Approximation of Central Limits

BY R. DAWSON

Unclassified

A development of asymptotic formulas for estimating an ordinate or tail area in the distribution of the sum of a large number of identically distributed random variables, based upon shifting the point of estimation.

1. INTRODUCTION

In accordance with the central limit theorem (Cramér, [1], Sec. 17.4, p. 215), the standardized sum of an increasing number of independent and identically distributed random variables is asymptotically normal; and when the number of variables is large enough, the assumption of normality is adequate. But the convergence is not uniform, being notoriously poor at extreme values of the sum, and it has been found helpful [2], [4] to develop asymptotic series to improve the fit. The object of this paper is to present a unified theory of estimating ordinates and tail areas, enlarging or simplifying known results, and adding new results. The unifying principle is to shift the point of estimation (whether of the frequency or cumulative distribution function) to the mean, or near the mean, so as to eliminate the leading term of the residual asymptotic series. (The remaining terms also tend to be reduced.)

2. THE ASYMPTOTIC EXPANSION OF CRAMÉR

Cramér ([1] 17.7) developed an Edgeworth series for the asymptotic frequency function $h(x)$ and the cumulative distribution function $H(x)$ of the standardized sum of a large number of independent random variables $\{y_i\}$ sharing a common c. d. f. $F(y)$. (To avoid difficulties, it will be assumed that $F(y)$ has finite moments of all orders.) A concise redevelopment, following Cramér except for changes made in order to educe the general term, is presented.

Let $\psi(t)$ be the characteristic function of χ, the standardized sum of c replications of y, and let $\psi_1(t)$ be the c. f. of $y - Ey$, "Ey" denoting the mean of y. The nth cumulant of y will be denoted by κ_n, and it will also be convenient to define

$$\begin{aligned} \lambda_n &= \kappa_n/\kappa_2^{n/2} = \kappa_n/\sigma^n \; [\textit{Cramér, p. 225}], \\ \nu_n &= \lambda_n/n! = \kappa_n/n!\,\sigma^n . \end{aligned} \tag{2.1}$$

Then ([1], pp. 224-226) ψ is related to ψ_1 by

$$\psi(t) = \left[\psi_1\left(\frac{t}{\sigma\sqrt{c}}\right)\right]^c = exp\left\{c \sum_n \frac{\kappa_n}{n!}\left(\frac{it}{\sigma\sqrt{c}}\right)^n\right\} \tag{2.2}$$

$$= exp\left\{c \sum_{n=1}^{\infty}\left(\frac{it}{\sqrt{c}}\right)^n \nu_n = \sqrt{2\pi}\ \phi(t)\ exp\left\{c \sum_{n=3}^{\infty}\left(\frac{it}{\sqrt{c}}\right)^n \nu_n\right\}\right.,$$

where

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

because

$$\nu_1 = \lambda_1 = 0 \text{ and } \nu_2 = \lambda_2/2! = 1/2.$$

Hence, using the inversion formula ([1], p. 94) for recovering the frequency from the characteristic function,

$$\begin{aligned} h(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\psi(t)\,dt \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt \cdot \prod_{n=3}^{\infty} exp\left\{(it)^n \nu_n / c^{n/2-1}\right\} \end{aligned} \tag{2.3}$$

Now any term of finite order in c arising upon expansion of the exponentials contains a product

$$\frac{\nu_3^{a_3}}{a_3!}\quad \frac{\nu_4^{a_4}}{a_4!} \cdot \frac{\nu_5^{a_5}}{a_5!}\quad \cdots$$

in which all but a finite subset of the $\{a_n\}$ are zero.
The corresponding power of $1/c$ is

$$\sum (n/2-1)\,a_n = (1/2)\sum na_n - \sum a_n,$$

and the corresponding power of it is $\sum na_n$, which integrates ([1], Sec. 10.1.2, p. 89) by the formula

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(it)^m e^{-itx}\phi(t)dt = (-D_x)^m \phi(x) = (-1)^m \phi^{(m)}(x), \tag{2.4}$$

