Reprinted from INFORMATION AND CONTROL, Volume 2, No. 1, March 1959
Copyright © by Academic Press Inc. *Printed in U.S.A.*

INFORMATION AND CONTROL **2,** 90–99 (1959)

A Note On a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H. A. Simon

BENOIT MANDELBROT

IBM Research Center, Yorktown Heights, New York

This note is a discussion of H. A. Simon's model (1955) concerning the class of frequency distributions generally associated with the name of G. K. Zipf. The main purpose is to show that Simon's model is analytically circular in the case of the linguistic laws of Estoup-Zipf and Willis-Yule. Insofar as the economic law of Pareto is concerned, Simon has himself noted that his model is a particular case of that of Champernowne; this is correct, with some reservation. A simplified version of Simon's model is included.

I. A GENERALIZATION OF THE LAW OF PARETO. LAWS OF TYPE (Z)

There is a wide class of phenomena, mostly found in social sciences, which follow laws very similar to that of Pareto. That is, by using a certain (rather unnatural) notation, one can express these laws in a single analytic form. To exhibit this expression, consider a (possibly infinite) discrete population of *items*, each of which carries a *label* chosen from a (also possibly infinte) discrete set. Let $f(i,k)$ be the number of different labels, each of which occurs exactly i times, in a sample of k items.[1]

One then finds in a number of situations that, for large values of k,

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

with some $\rho > 0$. One will say that such phenomena follow a law of type

[1] The "sampling" method need not be the same in all the examples, and more will need to be said about it, in each case. Anyway, in some cases one has no actual control over the sample size, and one cannot compare samples of different sizes. One cannot be sure, then, of the sampling method which has generated the closed collection which one observes. The present method of exhibiting the data for a fixed k has the advantage that the sampling method need not be known; but this also makes the similarity between the examples below quite questionable, *a priori*. This impression becomes stronger as the study of any such law is developed in greater detail. We hope to take up this topic on another occasion.

(Z) [in honor of G. K. Zipf (1949), who has been most active in studying various occurrences of such laws]. 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. The function $G(k)$, which must have positive values, need not be the same function of k in the different cases to be considered. In any case G also depends upon ρ and upon the form of f in the region where the law (Z) does not apply any more.

Alternative Form of (Z)

Frequently, one is presented with rough data in which the labels occurring between i' and i'' times each have already been grouped and counted together (the interval i''–i' may refer, for example, to an "income bracket"). Further, these i'–i'' may vary along the scale of i. In these cases, it is more convenient to plot the numbers of items, r, each of which occurs more than i times as a function of i. If i is large, one may approximate the sum $\sum i^{-(\rho+1)}$ by an integral, and one obtains:

$$r = \sum_{j=i}^{\infty} f(j, k) \sim G(k)\rho^{-1}i^{-\rho}$$

One may further order all labels by decreasing numbers of occurrences in a sample. In this ordering, the preceding $r(i)$ becomes the "rank" of a label occurring i times. It is then often convenient to read the above function as giving the number of occurrences, i, as a function of the rank, r. Obviously,

$$i = G(k)^{1/\rho}\rho^{-1/\rho}r^{-1/\rho} = G'(k)r^{-B} \text{ (with } B = 1/\rho)$$

The most obvious property of any of these alternative forms of (Z) is that, if plotted on log-log paper, they all are represented by straight graphs. This is exactly how all the instances of laws (Z) were first discovered and are best studied now. For example, one should find that:

$$\log f(i,k) = \log G(k) - (\rho + 1) \log i$$

$$\log if(i,k) = \log G(k) - \rho \log i$$

$$\log i = \log G'(k) - B \log r$$

II. THE THREE BEST EXAMPLES OF LAWS OF TYPE (Z)

Income Distribution and the Law of Pareto

One can make the items of the population be quantized units of money, and make the labels be the names of the persons by whom each unit of money was earned (or, alternatively, to whom each unit of money belongs at a given time; but this occurrence of the law (Z) is less well established). Then $f(i,k)$ will be the number of persons earning exactly i units of money, out of a total income equal to k. The law (Z) holds best for large incomes. This is the classical prototype of all laws of type (Z) and it due to V. Pareto (1897). One always has $\rho > 1$.

