RANDOM UNIT VECTORS II: USEFULNESS OF GRAM-CHARLIER AND RELATED SERIES IN APPROXIMATING DISTRIBUTIONS[1]

BY DAVID DURAND AND J. ARTHUR GREENWOOD

Massachusetts Institute of Technology and *Harvard University*

0. Summary. The distribution of the sum of n random coplanar unit vectors and of a given component of the sum has been discussed by many authors, who have shown that each distribution can be approximated in series that are asymptotically normal. But the difficult question of the usefulness of these approximations for finite n—in particular for small n—has not been exhaustively treated. Accordingly, this paper reexamines some analyses of Pearson's series for the vector sum, presents corresponding series for a component, and examines the accuracy of the latter series.

1. Basic formulas. Given a sample of random coplanar unit vectors [cos ξ_i, sin ξ_i], where all values of ξ_i $(i = 1, 2, \ldots n)$ between 0 and 2π are equally likely, we define the quantities

$$V = \sum \cos \xi_i, \qquad W = \sum \sin \xi_i, \qquad R = (V^2 + W^2)^{1/2}.$$

According to Kluyver [10], the probability that $0 \leqq R \leqq r$ is

$$P_R(r, n) = r \int_0^\infty [J_0(t)]^n J_1(rt)\, dt, \tag{1}$$

and the probability that $0 \leqq V \leqq v$ is

$$P_V(v, n) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty [J_0(t)]^n \frac{\sin vt}{t}\, dt. \tag{2}$$

Differentiating (2) yields a formula for the differential of probability, namely

$$dP_V(v, n) = \left[\frac{1}{\pi} \int_0 [J_0(t)]^n \cos vt\, dt\right] dv, \tag{3}$$

explicitly given by Lord [11].

2. Series approximation of the R-distribution. As an asymptotic approximation to (1), the method of steepest descent yields a formula originally due to Rayleigh [16]—namely,

$$1 - e^{-x}, \tag{4}$$

where $x = r^2/n$. It will be seen that (4) is the volume under the two-dimensional Gaussian bell

$$dP = \frac{1}{\pi n} e^{-\frac{v^2+w^2}{n}}\, dv\, dw$$

Received November 12, 1956.

[1] This is a revision of the paper "Approximation to the Distribution of the Sum of Cosines of Random Angles (Preliminary Report)," which was presented at the Ann Arbor meeting of the Institute of Mathematical Statistics, August 30, 1955.

inside the circle $v^2 + w^2 = nx$. As an approximation to the differential $(1/2\pi r)\, d/dr P_R(r, n)$, Pearson [14] derived an asymptotic series having the form

$$\frac{1}{2\pi r}\frac{d}{dr} P_R(r, n) = \frac{e^{-x}}{\pi n} \sum_{i=0}^{\infty} c_i\, L_i(x), \tag{5}$$

where the $L_i(x)$ are Laguerre polynomials, and the c_i are as follows:

$c_0 = 1$, $c_1 = 0$,

$2!c_2 = -1/2n$, $3!c_3 = -2/3n^2$,

$4!c_4 = (6n - 11)/8n^3$, $5!c_5 = (50n - 57)/15n^4$,

$6!c_6 = -(1892 - 2125n + 270n^2)/144n^5$.

Pearson believed that series (5) through c_6 would provide a satisfactory approximation for $n > 6$. Lord [13], however, writes: "These formulae have been very little tested for $s > 1$, but they would appear to behave rather like the Type A series $(s = 1)$[2] and to give satisfactory approximations for nearly normal distributions except at the tails. In the case of the distribution of the sum of n coplanar random vectors of equal magnitude, Pearson concluded that five terms [i.e., through c_6] of the series were enough to give four-decimal accuracy for $n \geqq 7$, but investigations which the author hopes to publish shortly suggest that he was rather optimistic." Lord supports his belief with some illustrative calculations for $n = 4, 6, 8$, and 10. On the basis of our own calculations we concur with Lord's view, although we realize that the suitability of an approximation depends upon the number of reliable decimal places that one demands to work with, and this in turn depends upon whether one wishes to approximate the central portion of the distribution or the tails. For significance tests one wants fairly accurate points in the tail.