$m = 0, 1, 2, \ldots$.
Expanding (2.3) in the light of these considerations gives ([1], 17.7.2, p. 228)

$$\begin{aligned} h(x) \sim\ & \phi(x) - c^{-1/2}\nu_3\phi^{(3)}(x) + \frac{1}{c}\left[\nu_4\phi^{(4)} + \frac{1}{2!}\nu_3^2\phi^{(6)}\right] \\ & - c^{-3/2}\left[\nu_5\phi^{(5)} + \nu_3\nu_4\phi^{(7)} + \frac{1}{3!}\nu_3^3\phi^{(9)}\right] \\ & + c^{-2}\left[\nu_6\phi^{(6)} + (\nu_3\nu_5 + \frac{1}{2!}\nu_4^2)\phi^{(8)} + \frac{1}{2!}\nu_3^2\nu_4\phi^{(10)} + \frac{1}{4!}\nu_3^4\phi^{(12)}\right] - \cdots, \end{aligned} \tag{2.5}$$

where the coefficient (for any r) of $c^{-r/2}$ has the sign $(-1)^r$ and contains all products of k of the ν's such that the sum of the subscripts (counting multiplicities and barring subscripts less than 3) is $r + 2k$; the corresponding derivative of $\phi(x)$ is of order $r + 2k$, and the numerical divisor is the product of the factorials of the exponents of the ν's.

The corresponding expansion ([1], p. 229) for the c.d.f. $H(x)$ is found by integration of the successive terms, in effect replacing $\phi(x)$ by

$$\Phi(x) = \int_{-\infty}^{x}\phi(t)\,dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

Thus

$$H(x) \sim \Phi(x) - c^{-1/2}\nu_3\phi^{(2)}(x) + \frac{1}{c}\left[\nu_4\phi^{(3)}(x) + \frac{1}{2!}\nu_3^2\phi^{(5)}(x)\right] - \cdots. \tag{2.6}$$

The asymptotic frequency function (2.5) is particularly simple and most rapidly convergent at zero. Now

$$\phi^{(2n)}(0) = \frac{1}{\sqrt{2\pi}}(-1)(-3)(-5)\cdots(1-2n) = \frac{(-1)^n}{\sqrt{2\pi}}(2n-1)!!, \tag{2.7}$$

where the double exclamation point denotes the skip-factorial, as in $5!! = 5 \cdot 3 \cdot 1$,
and

$$\phi^{(2n+1)}(0) = 0,$$

Hence, the substitution $x = 0$ in (2.5) gives

$$\begin{aligned} h(0)\cdot\sqrt{2\pi} =\ & 1 + \frac{1}{c}\left[3!!\,\nu_4 - \frac{5!!}{2!}\nu_3^2\right] \\ & + c^{-2}\left[-5!!\,\nu_6 + 7!!\left(\frac{1}{2!}\nu_4^2 + \nu_3\nu_5\right) - \frac{9!!}{2!}\nu_3^2\nu_4 + \frac{11!!}{4!}\nu_3^4\right] \\ & + c^{3}\left[7!!\,\nu_8 - 9!!\left(\nu_3\nu_7 + \nu_4\nu_6 + \frac{1}{2!}\nu_5^2\right) + 11!!\left(\frac{1}{2!}\nu_3^2\nu_6 + \nu_3\nu_4\nu_5 + \frac{1}{3!}\nu_4^3\right)\right. \\ & \left. - 13!!\left(\frac{1}{3!}\nu_3^3\nu_5 + \frac{1}{2!2!}\nu_3^2\nu_4^2\right) + \frac{15!!}{4!}\nu_3^4\nu_4 - \frac{17!!}{6!}\nu_3^6\right] + \cdots, \end{aligned} \tag{2.8}$$

where the rule for forming the successive terms is that the coefficient of c^{-n} $(n = 1, 2, \cdots)$ contains all products of ν's such that half the sum of the subscripts (counting multiplicities — that is, first multiplying each subscript by the corresponding exponent) minus the sum of the exponents equals n. The numerical coefficient of any product of ν's whose subscripts sum to $2r$ is

$$(-1)^r(2r-1)!! = (-1)^r(2r)!/2^r r!$$

divided by the product of the factorials of the exponents.

On the other hand $H(x)$ is most rapidly convergent at $x = \pm 1$, where the leading error term vanishes along with $\phi^{(2)}(x)$.

