On a Class of Skew Distribution Functions

Herbert A. Simon

Biometrika, Volume 42, Issue 3/4 (Dec., 1955), 425-440.

ON A CLASS OF SKEW DISTRIBUTION FUNCTIONS

By HERBERT A. SIMON†

Carnegie Institute of Technology

I. Introduction

It is the purpose of this paper to analyse a class of distribution functions that appears in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena. Its appearance is so frequent, and the phenomena in which it appears so diverse, that one is led to the conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms. The empirical distributions to which we shall refer specifically are: (A) distributions of words in prose samples by their frequency of occurrence, (B) distributions of scientists by number of papers published, (C) distributions of cities by population, (D) distributions of incomes by size, and (E) distributions of biological genera by number of species.

No one supposes that there is any connexion between horse-kicks suffered by soldiers in the German army and blood cells on a microscope slide other than that the same urn scheme provides a satisfactory abstract model of both phenomena. It is in the same direction that we shall look for an explanation of the observed close similarities among the five classes of distributions listed above.

The observed distributions have the following characteristics in common:

(*a*) They are J-shaped, or at least highly skewed, with very long upper tails. The tails can generally be approximated closely by a function of the form

$$f(i) = (a/i^k)\, b^i, \tag{1·1}$$

where a, b, and k are constants; and where b is so close to unity that in first approximation the final factor has a significant effect on $f(i)$ only for very large values of i. Thus, for example, the number of words that occur exactly i times in James Joyce's *Ulysses* is about a/i^k; the number of authors who published exactly i papers in *Econometrica* over a twenty-year period is approximately a/i^k; and so on.

(*b*) The exponent, k, is greater than 1, and in the cases of word frequencies, publication, and urban populations is very close to 2.‡

(*c*) In the cases of word frequencies, publications and biological genera, the function (1·1) describes the distribution not merely in the tail but also for small values of i. In these cases the ratio $f(2)/f(1)$ is generally in the neighbourhood of one-third, and almost never reaches one-half; while $f(1)/n$, where $n = \sum_1^\infty f(i)$, is generally in the neighbourhood of one-half.

Property (*a*) is characteristic of the 'contagious' distributions—for example, the negative binomial as it approaches its limiting form, Fisher's logarithmic series distribution. However, in the case of the negative binomial, k cannot exceed unity (and equals unity only in

† I have had the benefit of helpful comments from Messrs Benoit Mandelbrot, Robert Solow and C. B. Winsten. I am grateful to the Ford Foundation for a grant-in-aid that made the completion of this work possible.

‡ See Zipf (1949) for numerous examples of distributions with this property.

the limiting case of the log series); and if the distribution has a long tail, so that the convergence factor, b, is close to unity, $f(2)/f(1)$ cannot be less than one-half. Hence the negative binomial and Fisher's log series distributions do not provide a satisfactory fit for data possessing property (a) together with either (b) or (c).†

It is well known that the negative binomial and the log series distributions can be obtained as the stationary solutions of certain stochastic processes. For example, J. H. Darwin (1953) derives these from birth and death processes, with appropriate assumptions as to the birth- and death-rates and the initial conditions. In this paper we shall show that stochastic processes closely similar to those yielding the negative binomial or log series distributions lead to a class of functions having the three properties enumerated above. This class of functions is given by

$$f(i) = AB(i, \rho + 1), \tag{1·2}$$

where A and ρ are constants, and $B(i, \rho + 1)$ is the Beta function of $i, \rho + 1$:

$$B(i, \rho + 1) = \int_0^1 \lambda^{i-1}(1-\lambda)^\rho \, d\lambda = \frac{\Gamma(i)\,\Gamma(\rho+1)}{\Gamma(i+\rho+1)} \quad (0 < i;\ 0 < \rho < \infty). \tag{1·3}$$

Now it is a well-known property of the Gamma function (Titchmarsh, 1939, p. 58) that as $i \to \infty$, and for any constant, k,

$$\frac{\Gamma(i)}{\Gamma(i+k)} \sim i^{-k}. \tag{1·4}$$

Hence, from (1·3), we have, as $i \to \infty$:

$$f(i) \sim \Gamma(\rho + 1)\, i^{-(\rho+1)}. \tag{1·5}$$

Therefore, the distribution (1·2) approximates the distribution (1·1) in the tail (more precisely, through the range in which the convergence factor of the latter is close to one). Further, if ρ is positive, k will be greater than 1, as required by (b); and if ρ is equal to 1, k will be equal to 2. It is easy to see that in the latter case we will have

$$f(i) = \frac{1}{i(i+1)}, \quad \sum_{i=1}^{\infty} f(i) = 1, \tag{1·6}$$

so that $f(2)/f(1) = \frac{1}{3}$; and $f(1)/n = \frac{1}{2}$, as required by (c).

In the remainder of this paper I propose: (a) to describe a stochastic process that leads to the stationary distribution (1·2); (b) to discuss some generalizations of this process; and (c) to construct hypotheses as to why the empirical phenomena mentioned above can be represented, approximately, by processes of this general kind. Before proceeding, I should like to mention two earlier derivations, one of (1·2), the other of (1·1), that I have been able to discover in the literature.

Some thirty years ago, G. Udny Yule (1924) constructed a probability model, with (1·2) as its limiting distribution, to explain the distribution of biological genera by numbers of species. He also derived a modified form of (1·2), replacing the complete Beta-function of (1·3) by the incomplete Beta-function with upper limit of integration $\alpha < 1$. (This modification has the same effect as the introduction of the convergence factor, b^i, in (1·1)—it causes a more rapid decrease in $f(i)$ for very large values of i; cf. also Darwin (1953, p. 378).) It seems highly appropriate to call the distribution (1·2) the Yule distribution.