In considering the accuracy of a series approximation, one may be interested in two things: first, how many decimals a fixed number of terms will yield; and second, the most profitable term to stop at. We missed the second point in our previous paper, where we presented a table (Table 3 of Greenwood and Durand [6]) showing calculated values of (1) for $n = 7$ and $n = 14$ for comparison with Rayleigh's approximation (4), the integral of Pearson's approximation (5) through c_6, and a rearrangement of the integrated series through n^{-3}. On reexamining the calculations underlying this table, we find that the term in n^{-1} secures about three decimal accuracy; and that for $n = 7$, further terms are not very helpful, but for $n = 14$ the term in n^{-2} increases accuracy roughly from three decimals to four (see Table 1).

3. Series approximation of the V-distribution. Early writers on the random walk—including Einstein [4], Rayleigh [17], and Wiener [20]—recognized that the distribution of V is asymptotically normal. The mean is obviously zero and the

[2] Lord is considering the generalizations of the R- and V-distributions to spaces of s dimensions (cf. Watson [18], p. 420–421).

TABLE 1

Values of $P_R(r, 7)$ and $P_R(r, 14)$ and several approximations

r	$P_R(r, n)$ by quadrature	Approximation through terms in—				
		n^0	n^{-1}, c_2	n^{-2}	n^{-3}	c_4
				$n = 7$		
1.0	0.12500	0.13312	0.12491	0.12497	0.12500	0.12530*
2.0	0.41782	0.43528	0.41882	0.41817	0.41819	0.41819
3.0	0.71404	0.72355	0.71448	0.71349	0.71337*	0.71305
4.0	0.90039	0.89830	0.90067	0.90087	0.90082	0.90095
5.0	0.97864	0.97188	0.97752	0.97832	0.97846	0.97857
6.0	0.99788	0.99416	0.99753	0.99781	0.99785	0.99779
7.0	1.00000	0.99909	1.00023	1.00011	1.00006	1.00003
				$n = 14$		
1.0	0.06667	0.06894	0.06665	0.06666	0.06667	0.06668
2.0	0.24193	0.24852*	0.24195	0.24192	0.24193*	0.24195*
3.0	0.46583	0.47421	0.46602	0.46583	0.46583	0.46583
4.0	0.67524	0.68109	0.67551	0.67525	0.67524	0.67521
5.0	0.83105	0.83232	0.83118	0.83107	0.83105	0.83104
6.0	0.92570	0.92357	0.92558	0.92570	0.92570	0.92571
7.0	0.97285	0.96980	0.97263	0.97283	0.97285	0.97286
8.0	0.99197	0.98966	0.99183	0.99195	0.99196	0.99196
9.0	0.99815	0.99693	0.99813	0.99815	0.99815	0.99814
10.0	0.99969	0.99921	0.99973	0.99970	0.99969	0.99969
11.0	0.99997	0.99982	1.00000	0.99997	0.99997	0.99997
12.0	1.00000	0.99997	1.00002	1.00000	1.00000	1.00000

* These values correct erroneous entries in Table 3 of [6].

variance is easily shown to be $n/2$. Horner [8] recapitulated these results and gave (p. 153) the specific formulas:

$$
\begin{aligned}
dP_V(v, 1) &= \frac{dv}{\pi(1 - v^2)^{1/2}} \\
dP_V(v, 2) &= \frac{dv}{\pi^2} \int_{|v|-1}^{1} \frac{du}{[1 - u^2]^{1/2}[1 - (u - |v|)^2]^{1/2}},
\end{aligned} \tag{6}
$$

and showed that $dP_V(v, n)$ may be computed by convolving $dP_V(v, n - 1)$ with $dP_V(v, 1)$. Lord [11], more generally, showed that $dP_V(v, n)$ may be computed by convolving $dP_V(v, n - k)$ with $dP_V(v, k)$. Since Horner considered computation by convolution difficult, in which view we concur (see below), he derived a modified Pearson series and employed it to estimate $dP_V(v, 7)$. He evidently thought highly of this series approximation, since he used it as a standard against which to test the simple normal approximation for $n = 7$, and he even considered the normal approximation "very close." Slack ([15], p. 77) considered the distribution of V effectively normal "except when n is very small, i.e., <10." We do not share this optimism—though it is again a question of how many decimal places in what section of the curve are required to render the fit "very close."