3. ORDINATE APPROXIMATION

An equally-spaced discrete distribution over the real line may be regarded, without significant loss of generality, as assigning probabilities $\{p_n\}$ to the integers $\ldots, -2, -1, 0, 1, 2, \ \ , n, \ldots,$ such that $\sum p_n = 1$, and such that the indices n at which $p_n > 0$ have no common factor greater than unity. ([6], p. 868). If the distribution is represented by a power series (or polynomial)

$$f(z) = \sum_{p} p_n z^n, \tag{3.1}$$

then the probability that the sum of c independent observations will be

N is P_N where

$$\sum_N P_N z^N = [f(z)]^c.$$

Next, the abscissa at which the approximation is made will be shifted from N to 0. To achieve this shift, let

$$f_t(z) = f(tz)/f(t) = \sum_n p_n t^n z^n / f(t), \tag{3.2}$$

where t is a positive real parameter. The expression $f_t(z)$ may be regarded as the generating function of a new distribution which assigns the probability $p_n t^n/f(t)$ to drawing the integer n.

Then

$$[f_t(z)]^c = \sum_N \frac{t^N p_N z^N}{[f(t)]^c} \tag{3.3}$$

is the generating function for the sum (Y, say) of c independent observations from the new distribution, in the sense that the coefficient of z^N gives the probability that $Y = N$. Let $y_1, y_2, \ldots, y_c$, each independently distributed according to $f_t(z)$, be the new individual random variables, so that $Y = \Sigma\, y_i$. Then the mean and variance of any y are

$$Ey_1 = \frac{tf'(t)}{f(t)}, \quad Vy_1 = \frac{tf'(t) + t^2 f''(t)}{f(t)} = \sigma^2 \text{ (say)}$$

while

$$EY = \frac{ctf'(t)}{f(t)} \text{ and } VY = c\sigma^2. \tag{3.4}$$

Now, by proper choice of t, the means of Y can be placed at N, and, in view of (3.4), this choice of t must be some number ρ satisfying

$$c\rho f'(\rho) = N f(\rho). \tag{3.5}$$

It can be shown ([2], Theorem 6.2, p. 638) that this equation determines a unique positive real number ρ provided only that N is not the largest possible nor smallest possible (if such there be) value of Y.

Under the choice $t = \rho$, the probability that $Y = N$ is asymptotically $h(0)/\sigma\sqrt{c}$, where the ν's in the expansion of $h(0)$ are those of Y_1. Or, by (3.3),

$$\frac{\rho^N P_N}{[f(\rho)]^c} \sim \frac{h(0)}{\sigma\sqrt{c}},$$

i.e.,

$$P_N \sim \frac{[f(\rho)]^c}{\rho^N \sigma\sqrt{c}} \quad h(0) = \frac{[f(\rho)]^c}{\rho^N \sigma\sqrt{2\pi c}} \left\{ 1 + \frac{1}{c}\left[3!!\, \nu_4 - \frac{5!!}{2!}\, \nu_3^2 \right] + \cdots \right\} \tag{3.6}$$

where as many further terms as desired may be found in (2.8), σ and the ν's belonging to the distribution

$$Prob.\,(y = n) = p_n \rho^n / f(\rho).$$

Thus

$$\sigma = \sqrt{\kappa_2}, \quad \kappa_n = \frac{\partial^n}{\partial u^n} \log f(\rho e^u)\Big|_{u=0}, n > 0,$$

$$\text{and } \nu_n = \kappa_n / n!\, \sigma^n \text{ (Cf. [4], p. 868).} \tag{3.7}$$

It should be noted that for (3.6) to be asymptotic in powers of $1/c$ it is necessary that N do not run off toward one of the tails of Y; the ν's, regarded as functions of ρ, must be independent of c (or at least bounded as $c \to \infty$ (cf. [2], p. 640). If N/c approaches its least or greatest possible value, ρ goes to zero or to infinity, so it must be required that N/c be bounded away from its extremes. (When the series $f(z)$ is infinite in both directions, this requirement is tantamount to the condition given by Good [6], Theorem 6.1, that N/c be held inside a finite interval; but otherwise the latter condition does not ensure any sort of asymptotic convergence.) When N/c tends to an extreme value, and the ν's are unbounded as $c \to \infty$, there still exist circumstances under which (3.6) is asymptotic, even though the residuum is no longer asymptotic in powers of $1/c$; the orders of the ν's in c must then be considered.

