[192]

Reprinted from INFORMATION AND CONTROL, Volume 4, Nos. 2-3, September 1961
Copyright © by Academic Press Inc. *Printed in U.S.A.*

INFORMATION AND CONTROL 4, 217-223 (1961)

Reply to "Final Note" by Benoit Mandelbrot

HERBERT A. SIMON

Carnegie Institute of Technology,
Pittsburgh, Pennsylvania

Dr. Mandelbrot's original objections (1959) to using the Yule process to explain the phenomena of word frequencies were refuted in Simon (1960), and are now mostly abandoned. The present "Reply" refutes the almost entirely new arguments introduced by Dr. Mandelbrot in his "Final Note," and demonstrates again the adequacy of the models in (1955).

My reply can be quite brief, since it involves mostly showing that Dr. Mandelbrot's central assumptions are invalid. I have numbered my sections to correspond to his.

SECTION I

Since this section merely states claims, without proof, I need remark on only two points that will be important later:

1. In my 1955 paper, the Yule distribution was derived under an Assumption I, which is vastly weaker than the Assumption I′ that Dr. Mandelbrot uses throughout his note. His statement that "this generalization will not change any conclusion concerning the form of $f(i, k)$" is literally correct but thoroughly misleading, for the principal conclusions in Sections II and V of the "Final Note" are not conclusions about the form of $f(i, k)$, and are false when Assumption I′ is replaced by Assumption I.

2. Dr. Mandelbrot assumes that "$n'(k)$ decreases with $1/k$ sufficiently fast for $\sum_1^\infty n'(h)/h$ to be convergent." From examination of empirical data, I am persuaded that convergence is the exception, divergence the rule. In most of the cases I have looked at, $n'(k) = a\,(\log bk)^{-1}$ gives a better approximation to the data than $n'(k) = ak^{-c}$, particularly as k grows large. With the former function, the series does not, of course, converge.

SECTION II

The derivation of the Yule distribution is fundamental to what follows. I should like to present it in a form that makes more transparent why Dr. Mandelbrot's conclusions are wrong.

1. Taking k as a continuous variable, we consider the distribution, when the sample size reaches k, of the new word types that entered the sample when it was about size k_0. The equations for the process are:

$$f_k'(i,\ k) = [1 - n'(k)]k^{-1}[(i - 1)f(i - 1, k) - if(i, k)]. \qquad (1)$$

To solve these equations exactly, we replace k by a new variable, $\tau(k)$ [which, we shall see, is approximately equal to $\log (k/k_0)$] defined by:

$$d\tau = (1 - n'(k))k^{-1}\, dk, \qquad \tau(k_0) = \tau_0 = 0. \qquad (2)$$

With this replacement, Eq. (1) is transformed into

$$f_\tau'(i,\ \tau) = [(i - 1)f(i - 1,\ \tau) - if(i,\ \tau)]. \qquad (3)$$

The solution of Eq. (3), for the boundary conditions $f(i_0, 0) = 1$, $f(i, 0) = 0$ for $i \neq i_0$, is well known (Feller, 1957, p. 403). It is a special case of the negative binomial:

$$f(i,\ \tau) = \frac{(i - 1)!}{i_0\,!(i - i_0)!}\, e^{-i\tau}(1 - e^{-\tau})^{i-i_0}. \qquad (4)$$

The mean and variance are ix^{-1} and $i(1 - x)x^{-2}$, respectively, where $x = e^{-\tau}$.

In the particular case where $i_0 = 1$ (the word type has one occurrence when $k = k_0$), this reduces to the geometrical distribution:

$$f(i,\ \tau) = e^{-\tau}(1 - e^{-\tau})^{i-1}. \qquad (5)$$

The mean and variance of this distribution are x^{-1} and $(1 - x)x^{-2}$, respectively, where $x = e^{-\tau}$. We may call (5) the diffusion equation for the process, since it describes (under Assumption I′) the future history of word types that first enter the sample at $\tau = 0$.

Suppose the rate of entry of new words into the sample is given by

$$dn(k) = n'(k)\, dk, \qquad 1 \leqq k. \qquad (6)$$

Then, to find the aggregate distribution when the sample size is k, we multiply (5) by (6) and integrate over the range $1 \leqq r \leqq k$, obtaining:

$$f(i,\ k) = \int_{r=1}^{k} e^{-\tau(k,r)}[(1 - e^{-\tau(k,r)})]^{(i-1)} n'(r)\, dr. \qquad (7)$$

Up to this point I have introduced no approximations in solving (1). The central issue in dispute is under what conditions (7) is satisfactorily approximated, *for a given* k, by the Beta function $B(i, \rho(k) + 1)$, where $\rho(k)$ may depend on the shape of the function $n'(r)$ in the interval $1 \leqq r \leqq k$. Because of the presence of k in the limits of integration of (7), we are, strictly speaking, concerned with the incomplete Beta function, $B(i, \rho(k) + 1; k)$, but if k is large, this may, under most circumstances, be approximated very well for $i \ll k$ by the complete Beta function.

