpone further discussion of it to that time.

$$f(i,m) = A\lambda^i B(i + c, d - c + 1), \tag{4}$$

where

$$\lambda = \frac{(1 - \alpha)(k + dn_k)}{(k + cn_k)}$$

and B is the beta function. If we compare (4) with (2), we see that the latter has a convergence factor, λ, that is missing from the former, and that ρ has been replaced by $d - c = \rho^*$. In particular, if d is not much larger than c, we will have $\rho^* < 1$.

The process (3) has a number of interesting special and limiting cases. For example, if $c = d$, the steady state distribution is a generalization of Fisher's log series distribution: $f = A(1 - \alpha)^i/(i + c)$. On the other hand, as d approaches zero and c increases without limit, we obtain the limiting process:

$$f(i,m) = \frac{(1 - \alpha)}{m_k}[f(i - 1) - f(i)] - \frac{1}{k}[if(i) - (i + 1)f(i + 1)] = 0 \tag{5}$$

the steady state distribution for which is simply the Poisson distribution: $f(i) = A\lambda^i/i!$ The reader can verify these results, by substituting the solutions in Eqs. (3) and (5), respectively.

4. THE MEANING OF THE WORD FREQUENCY DISTRIBUTIONS

It appears from this analysis that the stochastic interpretation of the word frequency data proposed in "Yule Distribution" is decidedly more adequate than Dr. Mandelbrot allows. What is the relation of this interpretation to the alternative interpretations that Dr. Mandelbrot had proposed (1953, 1954)? Dr. Mandelbrot's models are of two types:

(1) Derivations of the distribution from various assumptions of efficient letter-by-letter coding of the language;

(2) Derivations of the distribution from various Markovian assumptions about the stochastic formation of words from strings of letters.

From a formal mathematical standpoint, Dr. Mandelbrot's efficient coding models and his stochastic models are substantially equivalent. The two types of derivations correspond, respectively, to derivations in classical statistical mechanics based on entropy maximization, on the one hand, and statistical equilibrium, on the other. Dr. Mandelbrot's

stochastic models are quite different from those of "Yule Distribution," since the latter rest on no assumptions whatsoever about the statistical properties of the alphabet in which the words are encoded.

It seems to me something more than a matter of taste and convenience whether certain empirical regularities can be explained as the products of stochastic processes arising from imitation and association, as proposed in "Yule Distribution"; whether we explain them by postulating a mechanism that maximizes the amount of information transmitted per symbol; or whether we explain them on the basis of statistical properties of the encoding process. My feeling that the teleological explanations are particularly to be avoided unless other evidence requires them is perhaps a prejudice, but it is a prejudice shared by others. Miller, Newman, and Friedman (1958) say, for example:

"This derivation [the one numbered (2) above] has the advantage that it does not assume optimization in terms of cost; it begins with the more palatable assumption that the human source is a stochastic process."

As between the two stochastic explanations, I confess also a preference for that developed in "Yule Distribution." First, unlike the stochastic derivation from coding considerations, it involves mechanisms of imitation and association that are consistent with what we know about social and psychological processes. Second, while all the data on the word frequency distribution show it to be extremely regular, the data on the variation of word frequency with word length show only a very rough relation. This suggests that very frequent words become abbreviated in use, and hence generally become short words. Use causes shortness, not shortness use. Common sense suggests the same thing. However, it would be nice to be able to choose between the two major types of stochastic models on the basis of clearcut evidence rather than these very crude considerations. The evidence remains to be discovered.

RECEIVED: July 1, 1959. Revised September 15, 1959.

REFERENCES

CHAMPERNOWNE, D. G., (1953). A model of income distribution. *Econ. J.* **63**, 318.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237.

MANDELBROT, B. (1953). An informational theory of the statistical structure of

language. *In* "Communication Theory" (Willis Jackson, ed.), pp. 486–502. Butterworths, London.

MANDELBROT, B. (1954). On recurrent noise-limiting coding, *Proc. Symposium on Information Networks*, pp. 205–222. Polytechnic Institute of Brooklyn, New York.

MANDELBROT, B. (1959). A note on a class of skew distribution functions. *Information and Control*, **2**, 90–99.

MILLER, G. A., NEWMAN, E. B. and FRIEDMAN, E. A., (1958). Length-frequency statistics for written english. *Information and Control*, **1**, 370–389.

RAPOPORT, A. (1957). Comment: The stochastic and the 'teleological' rationales of certain distributions and the so-called principle of least effort. *Behavioral Science*, **2**, 147–161.

SIMON, H. A. (1955). On a class of skew distribution functions. *Biometrika*, **42**, 425–440.

YULE, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. B*, **213**, 21.

ZIPF, G. K. (1949). "Human Behavior and the Principle of Least Effort." Addison Wesley, Reading, Massachusetts.