Horner did not present the modified Pearson series, and to our knowledge, no one else has. We therefore derived the series according to the method described by Lord ([13] p. 347), and we present it below, partly to support subsequent computations and partly because we think the term in n^{-1} may be useful. In this series, the substitution $z = v(2/n)^{1/2}$ makes the variance unity and simplifies the series so that

$$dP_V(v, n) = dz \sum_{i=0}^{\infty} c_i \phi^{2i}(z)/(-2)^i \tag{7}$$

Here, the notation

$$\phi^n(z) = (2\pi)^{1/2} \frac{d^n}{dz^n} e^{-z^2/2}$$

conforms to the Harvard tables [7], which are probably the best means for evaluating the series; and the c_i are identical with those in Pearson's series. Integration of (7), term by term, gives $P_V(v, n)$.

To establish limits of error for (7) or its integral is a problem for which we have found no simple, systematic solution. In the form given, (7) is a Gram-Charlier series of Type A. Theorem 4 of Cramér [3] indicates that it converges absolutely for $n \geqq 3$, and its integral converges absolutely for all n. It is sometimes convenient to rearrange the terms of (7) in decreasing powers of n, yielding an Edgeworth series. Theorems 2 and 3 of Cramér indicate that this Edgeworth series and its integral through $n^{-\alpha}$ is asymptotic with error $O(n^{-\alpha-1})$. Finally, a theorem by Esseen ([5], p. 43) establishes bounds to the discrepancy between $P_V(v, n)$ and its normal approximation. But none of these facts seems to have any great practical value. Esseen's inequality indicates that

$$\left| \int_{-\infty}^{z} \phi(t)\, dt - P_V(v, n) \right| < 9.003 n^{-1/2};$$

and this implies that a sample of some 10^8 observations is required to assure three decimal accuracy. This, of course, is absurd, since the distribution of V is symmetrical and the error is $O(n^{-1})$, not $O(n^{-1/2})$. But no one, to our knowledge, has worked out bounds for a symmetrical distribution.

In hopes of setting reasonable bounds for the error in the series approximations, we proceeded to ascertain certain values of (2) and (3) that were fairly easy to compute and to compare these with the approximations. The maximum ordinate

$$dP_V(0, n)/dz = \frac{1}{\pi} \left(\frac{n}{2}\right)^2 \int_0^{\infty} [J_0(t)]^n\, dt$$

was fairly easily calculated by quadratures with available equipment. To do this job, a punched card table of the Bessel function J_0 was involuted and summed on IBM machines. Then the sum was corrected for curvature at the upper end; and when necessary, the portion of the integral lying outside the limits of the

TABLE 2

Values of $dP_V(0, n)/dz$ and several approximations.

n	$dP(0, n)/dz$ by quadrature	Approximation through terms in—						
		n^{-1}, ϕ^4	n^{-2}	n^{-3}	ϕ^6	ϕ^8	ϕ^{10}	ϕ^{12}
3	.34948	.37401	.37386	.37418	.36477	.36831	.37582	.38110
4	.40637	.38024	.38016	.38029	.37505	.37782	.38147	.38287
5	.37928	.38398	.38393	.38400	.38066	.38273	.38475	.38515
6	.38947	.38648	.38644	.38648	.38417	.38574	.38697	.38706
7	.38742	.38826	.38823	.38825	.38656	.38779	.38859	.38858
8	.39002	.38959	.38957	.38959	.38829	.38928	.38983	.38979
9	.39048	.39063	.39061	.39063	.38960	.39041	.39080	.39075
10	.39152	.39146	.39145	.39146	.39063	.39130	.39159	.39154

punched card table was evaluated by integrating the asymptotic series for $[J_0]^n$. As a check the involution and summation was repeated on an entirely different series of IBM machines, and finally the integral

$$\int_0^{30.6346065} J_0(t)\, dt$$

—that is, from 0 to j_{10}—was compared with the value given by Watson ([18], p. 752) for half this integral.