2. In the important special case where $[1 - n'(k)] = \text{const.} = \rho^{-1}$, we can derive immediately from (2) that

$$x = e^{-\tau} = (k/k_0)^{-\rho}. \tag{8}$$

Transforming variables again in (7) to express the integral in terms of x, we get

$$\begin{aligned} f(i, x) &= \int_1^x y(1 - y)^{i-1}(1 - \rho^{-1})k\rho^{-1}y^{-1}\,dy \\ &= A \int_x^1 (1 - y)^{i-1}y^{\rho}\,dy, \qquad A = \text{const.} \end{aligned} \tag{9}$$

which is the usual form of the integral for the incomplete Beta function. My approximation in 1955, p. 431, obtained from a steady-state assumption, boils down to the observation that if $n'(k)$ is slowly decreasing instead of constant, we can still approximate the distribution for given k by (9), merely replacing the exponent in the integrand by

$$\sigma(k) = \frac{n'(k)k}{n(k)}\,\rho(k),$$

where $\rho(k) = [1 - n'(k)]^{-1}$.

3. In the next part of his Section II, Dr. Mandelbrot derives $f(i, k; i_0, k_0)$ approximately. We already have the exact result in Eq. (4) above; it is the negative binomial. I shall have something to say in Section V, 5 about the use Dr. Mandelbrot makes of it and its "gaussian approximation," which holds only in the limit, for large i_0.

4. In the remainder of this section Dr. Mandelbrot derives results, which he himself calls absurd, by using Assumption I′ where the weaker Assumption I should be used. But if the weaker assumption is used, the results do not follow, for then words, and cities, do *not* "behave as if

each had a well-defined probability." In particular, "diachronal" variation is quite consistent with Assumption I. (On city sizes, see also Simon, 1955, p. 437.) These paragraphs simply prove that I was right in insisting on Assumption I as the basis for the derivations in Simon (1955).

SECTION III

According to Dr. Mandelbrot, my model implies that if we pick a word that has a frequency of occurrence of i_0 when the sample size is k_0, then as k increases, the relative frequency i/k of that word will become vanishingly small. There are two things wrong with the use he makes of this argument.

1. The prediction that $E(r(p))$ tends to zero is far from absurd. The gradual change of topic as a sequence of prose unfolds (which also accounts for the continued introduction of new words) brings about a substantial regression, on the average, in the relative frequencies of the words that have already occurred, hence also in the number of words above any given relative frequency. There is nothing contrary to fact about this observable phenomenon.

This regression (Dr. Mandelbrot's "diachronal variation") occurs "in the small"—when we look at a continuous sample of an author's prose—and "in the large"—when we look at the historical stream of a language. For example, the *verb* "art" was once extremely common in English. It is precisely this regression that makes a market for "translations" of Chaucer and Shakespeare, and a model that does not account for it, or for the radical fluctuations in the frequencies of both common and uncommon words from one prose sample to another, is faulty. Hence, I should not wish to adopt Dr. Mandelbrot's proposal for "avoiding the unchecked decrease of i/k." This and related matters are discussed somewhat more fully on pages 433–435 of Simon (1955).

2. Dr. Mandelbrot's arithmetic examples greatly overestimate the empirically observable rate of decline of $E(r(p))$, for they are based on the quite unjustified assumption that ρ is a constant, independent of k. On the contrary (see Section II, 2 above), the data indicate that $\rho = 1/(1 - n')$ tends to unity as k increases. I nowhere make the assumption of constant elasticity in my model for slowly decreasing n'; we now have an additional reason for avoiding this assumption. To the best of my knowledge, moreover, $\rho = 1.2$ or 1.1 has not been observed for $k = 10^6$, much less $k = 10^9$.

SECTION IV

This section is simply irrelevant, since I do not assume constant elasticity in my model. As I shall now show, the promised proof of the italicized statement about "circularity" is not provided in Section V.

SECTION V

In this section Dr. Mandelbrot fails utterly to prove that the approximate derivation of the Beta function from the Yule process requires the assumption of constant elasticity.

1. By "very slow decrease in $n'(k)$" I mean rates of decrease like those encountered in the data (see Section I, 2). The purpose of Simon (1955) was to explain certain observed data. The approximation does that.

2. I have nowhere argued that the Beta function can be obtained independently of assumptions on $n'(k)$. Moreover, my results have nothing to do with the Laplace transform: in (1955) I approximated an integral $I(i, k, n)$, *for a particular value of* k, by a particular function $F(i, k, \sigma)$ where the parameter σ may now be a function of k and of the function n.

