Statistical Science
1991, Vol. 6, No. 3, 284–302

Closed Form Summation for Classical Distributions: Variations on a Theme of De Moivre

Persi Diaconis and Sandy Zabell

Abstract. De Moivre gave a simple closed form expression for the mean absolute deviation of the binomial distribution. Later authors showed that similar closed form expressions hold for many of the other classical families. We review the history of these identities and extend them to obtain summation formulas for the expectations of all polynomials orthogonal to the constants.

Key words and phrases: Binomial distribution, Stirling's formula, history of probability, Pearson curves, Stein's identity, mean absolute deviation.

1. INTRODUCTION

Let S_n denote the number of successes in n Bernoulli trials with chance p of success at each trial. Thus $P\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k} = b(k;\ n,\ p)$. In 1730, Abraham De Moivre gave a version of the surprising formula

$$(1.1)\quad E\{|S_n - np|\} = 2\upsilon(1-p)\,b(\upsilon;\ n,\ p),$$

where υ is the unique integer such that $np < \upsilon \leq np + 1$. De Moivre's formula provides a simple closed form expression for the mean absolute deviation (MAD) or L_1 distance of a binomial variate from its mean. The identity is surprising, because the presence of the absolute value suggests that expressions for the tail sum $\sum_{k \leq np} b(k;\ n,\ p)$ might be involved, but there are no essential simplifications of such sums (see, e.g., Zeilberger, 1989).

Dividing (1.1) by n, and using the result that the modal term of a binomial tends to zero with increasing n, it follows that

$$(1.2)\qquad E\left\{\left|\frac{S_n}{n} - p\right|\right\} \to 0.$$

De Moivre noted this form of the law of large numbers and thought it could be employed to justify passing from sample frequencies to population proportions. As he put it (De Moivre, 1756, page 242):

> COROLLARY. *From this it follows, that if after taking a great number of experiments, it should be perceived that the happenings and failings have been nearly in a certain proportion, such as of 2 to 1, it may safely be concluded that the probabilities of happening or failing at any one time assigned will be very near in that proportion, and that the greater the number of experiments has been, so much nearer the truth will the conjectures be that are derived from them.*

Understanding the asymptotics of (1.2) in turn led De Moivre to his work on approximations to the central term of the binomial. In Section 2, we discuss this history and argue that it was De Moivre's work on this problem that ultimately led to his proof of the normal approximation to the binomial.

De Moivre's formula is at once easy enough to derive that many people have subsequently rediscovered it, but also hard enough to have often been considered worth publishing, varying and generalizing. In Section 3, we review these later results and note several applications: one to bounding binomial tail sums, one to the Bernstein polynomial version of the Weierstrass approximation theorem and one to proving the monotonicity of convergence in (1.2).

Persi Diaconis is Professor of Mathematics at Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138. Sandy Zabell is Professor of Mathematics and Statistics at Northwestern University, Evanston, Illinois 60208.

In the second half of this article, we offer a generalization along the following lines: De Moivre's result works because $\sum_a^b (k - np)b(k; n, p)$ can be summed in closed form for any a and b. The function $x - np$ is the first orthogonal polynomial for the binomial distribution. We show that in fact *all* orthogonal polynomials (except the zeroth) admit similar closed form summation. The same result holds for many of the other standard families (normal, gamma, beta and Poisson). There are a number of interesting applications of these results that we discuss, and in particular, there is a surprising connection with Stein's characterization of the normal and other classical distributions.

De Moivre's formula arose out of his attempt to answer a question of Sir Alexander Cuming. Cuming was a colorful character whose life is discussed in a concluding postscript.

PART 1: DE MOIVRE'S FORMULA AND ITS DESCENDANTS

2. CUMING'S PROBLEM AND DE MOIVRE'S L_1 LIMIT THEOREM

Abraham De Moivre (1667–1754) wrote one of the first great books on probability, *The Doctrine of Chances*. First published in 1718, with important new editions in 1738 and 1756, it contains scores of important results, many in essentially their modern formulation. Most of the problems considered by De Moivre concern questions that arise naturally in the gambling context. Problem 72 of the third edition struck us somewhat differently:

> A and B playing together, and having an equal number of Chances to win one Game, engage to a Spectator S that after an even number of Games n is over, the Winner shall give him as many Pieces as he wins Games over and above one half the number of Games played, it is demanded how the Expectation of S is to be determined.

In a modern notation, De Moivre is asking for the expectation $E\{\,|\,S_n - n/2\,|\,\}$. In *The Doctrine of Chances*, De Moivre states that the answer to the question is $(n/2)E/2^n$, where E is the middle term of the binomial expansion of $(1 + 1)^n$, that is, $\binom{n}{n/2}$. De Moivre illustrates this result for the case $n = 6$ (when $E = 20$ and the expectation is 15/16).

Problem 73 of *The Doctrine of Chances* then gives equation (1.1) for general values of p (De Moivre worked with rational numbers). At the conclusion of Problem 73, De Moivre gives the Corollary quoted earlier. Immediately following this De Moivre moves on to the central limit theorem.

We were intrigued by De Moivre's formula. Where had it come from? Problem 73, where it appears, is scarcely a question of natural interest to the gamblers De Moivre might have spoken to, unlike most of the preceding questions discussed in the *Doctrine of Chances*. And where had it gone? Its statement is certainly not one of the standard identities one learns today.

2.1 The Problem of Sir Alexander Cuming

Neither the problem nor the formula appear in the 1718 edition of *The Doctrine of Chances*. They are first mentioned by De Moivre in his *Miscellanea Analytica* of 1730, a Latin work summarizing his mathematical research over the preceding decade (De Moivre, 1730). De Moivre states there (page 99) that the problem was initially posed to him in 1721 by Sir Alexander Cuming, a member of the Royal Society.

In the *Miscellanea Analytica*, De Moivre gives the solution to Cuming's problem (pages 99–101), including a proof of the formula in the symmetric case (given below in Section 2.3), but he contents himself with simply stating without proof the corresponding result for the asymmetric case. These two cases then appear as Problems 86 and 87 in the 1738 edition of the *Doctrine of Chances*, and Problems 72 and 73 in the 1756 edition.

As De Moivre notes in the *Doctrine of Chances* (1756, pages 240–241), the expectation of $|S_n - np|$ *increases* with n, but *decreases* proportionately to n; thus he obtains for $p = \frac{1}{2}$ the values in Table 1. (De Moivre's values for $E|S_n - np|$ are inaccurate in some cases (e.g., $n = 200$) in the third or fourth decimal place.)

A proof of monotonicity is given in Theorem 3 of Section 3.2 below. De Moivre does not give a proof in either the symmetric or asymmetric cases, and it is unclear whether he had one, or even whether he intended to assert monotonicity rather than simply limiting behavior.

Had De Moivre proceeded no further than this, his formula would have remained merely an

TABLE 1
Exact values of mean absolute deviation

n	$E\|S_n - np\|$	$E\|S_n - np\|/n$
6	0.9375	0.1563
12	1.3535	0.1128
100	3.9795	0.0398
200	5.6338	0.0282
300	6.9041	0.0230
400	7.9738	0.0199
500	8.9161	0.0178
700	10.552	0.0151
800	11.280	0.0141
900	11.965	0.0133

interesting curiosity. But, as we will now show, De Moivre's work on Cuming's problem led *directly* to his later breakthrough on the normal approximation to the binomial and here, too, the enigmatic Sir Alexander Cuming played a brief, but vital, role.

2.2 "... the hardest Problem that can be proposed on the Subject of Chance"

After stating the Corollary quoted earlier, De Moivre noted that substantial fluctuations of S_n/n from p, even if unlikely, were still possible and that it was desirable, therefore, that "the Odds against so great a variation ... should be assigned"; a problem which he described as "the hardest Problem that can be proposed on the Subject of Chance" (De Moivre, 1756, page 242).

But initially, perhaps precisely because he viewed the problem as being so difficult, De Moivre seems to have had little interest in working on the questions raised by Bernoulli's proof of the law of large numbers. No discussion of Bernoulli's work occurs in the first edition of the *Doctrine of Chances*; and, in its preface, De Moivre even states that, despite the urging of both Montmort and Nicholas Bernoulli that he do so, "I willing resign my share of that Task into better Hands" (De Moivre, 1718, page xiv).

What then led De Moivre to reverse himself only a few years later and take up a problem that he appears at first to have considered both difficult and unpromising? Surprisingly, it is possible to give a definitive answer to this question.