Essentially the same asymptotic formula (3.6 & 3.5) for the ordinate of the sum holds for a continuous random variable with frequency $f(x)$ and moments of all orders; it is convenient, before applying (3.6), to make the substitutions

$$\tau = \log \rho, \qquad f(\rho) = f(e^{\tau x}), \tag{3.8}$$

where τ is defined by

$$\frac{\partial}{\partial \alpha} \log \int_{-\infty}^{\infty} e^{\alpha x} f(x)\,\Big|_{\alpha=\tau} = N/c. \tag{3.9}$$

(Cf. Daniels [2.6, p. 633], who points out that the requirements on the existence of moments can be relaxed somewhat). To illustrate the difference in manipulation occasioned by a continuous underlying variable, continuity will be assumed in the ensuing development of the asymptotic formula for the tail area, although the results will be readily applicable to a discrete initial distribution.

4. APPROXIMATE QUADRATURE

By proper choice of scale, the individual random variable ξ may be supposed to have mean 0 and standard deviation 1. Let $f(x)$ be the frequency function of ξ, and assume that the corresponding moment-generating function

$$\mu(\alpha) = E\, e^{\alpha\xi} = \int_{-\infty}^{\infty} e^{\alpha x} f(x)\, dx \tag{4.1}$$

exists for all (or some workable range of) values of α. The tail area to be estimated is the asymptotic probability that the standardized sum

$$\Xi = (\xi_1 + \xi_2 + \cdots + \xi_c)/\sqrt{c} \tag{4.2}$$

will exceed some fixed constant M, where it may be assumed without real loss of generality that $M > 0$, said area being

$$Prob.\,(\Xi > M) = \int_M^{\infty} f_c(x)\, dx, \tag{4.3}$$

where $f_c(x)$ is the frequency of Ξ. Now if $M \leq 1$, formula (2.6) should be applied at once; the development which follows assumes $M > 1$.

Let the (conjugate) random variable η be defined by the frequency function

$$g(x) = e^{ax} f(x)/\mu(a), \tag{4.4}$$

where a is a real parameter. Then

$$E\eta = D_a \log \mu(a) = \mu'(a)/\mu(a) = m \text{ (say)}$$

and

$$V\eta = D_a^2 \log \mu(a) = \sigma^2, \text{ say,} \tag{4.5}$$

and the corresponding moments of the sum G of a sample of c independent observations from the distribution $g(x)$ are

$$EG = cE\eta = cm, \text{ and } VG = c\sigma^2. \tag{4.6}$$

To use G as a shifted version of Ξ, it is necessary to relate their frequency functions. If $f^*(x)$ is the f.f. of $\sum\xi_i$, then f^* is related to f_c by

$$f_c(x) = \sqrt{c}\, f^*(x\sqrt{c}), \tag{4.7}$$

and to $g^*(x)$, the $f.f.$ of G, by

$$f^*(x) = [\mu(a)]^c e^{-ax} g^*(x). \tag{4.8}$$

Hence, finally,

$$f_c(x) = \sqrt{c}\,[\mu(a)]^c e^{-ax\sqrt{c}} g^*(x\sqrt{c}). \tag{4.9}$$

Now

$$H = \frac{G - EG}{S.D.(G)} = \frac{G - cm}{\sigma\sqrt{c}} \tag{4.10}$$

is the standard form of G, with a f.f. $h(x)$ satisfying

$$g^*(x) = \frac{1}{\sigma\sqrt{c}}\, h\, \frac{x - cm}{\sigma\sqrt{c}}. \tag{4.11}$$