† The contrasting characteristics of distributions for which the log series provides a satisfactory fit and those, under consideration here, for which it does not are illustrated by examples (i) and (ii), respectively, in Good (1953).

Because Yule's paper predated the modern theory of stochastic processes, his derivation was necessarily more involved than the one we shall employ here. Moreover, while the assumptions he required are plausible for the particular biological problem he treated, the corresponding assumptions applied to the four other phenomena we have mentioned appear much less plausible. Our derivation requires substantially weaker assumptions than Yule's about the underlying probability mechanism.

More recently D. G. Champernowne (1953) has contructed a stochastic model of income distribution that leads to (1·1) and to generalizations of that function. Since the points of similarity between his model and the one under discussion here are not entirely obvious at a first examination, I shall consider their relation in a later section of this paper.

II. THE STOCHASTIC MODEL

For ease of exposition, the model will be described in terms of word frequencies. In a later section, alternative interpretations will be provided. Our present interest is in the kind of stochastic process that would lead to (1·2).

Consider a book that is being written, and that has reached a length of k words. We designate by $f(i, k)$ the number of *different* words that have occurred exactly i times in the first k words. That is, if there are 407 different words that have occurred exactly once each, then $f(1, k) = 407$.

Assumption I. The probability that the $(k+1)$*-st word is a word that has already appeared exactly* i *times is proportional to* $if(i, k)$—that is, to the total number of occurrences of all the words that have appeared exactly i times.

Note that this assumption is much weaker than the assumption (I′): that the probability a *particular* word occur next be proportional to the number of its previous occurrences. Assumption (I′) implies (I), but the converse is not true. Hence we leave open the possibility that, among all words that have appeared i times the probability of recurrence of some may be much higher than of others.

Assumption II. There is a constant probability, α*, that the* $(k+1)$*-st word be a new word*—a word that has not occurred in the first k words.

Assumptions (I) and (II) describe a stochastic process, in which the probability that a particular word will be the next one written depends on what words have been written previously. If this process correctly describes the selection of words, then the words in a book cannot be regarded as a random sample drawn from a population with a prior distribution. The reasonableness of the former, as compared with the latter type of explanation of the observed distributions, will be discussed in § IV.

From (I), it follows that

$$\mathscr{E}\{f(i, k+1)\} - f(i, k) = K(k)\{(i-1)f(i-1, k) - if(i, k)\} \quad (i = 2, \ldots, k+1), \tag{2·1}$$

for if the $(k+1)$st word is one that has previously occurred $(i-1)$ times, $f(i, k+1)$ will be increased over $f(i, k)$, and the probability of this, by assumption (I), is proportional to $(i-1)f(i-1, k)$; if the $(k+1)$st word is one that previously occurred i times, $f(i, k+1)$ will be decreased, and the probability of this, by assumption (I), is proportional to $if(i, k)$; while in all other cases, $f(i, k+1) = f(i, k)$.

From (I) and (II) we obtain similarly

$$\mathscr{E}\{f(1, k+1)\} - f(1, k) = \alpha - K(k)f(1, k) \quad (0 < \alpha < 1). \tag{2·2}$$

Since we will be concerned throughout with 'steady-state' distributions (as defined by equation (2·8) below), we replace the expected values in (2·1) and (2·2) by the actual frequencies. (Alternatively, we might replace frequencies on the right-hand side of the equation by probabilities.) That is, we write, instead of (2·1) and (2·2),

$$f(i,k+1)-f(i,k) = K(k)\{(i-1)f(i-1,k)-if(i,k)\} \quad (i=2,\ldots,k+1), \tag{2·3}$$

$$f(1,k+1)-f(1,k) = \alpha - K(k)f(1,k), \tag{2·4}$$

where the f's now represent expected values.

Now, we wish to evaluate the factor of proportionality $K(k)$. Since $K(k)\,if(i,k)$ is the probability that the $(k+1)$st word is one that previously occurred i times, we must have

$$\sum_{i=1}^{k} K(k)\,if(i,k) = K(k)\sum_{i=1}^{k} if(i,k) = 1-\alpha. \tag{2·5}$$

But $\sum_{i=1}^{k} if(i,k)$ is the total number of words up to the kth, hence

$$\sum_{i=1}^{k} if(i,k) = k, \tag{2·6}$$

and

$$K(k) = \frac{1-\alpha}{k}. \tag{2·7}$$

Substituting (2·7) in (2·3) and (2·4), we could solve these differential equations explicitly. We can avail ourselves, however, of a simpler—though non-rigorous—method for discovering the solutions, and can then test their correctness by substitution in the original equations. Consider the 'steady-state' distribution in the following sense. We assume

$$\frac{f(i,k+1)}{f(i,k)} = \frac{k+1}{k} \quad \text{for all } i \text{ and } k; \tag{2·8}$$

so that all the frequencies grow proportionately with k, and hence maintain the same relative size. (Since we must have $f(i,k) = 0$ for $i > k$, equation (2·8) cannot hold exactly for all i and k. But as explained above, we are concerned at the moment with heuristic rather than proof.)