De Moivre's solution to Cuming's problem requires the numerical evaluation of the middle term of the binomial. This is a serious computational drawback, for, as De Moivre himself noted, the direct calculation of the term for large values of n (the example that he gives is $n = 10{,}000$) "is not possible without labor nearly immense, not to say impossible" (De Moivre, 1730, page 102).

But this did not discourage the irrepressible Sir Alexander Cuming, who seems to have had a talent for goading people into attacking problems they otherwise might not. (Our concluding postscript gives another example.) Let De Moivre tell the story himself, in a passage from the Latin text of the *Miscellanea Analytica*, which has not, to our knowledge, been commented on before (De Moivre, 1730, page 102):

> Because of this, the man I praised above [*vir supra laudatus*; i.e., Cuming] asked me whether it was not possible to think of some method [*num possem methodum aliquam excogitare*] by which that term of the binomial could be determined without the trouble of multiplication or, what would come to the same thing in the end, addition of logarithms. I responded that if he would permit it, I would attempt to see what I could do in his presence, even though I had little hope of success. When he assented to this, I set to work and within the space of one hour I had very nearly arrived at the solution to the following problem [*intra spatium unius circiter horae, eò perduxi ut potuerim solutionem sequentis Problematis prope elicere*].

This problem was "to determine the coefficient of the middle term of a very large even power, or to determine the ratio which the coefficient of the middle term has to the sum of all coefficients"; and the solution to it that De Moivre found in 1721, the asymptotic approximation

$$\frac{1}{2^n}\binom{n}{n/2} \doteq 2\frac{21}{125}\left(1-\frac{1}{n}\right)^n/\sqrt{n-1}$$

to the central term of the binomial, was the first step on a journey that led to his discovery of the normal approximation to the binomial 12 years later in 1733 (Schneider, 1968, pages 266–275, 292–300; Stigler, 1986, pages 70–88; Hald, 1990, pages 468–495). The 1721 date for the initial discovery is confirmed by De Moivre's later statement regarding the formula, in his privately circulated note of November 12, 1733, the *Approximatio ad Summam Terminorum Binomii* $(a+b)^N$ *in Seriem Expansi* that "it is now a dozen years or more since I had found what follows" (De Moivre, 1756, page 243).

Thus De Moivre's work on Cuming's problem led him immediately to the L_1 law of large numbers for Bernoulli trials, and eventually to the normal approximation to the binomial distribution. He appears to have regarded the two as connected, the second a refinement of the first. But there is one feature about De Moivre's train of thought that is puzzling. How did he make the leap from

$$E\left\{\left|\frac{S_n}{n}-p\right|\right\}\to 0 \quad \text{to} \quad P\left\{\left|\frac{S_n}{n}-p\right|>\epsilon\right\}\to 0?$$

De Moivre certainly knew the second statement from his work on the normal approximation to the binomial, as well as from Bernouli's earlier work on the law of large numbers. But more than 120 years would have to elapse before Chebychev's inequality would allow one to easily reach the second conclusion from the first.

Of course, the currently recognized modes of convergence were not well delineated in De Moivre's time. One can find him sliding between the weak and strong laws in several places. His statement of

$\frac{mp+l}{fp-l+1} \times \frac{mp+1}{mp} \times \frac{fp}{mp}$, & $\frac{fp+l}{mp-l+1} \times \frac{fp+1}{fp} \times \frac{mp}{fp}$, quæ ſunt eæ ipſæ quantitates, minori quarum æquanda eſt quantitas q.

In ſpeciali applicatione ad numeros, poſitis ut prius $n=14000$, $mp=72000$, $fp=6800$, invenit priorem harum quantitatum $= 44\frac{74}{100}$, poſteriorem vero $=44\frac{58}{100}$, cui utpote minori ponit q æqualem, ideoque $q-1=43\frac{58}{100}$, unde demum concludit fore ut ſi fiant experimenta 14000, probabilius futurum ſit ad minimum in ea ratione quam habet $43\frac{58}{100}$ ad 1, eventum illum cujus contingentia in ſingulis experimentis eſt ad non-contingentiam ut 18 ad 17, neque ſæpius ſe oſtenſurum quam 7363 vices, neque rarius quam 7037.

CAPUT II.

CUM aliquando labente Anno 1721, Vir Clariſſimus *Alex. Cuming* Eq. Au. Regiæ Societatis Socius, quæſtionem infra ſubjectam mihi propoſuiſſet, ſolutionem problematis ei poſtero die tradideram.

PROBLEMA I.

Colluſores duo A *&* B *quorum dexteritates ponantur æquales, ſpectatori cuidam ita ſe obſtringant, ut poſt elapſum numerum* n *ludorum parem, uter victorem ſe præſtiterit, is ei tot nummos largiturus ſit quot plures ludos vicerit quam qui deſignentur per* $\frac{1}{2}n$; *quæritur quanti æſtimanda ſit Expectatio Spectatoris.*

SOLUTIO.

Denotet E medium terminum Binomii $a+b$ ad poteſtatem n evecti, poſitis ſigillatim a & $b=1$. tunc erit $\frac{\frac{1}{2}nE}{2^n}$, Expectatio quæſita.

INVESTIGATIO.

Sit E terminus medius, ſeu potius Coefficiens termini medii poteſtatis $\overline{a+b}^n$, D & F Coefficientes terminorum hinc inde medio proxime adſtantium, C & G Coefficientes terminorum binis intervallis a medio diſtantium, & ſic deinceps pergendo utrinque a medio ad utrumque extremorum.

Jam ex formatione Binomii, palam eſt terminum medium eos caſus deſignaturum quibus accidere poſſit, ut neuter Colluſorum alteri præpolleat, terminos huic proxime adſtantes eos caſus deſignaturos quibus poſſit evenire, ut alter alterum ſit ſuperaturus ludis binis, ſeu ut eorum alter numerum $\frac{1}{2}n$ ſit ſuperaturus ludo uno; terminos his deinde proximos deſignaturos hos caſus quibus accidere poſſit, ut eorum alter alterum ſit ſuperaturus ludis quaternis, ſeu ut numerum $\frac{1}{2}n$ ſuperaturus ſit ludis binis, & ſic deinceps progrediendo ad terminos extremos.

Erit igitur Expectatio ſpectatoris in ludorum numero pari $=E\times 0+\overline{D+F}\times 1+\overline{C+G}\times 2+\overline{B+H}\times 3+\overline{A+K}\times 4$ &c. ſive propter Æqualitatem Coefficientium hinc inde a Medio æqualiter diſtantium, erit Expectatio ſpectatoris $= E\times 0+2D+4C+6B+8A$ &c.

Sed ex Proprietate Coefficientium, invenietur eſſe

$\overline{n+2}\times D=nE$
$\overline{n+4}\times C=\overline{n-2}\times D$
$\overline{n+6}\times B=\overline{n-4}\times C$
$\overline{n+8}\times A=\overline{n-6}\times B$
$\overline{n+10}\times 0=\overline{n-8}\times A$
&c. &c.

Jam cum ſumma prioris Columnæ æqualis ſit ſummæ poſterioris, erit

$nD+nC+nB+nA$ &c. $=$ $nE+nD+nC+nB+nA$ &c.
$+2D+4C+6B+8A$ $-2D-4C-6B-8A$

tum deletis hinc inde terminis æqualibus, cæteriſque ad eandem partem tranſpoſitis, fiet

FIG. 1. *Pages 99 and 100 of De Moivre's Miscellanea Analytica de Seriebus et Quadraturis show the proof reproduced here in Section 2.3. By permission of the Houghton Library, Harvard University.*

the corollary: "the happenings and failings have been nearly in a certain proportion," has a clear element of fluctuation in it. In contrast, even today L_1 convergence has a distant, mathematical flavor to it. It is intriguing that De Moivre seemed to give it such a direct interpretation.

2.3 De Moivre's Proof

De Moivre's proof that $E[\,|S_n - n/2|\,] = (1/2)nE/2^n$ is simple but clever, impressive if only because of the notational infirmities of his day. Since it only appears in the Latin of the *Miscellanea Analytica* (Fig. 1) and is omitted from *The Doctrine of Chances*, we reproduce the argument here.

DE MOIVRE'S PROOF OF FORMULA (1.1), CASE $p = 1/2$. Let E denote the "median term" (*terminus medius*) in the expansion of $(a+b)^n$, D and F the coefficients on either side of this term, C and G the next pair on either side, and so on. Thus the terms are $\ldots, A, B, C, D, E, F, G, H, K, \ldots$.