Applying (4.9) and (4.11) to (4.3) gives

$$\begin{aligned} Prob.(\Xi > M) &= \frac{[\mu(a)]^c}{\sigma}\int_M^\infty e^{-ax\sqrt{c}}\, h\left(\frac{x - m\sqrt{c}}{\sigma}\right) dx \\ &= \mu^c e^{-acm}\int_{\frac{M-m\sqrt{c}}{\sigma}}^\infty e^{-a\sigma u\sqrt{c}} h(u)\, du, \end{aligned} \tag{4.12}$$

where $u = (x - m\sqrt{c})/\sigma$. But $h(u)$ plays the role of $h(x)$ in (2.5), the ν's becoming invariants of η, and so

$$\begin{aligned} Prob.(\Xi > M) &= \mu^c e^{-acm}\int_\lambda^\infty e^{-\kappa x}\left\{\phi(x) - \frac{\nu_3}{\sqrt{c}}\phi^{(3)}(x) + \cdots\right\} dx, \\ \lambda &= (M - m\sqrt{c})/\sigma, \quad \kappa = a\sigma\sqrt{c}, \quad \mu = \mu(a). \end{aligned} \tag{4.13}$$

The integration follows the formulas

$$\begin{aligned} \int_\lambda^\infty e^{-\kappa x}\phi(x)dx &= e^{\kappa^2/2}[1 - \Phi(\kappa + \lambda)], \\ \int_\lambda^\infty e^{-\kappa x}\phi^{(n)}(x)dx &= e^{-\kappa\lambda}\phi^{(n-1)}(\lambda) + \kappa\int_\lambda^\infty e^{-\kappa x}\phi^{(n-1)}(x)\,dx \end{aligned} \tag{4.14}$$

(the first is found by completing the square in the exponential — the

second by integrating by parts).

Selecting that value τ for a which makes

$$\int_\lambda^\infty e^{-\kappa x}\phi^{(3)}(x)\,dx = 0, \tag{4.15}$$

and carrying out the integration of the series in (4.13) gives

$$Prob.(\Xi > M) = [\mu(\tau)]^c e^{-\tau cm} e^{\kappa^2/2}[1 - \Phi(\kappa + \lambda)] + R, \tag{4.16}$$

where

$$R \sim \int_\lambda^\infty e^{-\kappa u}\left\{\frac{\nu_4}{c}\phi^{(4)}(x) + \frac{\nu_3^2}{2c}\phi^{(6)}(x) - \cdots\right\} du, \ i.e.$$

$$R \sim e^{-\kappa\lambda}\left(\frac{1}{c}\left\{\nu_4\phi^{(3)}(\lambda) + \frac{1}{2}\nu_3^2\left[\phi^{(5)}(\lambda) + \kappa\phi^{(4)}(\lambda) + \kappa^2\phi^{(3)}(\lambda)\right]\right\} - \frac{1}{c^{3/2}}\{\cdots\} + \cdots\right) \tag{4.17}$$

(with as many terms as desired from (2.5) and (4.14)).

$$\mu(\tau) = \int_{-\infty}^\infty e^{\tau x} f(x)\,dx, \quad \kappa = \tau\sigma\sqrt{c}, \quad \lambda = (M - m\sqrt{c})/\sigma,$$

and

$$m = K_1, \quad \sigma = \sqrt{K_2}, \quad \nu_n = K_n/n!\,\sigma^n,$$

where

$$K_n = D_a^n \log \mu(a)\big|_{a=\tau}.$$

The determination of τ may be facilitated by completing the square in the exponent of (4.15) and integrating:

$$\begin{aligned} \int_{-\infty}^\infty e^{-\kappa x}\phi^{(3)}(x)dx = e^{\kappa^2/2}\Big\{&\kappa^3[1 - \Phi(\kappa + \lambda)] - 3\kappa^2\phi(\kappa + \lambda) \\ &- 3\kappa\phi'(\kappa + \lambda) - \phi''(\kappa + \lambda)\Big\}. \end{aligned} \tag{4.18}$$

An upper approximation to τ is obtained by taking $a = \rho$, where ρ (cf. 3.5 and 3.9) satisfies

$$\mu'(\rho)/\mu(\rho) = M/\sqrt{c},$$

because one finds for this choice that the value of the integral (4.18) is small and positive – lying either between or very near one of $1/\sqrt{2\pi}\,(M^2 + 1)$ and $3/\sqrt{2\pi}\cdot(M^2 + 3)$ – whereas the choice $a = 0$ gives $\kappa = 0$, $\lambda = M$ and $-\phi''(M)$ for (4.18), which is negative under the assumption $M > 1$. Thus, by continuity, there is a τ, $0 < \tau < \rho$.

