REPRESENTATION OF PROBABILITY DISTRIBUTIONS BY CHARLIER SERIES*

By R. P. Boas, Jr.

Brown University

Summary. The paper describes some results concerning the representation of a function by linear combinations of the successive differences of the Poisson distribution, not necessarily the partial sums of the type B series of Charlier.

1. Introduction. For various purposes it is often desired to expand a probability distribution $f(x)$ in a series

(1) $$f(x) \sim \sum_{k=0}^{\infty} c_k \theta_k(x),$$

where the $\theta_k(x)$ are a given set of standard functions. Arguments of a heuristic nature led Charlier [4, 5, 6] to suggest that it would be useful to take the $\theta_k(x)$ in (1) to be either the successive derivatives or the successive differences of some fixed function; the two cases are often referred to as type A series and type B series, respectively. Charlier gave formulas for determining the coefficients in the two cases, but the question of whether the formal series represents the given function in any reasonable sense has to be investigated separately for each particular choice of the function generating the series. Only one special case of each type has been much used: for the A-series, $\theta_0(x)$ is the normal density function $(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2}$; for the B-series, $\theta_0(x)$ is the Poisson function $e^{-\lambda}\lambda^x/x!$ (when x is restricted to take only nonnegative integral values). We shall refer only to these special cases when we speak of A- and B-series in this paper.

There are two distinct problems (which have, however, often been confused) connected with the representation of a function $f(x)$ by a series (1); for convenience, we shall refer to them in this paper as the practical problem and the theoretical problem. In the *practical problem*, we have an empirical function $f(x)$, defined only for a finite number of values of x, which we suspect is representable by $c_0\theta_0(x)$ together with a small correction, so that we hope that a few (say three or four) terms of (1) may give a good representation of $f(x)$ in a relatively simple analytical form with a reasonable amount of computational labor. In some cases, and certainly with the classical A- and B-series which we are considering, we could represent, as closely as desired, any $f(x)$ (however irregular) which takes nonzero values at only a finite number of points; but there is no interest in doing this if the process involves finding too many terms of the series. (Neglect of this fact has led to ill-founded statements by mathematicians about the satisfactory nature of the A- or B-series; but see [27, pp. 38–39].)

Thus it would be of interest to know, if possible, under what circumstances a given empirical density can be represented fairly well by a few terms of a series of a given kind. If no simple criterion can be given, it is desirable to have a means

* Address delivered by invitation at the meeting of the Institute at Boulder, Colorado, on September 1, 1949.

of computing coefficients which will make a few terms of (1) give the best possible fit—best possible being defined in a way appropriate for the problem at hand.

In the *theoretical problem*, $f(x)$ is a function defined for all values of x, or at least for all of an infinite set of equally spaced values of x, arising from theoretical considerations which suggest $c_0\theta_0(x)$ as a reasonable first approximation to $f(x)$. For example, the central limit theorem states that under certain conditions the cumulative distribution function of the sum of a large number of independent random variables is approximately normal; then we might expect that this distribution function would be representable by a series (1) with $\theta_0(x)$ the normal distribution function. For such theoretical purposes we should like to have criteria for the representability of a sufficiently general $f(x)$ by a series (1), where representability is of course to be interpreted appropriately, as ordinary convergence, uniform convergence, convergence in mean square, asymptotic representation, etc., according to the requirements of the problem at hand. The larger the class of $f(x)$ for which we can prove a representation theorem, the larger is the possible domain of applicability of the series to theoretical problems.

2. The A-series. This paper is concerned with the B-series, but for comparison we first mention some properties of the A-series. In the case of the classical A-series, we have the attractive fact that the functions $\theta_n(x)$ are orthogonal with weight function $e^{\frac{1}{2}x^2}$, that is,

$$\int_{-\infty}^{\infty} \theta_n(x)\theta_m(x)e^{\frac{1}{2}x^2}\,dx = 0, \qquad m \neq n.$$

In fact, $e^{\frac{1}{2}x^2}\theta_n(x)$ is, except for a numerical factor, the nth Hermite polynomial. This orthogonality property enables one to compute the coefficients in a series (1) with great ease from

$$n!\,c_n = \int_{-\infty}^{\infty} f(x)\theta_n(x)e^{\frac{1}{2}x^2}\,dx, \tag{2}$$

or since $\theta_n(x)e^{\frac{1}{2}x^2}$ is a polynomial, from the moments of $f(x)$. By the classical theory of orthogonal functions, this means that if the c_n are so computed, and we take $N + 1$ terms of the series, we minimize

$$\int_{-\infty}^{\infty} e^{\frac{1}{2}x^2}[f(x) - F_N(x)]^2\,dx \tag{3}$$

for all possible sums

$$F_N(x) = \sum_{n=0}^{N} c_n\theta_n(x). \tag{4}$$

The convergence theory of Hermite series has been thoroughly investigated by mathematicians, so that it would appear that in theoretical problems, in which $f(x)$ is given for all values of x, we are in a position to find out everything about the representation of $f(x)$ by an A-series. Also in problems of practical curve-fitting, the fact that the closest approximation to $f(x)$ (in the sense (3)) by sums of the form (4) is given by choosing the coefficients according to (2) seems to leave no more to be said.

However, the formal elegance of the A-series seems to be somewhat misleading. Even when a series converges it by no means follows that its Nth partial sum is the best selection of N terms for representing a given function. Even though the partial sums do give the best fit in the sense of (3), it may not be desirable to measure the closeness of approximation by (3); some other measure of approximation may be better suited to the end in view. For example, it is known that the partial sums of Edgeworth's series (see [8]), which is a rearrangement of the A-series, are more satisfactory for some purposes than the partial sums of the A-series with the coefficients determined by (2). More precisely, Edgeworth's series furnishes an asymptotic expansion, with a remainder term whose order of magnitude can be estimated quite precisely, in circumstances where the series of orthogonal functions does not do this. Again, for practical purposes a few terms of the A-series sometimes exhibit undesirable properties (such as negative frequencies). If $f(x)$ is a function defined only for integral values of x, A. Fisher [10] has suggested and applied the idea of minimizing, not (3), but the sum $\sum_{-\infty}^{\infty} |f(x) - F_n(x)|^2$ in order to determine the coefficients of the approximating sums.