From (2·8) it follows that

$$\frac{f(i,k)}{f(i-1,k)} = \frac{f(i,k+1)}{f(i-1,k+1)} = \beta(i), \tag{2·9}$$

where $\beta(i)$ does not involve k. Hence, the *relative* frequencies, which we will designate by $f^*(i)$, are independent of k. Substituting (2·7), (2·8) and (2·9) in (2·3), we get

$$\left(\frac{k+1}{k}-1\right)f(i,k) = \frac{(1-\alpha)}{k}\left\{\frac{(i-1)}{\beta(i)}-i\right\}f(i,k). \tag{2·10}$$

Cancelling the common factor, and solving β for (i), we obtain

$$\beta(i) = \frac{(1-\alpha)(i-1)}{1+(1-\alpha)i} = \frac{f^*(i)}{f^*(i-1)} \quad (i=2,\ldots,k). \tag{2·11}$$

For convenience, we introduce

$$\rho = \frac{1}{1-\alpha} \quad (1<\rho<\infty). \tag{2·12}$$

Since $f^*(i) = \beta(i) f^*(i-1) = \beta(i).\beta(i-1)\dots\beta(2) f^*(1)$, we obtain from (2·11) and (2·12)

$$f^*(i) = \frac{(i-1)(i-2)\dots 2.1}{(i+\rho)(i+\rho-1)\dots(1+\rho)} f^*(1) = \frac{\Gamma(i)\,\Gamma(\rho+1)}{\Gamma(i+\rho+1)} f^*(1) = B(i,\rho+1) f^*(1) \quad (i = 2,\dots,k). \tag{2·13}$$

The second relation follows from the fact that

$$\Gamma(i+\rho+1) = (i+\rho)\,\Gamma(i+\rho) = (i+\rho)(i+\rho-1)\dots(1+\rho)\,\Gamma(\rho+1). \tag{2·14}$$

But (2·13) is identical with (1·2) if we take $A = f^*(1)$.

That (2·13) is in fact a solution of (2·3) can be verified by direct substitution. Moreover, it is in the following sense a stable solution. Suppose that (2·11) is *not* satisfied. Whatever be the values of the $f(i,k)$ for a given k, we may write without loss of generality

$$\frac{f(i,k)}{f(i-1,k)} = \frac{(1-\alpha)(i-1)}{(1-\alpha)i + 1 + \epsilon(i,k)}, \tag{2·15}$$

where $\epsilon(i,k)$ is some function of i and k. If we now divide both sides of (2·3) by $f(i,k)$ and substitute (2·15) in the right-hand side of the resulting equation, we obtain after simplification

$$\frac{f(i,k+1)}{f(i,k)} = \frac{k+1+\epsilon(i,k)}{k}. \tag{2·16}$$

Hence the ratio of $f(i,k+1)$ to $f(i,k)$ will be greater than $(k+1)/k$ if $\epsilon(i,k)$ is positive, and less than $(k+1)/k$ if $\epsilon(i,k)$ is negative. Since new words are introduced at a constant rate, $\sum_1^k f(i,k)$ must be proportional to k; therefore, by (2·16), we will have

$$\sum_{i=1}^{k+1} f(i,k+1) - \frac{k+1}{k} \sum_{i=1}^{k+1} f(i,k) = \frac{1}{k} \sum_{i=1}^{k} \epsilon(i,k) f(i,k) = 0. \tag{2·17}$$

We may interpret the three equations, (2·15)–(2·17), as follows. In an average sense, the frequencies will grow proportionately with k. If a particular frequency is 'too large' compared with the next lower frequency ($\epsilon(i,k)$ negative in (2·15)), it will grow at a rate slower than the average; if it is 'too small' ($\epsilon(i,k)$ positive), it will grow more rapidly than the average.

It remains to be shown that $f^*(i) = B(i,\rho+1) f^*(1)$ is a proper distribution function. In particular, we require that $\sum_{i=1}^{k} iB(i,\rho+1)$ converge as $k \to \infty$. Now, it is well known that $\sum_{i=1}^{\infty} i^{-a}$ converges for every $a > 1$. But by (1·4),

$$\sum_{i=1}^{\infty} iB(i,\rho+1) \sim i.i^{-(\rho+1)} = i^{-\rho}. \tag{2·18}$$

Hence, by the usual ratio comparison test, $\sum_{i=1}^{\infty} iB(i,\rho+1)$ converges for $\rho > 1$, as required.

From the definition of α the total number, n_k, of *different* words will be αk; while the total number of word occurrences is k. That is

$$n_k = \sum_{i=1}^{k} f(i,k) = \alpha k = \alpha \sum_{i=1}^{k} if(i,k). \tag{2·19}$$

Returning to (2·4), and using (2·8), we get

$$\left(\frac{k+1}{k}-1\right)f^*(1)=\alpha-\frac{1-\alpha}{k}f^*(1), \tag{2·20}$$

whence

$$f^*(1)=\frac{k\alpha}{2-\alpha}=\frac{n_k}{2-\alpha}. \tag{2·21}$$

From (2·12) and (2·21), and by successive application of (2·11), we can compute the values of $\rho, f^*(1)/n_k, f^*(2)/n_k, f^*(3)/n_k$, etc., for given values of α (Table 1).

Table 1

α	ρ	$f^*(1)/n_k$	$f^*(2)/n_k$	$f^*(3)/n_k$
0·0	1	0·500	0·167	0·083
0·1	1·11	0·527	0·169	0·082
0·2	1·25	0·556	0·171	0·080
0·3	1·43	0·588	0·171	0·077
0·5	2·00	0·667	0·167	0·067
0·7	3·33	0·769	0·144	0·046
0·9	10·00	0·909	0·076	0·012

Thus far we have considered the case where α, the rate at which new words are introduced, is independent of k. We can easily generalize to the case where α is a function of k by making the appropriate substitution in (2·4). The equations can then be solved directly, but the method employed to obtain a 'steady-state' distribution is not applicable, since it is not easy to define what is meant by the steady state in this more general case. We will content ourselves with some approximate results for two special cases. These special cases will give us insight as to how a distribution function may arise which, for small values of i, can be approximated by (1·2), with $0<\rho<1$.