The expectation of the spectator after an even number of games is

$$E\times 0 + (D+F)\times 1 + (C+G)\times 2 + (B+H)\times 3 + (A+K)\times 4 + \cdots .$$

Because the binomial coefficients at an equal distance from either side of the middle are equal, the expectation of the spectator reduces to

$$0E + 2D + 4C + 6B + 8A + \cdots .$$

But owing to the properties of the coefficients, it follows that

$$(n+2)D = nE$$
$$(n+4)C = (n-2)D$$
$$(n+6)B = (n-4)C$$
$$(n+8)A = (n-6)B$$
$$\cdots$$

Setting equal the sum of the two columns then yields

$$\begin{aligned} & nD + nC + nB + nA + \cdots \\ & +2D + 4C + 6B + 8A \cdots \\ & \quad = nE + nD + nC + nB + nA + \cdots \\ & \qquad -2D - 4C - 6B - 8A - \cdots . \end{aligned}$$

Deleting equal terms from each side, and transposing the remainder, we have

$$4D + 8C + 12B + 16A + \cdots = nE$$

or

$$2D + 4C + 6B + 8A + \cdots = \frac{1}{2}nE.$$

Since the probabilities corresponding to each coefficient result from dividing by $(a + b)^n$, here $(1 + 1)^n = 2^n$, De Moivre's theorem follows. □

REMARK. For a mathematician of his stature, surprisingly little has been written about De Moivre. Walker's brief article in *Scripta Mathematica* (Walker, 1934) gives the primary sources for the known details of De Moivre's life; other accounts include those of Clerke (1894), David (1962, pages 161–178), Pearson (1978, pages 141–146) and the *Dictionary of Scientific Biography*.

Schneider's detailed study (Schneider, 1968) provides a comprehensive survey of De Moivre's mathematical research. During the last two decades, many books and papers have appeared on the history of probability and statistics, and a number of these provide extensive discussion and commentary on this aspect of De Moivre's work; these include most notably, the books by Stigler (1986) and Hald (1990). Other useful discussions include those of Daw and Pearson (1972), Adams (1974), Pearson (1978, pages 146–166), Hald (1984, 1988) and Daston (1988, pages 250–253).

3. LATER PROOFS, APPLICATIONS AND EXTENSIONS

3.1 Later Proofs

De Moivre did not give a proof of his expression for the MAD in the case of the asymmetrical binomial (although he must have known one). This gap was filled by Isaac Todhunter (1865, pages 182–183) who supplied a proof in his discussion of this portion of De Moivre's work.

Todhunter's proof proceeds by giving a closed form expression for a sum of terms in the expectation, where the sum is taken from the outside in. We abstract the key identity in modern notation.

LEMMA 1 (Todhunter's Formula). *For all integers* $0 \le \alpha < \beta \le n$,

$$\sum_{k=\alpha}^{\beta} (k - np)b(k; n, p) = \alpha qb(\alpha; n, p) - (n - \beta)pb(\beta; n, p).$$

PROOF. Because $p + q = 1$,

$$\begin{aligned} \sum_{k=\alpha}^{\beta} (k - np)b(k; n, p) &= \sum_{k=\alpha}^{\beta} \{kq - (n - k)p\}b(k; n, p) \\ &= \sum_{k=\alpha}^{\beta} kqb(k; n, p) \\ &\quad - \sum_{k=\alpha}^{\beta} (n - k)pb(k; n, p). \end{aligned}$$

But $(k + 1)qb(k + 1; n, p) = (n - k)pb(k; n, p)$; thus every term in the first sum (except the lead term) is canceled by the preceding term in the second sum, and the lemma follows. □

We know of no proof for the $p \neq 1/2$ case prior to that given in Todhunter's book. Todhunter had an encyclopedic knowledge of the literature, and it would have been consistent with his usual practice to mention further work on the subject if it existed. He (in effect) proved his formula by induction.

Todhunter assumed, however, as did De Moivre, that np is integral (although his proof does not really require this); and this restriction can also be found in Bertrand (1889, pages 82–83). Bertrand noted that if $q = 1 - p$ and

$$F(p, q) =: \sum_{k > np} \binom{n}{k} p^k q^{n-k},$$

then the mean absolute deviation could be expressed as $2pq\left\{\frac{\partial F}{\partial p} - \frac{\partial F}{\partial q}\right\}$, and that term-by-term cancellation then leads to De Moivre's formula. The first discussion we know of giving the general formula without any restriction is in Poincaré's book (1896, pages 56–60; 1912, pages 79–83): if v is the first integer greater than np, then the mean absolute deviation is given by $2vqb(v; n, p)$. Poincaré's derivation is based on Bertrand's but is a curiously fussy attempt to fill what he apparently viewed as logical lacunae in Bertrand's proof. The derivation later appears in Uspensky's book as a problem (Uspensky, 1937, pages 176–177), possibly by the route Poincaré (1896) → Czuber (1914, pages 146–147) → Uspensky (1937).

De Moivre's identity has been rediscovered many times since. Frisch (1924, page 161) gives the Todhunter formula and deduces the binomial MAD formula as an immediate consequence. This did not stem the flow of rediscovery, however. In 1930, Gruder (1930) rediscovered Todhunter's formula, and in 1957 Johnson, citing Gruder, noted its application to the binomial MAD. Johnson's (1957) article triggered a series of generalizations. The MAD

formula was also published in Frame (1945). None of these authors connected the identity to the law of large numbers so it remained a curious fact.

REMARK. The formula for the mean absolute deviation of the binomial distribution can be expressed in several equivalent forms which are found in the literature. If v is the least integer greater than np and $Y_{n,p}$ is the central term in the expansion of $(p+q)^n$, then the mean absolute derivation equals

$2vqb(v; n, p)$ (Poincaré, 1896; Frisch, 1924; Feller, 1968)

$= 2npqb(v-1; n-1, p)$ (Uspensky, 1937)

$= 2npqY_{n-1}$ (Frame, 1945)

$= 2v\binom{n}{v}p^v q^{n-v+1}$ (Johnson, 1957).

In his solution to Problem 73, De Moivre states that one should use the binomial term $b(j; n, p)$ for which $j/(n-j) = p/(1-p)$; since this is equivalent to taking $j = np$, the solution tacitly assumes that np is integral. In this case $b(j; n, p) = b(j; n-1, p)$ and $j = v-1$, hence

$$2npqb(j; n, p) = 2npqb(v-1; n-1, p);$$

thus the formula given by De Moivre agrees with the second of the standard forms.

3.2 Applications

Application 1. As a first application we give a binomial version of Mills ratio for binomial tail probabilities.

THEOREM 1. *For $\alpha > np$, $n \geq 1$ and $p \in (0, 1)$,*

$$\frac{\alpha}{n} \leq \frac{1}{b(\alpha; n, p)} \sum_{k=\alpha}^{n} b(k; n, p) \leq \frac{\alpha(1-p)}{\alpha - np}.$$

PROOF. For the upper bound, use Lemma 1 to see that

$$\begin{aligned}\alpha \sum_{k=\alpha}^{n} b(k; n, p) &\leq \sum_{k=\alpha}^{n} kb(k; n, p) \\ &= np \sum_{k=\alpha}^{n} b(k; n, p) + \alpha qb(\alpha; n, p).\end{aligned}$$

The lower bound follows similarly. □

REMARK. The upper bound is given in Feller (1968, page 151). Feller gives a much cruder lower bound. Slightly stronger results follow from Markov's continued fraction approach, see Uspensky (1937, pages 52–56). As usual, this bound is poorest when α is close to np. For example, when $p = 1/2$, and $\alpha = [n/2] + 1$, the ratio is of order $\sqrt{n}$ while the lower bound is approximately $1/2$ and the upper bound is approximately $n/4$. The bound is useful in the tails. Similar bounds follow for other families which admit a closed form expression for the mean absolute deviation.

Application 2. De Moivre's formula allows a simple evaluation of the error term in the Bernstein polynomial approximation to a continuous function. Lorentz (1986) or Feller (1971, Chapter 8) give the background to Bernstein's approach.