3. The B-series. We can now see how the status of the B-series resembles or differs from that of the A-series. Here we deal principally with a function defined for integral values of x; $\theta_0(x) = \theta(x) = e^{-\lambda}\lambda^x/x!$, $\Delta\theta(x) = \theta(x) - \theta(x-1)$, $\Delta^k\theta(x) = \Delta(\Delta^{k-1}\theta(x))$ and $\theta_k(x) = \Delta^k\theta(x)$; $\theta(x)$ is taken to be 0 for negative integral x. We shall refer to this as the *discrete case* of the B-series. The literature of the subject contains a number of rather painful attempts to put the coefficients into usable form, persisting even after the simple formula

$$c_n = (1/n!) \sum_{l=0}^{n} \binom{n}{l} (-1)^l \lambda^{n-l} \mu_l \tag{5}$$

had been obtained, where μ_n is the nth factorial moment,

$$\mu_n = \sum_{k=n}^{\infty} f(k)k!/(k-n)!.$$

Formula (5) can be derived, for example, by using orthogonality properties of the $\theta_r(x)$. We have, in fact, that $\sum_{x=0}^{\infty} \theta_n(x)\theta_m(x)/\theta_0(x)$ is 0 or $n!\,\lambda^{-n}$ according as $n \neq m$ or $n = m$.

The parameter λ in the B-series is at our disposal, and can for example be chosen in such a way as to improve the convergence of the series. For purposes of practical curve-fitting, it has been customary to choose λ equal to the mean of the distribution $f(x)$, a choice which makes the coefficient c_1 of $\Delta\theta$ equal to zero. Charlier also suggested other methods in which c_1 and c_2, or c_1, c_2 and c_3 are zero [7]. Such choices, of course, may reduce the amount of computation needed to make use of a given number of differences in fitting a curve; aside from this consideration their use seems to depend on the belief that one improves the convergence of a series by adjusting any available parameters so that as many as possible of the initial terms of the series are zero. This belief does not always seem to be confirmed by the facts. (In particular, compare columns 2 and 5 of Table 1, columns 2 and 4 of Table 2, or columns 2 and 4 of Table 3.)

The theoretical problem of what $f(x)$ can be represented by convergent B-series has been studied by several authors [12, 13, 17, 19, 20, 21, 23, 24, 26, 28]; the study by Schmidt [24; see also 25 and 17] gives necessary and sufficient conditions for the representation in the case of a nonnegative $f(x)$, so that, at least in all cases of interest in statistics, the theoretical problem seems to be completely solved. However, one of the purposes of the present paper is to reopen this apparently closed problem.

There is also a *continuous* version of the B-series, which is suggested by the fact that

$$\theta(x) = (2\pi)^{-1} e^{-\lambda} \int_{-\pi}^{\pi} e^{-ixu} \exp (\lambda e^{iu})\, du \tag{6}$$

reduces to the Poisson function $e^{-\lambda}\lambda^x/x!$ for positive integral x (and to 0 for negative integral x). This form of the B-series has not been much used, and its use is subject to suspicion since it has rather peculiar properties. In particular, it cannot represent, in any reasonable sense, a positive function $f(x)$ or one which is too small as $x \to \infty$ [26, 3]; since the functions which present themselves for representation in practice are both positive and small at infinity, the continuous case of the B-series looks unpromising for applications. (See also [27a], 1a.) However, it has been applied [15].

The purpose of this paper is to describe some results on the B-series which have been obtained in a mathematical paper [3], devoted to what we have called the theoretical problem; some contributions to the practical problem will also be given in the present paper. The starting point of this investigation was the question of what happens if one tries to approximate a function, not by the partial sums of the series (1), but by some other combination of the first N functions $\theta_n(x)$, when approximation is taken in the sense of (unweighted) least-squares. This method of approximation seems well adapted to statistical problems, and leads to simpler mathematical work than ordinary point-by-point convergence of the partial sums. The B-series itself gives a least squares approximation with a weight function $1/\theta_0(x)$. We consider here only the classical B-series, when $\theta_0(x) = \theta(x) = e^{-\lambda}\lambda^x/x!$, $\theta_n(x) = \Delta^n\theta_0(x)$; the main results are substantially the same for rather more general cases [3; see also 14, 25]. In addition, here we consider only nonnegative $f(x)$, assumed zero for negative x. Functions which need not be zero for negative x are handled easily by generalizing the B-series to the form [3]

$$f(x) \sim \sum_{-\infty}^{-1} b_n \nabla^n \theta(x) + \sum_{0}^{\infty} a_n \Delta^n \theta(x), \tag{7}$$

where ∇ denotes the advancing difference: $\nabla\theta(x) = \theta(x) - \theta(x+1)$; there seems to be no particular reason (other than a historical one) for preferring one kind of difference to the other. The generalized series (7) might be useful for graduating symmetrical probability distributions, although it does not seem to have been considered in the literature (cf. [1a]).

4. Results: practical problem. Our question takes somewhat different forms in the two cases which we have described as the practical and the theoretical. In the former, we ask what the coefficients $a_n^{(N)}$ should be so that

$$(8) \qquad \sum_{x=0}^{\infty} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2$$

shall be a minimum, where $f(x)$ is an empirically given function and N is a given integer, in general not very large. If N is 0, 1 or 2, that is, if we use 1, 2 or 3 terms, the best choice of the $a_k^{(N)}$ in (8) can be calculated without difficulty.