Case I. Suppose the system to be in the steady state described by (2·13) with $k=k_0$, and that the flow of *new* words suddenly ceases, so that $\alpha(k)=0$ for $k>k_0$. We will now have $K(k)=1/k$ for $k>k_0$, and (2·4) becomes

$$f(1,k+1)=\left(1-\frac{1}{k}\right)f(1,k)=\frac{k-1}{k}f(1,k). \tag{2·22}$$

We define

$$\gamma(i)=\frac{f(i,k+1)}{f(i,k)} \quad (i=2,\ldots,k+1). \tag{2·23}$$

Since no new words are being introduced, we must have

$$\begin{aligned} n_k &= f(1,k)+\sum_{i=2}^{k}f(i,k)=f(1,k+1)+\sum_{i=2}^{k}f(i,k+1)\\ &=\frac{(k-1)}{k}f(1,k)+\sum_{i=2}^{k}\gamma(i)f(i,k), \end{aligned} \tag{2·24}$$

whence

$$\frac{\sum_{i=2}^{k}[\gamma(i)-1]f(i,k)}{\sum_{i=2}^{k}f(i,k)}=\frac{1}{k}\frac{f(1,k)}{\sum_{i=2}^{k}f(i,k)}. \tag{2·25}$$

Let us define next

$$\beta(i) = \frac{f(i,k)}{f(i-1,k)} = \frac{(i-1)}{(1+\rho_i)} \tag{2·26}$$

(where we suppose that ρ_i changes only slowly with k). Instead of (2·3), we have

$$f(i,k+1) - f(i,k) = \frac{1}{k}[(i-1)f(i-1,k) - if(i,k)]. \tag{2·27}$$

Substituting (2·23) and (2·26) in this, we get

$$\gamma(i) - 1 = \frac{1}{k}[(i+\rho_i) - i], \tag{2·28}$$

whence

$$\rho_i = k(\gamma(i) - 1), \tag{2·29}$$

and

$$\bar{\rho} = \frac{\sum_{i=2}^{k} k(\gamma(i)-1)f(i,k)}{\sum_{i=2}^{k} f(i,k)} = \frac{f(1,k)}{\sum_{i=2}^{k} f(i,k)} = \frac{f(1,k)}{n_k - f(1,k)}. \tag{2·30}$$

Define

$$p_1 = f(1,k)/n_k. \tag{2·31}$$

Then

$$\bar{\rho} = \frac{p_1}{1-p_1} \quad \text{and} \quad 0 < \bar{\rho} < \infty. \tag{2·32}$$

Proceeding heuristically, we can see that after α becomes zero, $f(1,k)$ will begin to decrease with k, and the value of ρ_i will be larger the larger is i. For small values of i, we will have $\rho(i) < \bar{\rho}$, and for large values, $\rho(i) > \bar{\rho}$. However, the tail of the distribution will be affected only slowly by the change in α. Hence, we may suppose that $\lim_{i \to k_0} \rho(i) = \rho_0$, where ρ_0 is $\rho(k_0)$. On the other hand, since the weighted average in (2·29) is heavily influenced by the large frequencies for small values of i, ρ_i will be only slightly less than $\bar{\rho}$. Hence we may expect the distribution to take the form of a slightly curved line on a double-log scale, with a slope of $-(\bar{\rho}+1)$ at the lower end, and a slope of $-(\rho_0+1)$ at the upper end. If $\rho_0 > 2$, then $\Sigma i f(i,k)$ will converge. An example of such a distribution will be given in § IV.

Case II. A second approximate solution can be obtained if we assume that α decreases with k, but very slowly. By definition, we have $\alpha = dn_k/dk = n'$. The condition for a steady state (all frequencies increasing proportionately) is now

$$f(i,k+1) = [1 + (n'/n_k)]f(i,k). \tag{2·33}$$

Substituting as before, (2·7) and (2·33) in (2·3), we again obtain (2·13), where ρ is now given by

$$\rho = \frac{n'k}{n_k} \frac{1}{(1-n')}. \tag{2·34}$$

The slope obtained in the derivation for constant α has now been multiplied by the factor $(n'k)/n_k$, which for monotonically decreasing α is less than one. Hence, the effect of a decrease in the rate of introduction of new words is to lengthen the tail of the distribution, as was also true in case I. If the new value of ρ is less than one, we do not have a proper distribution function (see equation (2·18)), hence the equation can hold only for small and moderate values of i, and there must be a curve (on a logarithmic scale) in the tail of the distribution.

III. An alternative formulation of the process

There are some alternative ways for deriving the relation (2·13). One of these will be useful to us when we come, in the next section, to a more specific discussion of word frequencies and frequencies of publications. Moreover, this derivation avoids the difficulties we have encountered in the definition of 'steady state'.

Equation (2·10) may be written

$$0 = (1-\alpha)[(i-1)f^*(i-1) - if^*(i)] - f^*(i) \quad (i = 2, \ldots, k), \tag{3·1}$$

where we have again written $f^*(i)$ for $f(i, k)$.

Similarly, from (2·4), we obtain

$$0 = 1 - (1-\alpha)f^*(1) - f^*(1). \tag{3·2}$$

These two equations may be interpreted as follows. We consider a sequence of k words. We add words to the sequence in accordance with assumptions (I) and (II) of § II, *but we drop words from the sequence at the same average rate*, so that the length of the sequence remains k. The method according to which we drop words is the following:

Assumption III. If one representative of a particular word is dropped, then all representatives of that word are dropped, and the probability that the next word dropped be one with exactly i representatives is $f^*(i)$.

This assumption would be approximately satisfied, for example, if the representatives of each word, instead of being distributed randomly through the sequence, were closely 'bunched'. This possibility is consistent with assumption (I).

Equation (3·1), in our new interpretation, may be regarded as the steady-state equilibrium of the stochastic process defined by

$$f(i, m+1) - f(i, m) = (1-\alpha)[(i-1)f(i-1, m) - if(i, m)] - f(i, m), \tag{3·3}$$

where m is now not the total number of words (which remains a constant, k), but the number of additions to (and withdrawals from) an initial arbitrary sequence of k words. Since the k of this process, unlike that of § II, remains constant, the ordinary proofs of the existence of a unique steady-state solution will apply (see Feller, 1950, p. 373), and we avoid the troublesome questions of rigour that confronted us in § II.