Let f be a continuous function on [0, 1]. Bernstein's proof of the Weierstrass approximation theorem approximates $f(x)$ by the Bernstein polynomial

$$B(x) = \sum_{i=0}^{n} f\left(\frac{i}{n}\right)\binom{n}{i}x^i(1-x)^{n-i}.$$

The quality of approximation is often measured in terms of the modulus of continuity:

$$\omega_f(\delta) = \sup_{|x-y| \leq \delta} |f(y) - f(x)|.$$

With this notation, we can state the following theorem.

THEOREM 2. *Let f be a continuous function on the unit interval. Then for any $x \in [0, 1]$*

$$|f(x) - B(x)| \leq \omega_f\left(\frac{1}{\sqrt{n}}\right)\left(1 + \frac{2v(1-x)}{\sqrt{n}}\, b(v; n, x)\right)$$

with $nx < v < nx + 1$.

PROOF. Clearly

$$|f(x) - B(x)| \leq \sum_{i=0}^{n} \left|f(x) - f\left(\frac{i}{n}\right)\right|\binom{n}{i}x^i(1-x)^{n-i}.$$

For any $\delta \in (0, 1)$, dividing the interval between x and i/n into subintervals of length smaller than δ shows

$$|f(x) - f(i/n)| \leq \omega_f(\delta)\left(1 + \frac{|x - i/n|}{\delta}\right).$$

Using this and De Moivre's formula gives the theorem, taking $\delta = 1/\sqrt{n}$. □

REMARK. (1) Lorentz (1986, pages 20, 21) gives $|f(x) - B(x)| \leq \frac{5}{4}\omega_f(1/\sqrt{n})$. Lorentz shows that the function $f(x) = |x - \frac{1}{2}|$ has $|f(x) - B(x)| \geq \frac{1}{2}\omega_f(1/\sqrt{n})$ so the $1/\sqrt{n}$ rate is best possible.

(2) To get a uniform asymptotic bound from Theorem 2, suppose n is odd. Then Blyth (1980) shows

that the mean absolute deviation (given by formula (1.1)) is largest for $p = \frac{1}{2}$. The upper bound in theorem 2 becomes

$$|f(x) - B(x)| \le \omega_f\left(\frac{1}{\sqrt{n}}\right) \cdot \left(1 + \frac{(n+1)}{2\sqrt{n}}\, b\left(\frac{n+1}{2}; n, \frac{1}{2}\right)\right).$$

By Stirling's formula the right hand side is asymptotic to $\omega_f\left(\frac{1}{\sqrt{n}}\right)\left(12 + \frac{1}{\sqrt{2\pi}}\right)$.

(3) Bernstein polynomials are useful in Bayesian statistics because of their interpretation as mixtures of beta distributions (see Dallal and Hall, 1983; Diaconis and Ylvisaker, 1985). The identities for other families presented in Section 4 can be employed to give similar bounds for mixtures of other familes of conjugate priors.

Application 3. As a final application, we apply the general form of De Moivre's formula (1.1) to show that the MAD of S_n is increasing in n, but that the MAD of S_n/n is decreasing in n. For S_n, let $v_n = [np + 1] = [np] + 1$, so that $np < v_n \le np + 1$.

THEOREM 3. *Let* $S_n \sim B(n, p)$ *and* $M_n =: E[|S_n - np|]$. *If* p *is fixed, then for every* $n \ge 1$,

(3.1) $M_n \le M_{n+1}$, *with equality precisely when* $(n+1)p$ *is integral*;

(3.2) $\frac{M_n}{n} \ge \frac{M_{n+1}}{n+1}$, *with equality precisely when* np *is integral.*

PROOF. It is necessary to consider two cases.

Case 1. $v_n = v_{n+1}$. Then by the general form of De Moivre's formula

$$\frac{M_{n+1}}{M_n} = \frac{(n+1)q}{n+1-v_n}$$

and

$$\frac{M_{n+1}/(n+1)}{M_n/n} = \frac{nq}{n+1-v_n}.$$

But $(n+1)p < [(n+1)p + 1] = v_{n+1} = v_n$, hence $n + 1 - v_n < (n+1)q$, so that $M_{n+1}/M_n > 1$. Similarly, $v_n \le np + 1$, hence $nq \le n + 1 - v_n$, and inequality (3.2) follows, with equality if and only if $np + 1$, hence np is integral.

Case 2. $v_n < v_{n+1}$. In this case, by De Moivre's formula,

$$\frac{M_{n+1}}{M_n} = \frac{(n+1)p}{v_n}$$

and

$$\frac{M_{n+1}/(n+1)}{M_n/n} = \frac{np}{v_n}.$$

Since $v_n < v_{n+1}$, clearly $v_n = v_{n+1} - 1 = [(n+1)p] \le (n+1)p$, and inequality (3.1) follows, with equality if and only if $(n+1)p$ is integral. Since $np < v_n$, inequality (3.2) follows immediately, and the inequality is strict.

Since np integral implies $v_n = v_{n+1}$, and $(n+1)p$ integral implies $v_n < v_{n+1}$, the theorem follows. □

REMARK. De Moivre's formula can be applied outside the realm of limit theorems. In a charming article, Blyth (1980) notes that the closed form expansion for the MAD has a number of interesting applications. If S_n is a binomial random variable with parameters n and p, the deviation $E|S_n/n - p|$ represents the risk of the maximum likelihood estimator under absolute value loss. As p varies between 0 and $\frac{1}{2}$, the risk is roughly monotone but, if $n = 4$, $p = \frac{1}{4}$, the estimate does better than for nearby values of p. Lehmann (1983, page 58) gives De Moivre's identity with Blyth's application.

3.3 Extensions to Other Families

De Moivre's identity can be stated approximately thus: For a binomial variate, the mean absolute deviation equals twice the variance times the density at the mode. It is natural to inquire whether such a simple relationship exists between the variance σ^2 and the mean absolute deviation μ_1 for families other than the binomial. This simple question appears to have been first asked and answered in 1923 by Ladislaus von Bortkiewicz. If $f(x)$ is the density function of a continuous distribution with expectation μ, von Bortkiewicz showed that the ratio $R =: \mu_1/2\sigma^2 f(\mu)$ is unity for the gamma ("De Forestsche"), normal ("Gaussche"), chi-squared ("Helmertsche") and exponential ("zufälligen Abstände massgebende") distributions ("Fehlergesetz"); while it is $(\alpha + \beta + 1)/(\alpha + \beta)$ for the beta distribution ("Pearsonsche Fehlergesetz") with parameters α and β.

Shortly after von Bortkiewicz's paper appeared, Karl Pearson noted that the continuous examples considered by von Bortkiewicz could be treated in a unified fashion by observing that they were all members of the Pearson family of curves (Pearson, 1924). If $f(x)$ is the density function of a continu-

ous distribution, then $f(x)$ is a member of this family if it satisfies the differential equation

$$(3.3) \qquad \frac{f'(x)}{f(x)} = \frac{x + a}{b_0 + b_1 x + b_2 x^2}.$$

Then, letting $p(x) = b_0 + b_1 x + b_2 x^2$, it follows that

$$(fp)'(x) = f(x)\{(1 + 2b_2)x + (a + b_1)\}.$$

If

$$(3.4) \quad b_2 \neq -\frac{1}{2}, \text{ and } f(x)p(x) \to 0 \text{ as } x \to \pm\infty,$$

then integrating from $-\infty$ to ∞ yields

$$\mu = -\frac{a + b_1}{1 + 2b_2},$$

so that

$$f(x)\{x - \mu\} = \frac{(fp)'(x)}{1 + 2b_2}$$

and

$$(3.5) \qquad \int_{-\infty}^{t} (x - \mu) f(x)\, dx = \frac{f(t)p(t)}{1 + 2b_2}.$$

This gives the following result.

PROPOSITION 1. *If f is a density from the Pearson family* (3.3) *with mean μ and* (3.4) *is satisfied, then*

$$\int_{-\infty}^{\infty} |x - \mu| f(x)\, dx = -2f(\mu)\left\{\frac{b_0 + b_1\mu + b_2\mu^2}{1 + 2b_2}\right\}.$$

REMARK. If $\beta_1 =: \mu_3^2/\mu_2^3$ and $\beta_2 =: \mu_4/\mu_2^2$ denote the coefficients of skewness and kurtosis, then, as Pearson showed, this last expression may be re-expressed as

$$\mu_1 = C2\sigma^2 f(\mu) \quad \text{where} \quad C = \frac{4\beta_2 - 3\beta_1}{6(\beta_2 - \beta_1 - 1)}.$$

The constant $C = 1 \Leftrightarrow 2\beta_2 - 3\beta_1 - 6 = 0$, which is the case when the underlying distribution is normal or Type 3 (gamma). We give further results for Pearson curves in the next section.