For $N = 0$, our question is that of finding the best least-squares fit to $f(x)$ by a Poisson distribution $a_0^{(0)} e^{-\lambda} \lambda^x / x!$; the best choice of $a_0^{(0)}$ is then

$$(9) \qquad a_0^{(0)} = \left\{ e^{\lambda} \sum_{x=0}^{\infty} f(x) \lambda^x / x! \right\} \Big/ J_0(2i\lambda),$$

where

$$J_0(iy) = 1 + y^2/(2!)^2 + y^4/(4!)^2 + \cdots$$

(J_0 denotes the Bessel function of order 0); on the other hand, the usual formula (5) gives the different coefficient

$$c_0 = \mu_0 = \sum_{x=0}^{\infty} f(x).$$

This, of course, is simpler than (9) to compute, although its use is based on the uncritical assumption that the first term of the series (1) is the best one to take if only one term is to be used. Charlier [7; see also 10, pp. 101–103] suggested a different formula in which one uses, not $\Delta^k \theta(x)$, but $\Delta^k \theta(px + q)$, the parameters p, q, λ being adjusted to make the terms of (1) in $\Delta\theta$, $\Delta^2\theta$, $\Delta^3\theta$ all zero; here $\theta(x)$ is defined when x is not an integer by interpreting $e^{-\lambda}\lambda^x/x!$ as $e^{-\lambda}\lambda^x/\Gamma(x+1)$, and *not* by using formula (6). Table 2 shows that in at least one numerical case (9) gives a better least-squares fit than Charlier's method (and without introducing gamma functions to take care of $\theta(x)$ for fractional x). However, it is not excluded that Charlier's method will give better results in other cases, since with the change of the functions $\theta_n(x)$ the results of this paper cease to apply.

For $N = 1$, we get the best least-squares approximation to $f(x)$ by

$$a_0^{(1)} \theta(x) + a_1^{(1)} \Delta\theta(x)$$

if

$$(10)\qquad \begin{aligned} a_0^{(1)} &= \frac{e^{2\lambda}}{\alpha+\beta}\left(\textstyle\sum_0 + \sum_1\right),\\ a_1^{(1)} &= \left\{\frac{\beta}{\alpha^2-\beta^2}\textstyle\sum_0 - \frac{\alpha}{\alpha^2-\beta^2}\sum_1\right\} e^{2\lambda}, \end{aligned}$$

where $\sum_0 = \sum_{x=0}^{\infty} f(x)\theta(x)$, $\sum_1 = \sum_{x=0}^{\infty} f(x)\theta(x-1)$, $\alpha = J_0(2i\lambda)$, $\beta = -iJ_1(2i\lambda)$, the J's again denoting Bessel functions. For $N = 2$, the corresponding formulas involve also $\gamma = -J_2(2i\lambda)$ and $\sum_2 = \sum_{x=0}^{\infty} f(x)\theta(x-2)$. They are:

$$(11)\qquad \begin{aligned} e^{-2\lambda} a_0^{(2)} &= \frac{\beta-\alpha}{2\beta^2-\alpha^2-\alpha\gamma}\textstyle\sum_0 + \frac{2\beta-\alpha-\gamma}{2\beta^2-\alpha^2-\alpha\gamma}\sum_1 + \frac{\beta-\alpha}{2\beta^2-\alpha^2-\alpha\gamma}\sum_2,\\ e^{-2\lambda} a_1^{(2)} &= \frac{\beta\gamma-\alpha\beta+2\beta^2-2\alpha\gamma}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\textstyle\sum_0 + \frac{\alpha+\gamma-2\beta}{2\beta^2-\alpha^2-\alpha\gamma}\sum_1\\ &\qquad + \frac{2\alpha^2-2\beta^2+\beta\gamma-\alpha\beta}{2\beta^2-\alpha^2-\alpha\gamma}\textstyle\sum_2,\\ e^{-2\lambda} a_2^{(2)} &= \frac{\alpha\gamma-\beta^2}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\textstyle\sum_0 + \frac{\beta}{2\beta^2-\alpha^2-\alpha\gamma}\sum_1\\ &\qquad + \frac{\beta^2-\alpha^2}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\textstyle\sum_2. \end{aligned}$$

The functions $i^n J_n(iy)$ are real for real y, and extensive tables are available [32].

Some numerical examples showing the comparison between graduation by these formulas and by the corresponding number of terms of the B-series are given in Tables 1–3. It will be noticed that (as the theory indicates) one gets a better least-squares fit by formulas (9), (10) or (11) than by a corresponding number of terms of the B-series using the coefficients (5). However, one may not get a better fit if goodness of fit is measured in some other way, e.g. by χ^2. Unfortunately the coefficients calculated by this method increase rapidly in complexity as the number of terms increases, and even the coefficients for $N = 3$ would involve very heavy algebra. Since numerical examples [2] indicate that it is often necessary to go to terms in $\Delta^4\theta$ for a satisfactory fit, it might be worth while to calculate the next few coefficients.

5. Results: theoretical problem. In the case of a theoretical distribution we ask how coefficients should be determined so that

$$(12)\qquad \sum_{x=0}^{\infty}\left| f(x) - \sum_{k=0}^{N} a_k^{(N)}\Delta^k\theta(x)\right|^2$$

will tend to 0 as $N \to \infty$. The convergence to 0 of (12) is a rather strong kind of convergence, since it implies convergence of the approximating sums to $f(x)$, not only for each x, but even uniformly for all x. Of course, the "best" choice of

$a_k^{(N)}$ as above would be expected to give convergence under the weakest hypotheses, but because of the complexity of these coefficients it seems desirable to make (12) only approximately a minimum; this actually makes no difference in the limit, although the approximation is not usually satisfactory for small values of N. To see the connection between the formulas used here and the "classical" formula (5) for the coefficients in (1), we note that (5) can be written

$$a_n = \frac{(-1)^n}{n!}\sum_{k=0}^{\infty} f(k) \frac{d^n}{dz^n} [z^k e^{\lambda(1-z)}]_{z=1}; \tag{13}$$