Sizes of Taxonomic Genera, and the Law of Willis

Let now the items be taxonomic species, and the label on a species be the name of the genus to which it belongs. Then, in a total taxonomic "family" of k species, $f(i,k)$ will be the number of genera with i species each. The law (Z), as interpreted in this fashion, was discovered by J. C. Willis (1922) in the context of biological taxonomies. His work was made known among statisticians by papers of G. U. Yule (1923) (referred to in Feller's book, 1957, p. 404). The same law was later found by Zipf (1949) to hold also for nonbiological taxonomies such as names of professions, business catalogs, etc. In all these cases ρ is always less than one, and usually it is close to $\frac{1}{2}$. The sampling process leading to (Z) is anything but clear, and one needs a delicate argument to see which quantity is a random variable here. [See B. Mandelbrot (1956).]

Word Frequencies and the Law of Estoup-Zipf

Let the items of the population be the words of a homogeneous running text of a single author, that is, typographical sequences of letters contained between successive "space" symbols. Two words will carry the same label, if they are identical sequences of letters. Then, $f(i,k)$ will be the number of different word-forms, each of which occurs exactly i times, in a total sample of k words (different or not). The law (Z), as applied in this context, was apparently first noted by J. B. Estoup (1916), but now—and even more than the other laws (Z)—it is mainly connected with the name of Zipf. The region where the law (Z) is best satisfied is that of rare words, that is, of large values of r, in (i,r) coordinates, and of small values of i. One finds, in general, that $\rho < 1$ for word frequencies; that is, $B = 1/\rho > 1$. The few cases where $\rho > 1$ are also quite exceptional in other respects (e.g., Modern Hebrew about 1935). The parameter ρ is a characteristic not only of the language used but chiefly of the author of the sample under study. One finds that ρ increases with the speaker's "intelligence" or "wealth of vocabulary."

In some of his earlier writings, Zipf claims that he has established empirically the stronger result that $\rho \equiv 1$ in all cases. This has not been vindicated. One may also point out that the law (Z) does not at all apply to words defined as the lexical units, or to nouns, verbs, etc., taken separately. This was strongly established by G. U. Yule (1944, p. 55).

III. THE CHALLENGE TO EXPLAIN THE LAWS (Z)

There are many more concrete examples of the type of analytic behavior for $f(i,k)$, which characterizes the family (Z), although the evidence is nowhere as strong as in the above three cases. These examples differ in all possible respects, even from the analytical viewpoint, since the function $G(k)$ may take different forms in different cases. However, the form of (Z) is so striking, and also is so very different from any classical distribution of statistics, that it is quite widely felt that it "should" have some basically simple reason, possibly as "weak" and as general as the reasons which account for the role of the Gaussian distribution. But, in fact, the laws (Z) turn out to be quite resistant to such an analysis. Thus, irrespective of any claim as to their practical importance, the "explanation" of their role has long been one of the best defined and most conspicuous challenges to those interested in statistical laws of nature. We have devoted several papers to what we hope to be positive contributions to this subject. The present paper is, on the contrary, of critical character, and will discuss H. A. Simon's attempt to find a single unified model for *all* the distributions (Z) by constructing a variant of the so-called birth (or birth-and-death) process.

IV. SIMON'S MODEL

The postulates of this model are made clearer by the following preliminary step. Start from a sample of k^* items, with the distribution $f^*(i,k^*)$. Assume then that the sample may be modified by letting k increase beyond k^*. (This is a fairly reasonable assumption in the case of word frequencies, since a text is indeed generated word by word. But a national income is surely *not* distributed dollar by dollar.) Look now for a chance process whereby $f^*(i,k^*)$ could be extrapolated to $k > k^*$, in a "stationary" fashion. Let $f^*(i,k)$ be the expected value of $f(i,k)$, given the initial condition $f^*(i,k^*)$.

