CHARACTERISTIC PROPERTIES OF SERIES DISTRIBUTIONS.

By D. D. Kosambi.

(*Received August 6 ; read October 1, 1948.*)

In what follows, I use the notation $\mu = E(x)$ = expectation (mean value) of a random variable x ; $\sigma^2 = E(x-\mu)^2$, its *variance* ; the letter p will be generally used to indicate probabilities.

1. In certain theoretical work on cosmic rays, it was found directly from unsolved differential equations that a distribution function (d.f.) should have the property $\mu = \sigma^2$. It is known that the Poisson distribution satisfies this ; the question then arose : does $\mu = \sigma^2$ suffice to exclude any other one-parameter d.f. ? The general answer is negative ; in the normal or the chi-square distribution, for example, it is immediately possible to specialise the two parameters so as to get $\mu = \sigma^2$, both being then functions of a single parameter ; the remark is valid in general for multiparametric d.f.'s. However, the problem required a discrete, not a continuously distributed stochastic variable. But even here, one can get $\mu = \sigma^2$ without the Poisson distribution. For example, let the variable x take on the values λ and 0 with probabilities p and $q = 1-p$. Then $\mu = \lambda p$, $\sigma^2 = \lambda^2 pq$, whence—excluding the trivial cases $p = 0$ and $\lambda = 0$—we still have $\mu = \sigma^2$ when $\lambda q = 1$. The physical problem required integral values 0, 1, 2, . . . for x, while $\lambda q = 1$ means $\lambda > 1$. This can be overcome by adding a unit to the values of x, which is now to take on the values $\lambda+1$, 1, with probabilities p, q. Then $\mu = \lambda p+1$, $\sigma^2 = \lambda^2 pq$, whence $\lambda p(\lambda q-1) = 1$ gives the required result. Finally, μ and σ^2 are additive when a number of independent random variables are added together, so that we can build up much more complex sets of integral values for x, with a d.f. that still has the property $\mu = \sigma^2$.

One general theorem, nevertheless, may be easily proved for the d.f.'s in question :

If a random variable takes on discrete values λ_n with probabilities proportional to powers z^{λ_n} of a single parameter z and $\mu = \sigma^2$, then the d.f. must be Poissonian.

2. First of all, consider only random variables that take on discrete integral values 0, 1, 2, . . . n . . . with probabilities proportional to 1, z, z^2, . . . z^n . . . , $z > 0$ being a real parameter of the distribution. The total probability must sum to unity, so that *the most general such distributions are necessarily represented by an analytic function of z with real non-negative coefficients :*

$$f(z) = a_0+a_1z+a_2z^2+ \ldots +a_nz^n+ \ldots ; \; a_n \geq 0 \text{ for all } n.$$

$$p(x = n) = \frac{a_nz^n}{f(z)}. \qquad \ldots \quad \ldots \quad \ldots \quad \ldots \qquad (2.1)$$

The convenience of such a distribution for our purposes, and for probability theory in general, derives from a property that can be proved without difficulty :

The kth moment of the d.f. is given by $(1/f)(z\,d/dz)^k \cdot f$, the kth factorial moment by z^kd^kf/fdz^k, and the kth semi-variant by $d^k \log f/d\,(\log z)^k$.

The kth moment is the expectation of x^k, i.e.,

$$\frac{1}{f(z)}(a_1z+2^ka_2z^2+ \cdots +a_nn^kz^n+ \cdots).$$

Differentiating $f(z)$ gives $a_1+2a_2z+\cdots$ so that $zf'(z)$ gives the mean, and it is seen at once that $z\,d/dz$ is the proper moment-building operator. For the semi-invariants, the quickest proof is by regarding the characteristic function, i.e. $E(e^{itx})$, which is

$$\frac{1}{f(z)}(a_0+a_1ze^{it}+2a_2z^2e^{2it}+\cdots) = f(ze^{it})/f(z). \quad .. \quad .. \quad (2.2)$$

As the semi-invariants are generated by derivatives to *it* of the log of the c.f., our final statement above follows immediately.