(5) results if we expand the derivative by Leibniz's rule and rearrange the sum. If we expand $e^{-\lambda z}$ in a power series before differentiating in (13), we obtain

$$a_n = (-1)^n \sum_{k=0}^{\infty} f(k) \sum_{l=\max(k,n)}^{\infty} \binom{l}{n} e^{\lambda}(-\lambda)^{l-k}/l! = e^{\lambda}(-1)^n \sum_{l=n}^{\infty}\binom{l}{n}\sum_{k=0}^{l}\frac{(-\lambda)^{l-k}}{(l-k)!}f(k).$$

If now we break this series off at $n = N$ to obtain

$$a_n^{(N)} = e^{\lambda}(-1)^n \sum_{l=n}^{N}\binom{l}{n}\sum_{k=0}^{l}\frac{(-\lambda)^{l-k}}{(l-k)!} f(k), \tag{14}$$

we obtain a sequence of approximations to $f(x)$ by sums $\sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x)$ which has, in general, much better convergence properties than the partial sums of the B-series with coefficients a_n given by (5). In particular, if $f(x) = 0$ for $x = -1, -2, \cdots$, this sequence of approximations converges to $f(x)$ whenever $\sum_{x=0}^{\infty} |f(x)|^2$ converges; on the other hand, for nonnegative $f(x)$ it is known [24] that the B-series converges if and only if $\lim_{x\to\infty} f(x)2^x x^k = 0$ for $k = 0, 1, 2, \cdots$, a much more restrictive condition. If we demand that the partial sums of the B-series converge in mean square, that is, that (12) tends to zero with $a_k^{(N)}$ independent of N, we have the even more restrictive condition [3] that $\limsup_{x\to\infty} \{f(x)\}^{1/x} \leqq \frac{1}{3}$.

The approximating sums with coefficients (14) have the additional property that they reproduce $f(x)$ exactly for $x = 0, 1, 2, \cdots, N$. One would expect that in general they would then tend to deviate rather widely from $f(x)$ for larger x, and so would not be satisfactory for practical curve-fitting. However, it seems possible that if we fit such a sum not to $f(x)$, but to $f(px + q)$, with suitable integers p and q, thus making the approximation agree with $f(x)$ at a set of values covering the whole range of definition of $f(x)$, it might give a satisfactory fit elsewhere. This possibility has not been investigated; a similar approach using the partial sums of the B-series was suggested by Charlier [7] and Fisher [10].

6. The continuous case of the B-series. In the continuous case we again ask, not when

$$f(x) = \sum_{n=0}^{\infty} a_n \Delta^n \theta(x) \tag{15}$$

with uniform convergence in every finite interval, but when

$$f(x) = \underset{N\to\infty}{\text{l.i.m.}} \sum_{n=0}^{N} a_n^{(N)} \Delta^n \theta(x), \tag{16}$$

which means that

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{n=0}^{N} a_n^{(N)} \Delta^n \theta(x) \right|^2 dx = 0. \tag{17}$$

For (15) the following negative results are known [26]: if $f(x) \geqq 0$, (15) cannot converge uniformly on every finite interval (unless $f(x) \equiv 0$); the series, if convergent uniformly on every finite interval, cannot converge to $f(x)$ unless the Fourier transform of $f(x)$ vanishes outside $(-\pi, \pi)$, a condition which

TABLE 1

Number of petals on buttercups. $\lambda = .631$

x	1 Observed frequency	2 Calculated 3 terms (formula 5)	3 Calculated 1 term (formula 9)	4 Calculated 2 terms (formula 10)	5 Calculated 3 terms (formula 11)	6 Calculated 3 terms (formula 14)
5	133	134.9	119.9	130.6	132.9	133.0
6	55	51.6	75.6	62.3	55.3	55.0
7	23	22.5	22.5	13.3	22.1	23.0
8	7	9.5	5.0	1.5	8.5	9.1
9	2	2.9	0.8	0.0	2.4	2.6
10	2	0.6	0.1	0.0	0.5	0.5
Total......	222	222.0	223.9	207.7	221.7	223.2

automatically excludes any $f(x)$ which vanishes for all large $|x|$ or even is too small as $x \to \infty$. Nevertheless, Jørgensen [15] applies the continuous case successfully to practical problems. A possible explanation of this apparent discrepancy is that if the $a_n^{(N)}$ in (16) are properly determined, (16) will be true under fairly general conditions. To be sure, the mean square difference in (17) cannot be made arbitrarily small unless the Fourier transform $g(x)$ of $f(x)$ vanishes outside $(-\pi, \pi)$, but if $|f(x)|^2$ is integrable the difference can be made small if $g(x)$ is itself small. If $g(x)$ does vanish outside $(-\pi, \pi)$, then (16) is true; and in fact the coefficients $a_k^{(N)}$ can be taken the same as in (14), so that the approximating sums depend only on the values of $f(x)$ for integral values of x; these values are known to determine $f(x)$ under our hypotheses on $g(x)$.