From the comparisons given in Table 2, one sees that the term $-\phi^4(z)/16n$, which is the first correction term in either the Gram-Charlier or the Edgeworth series, produces a substantial improvement over the simple normal approximation 0.39894. But the contribution of further terms is doubtful, to say the least. This is particularly true of the Edgeworth series, since the error through n^{-1} is positive for odd n and negative for even n so that inclusion of one or more further terms must improve half of the approximations at the expense of the other half. In effect, it appears that the Edgeworth series either does not converge, or converges to the wrong value, and that the Gram-Charlier series converges too slowly to be of great use, if refinements over the first correction term are required.

Table 2 gives a fair notion of the accuracy of the series approximation through n^{-1}, since the error

$$| \, dP_V(v, n)/dz - \phi(z) + \phi^4(z)/16n \, |$$

is bound to be large at $v = 0$; in fact, we are able to show that it achieves a local maximum there for $n = 5, 6, 7, 8$, and 9. The first derivative of this error is easily shown to be zero at $v = 0$. The second derivative

$$(-)^n \left[-\frac{1}{\pi} \left(\frac{n}{2}\right)^{3/2} \int_0^\infty t^2[J_0(t)]^n \cos vt\, dt + \phi^2(z) - \phi^6(z)/16n \right]$$

reduces to

$$(8) \qquad (-1)^n \left[-\frac{1}{\pi} \left(\frac{n}{2}\right)^{3/2} \int_0^\infty t^2[J_0(t)]^n\, dt + (1 - 15/16n)\phi(0) \right]$$

TABLE 3

Values of $P_V(v,\ 4)$ and several approximations

$z = v(2/n)^{1/2}$	$P(v, 4)$ by quadrature	Normal approximation	Approximation through terms in—				
			n^{-1}, ϕ^3	n^{-2}	n^{-3}	ϕ^5	ϕ^7
1.6	0.94452	0.94520	0.94398	0.94367	0.94355	0.94460	0.94410
1.8	0.96404	0.96407	0.96460	0.96461	0.96454	0.96545	0.96500
2.0	0.97820	0.97725	0.97894	0.97921	0.97922	0.97978	0.97947
2.2	0.98821	0.98610	0.98834	0.98878	0.98886	0.98902	0.98889
2.4	0.99482	0.99180	0.99412	0.99460	0.99472	0.99456	0.99458
2.6	0.99860	0.99534	0.99741	0.99782	0.99794	0.99763	0.99773
2.8	0.99998	0.99744	0.99912	0.99940	0.99949	0.99916	0.99929
2.832	1.00000	0.99769	0.99929	0.99955	0.99963	0.99931	0.99944

for $v = 0$. The quantity on the right is easily evaluated, and we were able to evaluate the Bessel-function integral by quadratures for $n = 5, 7, 8$, and 9; it is infinite for $n = 6$. (8) is indeed negative for $n = 5, 7, 8$, and 9; thus the error achieves a local maximum. For $n = 6$, (8) is negatively infinite.

Although convolution of the V-distribution is generally difficult, as Horner indicates, values in the tail of $dP_V(v,\ 4)/\ dv$ and $P_V(v,\ 4)$ are fairly easily obtained with modern computing equipment. We were able to evaluate (6) on the Harvard Mark IV computer by means of Bronwin's formula for numerical integration (cf. Whittaker and Robinson [19], p. 159) and then to obtain $dP_V(v,\ 4)/\ dv$ by quadrature. This operation was necessarily limited to the portion of $dP_V(v,\ 4)/\ dv$ unaffected by the singularity of $dP_V(v,\ 2)/\ dv$—that is, the portion outside $v = 2$. Finally, hand integration of $dP_V(v, 4)/\ dv$ produced values of $P_V(v,\ 4)$ for comparison with the series approximations in Table 3. Here, again, the first correction term effects a substantial improvement over the simple normal approximation, but additional terms contribute little. Note that the n^{-1} term provides almost three-decimal accuracy.