From the function-theoretic point of view, these distributions in series may be classified into two distinct types. The first consists of those with a finite radius of convergence, as for example in the ordinary geometric progression, $f(z) = 1/(1-z)$ which represents so many regular absorption phenomena, or R. A. Fisher's species[1] (and genes) distribution, $f(z) = -\log(1-z)$. The second class is of those $f(z)$ that have an infinite radius of convergence, being either rational integral functions as for the binomial (Bernoulli) distribution $f(z) = (1+z)^n$, where the parameter z is related to the usual probability p for success in a single trial by $p = z/(1+z)$, $q = 1/(1+z)$. The other class is of entire functions, of which the Poisson distribution $f(z) = e^z$ is most commonly used because of its limiting position and stability, which make it the analogue, for discrete integral-valued distributions, of the normal distribution for continuous variables.

From the point of view of probability theory, we make a different distinction:

In order to obtain a normal distribution in the limit (by shifting the origin to the mean value and taking σ as a new scale-unit), it is necessary that $f(z)$ be an entire function of finite order.

That the condition is necessary may be seen by the following considerations. For a finite radius of convergence we get a comparatively steady falling off in the values of consecutive terms. What is needed for a limiting value of the type desired is a term of high value which accounts for the greater part of $f(z)$ with rapid falling off in both directions. This excludes all except entire functions. But for our change of scale to be possible, σ^2 and μ as defined by the preceding theorem must be of comparable order of magnitude. Now it is well known that when the limit σ^2/μ exists, it is the *order* of the entire function $f(z)$, whence the restriction to a finite order. Finally, we should discuss the question of sufficiency of the condition, but this leads to the consideration of certain restrictions as to gaps in the coefficients a_n, and of possible restrictions in the way they approach zero, which remove the discussion to a level beyond that to which this note is restricted.

It is much easier to indicate one purely mathematical application of these parametric distributions in passing:

Every series distribution derived from a $f(z)$ gives a method of summability.

That is, to each member of a sequence $\{s_n\}$, we assign a *probability* of occurrence, and look for the expectation:

$$E(s) = \int s_n dxF(x, z) = \frac{\Sigma a_n s_n z^n}{f(z)} \quad . \; .. \quad .. \quad .. \quad (2.3)$$

The *sum* is the limit of this generalised mean value when z tends to the critical value ∞, or the finite positive real point on the circle of convergence which is necessarily a singularity of $f(z)$ as all the coefficients are non-negative. In either case, *the individual probabilities tend to extinction.*

There is no difficulty in generalising the Stieltjes integral of (2.3) to other types of probability distributions. For distributions with a parametric density function, we get a kernel $K(x, z)$

$$\left.\begin{array}{c}\text{(i) } K(x, z)\geq 0\,; \qquad \text{(ii) } \displaystyle\int K(x, z)dx = 1 \text{ for all } z\,; \\ \text{(iii) } K(x, z)\rightarrow 0 \text{ as } z\rightarrow\alpha, \text{ for all } x.\end{array}\right\} \quad .. \quad (2.4)$$

The limits for integration may be finite or infinite, as also the critical value $z = \alpha$. The *sum* of an integrable function $F(x)$ is then defined by lim. $\phi(z)$ as $z\rightarrow\alpha$ where

$$\phi(z) = \int F(x)K(x, z)dx. \quad .. \quad .. \quad .. \quad .. \quad (2.5)$$

It is curious that *the ordinary limit of* $\{s_n\}$ *or* $F(x)$ *appears as a singular method of summation,* for it is known that no such kernel $K(x, z)$ exists which represents the identity, i.e. transforms $\{s_n\}$ or $F(x)$ into itself. The Dirac δ-function of the physicist or some limiting process regularised by the introduction of another parameter is necessary.[2] In certain cases, a transformation is possible which simplifies the integral equations, as for example, in the rth Cesàro mean for continuous functions:

$$\phi_r(z) = \frac{r}{z^r}\int_0^z F(t)(z-t)^{r-1}\,dt \quad .. \quad .. \quad .. \quad (2.6)$$

taking $x = t/z$ transforms this into

$$\phi_r(z) = \int_0^1 F(zx)d(1-x)^r, \quad .. \quad .. \quad .. \quad (2.7)$$

so that we are using the kernel $F(zx)$ and fixed limits to evaluate the sum.

3. We now come to the proof of our original theorem, on the nature of the d.f. when $\mu = \sigma^2$. Before dealing with this, it is necessary to point out two corollaries to the foregoing section.