7. Discussion of some numerical results. Table 1. Column 2 gives the fit by two terms of the B-series (really three, since the coefficient of $\Delta\theta$ is zero when formula (5) is used), as calculated by Charlier [7] (that is, using terms through $\Delta^2\theta$). Column 3 gives the best least-squares fit by a single term, i.e., a Poisson distribution, calculated by formula (9); it is clear that this term alone does not represent the observations very well. Column 4 gives the best least-squares fit by terms through $\Delta\theta$. Column 5 gives the best least-squares fit by terms through $\Delta^2\theta$; the improvement over Charlier's fit by the same number of terms is evident by inspection. Column 6 gives, for comparison, the same number of terms calculated by formula (14), which gives an approximation to the best least-squares fit and necessarily reproduces the data exactly for the first three

TABLE 2

Failure of grains of barley. $\lambda = 2.757$

x	1 Observed frequency	2 Calculated 4 terms (Charlier)	3 Calculated 1 term (Formula 9)	4 Calculated 2 terms (Formula 10)	5 Calculated 3 terms (Formula 11)
0	53	63	47.3	49.9	48.4
1	131	139	130.4	134.7	133.4
2	180	174	179.8	181.6	182.3
3	170	151	165.3	163.2	164.3
4	111	111	113.9	110.0	109.8
5	50	60	62.7	59.3	58.1
6	22	32	28.8	26.5	25.2
7	22	14	11.4	10.2	9.3
8	7	6	3.9	3.4	2.9
9	2	2	1.1	1.0	0.8
10	1	0	0.3	0.2	0.2
Total......	749	752	744.9	740.0	734.7

values of x. The fact that (14) gives good results here is presumably connected with the small size of λ.

Table 2. Column 2 gives the values calculated by Charlier [7] for a fit after the linear transformation $x \to px + q$, with λ, p and q chosen to make the terms in $\Delta\theta$, $\Delta^2\theta$, $\Delta^3\theta$ all zero (the values were read to the nearest integer from Charlier's graph). Column 3 gives the best least-squares single-term fit calculated by formula (9); this is a considerable improvement for $x \leqq 6$, but for the remainder of the table it is rather poor. Column 4 gives the best least-squares fit by two terms; column 5, that by three. The χ^2-test indicates that the graduation is rather poor in all cases.

Table 3. Column 2 gives the classical calculation with terms through $\Delta^2\theta$; this was given by A. Fisher [10] and (more accurately) by Aroian [2]. Columns 3

and 4 give the best least-squares approximations by two and three terms; column 4 is better than column 2, in this sense, as expected. However, column 4 is a poorer fit when tested by χ^2, chiefly because of the poor fit at $x = 0$. It should be noted that two more terms of the B-series give a more satisfactory fit [2].

TABLE 3

α-particles from a bar of polonium. $\lambda = 3.87155$

x	1 Observed frequency	2 Calculated 3 terms (formula 5)	3 Calculated 2 terms (formula 10)	4 Calculated 3 terms (formula 11)
0	57	49.5	51.3	45.2
1	203	201.3	213.3	190.9
2	383	403.4	399.0	393.5
3	525	532.3	524.8	529.8
4	532	520.6	517.2	525.4
5	408	402.6	407.7	409.7
6	273	254.8	267.7	261.9
7	139	137.1	150.6	141.1
8	45	64.0	74.1	65.3
9	27	26.1	32.4	26.3
10	10	9.4	12.8	9.3
11	4	3.0	4.6	2.9
12	0	0.9	1.5	0.8
13	1	0.2	0.5	0.2
14	1	0.0	0.1	0.0
Total.......	2608	2605.2	2657.6	2602.3
		$\chi^2 = 10.2$ $n = 7$	$\chi^2 = 16.2$ $n = 8$	$\chi^2 = 11.4$ $n = 7$

8. Proofs: theoretical problem. We now outline the proofs of the results which we have stated. They depend on the fact that the numbers $\theta(x)$ $(x = 0, \pm 1, \pm 2, \cdots)$ (where $\theta(x) = 0$ when x is a negative integer) are the Fourier coefficients of the function $\varphi(u) = e^{-\lambda} \exp(\lambda e^{iu})$, i.e.

$$\theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u) e^{-ixu}\, du, \qquad x = 0, \pm 1, \pm 2, \cdots$$

Furthermore,

$$\Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)(1 - e^{iu})^k e^{-ixu}\, du.$$

If we then assume the condition $\sum_{-\infty}^{\infty} | f(x) |^2 < \infty$, with $f(x) = 0$ for $x = -1, -2, \cdots$, the numbers $f(x)$ are the Fourier coefficients of a function $g(x)$ of integrable square, by the Riesz-Fischer theorem from the theory of Fourier series [31, p. 74]:

$$f(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} g(u) e^{-ixu}\, du, \qquad x = 0, \pm 1, \pm 2, \cdots.$$

Thus

$$(18) \quad f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{-ixu} \left[g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{iu})^k \right] du,$$

and so the expressions on the left appear as the Fourier coefficients of the expressions in square brackets on the right. By Parseval's theorem for Fourier series [31, p. 76], then, we have

$$(19) \quad \sum_{x=-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2 = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{iu})^k \right|^2 du.$$

Thus we have reduced the problem of minimizing the mean-square difference on the left of (19) to that of minimizing the integral on the right of (19). By rearranging the sum in the integrand, we see that an equivalent problem is to minimize

$$(20) \qquad D = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u) - \varphi(u) \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du,$$

where the $c_k^{(N)}$ and $a_k^{(N)}$ are readily expressed in terms of each other; in fact,

$$(21) \qquad a_k^{(N)} = (-1)^k \sum_{l=k}^{N} \binom{l}{k} c_l^{(N)}.$$

Since $| \varphi(u) | = e^{-\lambda + \lambda \cos u} \geqq e^{-2\lambda} > 0$, we can write D in the form

$$D = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 | \varphi(u) |^2\, du,$$

so that

$$(22) \quad \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du \geqq 2\pi D \geqq e^{-4\lambda} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du,$$

since $e^{-2\lambda} \leqq |\varphi(u)| \leqq 1$. Thus we can make D arbitrarily small if and only if we can make

$$(23) \qquad D^* = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du$$

arbitrarily small. Now the Fourier coefficients of $g(u)$ are $f(x)$; those of $1/\varphi(u)$ are $e^{\lambda}(-\lambda)^x/x!$ for $x \geqq 0$, 0 for $x < 0$; by the convolution theorem for Fourier coefficients [31, p. 90] the nth Fourier coefficient of $g(u)/\varphi(u)$ is

$$(24) \qquad \sum_{k=0}^{n} f(n-k)e^{\lambda}(-\lambda)^k/k!, \qquad n = 0, 1, 2, \cdots,$$

and zero for $n < 0$. Furthermore, it is well known from the theory of Fourier series that D^* is a minimum if $c_k^{(N)}$ are chosen as the first $N + 1$ Fourier coefficients of $g(u)/\varphi(u)$, and that this minimum is arbitrarily small for large enough N if and only if the Fourier coefficients of $g(u)/\varphi(u)$ are zero for negative indices —which is in fact the case. If we then take the values (24) for $c_k^{(N)}$, $k = 0, 1, \cdots, N$, and express $a_k^{(N)}$ in terms of $c_k^{(N)}$ by (21), we arrive at the formula (14).