The solution of (3·1) and (3·2) is, of course, again given by

$$\frac{f^*(i)}{f^*(i-1)} = \frac{(1-\alpha)(i-1)}{1+(1-\alpha)i}. \tag{2·11}$$

If we were to replace the last terms of (3·1) and of (3·2), respectively, by terms corresponding to the usual form of the death process, we would have (cf. Darwin, 1953, p. 375; and Kendall, 1948)

$$0 = (1-\alpha)[(i-1)f^*(i-1) - if^*(i)] - [if^*(i) - (i+1)f^*(i+1)] \quad (i = 2, \ldots, k-1), \tag{3·4}$$

$$0 = 1 - (1-\alpha)f^*(1) - [f^*(1) - 2f^*(2)]. \tag{3·5}$$

The solution of this system of equations is

$$\frac{f^*(i)}{f^*(i-1)} = \frac{(1-\alpha)(i-1)}{i}, \tag{3·6}$$

which is Fisher's logarithmic series distribution.

Since the log series distribution is a limiting case of the negative binomial, we may ask whether there is a distribution that stands in the same relation to the latter as (2·11) stands in relation to (3·6). We can obtain such a distribution by a modification of the birth process in (3·1). We assume now that the birth-rate is the sum of two components—one proportional to $if(i)$, the other proportional to $f(i)$. In place of (3·1) we have

$$0 = \frac{(1-\alpha)k}{k+c}[(i-1+c)f^*(i-1)-(i+c)f^*(i)]-f^*(i) \quad (c \text{ a constant}), \tag{3·7}$$

the solution of which is

$$\frac{f^*(i)}{f^*(i-1)} = \frac{\lambda(i-1+c)}{\lambda(i+c)+1} = \frac{(i-1+c)}{(i+c+1/\lambda)}, \tag{3·8}$$

where

$$\lambda = k(1-\alpha)/(k+c).$$

A rather remarkable property of (3·8) is that in the tail it still has the limiting form (1·1) with $b = 1$. Hence for α and c small, this generalized Yule distribution will still possess the three properties listed in the introduction. The fact that a reasonably wide range of variation in the assumptions underlying the stochastic model does not alter greatly the form of the distribution adds plausibility to the use of such stochastic processes to explain the observed distributions. Our next task is to consider these explanations in more detail.

IV. THE EMPIRICAL DISTRIBUTIONS

In this section I shall try to give theoretical justifications for the observed fit of the Yule distribution to a number of different sets of empirical data.

A. *Word frequencies*

A substantial number of word counts have been made, in English and in other languages (see Hanley, 1937; Thorndike, 1937; Yule, 1944; Zipf, 1949; and Good, 1953). Equation (1·6) provides a good fit to almost all of them. When the more general function, (1·2), is used, the estimated value of ρ is always close to 1. When a convergence factor, b^i, is introduced to account for the deficiency in frequencies for very large values of i, the estimated value of b is also very close to 1. Good (1953), for instance, applies (1·6) multiplied by a convergence factor to the Eldridge count, and obtains $b = 0{\cdot}999667$.

These regularities are the more surprising in that the various counts refer to a quite heterogeneous set of objects. In the Yule and Thorndike counts, inflected forms are counted with the root word; in most of the other counts each form is regarded as a distinct word. The Yule counts include only nouns; the others, all parts of speech. The Dewey, Eldridge and Thorndike counts are composite—compiled from a large number of separate writings; most of the others are based on a single piece of continuous prose. I would regard this heterogeneity as further evidence that the explanation is to be sought in a probability mechanism, rather than in more specific properties of language; but at the same time, the heterogeneity complicates the task of specifying the probability mechanism in detail. I shall avoid questions of 'fine structure'—which would require an expertness in linguistics that I do not possess—and confine myself to three broad problems: (1) the distribution of word frequencies in the whole historical sequence of words that constitutes a language; (2) the distribution of word frequencies in a continuous piece of prose; (3) the distribution of word frequencies in a sample of prose assembled from composite sources.

(1) For obvious reasons, we do not have any empirical data on the cumulated word frequencies for a whole language. On *a priori* grounds, it does not appear unreasonable to postulate that these frequencies are determined by a process like that described in § II. The parameter α is then the rate at which neologisms appear in the language as a fraction of all word occurrences—and hence α can be assumed to be very close to zero.

(2) The process of § II might also describe the growth of a continuous piece of prose—for example, Joyce's *Ulysses*. But there are some serious objections to this hypothesis. An author writes not only by processes of *association*—i.e. sampling earlier segments of the word sequence—but also by processes of *imitation*—i.e. sampling segments of word sequences from other works he has written, from works of other authors, and, of course, from sequences he has heard. The model of § II apparently allows only for association, and excludes imitation.

The word frequencies in *Ulysses* provide obvious evidence of the importance of both processes. The fact that the proper noun 'Bloom' occurs 926 times and ranks 30th in frequency must be attributed to association. If Joyce had named his hero 'Smith', that noun, instead of 'Bloom', would have ranked 30th. On the other hand, 'they', which occurs 1010 times in *Ulysses* and ranks 27th, has very nearly the same rank—the 28th—in the Dewey count. In fact, of the 100 most frequent words in *Ulysses*, 78 are among the top 100 in the Dewey count. This similarity in ranking of 'common' words argues for imitation rather than association. Even for the common words, however, the variations in frequency from one count to another are far too great to be explained as fluctuations resulting from random sampling from a common population of words. The imitative process must involve stratified sampling, and imitation must be compounded with association.