4. Normalization of the V-distribution. The method of Cornish and Fisher [2] (cf. Kendall [9], Secs. 6.32 and 6.33) provides the means of deriving an approximately normal variate y (with unit variance) as a series expansion in $z = v(2/n)^{1/2}$ and n. This series through n^{-3} is

$$y = z + (z^3 - 3z)/16n + (71z^5 - 224z^3 - 15z)/4608n^2 + (385z^7 - 1323z^5 - 981z^3 + 1575z)/73728n^3.$$

It may be reverted to give z as a function of y, with the following result for terms through n^{-3}:

$$z = y - (y^3 - 3y)/16n - (17y^5 - 8y^3 - 177y)/4608n^2 - (33y^7 + 165y^5 - 1989y^3 + 999y)/73728n^3. \tag{9}$$

A series of this sort is of particular interest to statisticians for approximating percentage points of a distribution. Table 4 compares various approximations

TABLE 4

Percentage points of $P(v, 4)$ an several approximations

$P(v, 4)$	$z = v(n/2)^{1/2}$	Normal approximation	Approximation through terms in—		
			n^{-1}	n^{-2}	n^{-3}
.95	1.65010	1.64485	1.65242	1.65408	1.65496
.975	1.94862	1.95996	1.93419	1.93305	1.93402
.99	2.24588	2.32635	2.23868	2.22992	2.22977
.995	2.40712	2.57583	2.42953	2.41143	2.40886
.999	2.63436	3.09023	2.77399	2.71963	2.70274

derived from (9) with percentage points of $P_V(v, 4)$. The latter were obtained by inverse interpolation of values of $P_V(v, 4)$, not all shown in Table 3. As with previous comparisons, the term in n^{-1} provides a substantial improvement over the simple normal approximation; the terms in n^{-2} and n^{-3}, moreover, provide additional improvement for the extreme points $P_V = 0.995$ and 0.999. Even with this improvement, however, accuracy is less than two decimals.

5. Other possible series for approximating R- and V-distributions. Bennett [1 has proposed the use of Fourier-Bessel series for computing the distribution of R and Fourier sine series for V. We have not tested his claim that these series are more effective than quadratures.

6. Conclusion. For very large n, Rayleigh's approximation to the distribution of the vector sum or the normal approximation to the distribution of a component is clearly satisfactory. For very small n—less than 6 for the sum or less than 4 for a component—neither approximation is remotely satisfactory; but convolution is feasible even though laborious. For intermediate n, both the Rayleigh approximation for the sum and the normal approximation for a component can be substantially improved by inclusion of a single term in n^{-1}; and the additional computations are not excessive.

The inclusion of terms beyond n^{-1} appears, for the most part, not to be worth the trouble. For small n the improvement is hardly noticeable; for larger n the improvement may be appreciable, but it is probably not needed, since the n^{-1} term will give fair accuracy by itself. However, the use of (9) to approximate percentage points in the extreme tail may provide an exception.

As for accuracy, we believe that our series (7) through n^{-1} affords a most unsatisfactory approximation for $n = 3$, and a glance at Horner's Fig. 7 should convince anyone. For $n = 4$, we believe the approximations are still short of satisfactory, and we suggest use of the exact values in Tables 3 and 4 whenever these are appropriate. For $n = 5$, the approximation should be substantially better than for $n = 4$; indeed, since $dP_V(0, 5)/dz$ is approximated to within 0.00470 against 0.02613 for $dP_V(0, 4)/dz$, and since $P_V(v, 4)$ is approximated to nearly three decimals in the tails, we surmise that $P_V(v, 5)$ is approximated to at least three decimals, possibly approaching four in the tail.