Just as with De Moivre's calculation of the MAD for the binomial, the von Bortkiewicz–Pearson formulas were promptly forgotten and later rediscovered. Ironically, this would happen in Pearson's own journal. After the appearance in 1957 of Johnson's *Biometrika* paper on the binomial, a series of further papers appeared over the next decade which in turn rediscovered the results of von Bortkiewicz and Pearson: Ramasubban (1958) in the case of the Poisson distribution and Kamat (1965, 1966a) in the case of the Pearson family; see also the articles by Johnson (1958) Bardwell (1960) and Kamat (1966b).

PART 2: CLOSED FORM SUMMATION FOR CLASSICAL DISTRIBUTIONS

4. DE MOIVRE'S IDENTITY AND ORTHOGONAL POLYNOMIALS

De Moivre's identity follows from a closed form expression for the sum $\sum_{k=0}^{a} (k - np)b(k; n, p)$. The function $k \to k - np$ is the first orthogonal polynomial for the binomial distribution. In this part, we show that *all* of the orthogonal polynomials, except the zeroth, admit similar closed form partial sums and that the same holds true for the other classical distributions as well.

Passage to the limit shows such identities must hold for the orthogonal polynomials associated to the normal distribution (the Hermite polynomials). The arguments are clearest here, so we begin with this case in Section 4.1. A variety of applications are presented. Most notably, the identities give a singular value decomposition for an operator associated to "Stein's method" for proving limit theorems and finding unbiased estimates of risk. In Sections 4.2 and 4.3, we then show how very similar arguments permit the derivation of corresponding results in the case of the gamma and beta distributions, where the appropriate orthogonal polynomials are the Laguerre and Jacobi polynomials, respectively.

The occurrence of these three special families of orthogonal polynomials and distributions is not an accident. The Hermite, Laguerre and Jacobi polynomials form the three classical families of orthogonal polynomials, known to satisfy many important and special properties; and the normal, gamma and beta families are precisely those members of the Pearson family for which orthogonal polynomials of all orders exist. This connection is spelled out in Section 5.

Finally, corresponding results are derived for two families of discrete distributions: in Section 6.1, we discuss the Poisson distribution, and then, in Section 6.2, we finally return to where we began: the binomial.

4.1 The Normal Density and Hermite Polynomials

A familiar theorem says that the integral

$$\int_{-\infty}^{a} e^{-x^2/2}\, dx$$

cannot be written as an elementary function of a. Rosenlicht (1976) gives an accessible account in modern language. Of course, certain indefinite nor-

mal integrals can be simply evaluated, for example $\int_{-\infty}^{a} xe^{-x^2/2}dx$. The following lemma determines all polynomials whose integral can be so evaluated.

Recall first that the Hermite polynomials are the orthogonal polynomials on $\mathbb{R}$ with respect to the kernel $e^{-x^2/2}\,dx$ They are given by the explicit formula ($n = 0, 1, 2, \cdots$)

$$(4.1)\quad H_n(x) = n! \sum_{k=0}^{[\frac{n}{2}]} \frac{(-1)^k(\sqrt{2}\,x)^{n-2k}}{(n-2k)!\,k!}.$$

Thus $H_0 = 1$, $H_1 = \sqrt{2}\,x$, $H_2 = 2(x^2 - 1)$, $H_3 = 2\sqrt{2}\,(x^3 - 3x), \ldots$. They satisfy the relation

$$(4.2)\quad \int_{-\infty}^{\infty} H_r(x)H_s(x)e^{-x^2/2}dx = \sqrt{2\pi}\,2^r r!\delta_{rs}.$$

Background and standard properties of Hermite polynomials can be found in Chihara (1978), an excellent introduction to the subject of orthogonal polynomials.

The basic identity needed is the following.

LEMMA 1. *For $n \geq 1$ and any real a,*

$$\int_{-\infty}^{a} H_n(x)e^{-x^2/2}\,dx = -\sqrt{2}\,H_{n-1}(a)e^{-a^2/2}.$$

PROOF. The Hermite polynomials can be represented by the Rodrigues formula as

$$H_n(x) = (-\sqrt{2})^n e^{x^2/2}D^n e^{-x^2/2}$$

with D^n denoting n-fold differentiation. From this

$$\int_{-\infty}^{a} H_n(x)e^{-x^2/2}\,dx = (-\sqrt{2})^n \int_{-\infty}^{a} D^n e^{-x^2/2}\,dx$$
$$= -\sqrt{2}\,H_{n-1}(x)e^{-x^2/2}\Big|_{-\infty}^{a}. \quad \square$$

COROLLARY 1. *A polynomial $p(x)$ can be integrated against $e^{-x^2/2}$ in finite terms if and only if $p(x)$ is orthogonal to the constants in $L^2(e^{-x^2/2})$.*

EXAMPLE 1. The analog of De Moivre's identity for the standard normal distribution takes the form

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x|\, e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}};$$

this follows from Lemma 1, with $n = 1$ and $a = 0$.

EXAMPLE 2. In deriving a total variation bound on the binomial approximation to the hypergeometric distribution, Diaconis and Freedman (1980) encountered the identity

$$\int_{-\infty}^{\infty} |x^2 - 1|\, e^{-x^2/2}\,dx = 4e^{-1/2}.$$

It seemed surprising that such a normal integral could be evaluated. The Corollary clarifies the reason why such a formula exists: the polynomial part of the integrand involves $H_2(x)/2$.

EXAMPLE 3 (Stein's Method). Charles Stein has used the following characterization of the normal distribution as part of his approach to proving limit theorems and in deriving his unbiased estimate of risk in statistics.

LEMMA 2 (Stein, 1986). *A random variable Z has a standard normal distribution if and only if*

$$E(f'(Z)) = E(Zf(Z))$$

for every smooth function f of compact support.

If Z is normal, integration by parts shows that the identity is satisfied for all f such that either side makes sense. Stein's argument for the converse involves the operator U defined by

$$(4.3)\quad (Ug)(x) = e^{x^2/2} \int_{-\alpha}^{x} g(t)e^{-t^2/2}\,dt$$

This satisfies $(Ug)'(x) - x(Ug)(x) = g(x)$ for all g having mean 0 under the normal density. Now suppose P is a probability such that

$$\int (f'(x) - xf(x))P(dx) = 0$$

for a large class of functions f. Set $g(x) = I_A(x) - \Phi(A)$ where A is a fixed Borel set, I_A is the indicator function of A and $\Phi(A)$ is the standard normal probability of A. Then, set $f = Ug$ to see $P(A) = \Phi(A)$. A careful study of the properties of U plays an important part in Stein's method for proving limit theorems.

The operator U mapping g to f sends $L_0^2(e^{-x^2/2})$ into $L^2(e^{-x^2/2})$, where the subscript 0 denotes that part of L^2 orthogonal to the constants. Lemma 1 can be employed to give a singular value decomposition for the operator U. By (4.2), the functions $e_n(x) = H_n(x)/(\sqrt{2\pi}\,2^n n!)^{1/2}$ are an orthonormal basis for $L^2(e^{-x^2/2})$; Lemma 1 then yields Corollary 2.

COROLLARY 2. *The operator U defined by* (4.3) *is a bounded linear operator from $L_0^2(e^{-x^2/2})$ into $L^2(e^{-x^2/2})$. If $\{e_n\}_{n=1}^{\infty}$ and $\{e_n\}_{n=0}^{\infty}$ are taken as orthonormal bases of these spaces, then U satisfies*

$$Ue_n = \frac{-1}{\sqrt{n}}e_{n-1}.$$

REMARK. The bases $\{e_n\}_{n\geq 1}$ and $\{e_n\}_{n\geq 0}$ give a singular value decomposition of U with singular values $-1/\sqrt{n}$. We hope to use this decomposition to study the stability of Stein's characterization of the normal distribution. Corollary 2 may be used in conjunction with Stein's approach to give bounds and approximations for moments. Indeed, Stein

(1986, page 13, equation 33) gives the expression

$$Eh = E_0 h + E\big((T\circ\alpha - iT_0)\circ U_0\big)h,$$

with the expectation on the left being the basic object of study, E_0 the normal expectation and $(T\circ\alpha - iT_0)$ a simple operator. The operator U_0 is our U above. Taking h as the Hermite polynomials, e_n gives explicit identities for moments. As will be seen shortly, virtually identical interpretations hold for the characterizations of the other classical distributions.