It is worth emphasizing again at this point that assumption (I) does not require that the choice of the next word from among those previously written be completely random. Suppose, for example, that a writer were to assign to each page he has already written a number, p_j, $\Sigma p_j = (1-\alpha)$, the size of p_j varying with the 'affinity' of the subject discussed on the jth page to the subject next to be discussed. If his next word were selected by a stratified sampling of the previous pages, with probability p_j for each page, then equation (2·1) would generally be satisfied. For although individual words would be distributed unevenly through the preceding pages, the totality of words having a given frequency, i, in all the previous pages taken together would be distributed almost evenly through these pages. Hence, the various frequency strata would have proportionate probabilities of being sampled, for most choices of the p_j. This is all that is required for equation (2·1). This same comment applies to the assumption we shall subsequently make regarding imitative sampling from other works.

Let us now reconsider the problem of a piece of continuous prose. Since both the processes of association and imitation are involved, the sequence that is counted is to be regarded as a 'slice', of length k, of the entire sequence of words in the language, or of the entire sequence written by the author. Hence the word count is better described by the stochastic process of § III than by the process of § II.

In determining the probability that a word selected in such a sequence be one that has occurred exactly i times, we must consider separately the process of imitation and association. Assume that, on the average, a fraction, β, of the words added is selected by imitation, and the remaining fraction, $(1-\beta)$, by association. Since no new words can be introduced by association, the joint probability that the next word will be selected by association and will be a word that has already occurred i times is $(1-\beta)\,if(i,k)/k$.

The words selected by imitation present a more difficult problem, and we shall have to content ourselves with a reasonable assumption that has no rigorous justification. On the average, a word that has occurred i times will have a chance less than i/k of being the next one chosen by imitation, because in the sequence that is being sampled there are words that have not yet been chosen at all, and because with progressive change of subject, different strata of the language will be sampled. Since words with large i will generally be 'common' words, fairly uniformly distributed through all strata of the language, the deficiency may be expected to be proportionately greater for small i than for large i. As a rough, but reasonable, approximation let us assume that: the joint probability that the next word will be selected by imitation and will be a word that has already occurred i times is $\beta(i-c)f(i,k)/k$, where $0<c<1$. (Our result would not be essentially altered if we wrote $c(i)$ instead of c, provided only that $c(i)$ does not vary a great deal.)

Adding the two joint probabilities—for association and imitation, respectively—we find that the total probability that the next word be one that has occurred i times is $(i-\beta c)f(i,k)/k$. By summing this probability over i and subtracting from 1, we find that the probability that the next word be a new word is $\beta c(n_k/k)$.

If the method of dropping words from the sequence satisfies assumption (III) of § III, we set the difference between the birth-rate and the death-rate equal to zero, and obtain the steady-state equation

$$0 = (i-c\beta-1)f^*(i-1)-(i-c\beta)f^*(i)-f^*(i), \tag{4·1}$$

which has as its solution

$$\frac{f^*(i)}{f^*(i-1)} = \frac{(i-c\beta-1)}{(i-c\beta+1)}. \tag{4·2}$$

Again, we obtain a distribution with the required properties.

(3) The distribution of word frequencies in a sample of prose assembled from composite sources can be explained along the same general lines. Again, we may regard the sample as a 'slice' from a longer sequence, but we might expect the parameters c and β to be somewhat larger than in a comparable piece of continuous prose. The qualification 'comparable' is important, for c may be expected to be smaller for homogeneous prose using a limited vocabulary of common words than for prose with a large vocabulary and treating of a variety of subjects. Hence c might well be larger for the continuous *Ulysses* count than for the Eldridge count, which is drawn from newspaper sources. Indeed, the empirical evidence suggests that this is the case.

There is no point in elaborating the explanation further. What has been shown is that the observed frequencies can be fitted by distributions derived from probability assumptions that are not without plausibility.

A very different and very ingenious explanation of the observed word-frequency data has been advanced recently by Dr Benoit Mandelbrot (1953). His derivation rests on the assumption that the frequencies are determined so as to maximize the number of bits of information, in the sense of Shannon, transmitted per symbol. There are several reasons why I prefer an explanation that employs averaging rather than maximizing assumptions. First, an assumption that word usage satisfies some criterion of efficiency appears to be much stronger than the probability assumptions required here. Secondly, numerous doubts, which I share, have been expressed as to the relevance of Shannon's information measure for the measurement of semantic information.

Before leaving the subject of word frequencies, it may be of interest to look at some of the empirical data. Good (1953, pp. 257–60), has obtained good fits to the Eldridge count and to one of Yule's counts by the use of equation (1·6). Table 2, summarizes a few of the data on two word counts, and compares the actual frequencies, $f(1)$, $f(2)$ and $f(3)$ with the frequencies estimated from equation (1·3). The actual values of k and n_k are used to estimate $\alpha = n_k/k$, and (2·11) and (2·21) to obtain the expected frequencies. In both cases the observed value of n_k/k leads to an estimate of ρ in the neighbourhood of 1·1 to 1·2. An empirical fit to the whole distribution of a function of the form $f(i) = Ka^{-(\rho+1)}$ gives an estimated value of ρ, in both cases, of about one—in reasonable agreement with the first estimate. A good fit to both the *Ulysses* and the Eldridge counts can also be obtained from (4·2), with c equal to about 0·2 in the former case, and close to zero in the latter.

In the case of Thorndike's count of $4\frac{1}{2}$ million words in children's books (Thorndike, 1937), we may assume that the supply of new words was virtually exhausted before the end of the count. In his count $f(1)$ is substantially below $0{\cdot}5n_k$ (about $0{\cdot}34n_k$), as we would expect under these circumstances (see case I of § II). Thorndike estimated the empirical value of our $\bar{\rho}$ at 0·45, which is entirely consistent with the observed value of $0{\cdot}34n_k$ for $f(1)$. For, by (2·32), $p_1 = \bar{\rho}/(\bar{\rho}+1) = 0{\cdot}31$.