The most obvious procedure is to "estimate" that the probability that the next label will be one of those which has already occurred i times, is exactly i/k. This determines, at every step, the population from which the next label is drawn. If so, the increase $f(i,k+1) - f(i,k)$ is made out of the difference between (a) those cases where one draws a label which had previously occurred $i - 1$ times (and which now comes into the class of labels occurring i times) and (b) those cases where one draws a label which had previously occurred i times (and which now comes out of the class of labels occurring $i + 1$ times). Thus, $f(i,k)$ and also $f^*(i,k)$ satisfy the difference equation:

$$f^*(i,k+1) - f^*(i,k) = (1/k)[(i-1)f^*(i-1,k) - if^*(i,k)]$$

Let us approximate this by a differential equation.[2] One then obtains after a few steps

$$\frac{\partial \log f^*}{\partial \log k}\left(\equiv \frac{\partial \log (if^*)}{\partial \log k}\right) = -\frac{\partial \log (if^*)}{\partial \log i}$$

Hence, obviously , there must exist some function $F'(x)$, such that

$$if^*(i,k) = F'(i/k)$$

$$f^*(i,k) = (1/k)\, F(i/k) \quad \text{with } F(x) = (1/x)\, F'(x)$$

The function F is determined by the initial conditions $f^*(i,k^*)$, and it is defined only for $(i/k) > 1/k^*$. With different initial conditions, one may get any function $f^*(i,k)$ whatsoever, a quite obvious result.

However, the "estimation-theoretical" extrapolation does not allow for the possibility of sometimes drawing some entirely new label. Simon deals with this difficulty by postulating that there is a well-determined probability, $\alpha(k)$, for such an event. His model therefore strongly depends upon the possibility of generating his sample item by item, and (*a priori*) it is more appropriate for word frequencies than for income distributions.

If $\alpha(k)$ is known, the probability that the next item will carry a label which has already occurred i times could then reasonably be taken to be:

$$(1-\alpha)(i/k) \quad (\text{instead of } i/k)$$

[2] Simon does not do this, but at the end of his argument he approximates an eulerian beta function by a power, which amounts to the same thing. Note also that Simon's function $f^*(i)$ is defined (between formulas 2.9 and 2.10) to be independent of k. Then, (in formula 2.21) it is derived to have the form $f^*(i) = ka B(i, \rho + 1)/(2 - a)$. Further (between formulas 2.17 and 2.18), it is stated that $f^*(i)$ should be a "proper distribution function," which would require that $\sum_1^k iB(i, \rho + 1)$ converge as $k \to \infty$. This restriction actually excludes any probability distribution function with an infinite mean value.

This is easy to visualize (at least in the case of word frequencies) and it is sufficient to derive $f(i,k)$. But it is in contradiction with other experimental facts concerning possible stochastic processes generating the data following the law (Z). To avoid this difficulty, Simon puts together all the labels which had already occurred i times and he assumes their *joint* probability to be $(1 - \alpha)(i/k)f(i,k)$. Therefore the stochastic model with which he works from then on remains compatible with a great many actual processes. The fundamental differential equation finally becomes

$$\frac{\partial \log (if^*)}{\partial \log k} = -(1 - \alpha) \frac{\partial \log (if^*)}{\partial \log i}$$

Case Where $\rho > 1$

First, Simon assumes $\alpha(k)$ to be a constant α_0 (independent of k). Then the solution of the fundamental equation is:

$$if^*(i,k) = F(ki^{-\rho}) \quad \text{with } \rho = (1 - \alpha_0)^{-1}$$

Further, $\alpha(k) = dn(k)/dk$, where $n(k) = \sum_{i=1}^{k} f^*(i,k)$ is the total number of labels in a sample. Therefore, $n(k) = \alpha_0 k$. This requirement may in particular be satisfied by picking the so-called "steady-state" solution, in which *each* f^* is already proportional to k. This gives

$$F(ki^{-\rho}) = (\text{constant}) \cdot ki^{-\rho}$$

or

$$if^*(i,k) = (\text{constant}) \cdot ki^{-\rho}$$

This is, indeed, the form (Z) *with the restrictions that* $\rho > 1$ *and* $G(k) = k$. Actually, $f^* \sim k$ *cannot* be considered as being a steady-state requirement, and if this condition is dropped, f^* becomes (roughly speaking) undeterminate. We hope to take up this topic on another occasion, and wish only to show here that even the presant approach is undoubtedly inadequate for $\rho < 1$.