It will be observed that the minimum D is connected with the minimum D^* by

$$\min D \leqq \max |\varphi(u)| \cdot \min D^* \leqq \min D^* \leqq \frac{\min D}{\min |\varphi(u)|} \leqq e^{2\lambda} \min D,$$

so that all that we can say about the approximation given by (14) with a small N is that it is an upper bound for the best possible mean-square approximation by sums (18), and that the best mean-square approximation is at worst $e^{-2\lambda}$ times it. This means that if D^* is small, so is D; but D^* is not necessarily small even if D is. Hence we cannot in general expect the coefficients (14) to be suitable for practical curve-fitting, since they may increase the mean-square error by a factor of as much as $e^{2\lambda}$; we may, however, expect (14) to be better when λ is small.

Now, as we have already observed,

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x)$$

is the xth Fourier coefficient of

$$g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{it})^k ;$$

if we write (18) in the form

$$(25) \quad f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = \int_{-\pi}^{\pi} e^{-ixt} \left[g(t)/\varphi(t) - \sum_{k=0}^{N} a_k^{(N)}(1 - e^{it})^k \right] \varphi(t)\, dt,$$

and choose the $a_k^{(N)}$ as specified above, the expression in square brackets is $g(t)/\varphi(t)$ minus the first $N + 1$ terms of its Fourier series, and so the Fourier series of $[\cdots]$ involves no e^{ikt} with $k < N + 1$. Since the Fourier series of $\varphi(t)$ involves no e^{ikt} with $k < 0$, the product $\varphi(t)[\cdots]$ also involves no e^{ikt} with $k < N + 1$, and therefore the integral in (25) is zero for $x = 0, 1, 2, \cdots, N$ (since it represents the xth Fourier coefficient of $\varphi(t)[\cdots]$). In other words,

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = 0, \qquad x = 0, 1, 2, \cdots, N.$$

Furthermore, we can compute $f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x)$ for $x > N$ by the convolution formula from the Fourier series of $\varphi(t)$ and $[\cdots]$; for $n > N$, the nth Fourier coefficient of $[\cdots]$ is just that of $g(t)/\varphi(t)$, given by (24), and that of $\varphi(t)$ is $e^{-\lambda}\lambda^n/n!$, so for $x > N$

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = \sum_{l=N+1}^{x} \left(\sum_{k=0}^{l} f(l - k) e^{\lambda} (-\lambda)^k / k! \right) \theta(x - l)$$

and in particular

$$f(N + 1) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(N + 1) = \sum_{k=0}^{N+1} f(N + 1 - k)(-\lambda)^k/k!.$$

9. Proofs: practical problem. We have so far obtained only an estimate for the minimum of D, by obtaining the minimum of D^*; this estimate is satisfactory for large N and so for theoretical purposes. However, to obtain precisely the best mean-square approximation to $f(x)$ by a small number N of terms of the sum in (18), we have to choose $a_k^{(N)}$ so that

$$\sum_{k=0}^{N} a_k^{(N)} (1 - e^{it})^k \varphi(t)$$

is the first $N + 1$ terms of the expansion of $g(t)$ in terms of the set of functions obtained by replacing $(1 - e^{it})^k \varphi(t)$, $k = 0, 1, 2, \cdots$, by an equivalent orthonormal set. The process for obtaining this orthonormal set is well known; it turns out that the integrals that have to be evaluated are expressible in terms of Bessel functions of imaginary argument; the result is that the first orthonormal functions are

$$\psi_0(t) = (2\pi)^{-\frac{1}{2}} \alpha^{-\frac{1}{2}} \exp (\lambda e^{it}),$$

$$\psi_1(t) = (2\pi)^{-\frac{1}{2}} \frac{\alpha_1 - \alpha_0 e^{it}}{[\alpha_0(\alpha_0^2 - \alpha_1^2)]^{\frac{1}{2}}} \exp (\lambda e^{it}),$$

$$\psi_2(t) = (2\pi)^{-\frac{1}{2}} \frac{\alpha_1^2 - \alpha_0 \alpha_2 - \alpha_1(\alpha_0 - \alpha_2) e^{it} - (\alpha_1^2 - \alpha_0^2) e^{2it}}{[(\alpha_1^2 - \alpha_0^2)(\alpha_0 - \alpha_2)(2\alpha_1^2 - \alpha_0^2 - \alpha_0 \alpha_2)]^{\frac{1}{2}}} \exp (\lambda e^{it}),$$

where $\alpha_0 = J_0(2i\lambda)$, $\alpha_1 = -iJ_1(2i\lambda)$, $\alpha_2 = -J_2(2i\lambda)$. It is then a simple matter, first to express ψ_0, ψ_1, ψ_2 in terms of $\varphi(t)$, $\varphi(t)(1 - e^{it})$, $\varphi(t)(1 - e^{it})^2$, and then to determine $a_0^{(0)}$; $a_0^{(1)}$, $a_1^{(1)}$; and $a_0^{(2)}$, $a_1^{(2)}$, $a_2^{(2)}$. For example, the best two-term approximation for $g(u)$ in terms of $\psi_0(u)$, $\psi_1(u)$ is