Table 2

Word count	$\alpha = \frac{n_k}{k}$	$f(1)$		$f(2)$		$f(3)$	
		Actual	Estimate	Actual	Estimate	Actual	Estimate
Ulysses (Hanley, 1937)	0·115	16,432	15,850	4,776	4,870	2,194	2,220
Eldridge (Good, 1953)	0·136	2,976	3,220	1,079	977	516	400

B. *Scientific publications*

At least four sets of data are available on the number, $f(i)$, of authors contributing a given number, i, of papers each to a journal or journals (Davis, 1941; Leavens, 1953). These are counts of (*a*) papers written by members of the Chicago Section of the American Mathematical Society over a 25-year period; (*b*) papers listed in *Chemical Abstracts* (under A and B) over 10 years; (*c*) papers referred to in a history of physics; and (*d*) papers and abstracts in *Econometrica* over a 20-year period.

We may postulate a mechanism like that of § III, equation (3·1). The authorship of the next paper to appear is 'selected' by stratified sampling from the strata of authors who have previously published $1, 2, \ldots,$ papers, the probability for each stratum being proportional to $if(i)$. Again, the probabilities for individual authors need not be proportional to i, but only the probabilities for the aggregates of authors with the same i. For example (as in the case of words), the probability for a particular author may be higher if he has published recently than if he has not. The gradual retirement of authors corresponds to assumption (III).

A comparison of the actual frequencies, for i from 1 to 10, with the estimated frequencies, derived from (2·11) and (2·21), is shown in Table 3. The fit is reasonably good, when it is remembered that only one parameter is available for adjustment. However, it should be

noted that the estimated frequencies tend to be too high for $i = 1, 2$ and too low for $i = 3, \ldots, 10$. In three of the four cases, they are again too high for the tails of the distributions. A further refinement of the model is apparently needed to remove these discrepancies.

Table 3. *Number of persons contributing*

No. of contributions	Chicago Math. Soc.		*Chem. Abstracts*		Physicists		*Econometrica*	
	Actual†	Estimate	Actual†	Estimate	Actual†	Estimate	Actual‡	Estimate
1	133	—	3,991	4,050	784	824	436	453
2	43	46	1,059	1,160	204	217	107	119
3	24	23	493	522	127	94	61	51
4	12	14	287	288	50	50	40	27
5	11	10	184	179	33	30	14	16
6	14	7	131	120	28	20	23	11
7	5	5	113	86	19	14	6	7
8	3	4	85	64	19	10	11	5
9	9 } 13	3 } 10	64	49	6 } 32	8 } 24	1 } 12	4 } 12
10	1	3	65	38	7	6	0	3
11 or more	23	30	419	335	48	52	22	25
Estimated α	0§		0·30		0·39		0·41	
Estimated ρ	0·916§		1·43		1·64		1·69	
k	1,124		22,939		3,396		1,759	
n_k	278		6,891		1,325		721	
ρ'	1·07		—		—		—	

† Davis (1941).
‡ Leavens (1953).
§ $\rho = \bar{\rho}$ estimated in this case from (2·31) to (2·32).

C. *City sizes*

It has been observed, for every U.S. Census since the early nineteenth century, and for most other Western countries as well, that if $F(i)$ is the number of cities of population greater than i, then

$$F(i) \sim Ai^{-\rho}, \tag{4·3}$$

where ρ is close to 1 (see Zipf, 1949, chs. 9, 10).

Again, we would expect such a distribution if the underlying mechanism were one describable by equations like (2·3) and (2·4). Such a mechanism is not hard to conceive. First, equation (2·3) would hold if the growth of population were due solely to the net excess of births over deaths, and if this net growth were proportional to present population. This assumption is certainly satisfied at least roughly. Moreover, it need not hold for each city, but only for the aggregate of cities in each population band. Finally, the equation would still be satisfied if there were net migration to or from cities of particular regions, provided the net addition or loss of population of individual cities *within any region* was proportional to city size. That is, even if all California cities were growing, and all New England cities declining, the equation would hold provided the percentage growth or decline in each area were uncorrelated with city size.

In the case of cities, equation (4·3) could only be expected to hold down to some minimum city size—say, 5000 or 10,000. The constant α would then be interpreted as the fraction of the total population growth in cities above the minimum size that is accounted for by the new cities that reach that size.

D. *Income distribution*

Vilfredo Pareto is generally credited with the discovery that if personal incomes are ranked by size, the number of persons, $F(i)$, whose incomes exceed i can be approximated closely, for the upper ranges of income, by equation (4·3) with ρ usually in the neighbourhood of 1·5 (Davis, 1941; Champernowne, 1953). Hence, the income distributions bear a family resemblance in their upper ranges to those we have already considered, although the parameter, ρ, is substantially larger than 1—its characteristic value in the case of word frequencies and city size distributions.

A stochastic mechanism similar to those described in § III would again produce steady-state distributions closely resembling the observed ones. We picture the stream of income as a sequence of dollars allocated probabilistically to the recipients. If the total annual income of all persons above some specified minimum income is k dollars, the segment of this sequence running from the mth to the $(m+k)$th dollar is the income for the year beginning at time m. We assume that the probability that the next dollar will be allotted to some person with an annual income of i dollars is proportional to $(i+c)f(i)$, with c positive but small. This represents a modification of assumption (I) that decreases the proportion of the total stream going to persons of high income relative to the proportion going to persons with incomes close to the minimum. We assume that a fraction of the dollars is assigned to new persons—i.e. persons reaching the minimum income to which the assumptions apply (assumption (II)). We assume that there is considerable variance among persons within each income class in the probability of receiving additional income, so that the rate at which dollars are dropped from any income class as m increases satisfies assumption (III). Then we obtain again equation (3·8), which now holds for i greater than the minimum income. For large i, this distribution has the required properties with $1/\lambda = \rho$.