4.2 The Gamma Density and Laguerre Polynomials

For $\alpha > 0$, the gamma distribution with parameter α has density $\gamma_\alpha(x) = e^{-x}x^{\alpha-1}/\Gamma(\alpha)$ on $(0, \infty)$. The orthogonal polynomials for this density are called Laguerre polynomials. They have the explicit representation

$$(4.4)\quad L_n^{\alpha-1}(x) = \sum_{k=0}^{n}\binom{n+\alpha-1}{n-k}\frac{(-x)^k}{k!}.$$

Thus $L_0^{\alpha-1} = 1$, $L_1^{\alpha-1} = \alpha - x$, and $L_2^{\alpha-1} = \frac{1}{2}(\alpha+1)\alpha - (\alpha+1)x + \frac{1}{2}x^2$. The identity here is:

LEMMA 1. *Let $L_n^{\alpha-1}$ be defined by* (4.4). *Then, for $n \geq 1$ and $a > 0$,*

$$\int_0^a L_n^{\alpha-1}(x)\gamma_\alpha(x)\,dx = \frac{a}{n}\gamma_{\alpha+1}(a)L_{n-1}^{\alpha}(a).$$

PROOF. Chihara (1978, page 145) gives the Rodrigues type formula

$$L_n^\alpha = \frac{1}{n!}x^{-\alpha}e^x D^n\big[x^{n+\alpha}e^{-x}\big]. \qquad \square$$

REMARK. For integer values of α the integrals can be evaluated by elementary techniques, even when $n = 0$.

EXAMPLE 1. The analog of De Moivre's identity is

$$\frac{1}{\Gamma(\alpha)}\int_0^\infty |x-\alpha|\,x^{\alpha-1}e^{-x}\,dx = 2\alpha\gamma_{\alpha+1}(\alpha).$$

Here, as earlier, the mean absolute deviation is twice the variance times the density at its mode.

EXAMPLE 2 (Stein's Method). The gamma density can be characterized as follows: a random variable X has a γ_α density if and only if

$$E\big(Xf'(X)\big) = E\big((X-\alpha)f(X)\big)$$

for every smooth function f of compact support. This can be used to prove limit theorems for exponential and chi-squared variables via analogs of Stein's method. The formalism involves a study of the equation

$$xf'(x) + (\alpha - x)f(x) = g(x).$$

This can be solved explicitly as

$$(4.5)\quad Ug(x) = x^{-\alpha}e^x\int_0^x g(t)t^{\alpha-1}e^{-t}\,dt.$$

Lemma 1 gives Corollary 1.

COROLLARY 1. *The operator U defined in* (4.5) *is a bounded linear operator from $\mathcal{L}_0^2(\gamma_\alpha)$ into $\mathcal{L}^2(\gamma_{\alpha+1})$. If $\{L_n^{\alpha-1}\}_{n=1}^\infty$ and $\{L_n^\alpha\}_{n=0}^\infty$ are taken as orthogonal bases of these spaces, then*

$$U\big(L_n^{\alpha-1}\big) = \frac{1}{n}L_{n-1}^\alpha.$$

4.3 The Beta Distribution and Jacobi Polynomials

For $\alpha, \beta < 0$, the beta distribution with parameters α, β has density

$$\beta(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

on [0, 1]. The corresponding orthogonal polynomials are called Jacobi polynomials. They are given explicitly as

$$(4.6)\quad p_n^{\alpha-1,\beta-1}(x) = \sum_{k=0}^{n}\binom{n+\alpha-1}{k}\binom{n+\beta-1}{n-k}(x-1)^k x^{n-k}.$$

Well-known special cases include the Legendre polynomials (orthogonal polynomials for the uniform distribution on [0, 1]) and the Chebyshev polynomials of the first and second kind. The identity here becomes:

LEMMA 1. *Let p_n be defined by* (4.6). *For $n \geq 1$,*

$$\int_0^a p_n^{\alpha-1,\beta-1}(x)\beta(x;\alpha,\beta)\,dx = \frac{(-\alpha\beta)\beta(a;\alpha+1,\beta+1)}{n(\alpha-\beta+1)(\alpha+\beta)}p_{n-1}^{\alpha,\beta}(a).$$

PROOF. Chihara (1978, page 143) gives a Rodrigues type formula, which may be rewritten as

$$p_n^{\alpha,\beta}(x) = \frac{(-1)^n}{n!}x^{-\alpha}(1-x)^{-\beta}D^n\big(x^{n+\alpha}(1-x)^{n+\beta}\big).$$

The result follows after elementary manipulation. $\square$

EXAMPLE 1 (Dirichlet Distribution). For $m \geq 1$, the standard m-simplex is denoted

$$\Delta_m = \{x \in \mathbb{R}^m : x_i \geq 0,\ x_1 + \cdots + x_m = 1\}.$$

The symmetric Dirichlet distribution on Δ_m has density

$$D_k(x_1,\ldots,x_m) = \frac{\Gamma(km)}{\Gamma(k)^m}\Pi_{i=1}^m x_i^{k-1}.$$

This has been extensively used as a prior density for Bayesian calculations by I. J. Good.

For k large, D_k converges to a point mass at the center of the simplex $x^* = (1/m, 1/m, \ldots, 1/m)$. The rate of convergence of D_k to x^* can be studied as an application of Lemma 1. If P is a random choice from D_k, let $E_k = E_k \| P - x^* \|$, where

$$\| P - x^* \| = \frac{1}{2} \sum_{i=1}^{m} | P_i - \frac{1}{m} |$$

denotes total variation. Thus E_k is a subjectivist measure of the expected distance of a typical pick from D_k to the uniform measure x^*.

COROLLARY 1.

$$E_k = \frac{1}{k} \frac{\Gamma(mk)}{\Gamma(k)\Gamma(k(m-1))} \left(\frac{1}{m}\right)^k \left(1 - \frac{1}{m}\right)^{k(m-1)}.$$

PROOF. The proof follows from linearity using the mean absolute deviation formula for the beta: If $X \in [0, 1]$ has a $\beta(v; \alpha, \beta)$ density with mean $\mu = \alpha/(\alpha + \beta)$ then

$$E_{\alpha,\beta} | X - \mu | = \frac{2\mu(1-\mu)}{(\alpha + \beta)} \beta(\mu; \alpha, \beta).$$

This in turn follows easily from Lemma 1. □

REMARK. (1) If $k = 1$, D_k becomes the uniform distribution on Δ_m. Then,

$$E_1 = \frac{m-1}{m} \left(1 - \frac{1}{m}\right)^{m-1} \to \frac{1}{e}.$$

Thus a point chosen at random on Δ_m is not too close to the uniform distribution if m is large.

For m fixed, as $k \to \infty$,

$$E_k \sim \frac{1}{\sqrt{2\pi k}} \left(1 - \frac{1}{m}\right)^{3/2}.$$

Thus for k large, a typical pick from D_k is close to the center in total variation. Using Markov's inequality gives convergence of D_k to x^* in probability.

(2) A Stein-like characterization of the beta distribution appears in Section 5.

5. THE PEARSON FAMILY OF CURVES

The results derived for the normal, gamma and beta distributions can be generalized to other members of the Pearson family. Moreover, in a sense to be made precise, Pearson families are the only families of continuous probability densities for which the particular argument employed works. In this section, we present background, show that the orthogonal polynomials associated to a Pearson family admit closed form integrals and prove that this characterizes the Pearson families.

5.1 Pearson Curves

In 1895, the English statistician Karl Pearson introduced his famous family of frequency curves. As noted in Section 3.3, the elements of this family arise by considering the possible solutions to the differential equation

$$\frac{f'(x)}{f(x)} = \frac{a_0 + a_1 x}{b_0 + b_1 x + b_2 x^2} = \frac{q(x)}{p(x)}. \tag{5.1}$$

(Strictly speaking, Pearson took $a_1 = 1$, but it is more natural to include the coefficient and permit the possibility $a_1 = 0$.) The Pearson family has a simple structure. There are in essence five basic solutions, depending on whether the polynomial $p(x)$ in the denominator is constant, linear or quadratic and, in the latter case, on whether the discriminant of $p(x)$ is positive, negative or zero.

It is easy to show that the Pearson family is closed under location and scale change. Thus the study of the family can be reduced to the differential equations that result after an affine transformation of the independent variate.