$$g(u) \backsim \psi_0(u) \int_{-\pi}^{\pi} g(u)\bar{\psi}_0(u)\, du + \psi_1(u) \int_{-\pi}^{\pi} g(u)\bar{\psi}_1(u)\, du,$$

and the integrals $\int_{-\pi}^{\pi} g(u)\bar{\psi}_k(u)\, du$ are combinations of terms of the form

$$(2\pi)^{-1} \int_{-\pi}^{\pi} g(u)e^{iku} \varphi(u)\, du;$$

these in turn are Fourier coefficients of $g(u)\varphi(u)$ and so are expressible, by the Parseval formula, as products of the Fourier coefficients of $g(u)$ (namely, $f(n)$) and of $\varphi(u)$ (namely, $\theta(n)$). We omit the algebraic work; the results are given in formulas (9), (10), (11).

10. Proofs: continuous case. In the continuous case of our approximation problem we assume that $|f(x)|^2$ is integrable on $(-\infty, \infty)$ and look for coefficients $a_k^{(N)}$ that will minimize

$$D = \int_{-\pi}^{\pi} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2 dx,$$

where

$$\theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)e^{-ixu}\, du,$$

$$\Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)e^{-ixu}(1 - e^{iu})^k\, du.$$

Let $f(x)$ be the Fourier transform of $g(u)$; we can regard $\theta(x)$ as the Fourier transform of $\varphi(u)$, $\varphi(u)$ being defined as zero outside $(-\pi, \pi)$. Then by Parseval's theorem for Fourier transforms we have

$$2\pi D = \int_{|t|>\pi} |g(t)|^2\, dt + \int_{-\pi}^{\pi} \left| g(t) - \varphi(t) \sum_{k=0}^{N} a_k^{(N)}(1 - e^{it})^k \right|^2 dt.$$

Clearly, then, D cannot be made arbitrarily small unless $g(t) = 0$ almost everywhere outside $(-\pi, \pi)$; and if this condition is satisfied, D reduces to the same form which it had in the discrete case—see (19). Thus the problem of mean-square approximation in the continuous case reduces, if it can be solved at all, to the corresponding problem in the discrete case.

11. Representation by a series. We consider the representation of a given $f(x)$ by the B-series with the classical coefficients (5), but with mean-square convergence of the series. Here we assume that $f(x) \geqq 0$, $f(x) = 0$ for $x = -1, -2, \cdots$, and $\sum_{x=0}^{\infty} [f(x)]^2 < \infty$, ask whether we can have

$$\lim_{n\to\infty} \sum_{x=-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k \Delta^k \theta(x) \right|^2 = 0, \tag{26}$$

where here the a_k do not depend on n (but are not, in principle, required to have the form (5)). From our previous discussion this is equivalent to

$$\lim_{n\to\infty} \int_{-\pi}^{\pi} \left| g(t) - \varphi(t) \sum_{k=0}^{n} a_k(1 - e^{it})^k \right|^2 dt = 0,$$

and this implies that

$$\lim_{n\to\infty} | a_n |^2 \int_{-\pi}^{\pi} | \varphi(t) |^2 \, | 1 - e^{it} |^{2n} \, dt = 0.$$

From this it follows easily that

$$\sum_{n=0}^{\infty} a_n(1 - e^{it})^n$$

converges for $| t | < \pi$, or in other words that

$$H(z) = \sum_{n=0}^{\infty} a_n(1 - z)^n$$

converges on $| z | = 1$ except perhaps for $z = -1$, and hence converges in $| 1 - z | < 2$. By analytic continuation it is easy to identify $H(z)$ with $F(z)\Phi(z)$, where for $| z | < 1$,

$$F(z) = \sum_{n=0}^{\infty} f(n)z^n, \qquad \Phi(z) = \sum_{n=0}^{\infty} \theta(n)z^n = e^{\lambda(1-z)}.$$

Since $1/\Phi(z)$ has no singular points, $F(z)$ is analytic in $| 1 - z | < 2$ and hence in particular in $0 \leqq x < 3$; since $F(z)$ is a power series with nonnegative coefficients, it has a singular point at the positive real point on its circle of convergence [30, p. 214], and so it must be analytic at least in $| z | < 3$. This gives the restriction $\limsup_{n\to\infty} | f(n) |^{1/n} \leqq \frac{1}{3}$. Nevertheless, as we know, $f(x)$ is represented in mean-square by a sequence of sums of terms $a_k^{(N)} \Delta^k \theta(x)$ even if we assume only that $\Sigma \, | f(n) |^2$ converges.

In the continuous case, if $f(x) \geqq 0$ and we have

(27) $$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k \Delta^k \theta(x) \right|^2 dx = 0,$$

we must have $g(x) = 0$ almost everywhere outside $(-\pi, \pi)$ and then, as we saw previously, (26) holds also. Now since $f(x) \geqq 0$, $g(t)$ has derivatives of all orders if it has derivatives of all orders at $t = 0$ [29, p. 90] and it is easily seen from this that $g(t)$ is analytic for all real t if it is analytic at $t = 0$. Now on the one hand, unless $f(x) \equiv 0$, $g(t)$ cannot be analytic for all real t if (as we are supposing) $g(t)$ vanishes outside $(-\pi, \pi)$. On the other hand, $H(e^{it}) = g(t)/\varphi(t)$ for real values of t close to 0 and so, if t is regarded as a complex variable, for complex values of t near 0. Since $1/\varphi(t)$ is analytic everywhere, $g(t)$ is analytic at $t = 0$. From this contradiction we infer that a nonnegative $f(x)$ can never be represented in the form (27), although it may perfectly well be represented by

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k^{(n)} \Delta^k \theta(x) \right|^2 dx = 0.$$

REFERENCES

I. Charlier series

[1] A. C. Aitken, *Statistical Mathematics*, Oliver and Boyd, London, 1939, pp. 58–59, 66–67, 76–79.