The function $f(z)$ may be put into the canonical form $a_0 = 1$:

$$F(z) = 1+a_1\,z+a_2z^2+\ldots \quad .. \quad .. \quad .. \quad (3.1)$$

In the first place, if $a_0\neq 0$, we can cancel it out from each individual probability $a_rz^r/f(z)$. Secondly, suppose the first non-vanishing coefficient is a^k. This means that x does not take on the values $0, 1, 2, \ldots k-1$ at all. We then take the new random variable $x' = x-k$ and the $f(z)$ that describes its distribution has the coefficient $a_0'\neq 0$. Finally, *if $a_1\neq 0$ we may take a new parameter a_1z and the canonical form becomes*

$$F(z) = 1+z+b_2z^2+\ldots \quad .. \quad .. \quad .. \quad (3.2)$$

The second remark is that

The entire discussion for series distributions may be carried over to the case where x takes on discrete real values $\lambda_1, \lambda_2, \ldots \lambda_n \ldots$, the function $f(z)$ being then represented by a Dirichlet series with real non-negative coefficients and exponents $\{\lambda_n\}$.

$$f(z) = a_0+a_1e^{\lambda_1 s}+a_2e^{\lambda_2 s}+\cdots;\ a_k\geq 0. \quad .. \quad .. \quad (3.3)$$

The formal transformation $s = \log z$ is fully justified under the circumstances and shows the equivalence of the two cases, without further argument.

We have, from section 2,

$$\mu = \phi'(s)\,;\ \sigma^2 = \phi''(s), \text{ where } \phi = \log f,\ s = \log z. \quad .. \quad (3.4)$$

$$\therefore\ \mu = \sigma^2\rightarrow\phi''(s)-\phi'(s) = 0\rightarrow\phi = \alpha e^s+\beta\rightarrow f = ae^{bz}.$$

In the canonical form, we have $f = e^z$ which is the Poisson distribution. Moreover, as 3.4 is valid for an $f(z)$ defined by a Dirichlet series, we see that the random variable x must necessarily take on the discrete values 0, 1, 2, ... n ... inasmuch as the solution $f = ae^{bz}$ is unique.

4. In actual practice, the most important problems are the estimation of z, for a given distribution, from an observed random sample. A second problem is the choice of a proper $f(z)$ to fit given data, but this can obviously not be solved without further argument of some sort. To the former question, a simple theorem gives the answer. We restrict ourselves to the case of integral-valued distributions, which represents the general facts as well, as has been pointed out before. Let the sample frequencies in categories of value 0, 1, 2, .. n .. be $b_0, b_1, \ldots b_n \ldots$; further, indicate by $N = \Sigma b_k$ the total number, and by $m = \Sigma k b_k$ the observed sample mean. Then:

The maximal likelihood estimate of z from an observed sample is given for all distributions in series by equating the observed and theoretical means, $m = \mu = zf'/f$; the sampling variance of this estimate in large samples is given by $V(z) = z^2/N\sigma^2$.

The 'likelihood' is the compound probability[3] of getting just the sample observed, its logarithm being

$$\sum b_k \{\log a_k + k \log z - \log f(z)\} = \sum b_k \log a_k + N\{m \log z - \log f(z)\}. \qquad (4.1)$$

Differentiation gives $N(m/z - f'/f)$ which, equated to zero, gives our first result at once. It is known that by a second differentiation to z and substitution of the expected values for b_k we get $-1/V(z)$. Here, we have to differentiate $N(m - zf'/f)/z$, and then replace m by μ. This gives the second result of the theorem.

In some cases, however, the distribution is truncated by the very nature of the experiment, and the above method of estimation leads to rather tedious calculations. To take one important case, that of steady absorption phenomena in discrete steps, consider the observed frequencies b_k as occurring with an expected distribution in geometric progression $p_k = z^k/f(z)$, but breaking off at the class of n, the total number of terms being $n+1$. Here, we have

$$f(z) = 1 + z + z^2 + \ldots + z^n = \frac{1 - z^{n+1}}{1 - z}. \quad \ldots \quad \ldots \quad (4.2)$$

The maximal likelihood estimate of z means solving

$$m = \mu = \frac{z + 2z^2 + \ldots + nz^n}{1 + z + \ldots + zn} = \frac{z}{1 - z} - \frac{(n+1)z^{n+1}}{1 - z^{n+1}} \quad \ldots \quad (4.3)$$

while

$$\sigma^2 = \frac{z}{1 - z^2} - \frac{(n+1)^2 z^{n+1}}{(1 - z^{n+1})^2}. \quad \ldots \quad \ldots \quad \ldots \quad (4.4)$$