Case Where $\rho < 1$

To derive (Z) within Simon's framework, one is now obligated to assume that $\alpha(k)$ varies with k. In fact, if (Z) is to hold one must somehow find that

$$\frac{\partial \log (if^*)}{\partial \log i} = -\rho$$

Hence the requirement:

$$\frac{\partial \log (if^*)}{\partial \log k} = [1 - \alpha(k)]\rho$$

Since $\alpha < 1$ and $\rho < 1$, then $(1 - \alpha)\,\rho < 1$. This means that the numbers if^* increase less than linearly with k, and so does the total number of labels $n(k) = \sum f^*(i,k)$. Finally, $\alpha(k) = dn(k)/dk$ must tend to zero for $k \to \infty$, so that, for $k \gg 1$, it is sufficient to approximate if^* by the following expression:

$$if^*(i,k) = (\text{constant in } k)\cdot k^\rho$$

But it has been postulated that

$$if^*(i,k) = (\text{constant in } i)\cdot i^{-\rho}$$

Therefore, one must have (that is, one wishes to obtain as a result):

$$if^*(i,k) = (\text{constant})\cdot(i/k)^{-\rho}$$

This requires that

$$n(k) = \sum_1^k f^*(i,k) \sim \sum_1^\infty f^*(i,k) \sim k^\rho \sum i^{\cdot(\rho+1)}$$

$$n(k) \sim (\text{constant})\cdot k^\rho$$

and

$$\alpha(k) = dn(k)/dk = (\text{constant})\cdot k^{\rho-1}$$

To sum up: *if one wishes to obtain the law* (Z) *with* $\rho < 1$, *one must postulate explicitly that* $\alpha(k) = (\mathit{constant})\cdot k^{\rho-1}$. *No other* $\alpha(k)$ *would lead to* (Z). This result was unfortunately not explicitly written down in Simon's paper.

V. CONCLUSION

Simon's model is not adequate as an explanation of the whole of the family (Z). *It may conceivably be made acceptable if* $\rho > 1$, (*if the steady-state requirement may be motivated, or is added as a hypothesis*). *But the model is certainly to be abandoned if* $\rho < 1$.

The point is as follows: Simon showed that the law (Z) with $\rho > 1$, may be derived from $\alpha(k) = \alpha_0$. This criterion is a most difficult one to comprehend in the Pareto case, since this case is also precisely one for which the generation of the sample item by item is least justified. But this form of $\alpha(k)$ is still far simpler than that of (Z), to be explained. Thus, the reduction of (Z) to $\alpha(k)$ fits into one of the universal aims of scientific explanation, which is to reduce the complicated to the simple—even if this simplicity is at a level difficult to comprehend. (For example, the physicist finds it quite acceptable to base statistical thermodynamics upon the equiprobability of all microcanonical configurations, a conjecture which nobody can possibly ever check). *Simon's model, then, should be followed up and improved in the cases where $\rho > 1$, such as Pareto's case.* Such a study will be practically identical to that of the more general model of Champernowne (1953).