[1a] W. Andersson, "Short notes on Charlier's method for expansion of frequency functions in series," *Skandinavisk Aktuar. tids.*, Vol. 27 (1944), pp. 16–31.

[2] L. A. Aroian, "The type B Gram-Charlier series," *Annals of Math. Stat.*, Vol. 8 (1937), pp. 183–192.

[3] R. P. Boas, Jr., "The Charlier B-series," *Am. Math. Soc. Trans.* (to appear).

[4] C. V. L. Charlier, "Über das Fehlergesetz," *Ark. Mat. Astr. Fys.*, Vol. 2, no. 8 (1905).

[5] C. V. L. Charlier, "Die zweite Form des Fehlergesetzes," *ibid.*, no. 15 (1905).

[6] C. V. L. Charlier, "Über die Darstellung willkürlicher Funktionen," *ibid.*, no. 20 (1905).

[7] C. V. L. Charlier, "Researches into the theory of probability," *Lund Universitets Årsskrift N. F.*, Afd. 2, Vol. 1, no. 5 (1906).

[8] H. Cramér, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics and Mathematical Physics, no. 36, Cambridge University Press, 1937.

[9] W. P. Elderton, *Frequency Curves and Correlation*, 3d ed., Cambridge University Press, 1938, pp. 131–132.

[10] A. Fisher, *An Elementary Treatise on Frequency Curves and their Application in the Analysis of Death Curves and Life Tables*, Macmillan, New York, 1922.

[11] A. Fisher, *The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods*, Vol. 1, 2d ed., Macmillan, New York, 1922, pp. 271–279.

[12] M. Jacob, "Über die Charlier'sche B-Reihe," *Skandinavisk Aktuar. tids.*, Vol. 15 (1932), pp. 286–291.

[13] C. Jordan, "Sur la probabilité des épreuves répétées, le théorème de Bernoulli et son inversion," *Bull. Soc. Math. France*, Vol. 54 (1926), pp. 101–137.

[14] N. R. Jørgensen, "Note sur la fonction de répartition de type B de M. Charlier," *Ark. Mat. Astr. Fys.*, Vol. 10, no. 15 (1915).

[15] N. R. Jørgensen, "Undersøgelser over Frequensflader og Korrelation," *Copenhagen thesis*, 1916.

[16] M. G. Kendall, *The Advanced Theory of Statistics*, Vol. 1, 3d ed., Griffin, London, 1947, pp. 154–156.

[17] S. Kullback, "On the Charlier type B series," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 574–581; Vol. 19 (1948), p. 127.

[18] R. von Mises, *Vorlesungen aus dem Gebiete der angewandten Mathematik*, Vol. 1, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, Deuticke, Leipzig and Vienna, 1931, pp. 265–269.

[19] N. Obrechkoff, *Sur la loi de Poisson, la série de Charlier et les équations linéaires aux différences finies de premier ordre à coefficients constants*, Actual. Sci. Ind., no. 740, Hermann et Cie., Paris (1938), pp. 35–64.

[20] H. Pollaczek-Geiringer, "Die Charlier'sche Entwicklung willkürlicher Verteilungen," *Skandinavisk Aktuar. tids.*, Vol. 11 (1928), pp. 98–111.

[21] H. Pollaczek-Geiringer, "Über die Poissonsche Verteilung und die Entwicklung willkürlicher Verteilungen," *Zeits. f. Angew. Math. und Mech.*, Vol. 8 (1928), pp. 292–309.

[22] H. L. Rietz, *Mathematical Statistics*, Carus Mathematical Monographs, no. 3, Chicago, 1927, pp. 60–68, 170–177.

[23] E. Schmidt, *Über die Charliersche Entwicklung einer "arithmetischen Verteilung" nach den sukzessiven Differenzen der Poissonschen asymptotischen Darstellungsfunktion*

für die Wahrscheinlichkeit seltener Ereignisse, Sitzungsberichte der Preussischen Akademie der Wissenschaften, Phys.-Math. Kl. 1928, p. 148.

[24] E. SCHMIDT, "Über die Charlier-Jordansche Entwicklung einer willkürlichen Funktion nach der Poissonschen Funktion und ihren Ableitung," *Zeits. f. Angew. Math. und Mech.*, Vol. 14 (1933), pp. 139–142.

[25] H. L. SELBERG, "Über die Darstellung willkürlicher Funktionen durch Charliersche Differenzreihen," *Skandinavisk Aktuar. tids.*, Vol. 25 (1942), pp. 228–246.

[26] H. L. SELBERG, "Über die Darstellung der Dichtefunktion einer Verteilung durch eine Charliersche B-Reihe," *Archiv for Mathematik og Naturvidenskab*, Vol. 46 (1943), pp. 127–138.

[27] J. F. STEFFENSEN, *Some Recent Researches in the Theory of Statistics and Actuarial Science*, Cambridge University Press, 1930, pp. 35–48.

[27a] J. F. STEFFENSEN, "Note sur la fonction de type B de M. Charlier," *Svenska Aktuar. Tids.*, Vol. 3 (1916), pp. 226–228.

[28] J. V. USPENSKY, "On Charles Jordan's series for probability," *Annals of Math.* Series 2, Vol. 32 (1931), pp. 306–312.

II. Other references

[29] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[30] E. C. TITCHMARSH, *The Theory of Functions*, Oxford, 1932.

[31] A. ZYGMUND, *Trigonometrical Series*, Warszawa-Lwów, 1935.

[32] *Table of the Bessel Functions $J_0(z)$ and $J_1(z)$ for Complex Arguments*, prepared by the Mathematical Tables Project, National Bureau of Standards, 2d ed., Columbia University Press, New York, 1947.