Even for moderate values of n, the solution of the resulting algebraic equation of degree n for its single positive root involves a good deal of labour. A much quicker method, obviously, is to fit a linear regression to the logarithms of the observations, i.e. to $\log b_0, \log b_1, \ldots \log b_n$. We are interested only in the slope of the line, which should, were the sample exactly in accordance with theory, be $\log z$. Our estimate (essentially a geometric mean), therefore, is

$$\log z = \frac{1}{\nu} \sum \left(k - \frac{n}{2}\right) \log b_k, \text{ where } \nu = \frac{n(n+1)(n+2)}{12}. \quad \ldots \quad (4.5)$$

The statistic is seen to be consistent, but the question is of its efficiency, that is the size of its large-sample variance as compared to that of the previous estimate. As the total number N does not enter explicitly, the sampling variance in question is to be obtained by summing $p_k(\partial z/\partial b_k)^2$, with the substitution of Np_k for b_k in the result. This leads to

$$V(z) = \frac{z^2}{\nu^2 N^2} f(z) \sum \left(k-\frac{n}{2}\right)^2 z^{-k}. \qquad .. \qquad .. \qquad (4.6)$$

Recalling the specific form of the $f(z)$ to which we have restricted ourselves here, a little manipulation gives

$$V(z) = \frac{z^{2-n} f^2(z)}{N\nu^2} \left\{ \sigma^2 + \left(\mu - \frac{n}{2}\right)^2 \right\}. \qquad .. \qquad .. \qquad (4.7)$$

The method is efficient only when n is small as compared with μ and σ^2 for z being less than unity in absorption phenomena, the factor z^{-n} increases $V(z)$ enormously in comparison with the maximal likelihood variance. In most experimental work, the smallest frequencies observed are still large, so that the method of fitting a logarithmic regression gives a reasonable estimate of the absorption coefficient. For further refinements, it could always be used as a first approximation in the equation 4.3.

Events of the type discussed in the foregoing are generally regarded as having been compounded from many small, independent random effects. If a single elementary even be equated to a variate assuming values 0, 1 with probabilities q, p, its characteristic function will be $q+p$. exp. i . t. Following the classical procedure, we obtain the binomial distribution with c.f. $(q+e^{it})^n$ for repeated trials with the same probability. If $n \rightarrow \infty$ and we do not change scale, a limit exists if and only if $n . p \rightarrow z$, which leads to the Poisson distribution with z as a parameter. For independent events that differ in probability, the c.f. is

$$\Pi(q_r + pre^{it}) = \Pi q_r \Pi \left(1 + \frac{p_r}{q_r} e^{it}\right). \qquad .. \qquad .. \qquad .. \qquad (4.8)$$

This product converges if and only if $p_1+p_2+ \ldots +p_n+ \ldots.$ converges. Therefore *if the characteristic function of the compound event be expressible as an analytic function of a single parameter, the distribution must necessarily be in series.* For, putting $z = p_1 + .. + p_n + \ldots$, the parameter can only be z exp. i. t, except for a constant factor.

This justifies to some extent the assumption that the distributions with which physicists are concerned are in series. Other procedures are also possible. For example, in analysing the expansion of 4.8 it is seen that $f(z)$ corresponds to the reciprocal of Πq_r, and the coefficients are built up out of symmetric functions of p_r/q_r. A single term $p/q = p/(1-p) = p+p^2+ \ldots p^n + \ldots$ can be interpreted as made up of probabilities of the event occurring once, twice, thrice, etc., and summing these implies that no matter how often the event is repeated the value assigned to the event is always unity. That is, we have in effect adopted a sort of Boolean algebra $x^2 = x$, equivalence here meaning the value-category to which the event is to be ascribed. But it would have been as simple to assign p^2 to the value 2, i.e. to the coefficient of exp. z it, p^3 to 3, etc. In that case, we should have been led merely to the series distribution in geometric progression. In the general case, it is easily seen that, taking $z = p_1 + \ldots + p_n + \ldots$, the successive coefficients are always closely approximated by $a_2 z^2$, $a_3 z^3 \ldots a_n z^n \ldots$ and this provides sufficient justification for assuming a series distribution.

References.

[1] R. A. Fisher: Genetical Theory of Natural Selection (Oxford, 1930), 90.

[2] But in this connection, see the recent work of Laurent Schwartz: *Annales dés Telecommunications*, 1948, III, 135–140.

[3] R. A. Fisher: Statistical Methods for Research Workers, §53, §55.