A NOTE ON FREQUENCY DISTRIBUTION IN SERIES

BY

D. D. KOSAMBI.

For a random variable that can only assume discrete values, the method of generating functions is in common use as an aid to formulating moments.[1] The probability of the value x_j being p_j, the probability generating function is defined by $\sum p_j z^{x_j}$, and moments, factorial moments, and semi-invariants are obtained as coefficients in the various power-series expansions if they exist, of the function after substitution of e^a, $1+a$, for the parameter z, the third being obtained by expanding the logarithm of the function after the first change of variable. But here, the parameter is introduced "for the sole purpose of preventing the terms from being merged together",[2] and is, so to speak, without any value of its own.

I suggest, in what follows, a similar method of generating the probabilities and the moments, but making use of the variable z as an essential part of the distribution. The difference is that one obtains formulae involving the differential calculus in place of algebra, and a connection is to be seen at once with the formal apparatus of the theory of functions of a complex variable, summability, and Tauberian theorems.

1. Let $f(z)$ be an analytic function in the neighbourhood of the origin, with real non-negative derivatives at the origin. That is, $f(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n + \cdots$, a_i real, $\geqslant 0$ for all i, and let the series converge to f for some region beyond just the origin itself, say $|z| < r$. Then, we shall say that f gives a distribution, the probability of the value n being given by $a_n z^n/f$. For the present, the variate x is assumed to take on only the values $0, 1, 2, 3, \cdots n \cdots$

The kth moment, the expectation of n^k, is given by $\frac{\theta^k f}{f}$, where θ is the operator $z\frac{d}{dz}$ and the exponent k indicates a k-fold application. The corresponding factorial moment is given by $\frac{z^k d^k f}{f\, dz^k}$. The operator θ and the occurrence of the factor $1/f$ suggest

immediately, the transformation $\phi = \log f$, $s = \log z$. Then the successive derivatives of $\phi(s)$ are the semi-invariants λ_i; $\phi' = m$, $\phi'' = \sigma^2$ etc.

By merely specifying the function, various distributions can be characterized. The Bernoulli distribution is given by $f = [1+z]^n$. In the usual terminology, $p = z/(1+z)$, $q = 1/(1+z)$ gives the probabilities for success and failure in a single trial. The mean value is $m = zf'/f = nz/(1+z)$ $1 = np$, the variance being $\sigma^2 = nz/(1+z)^2 = npq$. Here, as in general, we shall take only real non-negative values of z for statistical purposes.

For the Poisson distribution, $f = e^z$, $m = \sigma^2 = \lambda_i = z$. Distributions having various properties can be built up by solving the corresponding differential equations. But the use of $\phi(s)$ shows that a non-vanishing multiplicative arbitrary constant in f as well as in z can be absorbed, or ignored, without any essential change. Thus, $\sigma^2 = km$ leads to $f = e^{z^k}$. For $\sigma^2 = m - k$ we have $f = z^k e^z$. The first would correspond to a Poisson distribution of twins, triplets, etc. for $k = 2, 3, \ldots$; the second is essentially the generalised Poisson distribution for a variate that cannot take any value less than k. But a variate occurring in even or odd values only where m tends to equality with σ^2 as both increase would be distributed, say, according to cosh z, or sinh z.

2. The direct determination of f would be useful for some properties. The distribution of a sum of two variates distributed according to $A(z) = a_0 + a_1 z + \cdots$ and $B(t) = b_0 + b_1 t + \cdots$ would be given by the probability for the value n of the sum equal to $\sum_{i+j=n} a_i z^i b_j t^j / AB$. For two Bernoulli distributions with the same n the function is $(1+z+t+zt)^n$, corresponding to a new parameter $z+t+zt$, or to the formula for compound probability $p = p_1 + p_2 - p_1 p_2 = p_1 q_2 + p_2 q_1 + p_1 p_2$. The Poisson distribution is essentially the only one where the sum is distributed according to the same law with the new parameter a sum of the two original parameters: $f(z) f(t) = f(z+t)$.

In some observations, even for continuous distributions, the quantities ϕ and s are observed directly, not f and z. Fechner's law naturally comes to mind here. One example that suggests itself is that of photo-sensitometry[3], where the logarithm of the darkening is plotted against the logarithm of the exposure for a given emulsion and time of development. The "gamma" value usually published is merely the slope of the curve, i.e. $\phi'(s)$ for the central portion,

where the curve has a point of inflection, $\phi'' = 0$, and approximates very closely to its tangent. According to our theory, this is a mean value, and in fact, if the darkening depends in a certain way upon the number of elementary particles affected (the grains of the emulsion), this interpretation would be entirely accurate. The curve of sensitivity usually published looks uncommonly like half a Gaussian curve of normal frequency, but this cannot be verified unless very accurate measurements are made in the range of solarization; these last have been neglected, because they appear to have no practical use in photography, the useful portion of the curve being only in the neighbourhood of the point of inflection.

The transformation to s, ϕ, suggests that for a denumerable set of discrete values of the variable, the proper generalisation of our distribution is not in the form of power series with fractional exponents, but Dirichlet series for f or ϕ. In the usual notation, these would be given, substituting $-s$ for s by $\sum a_n e^{-\lambda_n s}$. The connection with the theory of numbers is obvious, and need not be considered in detail here.