3. A ρ that decreases more than proportionately with the increase in k is not an example of "very slowly decreasing n'" by anyone's definition. The values of ρ in the spurious "counterexample" resemble none in the data (See Section I, 2). The discussion of the case where $n'(k) = 0$ for $k > K$ is simply irrelevant to the issues before us. [See pp. 430 and 436 of Simon (1955) for this case.]

4. The sentence, "From Assumption I′ it follows that ···" is true and completely irrelevant. Precisely to avoid such an untenable hypothesis, I replaced Assumption I′ by Assumption I; hence, Assumption I′ is not *my* assumption, but *Dr. Mandelbrot's*. Again, see pages 433–435 of Simon (1955). Since my derivation nowhere depends on Assumption I′, the paragraph that follows this sentence in Dr. Mandelbrot's *Final Note* is totally irrelevant.

5. The argument of the following paragraph is therefore also incorrect. First, contrary to Dr. Mandelbrot's unsupported assertions, the Yule distribution generally fits the data well for *small* i (e.g., Simon, 1955, p. 436); it is with *large* i that it breaks down. Secondly, one obviously cannot start from $f(4, k)$, considered as the basic point of departure, for the new words that enter around k_0 at no subsequent time cluster around the value $i = 4$. When $k = 4k_0$, the expected value of i

for these words is 4, but the distribution of i is still given by (5) with $x = e^{-\tau} = \frac{1}{4}$. The variance of this distribution is 12. The true relations among $f(i, k; 1, 0)$, $f(i, k; i_0, k_0)$, and $f(i_0, k_0; 1, 0)$ are given by the Chapman-Kolmogoroff equations (Feller, 1957, p. 424), thus:

$$\begin{aligned} f(i, k) = f(i, k; 1, 0) &= \sum_{i_0=1}^{k_0} f(i, k; i_0, k_0) f(i_0, k_0; 1, 0) \\ &= e^{-\tau}(1 - e^{-\tau})^{i-1}. \end{aligned} \tag{10}$$

This is again the geometric distribution, as we would expect. It has a variance proportional to x^2, hence is not "increasingly more concentrated around its maximum" as Dr. Mandelbrot claims. Incidentally, this maximum does not occur for $i = k/k_0$, the mean, but for $i = 1$, no matter how large k becomes! Monotonic decreasing functions are imperfectly approximated by bell-shaped curves. Heuristic arguments that assume the observations are tightly clustered around their expected value simply bear no weight when we are dealing with this kind of function. Hence, phrases like "as soon as i_0 becomes 4" and "becomes increasingly sharper as i_0 grows" have no meaning.

Thus we come to the end of the list of Dr. Mandelbrot's objections to my approximation without finding a single one that is valid.

SECTION VI

In this section of the "Final Note" a number of points are raised, most of which I dealt with in Simon (1960).

1. I have discussed mathematical errors at several points above, and in (Simon, 1960), particularly at page 84.

2. That Dr. Mandelbrot understands what is meant by "improper" is shown by his paragraph beginning: "Naturally, we do not deny . . . ." This paragraph is entirely consistent with Simon (1960, pp. 84–85), where I discuss this point. The divergence difficulties Dr. Mandelbrot mentions for $i < 1$ are illusory, since they arise only in the continuous approximation (in i) to the discrete model.

3. Dr. Mandelbrot now appears largely to agree with me about the empirical values of ρ for word frequency data. As I stated in Simon (1960), they are sometimes larger than 1, sometimes smaller, and usually very close indeed. Values of ρ close to 1 are characteristic not only of "literary" samples, but of the longer samples in general (see Section I, 2 and III, 2 above). We have never disagreed (e.g., Simon, 1955, p. 430, par. following Table 1) that we need models to handle all cases.

4. Dr. Mandelbrot's observations on the species distribution simply repeat what I have said on page 83 of Simon (1960). Regarding the inset of Figure 9-7 of Zipf, I recommend that the reader plot for himself the cumulative, or rank-frequency, distribution from the data given in the graph. He will see that Dr. Mandelbrot's assertion about the slope of this distribution is clearly false.

5. I do not agree that data on the type-token relations "invariably take the form" k^b, with b less than unity. Again, see Sections I, 2 and III, 2. Moreover, my approximation [Simon, 1955, Eq. (2.34)] shows that the exponent in the rank-frequency distribution will be larger than the exponent b by the factor $[1 - n'(k)]^{-1}$, hence the latter does not give an unbiased estimate of the former.

References

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

Mandelbrot, B. (1959). A note on a class of skew distribution functions. Analysis and critique of a paper by H. Simon. *Information and Control* **2,** 90.

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

Simon, Herbert A. (1960). Some further notes on a class of skew distribution functions. *Information and Control* **3,** 80.