5. APPLICATIONS

Good [6] discusses several applications of the ordinate approximation (3.6). One of the most important of these is the distribution of the multinomial maximum, by which is meant the largest category-frequency arising when a sample of fixed size is drawn from a discrete universe of equally probable categories. Curiously, measuring a suitable ordinate in this application serves to estimate the cumulative distribution function, since the coefficient of z^N in

$$\frac{N!}{c^N}\left(1+z+\frac{z^2}{2!}+\cdots+\frac{z^m}{m!}\right)^c \tag{5.1}$$

is the probability that, in a sample of N from c categories, the maximum will be less than or equal to m ([6], p. 865). Hence, regarding

$$f(z) = 1+z+\cdots+z^m/m! \tag{5.2}$$

as the generating function for a discrete probability distribution over $0,1,\ldots, m$ (logically the polynomial should be divided by the sum of its coefficients, but the extra factor takes care of itself), the coefficient of z^N may be approximated, for large c, by (3.6).

A recurrent statistical problem is to distinguish between two sources (simple hypotheses) by drawing a sample (of fixed size) consisting of independent observations, which contribute log Bayes factors to the odds on the alternative hypothesis over the null hypothesis. The test — in accordance with classical theory — consists of adding up the log Bayes factors applied by the individual observations, selecting the alternative hypothesis when and only when the sum exceeds a preassigned threshold. In repeated testings where the null hypothesis is almost always true, and where *any* occurrence of the alternative is looked for, the threshold is set far out in the tail of the null distribution (the distribution of the sum of scores under the null hypothesis) to prevent a flood of "wrong" answers. Since the proportion of wrong answers that will pass such a threshold cannot be estimated reliably by assuming that the wrong scores are normally distributed, the suggestion was made in [4] that when the threshold falls nearer the "right" mean (the mean total score under the alternative hypothesis) than the wrong mean, it be assumed rather that the *right* scores are normally distributed; the right distribution can be used to integrate the tail of the wrong distribution because the antilog of the score is the ratio of the right to the wrong score. The formula arrived at was

$$Prob.\,(W>x) = exp\left(\frac{1}{2}\sigma^2 - m\right)\left[1-\Phi\left(\frac{x-m}{\sigma}+\sigma\right)\right],$$

where m and σ are the mean and standard deviation of the (total) right score, and W is the score, viz. the sum of the natural logarithms of the Bayes factors in favor of the alternative contributed by the individual observations. While this formula proved adequate for most applications (see [5]), particularly for thresholds near the right mean, the method of this paper is more exact; indeed, the older method is tantamount to a shift of the point of estimation to an unscientifically selected site nearer the mean. A noteworthy third possibility for estimating the reduction is to follow the ingenious empirical procedure suggested by Rudolph McShane in this issue of the *NSA Technical Journal* [7].

6. ACKNOWLEDGMENT

The basic idea — shifting the point of estimation — was suggested informally to the author by A. M. Gleason (ca. 1945), but made little impression until the author saw Good's paper.

BIBLIOGRAPHY

1. Cramer, H. *Mathematical Methods of Statistics*, Princeton University Press, 1946.
2. Daniels, H. E. "Saddlepoint Approximations in Statistics", *Annals of Mathematical Statistics*, 25(1954) 631-650.
3. Dawson, R.
   I. "The Gleason Crutch," NSA Catalog C-1271, 1945.
4. II. "Improvement in the Accuracy of Gleason Crutch Estimates," NSA Catalog C-1282, 1945
5. III. "Empirical Tests of . . .Reduction . . .," NSA Catalog S-2605 1945 *(SECRET)*.
6. Good, I. J., "Saddle-Point Methods for the Multinomial Distribution," *Annals of Mathematical Statistics*, 28(1957) 861-881.
7. McShane, R., "Extreme Values in Cryptanalysis," *NSA Technical Journal*, Vol. IV, No. 4 (this issue) *(CONFIDENTIAL)*.