7. Acknowledgments. The work on this study was supported in part by a grant from the Sloan Research Fund of the School of Industrial Management at M.I.T. and in part by the Army Office of Ordnance Research under Contract DA-19-020-ORD-2684. The computation of $dP_V(v, 2)/dv$ and $dP_V(v, 4)/dv$, which underlie Tables 3 and 4, was performed by Greenwood and Dennis C. Marshall on the Harvard Mark IV calculator as a joint laboratory exercise under the direction of Dr. Anthony G. Oettinger. Integration of $[J_0(t)]^n$, leading to $dP_V(0, n)/dz$ in Table 2, was performed on punched-card equipment furnished by the National Bureau of Economic Research, and it was checked on the I.B.M. 650 calculator at the Office of Statistical Services, M.I.T. A substantial portion of the required hand calculation was performed by Lena Fishman at M.I.T.

Note added in proof. We are embarrassed to find that the notation in this paper disagrees with that in [6]. A table of concordance follows:

notation	symbol in [6]	symbol in this paper
r^2/n	z	x
$v(2/n)^{\frac{1}{2}}$	not used	z
exponentially distributed transform of r^2/n	y	not used
normally distributed transform of $v(2/n)^{\frac{1}{2}}$	not used	y

REFERENCES

[1] W. R. Bennett, "Distribution of the sum of randomly phased components," *Quart. Appl. Math.*, Vol. 5 (1947, i.e., January, 1948), pp. 385–393.

[2] E. A. Cornish and R. A. Fisher, "Moments and cumulants in the specification of distributions," *Revue de l'Inst. Int. Stat.*, Vol. 5 (1937, i.e. January, 1938), pp. 307–320.

[3] H. Cramér, "On the composition of elementary errors," *Skand. Aktuarietids.*, Vol. 11 (1928), pp. 13–74.

[4] A. Einstein, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," *Ann. Physik*, Vol. 17 (1905), pp. 549–560.

[5] C. G. Esseen, "Fourier analysis of distribution functions," *Acta Math.*, Vol. 77 (1945), pp. 1–125.

[6] J. A. Greenwood and D. Durand, "The distribution of length and components of the sum of n random unit vectors," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 233–246.

[7] Harvard University, Annals of the Computation Laboratory 23, *Tables of the Error Function and Its First Twenty Derivatives*, Harvard University Press, Cambridge, Mass., 1952.

[8] F. Horner, "A problem on the summation of simple harmonic functions of the same amplitude and frequency but of random phase," *Philos. Mag.*, Vol. 37 (1946), pp. 145–162.

[9] M. G. Kendall, *The Advanced Theory of Statistics*, Vol. 1, Charles Griffin, London, 1943, Secs. 6.20 ff.

[10] J. C. Kluyver, "A local probability problem," *Proc. Nederl. Akad. Wetensch.*, Vol. 8 (1906), pp. 341–350.

[11] R. D. Lord, "A problem on random vectors," *Philos. Mag.*, Vol. 39 (1948), pp. 66–71.

[12] R. D. Lord, "The use of the Hankel transform in statistics. I. General theory and examples," *Biometrika*, Vol. 41 (1954), pp. 44–55.

[13] R. D. LORD, "The use of the Hankel transform in statistics. II. Methods of computation," *Biometrika*, Vol. 41 (1954), pp. 344–350.

[14] K. PEARSON, "A mathematical theory of random migration," *Drapers Company Research Memoirs*, Biometric Series, No. 3, 1906.

[15] MARGARET SLACK, "The probability distributions of sinusoidal oscillations combined in random phase," *J. Inst. E. E.*, Vol. 93 (1946), Part 3, pp. 76–86.

[16] J. W. STRUTT, LORD RAYLEIGH, "On the resultant of a large number of vibrations of the same pitch and of arbitrary phase," *Philos. Mag.*, Vol. 10 (1880), pp. 73–78.

[17] J. W. STRUTT, LORD RAYLEIGH, "On the problem of random vibrations and of random flights in one, two, and three dimensions," *Philos. Mag.*, Vol. 37 (1919), pp. 321–347.

[18] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, 1944.

[19] E. T. WHITTAKER AND G. ROBINSON, *The Calculus of Observations*, 4th ed., Blackie and Son, London, 1944.

[20] N. WIENER, "The average of an analytic functional and the Brownian movement," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 7 (1921), pp. 294–298.