If deg $p(x) = 0$, then after change of variable the differential equation reduces to $f'(x)/f(x) = \pm x$; if $f(x)$ is assumed to be defined on the maximal interval possible (here $-\infty < x < \infty$), then in order for $f(x)$ to be integrable only the negative sign is permissible and $f(x)$ is seen to be the standard normal density. If deg $p(x) = 1$, then (up to change of location and scale) the resulting maximal solutions may similarly be seen to be the family of gamma distributions; this corresponds to Pearson's type 3.

If deg $p(x) = 2$, then the situation is somewhat more complex.

(1) If the discriminant $\Delta =: b_1^2 - 4b_0 b_2$ of the polynomial $p(x) = b_0 + b_1 x + b_2 x^2$ is negative, then $p(x)$ has no real roots, and after an affine change of variable the density $f(x)$ can be brought into the form

$$f(x) = C(1 + x^2)^{-\alpha} \exp\{\beta \arctan x\},$$

where C is the appropriate normalizing constant. If it is assumed that $f(x)$ is defined on the maximal possible interval—here $(-\infty, \infty)$—then $\alpha > 1/2$ and $-\infty < \beta < \infty$ ensure that $f(x)$ is integrable. Except for special values of α and β, this corresponds to Pearson's type 4; in particular, the t-distributions are a (rescaled) subfamily of this class.

(2) If the discriminant Δ is zero, then $p(x)$ has a single real root, and after an affine change of variable the density $f(x)$ can be brought into the form

$$f(x) = Cx^{-\alpha} \exp\left\{-\frac{\beta}{x}\right\}.$$

Here there are two maximal intervals, $(-\infty, 0)$ and $(0, \infty)$, but by the further change of variable $y = -x$, every such maximal density can be thought of as defined on the positive reals. In this case, $\alpha > 1$, $\beta \geq 0$ ensure that $f(x)$ is integrable. Except for special values of α and β, this corresponds to Pearson's type 5; in particular, the inverse Gaussian distributions are a (rescaled) subfamily of this class.

(3) If the discriminant Δ is positive, then $p(x)$ has two distinct real roots, and after an affine change variable the density $f(x)$ can be brought into the form

$$f(x) = C(1 - x)^{\alpha}(1 + x)^{\beta}.$$

Here there are three maximal intervals, $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$, but after a further change of variable these can be taken to be either $(0, 1)$ or $(0, \infty)$. If the maximal interval is $(0, 1)$, then α, $\beta > -1$ ensure that $f(x)$ is integrable; these are the beta densities and, except for special values of α and β, correspond to Pearson's type 1 (the asymmetric beta) and type 2 (the symmetric beta). If the maximal interval is $(0, \infty)$, then $\alpha > -$, $\alpha + \beta < -$ ensure that $f(x)$ is integrable; in particular, the F-distributions are a subfamily of this class.

5.2 Basic Summation Formula

This section shows that a natural family of polynomials, admitting closed form summation, can be associated to each Pearson density. These polynomials are orthogonal provided sufficiently many moments exist (we admit densities like the t). The proofs draw heavily on two unjustly neglected papers by Hildebrandt (1931) and Beale (1941).

THEOREM 1. *Let J be an open interval, and let $f: J \to \mathbb{R}^+$ be a positive differentiable function on J. If $f(x)$ satisfies the Pearson differential equation* (5.1) *on J, and $p(x) =: b_0 + b_1 x + b_2 x^2$, then:* (1) *For each $n \geq 1$, the function*

$$(5.2)\quad P_n(x) =: \frac{1}{f(x)} \frac{d^n}{dx^n} \{f(x)[p(x)]^n\}$$

is a polynomial of degree at most n; (2) for each $n \geq 1$, the polynomial $P_n(x)$ satisfies the self-adjoint, second-order Sturm-Liouville differential equation

$$(5.3)\quad \frac{d}{dx}\left[f(x)p(x)\frac{dy}{dx}\right] - \lambda_n f(x) y = 0,$$

on J, with $\lambda_n = n[a_1 + (n + 1)b_2]$; and (3) *for every $n \geq 1$ such that $\lambda_n \neq 0$ and every $\alpha, \beta \in J$, the integral*

$$\int_{\alpha}^{\beta} P_n(x) f(x)\, dx$$

is equal to

$$(5.4)\quad \frac{1}{\lambda_n}\left[f(\beta)p(\beta)\frac{dP_n}{dx}(\beta) - f(\alpha)p(\alpha)\frac{dP_n}{dx}(\alpha)\right].$$

PROOF. The observation that the functions in (5.2) are polynomials of degree $\leq n$ is due to Hildebrandt (1931, page 401). Hildebrandt also shows (1931, pages 404–5) that for each $n \geq 1$, the polynomial $P_n(x)$ satisfies the second-order differential equation

$$p(x)\frac{dy^2}{dx^2} + (a_0 + a_1 x + p'(x))\frac{dy}{dx} - c_n y = 0,$$

where $c_n = n[a_1 + (n + 1)b_2]$; the self-adjoint differential equation (5.3) is easily seen to follow from this. Finally, the summation formula (5.4) is an immediate consequence of (5.3). □

It is an important observation of Beale (1941, pages 99–100) that the coefficient of x^n in the polynomial $P_n(x)$ is

$$\prod_{j=0}^{n-1}\left[a_1 + (n + 1 + j)b_2\right].$$

Thus $P_n(x)$ is of degree n precisely when the *Beale condition* is satisfied:

(5.5) Neither of the following occurs:

(i) $a_1 = b_2 = 0$ (in which case $P_n(x)$ is constant);

(ii) $b_2 \neq 0$ and $-\dfrac{a_1}{b_2} \in \{n + 1, n + 2, \ldots, 2n\}$.

Let $P_0(x) \equiv: 1$, and $q(x) =: a_0 + a_1 x$. If one assumes that (5.5) holds for any $n \geq 1$, then the polynomials $\{P_0, P_1, P_2, P_3, \ldots\}$ are linearly independent; and this important case thus arises precisely when either (1) $p(x)$ is constant or linear or (2) $p(x)$ is quadratic and $-(a_1/b_2) - 1$ is not a positive integer. The various systems of polynomials that then arise have been classified by Beale (1937, 1941). After an appropriate affine transformation, the only possibilities are as follows: Let $\delta_1 = \deg\ q(x)$, $\delta_2 = \deg\ p(x)$; and if $\delta_2 = 2$, let $\Delta =$ discriminant $p(x)$. (Note that because $a_1 = b_2 = 0$ is assumed not to occur, the case $\delta_1 = 0$ and $\delta_2 = 0$ or 1 are excluded.) Table 2 collects these results.

The Hermite, Laguerre and Jacobi polynomials were discussed in the preceding section. For discussion and references concerning the basic properties of the Bessel polynomials, see generally Chihara (1978, pages 181–183). The last class of poly-

TABLE 2
Orthogonal polynomials for Pearson curves

δ_1	δ_2	Δ	Polynomial
1	0		Hermite
1	1		Laguerre
≤ 1	2	> 0	Jacobi
≤ 1	2	$= 0$	Bessel (Krall and Frink, 1949)
≤ 1	2	< 0	No accepted name (Romanovsky, 1929)

nomials were first described and discussed by Romanovsky (1929); see also Beale (1941).

5.3 A Converse Theorem

The key to the above argument is the Rodrigues formula (5.2). The next argument shows that the only probability densities admitting such a representation are Pearson families. Fix a positive integer N. Let $f(x)$ be a C^N probability density, defined and everywhere positive on an open interval J. A sequence of polynomials $\{P_n;\ 0 \leq n \leq N\}$ such that deg $P_n \leq n$ is said to satisfy a Rodrigues formula with respect to $f(x)$ if there exists a polynomial $g(x) > 0$ on J, and a sequence of nonzero constants C_n such that

$$(5.6)\quad P_n(x) = \frac{1}{C_n}\frac{1}{f_x}\frac{d^n}{dx^n}\left[f(x)g(x)^n\right], 0 \leq n \leq N.$$

The following result, which is based on a theorem of Cryer (1970, page 3), shows that having a Rodrigues formula for $N = 2$ implies that f is a member of the Pearson family.

THEOREM 1. *Let P_1 and P_2 be two polynomials such that* deg $P_1 \leq 1$ *and* deg $P_2 \leq 2$. *If* (5.6) *is satisfied for $n = 1, 2$ then $f(x)$ is a member of the Pearson family of probability distributions* (5.1).