Suppose now that $\rho < 1$ (this includes the word frequency case, for which the generation word by word is sensible). In this case, Simon at best reduces $f(i,k)$ (an easily observed fact) to $\alpha(k)$, which is quite conjectural and difficult to check experimentally, and besides is analytically identical to $f(i,k)$ and therefore altogether exactly as untractable. Anyway, he makes no attempt to explain this $\alpha(k)$, and does not even write it down explicitly. From reading his paper casually, the impression could be derived that if $\rho < 1$, *any* smooth and slowly decreasing function $\alpha(k)$ could explain the law (Z) just as well. Actually this is not so, and *the model is circular, from the viewpoint of the analytic form of the premises and of the conclusion.* This invalidates Simon's model, insofar as it concerns the Willis-Yule law and the usual ($\rho < 1$) case of the Estoup-Zipf law.

The objection does not apply to those Estoup-Zipf data for which $\rho > 1$. But, even there it is found experimentally that in a first approximation the rank order of words varies little with k and that i/k must tend to some limit for each r. On the contrary, Simon finds that

$$i/k = (\text{constant}) \cdot k^{1/\rho - 1} r^{-1/\rho}$$

which tends to zero as $k \to \infty$. Therefore the "steady-state" condition, from which this result follows, is in contradiction with a well-established experimental fact.

It is true, of course, that this first approximation is rather crude. In every text there are some words occurring with an exceptionally high frequency which cannot be explained by chance fluctuations only. In fact, these words carry most of the information concerning the "topic" of a text. But they are so few in number (see the work of H. P. Luhn) and their behavior is so little known experimentally, that it seems unlikely that any model could, at the present time, have the desirable feature of accounting for their behavior.

VI. CONCERNING MANDELBROT'S THEORY OF THE ESTOUP-ZIPF LAW

Against our theory of this linguistic law (1953, 1957a, 1957b) Simon presents two main objections, both of which appear to be ill-founded.

(a) He objects to the use of the maximizing procedure to show that the state of a text in which the Estoup-Zipf law is true is the "most probable" state, or the state of greatest information. He states that thermodynamics (which is the original model of our theory) "prefers averaging procedures." This is undoubtedly so, but it is only a matter of taste and of convenience and, for large systems, both methods are known to lead to the same result. Actually, in our paper (1957b), we have used the average-state argument, instead of a maximization.

However, one advantage of maximization is that the logarithm of the probability of a state may be interpreted as an information, and the most probable state is then also interpreted as being the state of largest information. This is a most interesting property, even if the maximization of information is not taken more literally than the maximization of entropy in the stationary state of thermodynamics.

(b) But, precisely, Simon objects *a priori* to our use of the concept of information, stating that "numerous doubts (which he shares) have been expressed as to the relevance of Shannon's measure of information for the measurement of semantic information." We may say that, in our eyes, there should be *no doubt* on this account: *"information" is utterly irrelevant to "semantics,"* and its use in linguistics only shows that *some* matters in that field may be explained without any semantics whatsoever.

RECEIVED: November 17, 1958.

REFERENCES

CHAMPERNOWNE, G. (1953). *Econ. J.* **63** 318.

ESTOUP, J. B. (1916). "Gammes Sténographiques," 4th ed. Paris.

FELLER, W. (1957). "Probability Theory," 2nd ed. Wiley, New York.

MANDELBROT, B. (1953). *In* "Communication Theory" (W. Jackson, ed.). Academic Press, New York.

MANDELBROT, B. (1956). *In* "Information Theory" (C. Cherry, ed.). Academic Press, New York.

MANDELBROT, B. (1957a). *In* "Logique, langage et théorie de l'information." Presses Universitaires de France, Paris.

MANDELBROT, B. (1957b). "Théorie mathématique de la loi d'Estoup-Zipf." Institut de Statistique, Paris.

PARETO, V. (1897). "Cours d'économie politique." Lausanne et Paris.

SIMON, H. A. (1955). *Biometrika* **42,** 425. This paper was also published without change in "Models of Man," p. 145. Wiley, New York, 1957.

WILLIS, J. C. (1922). "Age and Area." Cambridge Univ. Press, London and New York.

YULE, G. U. (1923). *Phil. Trans. Roy. Soc. London.* **B213,** 21.

YULE, G. U. (1944). "Statistical Study of Literary Vocabulary." Cambridge Univ. Press, London and New York.

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