The same result has been reached by D. G. Champernowne (1953), following a somewhat different route. He divides income recipients at time t_1 into classes of equal proportionate width. That is, if i_m is the minimum income considered, then the first class contains persons with incomes between i_m and ri_m, the second class, persons with incomes between ri_m and $r^2 i_m$, and so on. Next he introduces transition probabilities p_{gh}, that a person who is in class g at time t_1 will be in class h at time t_2. He assumes that p_{gh} is a function only of $(g-h)$. Now, by his definition of the income classes, the average income of persons in class g will be about $r^{(g-h)}$ times the average income of persons in class h. Hence, the expected income at t_2 of a person who was in class g at t_1 will be

$$\sum_h p_{gh} i_h = \sum_h p_{(g-h)} r^{(g-h)} i_g = \alpha i_g \quad (\alpha \text{ a constant}), \tag{4·4}$$

where i_g is the average income in class g. Prof. Champernowne assumes explicitly that $\alpha < 1$. From this it is clear that his model satisfies our assumptions (I) (in its original form) and (II). Further, since he assumes a substantial variance in income expectations among persons in a given class, our assumption (III) is also approximately satisfied. Hence, in spite of the surface differences between his model and those developed here, the underlying structure is the same.

E. *Biological species*

We conclude this very incomplete list of phenomena exhibiting the Yule distribution by mentioning the example originally analysed by Yule himself (1924). It was discovered by Willis that the number, $f(i)$, of genera of plants having i species each was distributed approximately according to (4·3), with $\rho < 1$. Yule explained these data by a probability model in which the probability, s, of a specific mutation occurring in a particular genus during a short time interval was proportional to the number of species in the genus; while the probability, r, of a generic mutation during the same interval was proportional to the number of genera. Starting at t_0 with a single genus of one species, he computed the distribution $f(i, t)$ for $t_1, t_2, \ldots$, and found the limit as $t \to \infty$. This limiting distribution corresponds to (2·13) with $\rho = r/s$. Yule observed that for $r < s$ (as required to fit the empirical data), this was not a proper distribution function, and obtained the approximate distribution for $t = T$. His procedure was equivalent to replacing the complete Beta function in (2·13) by the incomplete Beta function, taking as the upper limit of integration an appropriate function of T.

If, in the process of § II, we define k as the total number of different species and $f(i, k)$ as the number of genera with exactly i species, we see that our k is a monotonic increasing function of Yule's t (specifically, $k = e^{st}$). Making the appropriate transformation of variables, we find that Yule's assumption with respect to the rate of specific mutation corresponds to our assumption (I′) (and hence is considerably stronger than the assumption we employed in § II). Making the same transformation of variables with respect to his assumption of a constant rate of generic mutation, we find that $n_k = e^{rt}$. We can then compute $\alpha(k)$ (which will now vary with k) by taking the derivative of n_k with respect to k. We obtain

$$\alpha(k) = r e^{(r-s)t}/s. \tag{4·5}$$

If we substitute these values in equation (2·34) of case II, where we assumed slowly changing α, we find in the limit, as $t \to \infty$, $\rho = r/s$, as required. Hence, we see that the process of § II is essentially the same as the one treated by Yule.

It is interesting and a little surprising that when Yule, some twenty years after this discovery, examined the statistics of vocabulary, he did not employ this model to account for the observed distributions of word frequencies. Indeed, in his fascinating book on *The Statistical Study of Literary Vocabulary* (1944) he nowhere refers to his earlier paper on biological distributions.

V. CONCLUSION

This paper discusses a number of related stochastic processes that lead to a class of highly skewed distributions (the Yule distribution) possessing characteristic properties that distinguish them from such well-known functions as the negative binomial and Fisher's logarithmic series. In § I, the distinctive properties of the Yule distribution were described. In §§ II and III several stochastic processes were examined from which this distribution can be derived. In § IV, a number of empirical distributions that can be approximated closely by the Yule distribution were discussed, and mechanisms postulated to explain why they are determined by this particular kind of stochastic process. In the same section, the derivations of §§ II and III were compared with models previously proposed by Yule (1924) and Champernowne (1953) to account for the data on biological species and on incomes, respectively.

The probability assumptions we need for the derivations are relatively weak, and of the same order of generality as those commonly employed in deriving other distribution functions—the normal, Poisson, geometric and negative binomial. Hence, the frequency with which the Yule distribution occurs in nature—particularly in social phenomena—should occasion no great surprise. This does not imply that all occurrences of this empirical distribution are to be explained by the process discussed here. To the extent that other mechanisms can be shown also to lead to the same distribution, its common occurrence is the less surprising. Conversely, the mere fact that particular data conform to the Yule distribution and can be given a plausible interpretation in terms of the stochastic model proposed here tells little about the underlying phenomena beyond what is contained in assumptions (I) through (III).

REFERENCES

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

Darwin, J. H. (1953). Population differences between species growing according to simple birth and death processes. *Biometrika*, **40**, 370.

Davis, Harold T. (1941). *The Analysis of Economic Time Series*. Principia Press.

Feller, William (1950). *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley.

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

Hanley, Miles L. (1937). *Word Index to James Joyce's Ulysses*. University of Wisconsin Press.

Kendall, David G. (1948). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika*, **35**, 6.

Leavens, Dickson H. (1953). *Econometrica*, **21**, 630.

Mandelbrot, Benoit (1953). An informational theory of the statistical structure of language. In *Communication Theory* (ed. by Willis Jackson). Butterworths.

Thorndike, Edward L. (1937). On the number of words of any given frequency of use. *Psychol. Rec.* **1**, 399.

Titchmarsh, E. C. (1939). *The Theory of Functions*, 2nd ed. Oxford University Press.

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

Yule, G. Udny (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.