3. Another obvious connection with analysis is the notion of summability given by each of our distributions. The Poisson distribution, for instance, leads to Borel summability. Let us interpret n as the number of observations, or the number attached to an observation or measurement, of some quantity s; let $s_1, s_2, s_3 \cdots$ be the successive values obtained in this way. The probability of n observations (or the nth observation) having taken place being distributed in a Poisson series, the expectation of s is $\sum e^{-z} s_n z^n/n!$ The question: when are we justified in taking the limit of this as $z \to \infty$ as the limit of s_n, amounts to the Tauberian problem: when does a sequence summable-B converge? A better question would be: when can we take the limiting expectation as limit of the arithmetic means, that is when is B-summability equivalent to (C, 1) summability? Inasmuch as, in practice, the limit can never be attained, the important question is not only the approach to a limit, but actually the behaviour and distribution in the vicinity of the limit. That is, the Tauberian idea must be generalized. In view of the fact that a connection is known between Tauberian theorems and probability [5] through the Fourier-Stieltjes integral, I shall merely point out the phenomenon here.

4. Of the infinite series given, there naturally exist two classes: those with a finite and those with an infinite radius of convergence. For the first, we can take the radius as unity, since

a factor in the parameter z makes no real difference. The second class contains polynomials and integral functions, and is more interesting in a way because the distributions can be made to tend to the normal, under certain restrictions. For a polynomial, this is a matter of adjusting n, the degree ($=\lim. zf'/f$) and the parameter z carefully, letting n tend to infinity. For integral functions of infinite order, there seems to be no way of getting the normal distribution as a limit, because m and σ^2 are not of comparable magnitudes as $z\to\infty$. When the limit σ^2/m exists, it is the order of the integral function and the distribution for positive finite order approaches the normal under fairly general restrictions on a_n for the types of functions we use. The reason is that $f(z)$ does not then differ by much, relatively, from the greatest term in its expansion, [4] and the values of $a_n z^n/f$ taper off rapidly to zero on either side. Transferring the origin to the greatest term (relatively near the mean value, which would otherwise recede infinitely far from the origin) and changing the scale, it is easy to show, (at least when there are limited gaps in the coefficients and the coefficients themselves decrease monotonically,) that the distribution approaches a continuous one obeying the differential equation $d\mathrm{F}/\mathrm{F}dx = -cx$, which characterizes our normal distributions.

Series with a unit radius of convergence might also be useful, as the Lambert series distributions again permit results of the theory of numbers to be translated bodily into statistics. But in general, it would seem impossible to get the normal distribution as a limiting case, as is shown for example by the behaviour of the Abel-summability function $f(z) = 1/(1-z)$, $p_n = z^n(1-z)$, as $z\to 1$.

5. As an illustration of all the foregoing, we take a problem usually solved by other methods, and show that its fundamentals are made simpler to the understanding by use of probability. Arrange a large number of urns in serial order, labelled 1, 2, 3...n..., and let each urn contain as many balls as there are prime factors in its label-integer. Each factor is to be counted according to its multiplicity and unity may be taken as a prime only for the first urn. Now the exact probability of there being k balls in an urn chosen at random is not known, but an approximation is obtained from an asymptotic value for the number $\pi_k(x)$ of integers $\leqslant x$ containing exactly k prime factors, $\pi_k \backsim \dfrac{x}{\log x}\dfrac{(\log\log x)^{k-1}}{(k-1)!}$

Take the value of the probability (the ratio π_k/x) obtained from this as exact. This can be transformed by taking $z = \log\log x$, $\log x = e^z$. The distribution function is seen to be ze^z; every number has at least

one prime factor, and *we have a modified Poisson distribution* as explained in section 1. The expectation of the number of balls in an urn is $\frac{\theta z e^z}{z e^z} = z+1$. Thus we should have the total number of prime factors of all integers $\leqslant x$ as $x(\log\log x+1)$. Now the known [6] formula for this is $x(\log\log x+\mathrm{B})+\mathrm{O}(x)$. The $\mathrm{O}(x)$ term is naturally due to the neglected remainder, as we have taken an approximation as the exact value, and summed to infinity instead of stopping at x terms of the series. The value of the constant B is a little less than unity, due not only to the two causes given above, but also to the fact that there is a linkage between the total number of urns, and the probability; the parameter z is not independent of x. As the value of B is known, we do not investigate this further.

REFERENCES:

1. A. C. Aitken: *Statistical Mathematics*, (Edinburgh, 1939) sections 7-10.

2. *Ibid.* p. 19.

3. J. M. Eder: *Rezepte, Tabellen u. Arbeitsvorschriften f. Photographie* (Halle, 1927) pp. 251-258.

4. E. C. Titchmarsh: *Theory of Functions* (Oxford, 1932) chap. VII, particularly p. 253.

5. N. Wiener: *The Fourier Integral* (Oxford, 1933).

6. G. H. Hardy and E. K. M. Wright *Introduction to the Theory of Numbers* (Oxford 1938) Th. 435, p. 355.