PROOF. By assumption there exists a polynomial $g(x)$ and nonzero constants C_j, $1 \leq j \leq 2$, such that $C_j P_j f = D^j\{fg^j\}$ for $j = 1,2$. Thus

$$\begin{aligned} C_2P_2f &= D^2(fg^2) = D[(fg)Dg + gD(fg)] \\ &= fgD^2g + D(fg)Dg + D[g(C_1P_1f)] \\ &= fgD^2g + C_1P_1fDg + fgD(C_1P_1) \\ &\quad + C_1P_1D(fg) \\ &= fgD^2g + f\{C_1P_1Dg + gD(C_1P_1)\} \\ &\quad + (C_1P_1)^2f \\ &= f\{gD^2g + D(C_1P_1g) + (C_1P_1)^2\}; \end{aligned}$$

hence $C_2P_2 = gD^2g + D(C_1P_1g) + (C_1P_1)^2$. But if deg $g = m \geq 3$, then the degree of the right hand side would be $2m - 2 \geq 4$, which is impossible. Thus deg $g \leq 2$. But $C_1P_1f = D(fg) = fDg + gDf$, hence

$$\frac{Df}{f} = \frac{C_1P_1 - Dg}{g}.$$

It follows that

$$\frac{f'(x)}{f(x)} = \frac{a_0 + a_1x}{b_0 + b_1x + b_2x^2},$$

for appropriate constants a_0, a_1, b_0, b_1, and b_2; and thus, by definition, $f(x)$ is a member of the Pearson family. □

REMARK. De Moivre's MAD identity was given explicitly in Section 3.3. Stein (1986, Chapter 6) gives appropriate versions of his identity for general densities and explicitly specializes to Pearson curves $f(x)$. Suppressing regularity conditions, a random variable X has density $f(x)$ as at (5.1) if and only if for every smooth h of compact support

$$E[Xh(X) - p(X)h'(X)] = 0,$$

with $p(x)$ as in (5.1). Stein proves this by introducing an operator U, just as in the normal case. We presume that the orthogonal polynomials give a singular value decomposition for this operator to the extent that this makes sense (e.g., existence of moments).

5.4 Some Examples

The normal, gamma and beta families discussed earlier emerge as particularly important members of the Pearson family by a three stage process. (1) They satisfy the *Beale condition* (5.5) for every $n \geq 1$, so that their associated polynomials are linearly independent; (2) they represent solutions to the Pearson differential equation, which are integrable over the maximal permissible interval (i.e., up to the singular points of the differential equation, the zeros of the denominator polynomial $p(x)$); (3) they have moments of all orders, so that by basic Sturm–Liouville theory, their polynomials are in fact orthogonal (see, e.g., Simmons, 1972, pages 133–138, for the case of a bounded interval, applicable to the case where the P_n are the Jacobi polynomials, and $f(x)$ is a beta density).

EXAMPLES. (1) If $f(x) = t_N(x)$, the density of Student's t-distribution on N degrees of freedom, then $f(x) = C(1 + x^2/N)^{-(N+1)/2}$, and

$$\frac{f'(x)}{f(x)} = \frac{-(N+1)x}{N + x^2},$$

so that $-a_1/b_2 = N + 1$, and deg $P_n(x) < n$ for n in the range $(N - 1)/2 \leq n \leq N$.

(2) If $f(x) = x^N$, $x > 0$, then

$$\frac{f'(x)}{f(x)} = \frac{Nx^{N-1}}{x^N} = \frac{N}{x} = \frac{Nx}{x^2}.$$

Thus $f(x)$ is the solution to two versions of the Pearson differential equation, one with $p_1(x) = x$, the other with $p_2(x) = x^2$. But

$$P_n^1(x) =: \frac{1}{x^N} D^n\left[x^N(x^n)\right] = \frac{(N+n)!}{N!}, \qquad n \geq 1,$$

consistent with Beale's theorem (1937, page 209) that the polynomials are constant when the denominator of (5.1) is linear and the numerator constant; while

$$P_n^2(x) =: \frac{1}{x^N} D^n\left[x^N(x^{2n})\right] = \frac{(N+2n)!}{(N+n)!} x^n, \qquad n \geq 1,$$

so that in this case the family $\{P_0^2, P_1^2, P_2^2, \ldots\}$ is indeed a basis for the space of polynomials, consistent with the fact that $-a_1/b_2 = -N$ is never a positive integer.

The function $f(x)$ is not integrable, however, when viewed as a function over the entire positive axis, and so its domain of definition must be truncated in order for it to be normalizable. If we take $f(x) = x^N$, for $0 < x < x_0$, and $f(x) = 0$ otherwise, then the family $\{P_0^2, P_1^2, P_2^2, \ldots\}$ remains a basis for the polynomials but is not orthogonal with respect to $f(x)$.

(3) Consider finally the density $f(x) = x^{-3/2}e^{-1/x}$ on $(0, \infty)$. Then

$$\frac{f'(x)}{f(x)} = \frac{1 - 3x/2}{x^2}.$$

Because $-a_1/b_2 = 3/2$, the polynomials $\{P_1, P_2, P_3, \ldots\}$ in this case are linearly independent, and the function $f(x)$ is integrable. Because $f(x)$ has no moments, the polynomials obviously cannot be orthogonal with respect to $f(x)$. (They are, however, a "quasi-definite" system, orthogonal with respect to a complex measure; see Krall and Frink, 1949.)

6. TWO DISCRETE EXAMPLES

6.1 The Poisson Distribution and Charlier Polynomials

For $\lambda > 0$, let $q_\lambda(j) = e^{-\lambda}\lambda^j/j!$ denote the Poisson density on $0, 1, 2, \ldots$. The orthogonal polynomials are called Charlier polynomials. Chihara (1978, pages 170–172) gives background and details. A monic form of the polynomials can be given explicitly as

$$C_n^\lambda(x) = \sum_{k=0}^{n} \binom{n}{k}\binom{x}{k} k!(-\lambda)^{n-k}. \tag{6.1}$$

Then $C_0 = 1$, $C_1 = x - \lambda$, $C_2 = x(x-1) - 2\lambda x + \lambda^2$. The identity becomes:

LEMMA 1. *Let C_n be defined by* (6.1). *For $n \geq 1$, and $0 \leq a \leq \infty$ an integer,*

$$\sum_{k=0}^{a} C_n(k) q_\lambda(k) = -\lambda q_\lambda(a) C_{n-1}(a).$$

PROOF. The polynomials satisfy the recurrence relation

$$C_{n+1}^\lambda(x) = (x - n - \lambda)C_n^\lambda(x) - \lambda n C_{n-1}^\lambda(x).$$

The Christoffel–Darboux identity for polynomials P_n satisfying $P_n = (x - c_n)P_{n-1} - \lambda_n P_{n-2}$, is

$$\sum_{k=0}^{n} \frac{P_k(x)P_k(y)}{\lambda_1 \cdots \lambda_{k+1}} = \frac{P_{n+1}(x)P_n(y) - P_n(x)P_{n+1}(y)}{(\lambda_1 \cdots \lambda_{n+1})(x - y)}.$$

This holds for any n, x and y. We specialize this to the Charlier case, take $y = 0$ and use the duality relation

$$C_n(k) = (-\lambda)^{n-k} C_k(n).$$

The left side of the Christoffel-Darboux identity becomes

$$\sum_{k=0}^{n} \frac{C_k(x)C_k(0)}{\lambda_1 \cdots \lambda_{k+1}} = \sum_{k=0}^{n} \frac{(-\lambda)^{k-x} C_x(k)(-\lambda)^k}{\lambda^k k!} = (-\lambda)^{-x} \sum_{k=0}^{n} \frac{C_x(k)\lambda^k}{k!}.$$

The right side is easily seen to be

$$\frac{-\lambda^{n+1}(-\lambda)^{-x}}{n!} C_{x-1}(n),$$

where the identities $\Delta C_n(x) = nC_{n-1}(x)$ (Chihara 1978, page 171) were useful.

Equating the two sides then gives the stated result. □

REMARK. There is a Rodrigues-type formula involving finite differences that is available:

$$C_n(x) = \lambda^{-x}\Gamma(x+1)\Delta^n\left[\frac{\lambda^x}{\Gamma(x-n+1)}\right] \quad \text{with} \quad \Delta f(x) = f(x+1) - f(x).$$

Here x is treated as a variable and one easily sees that $C_n(x)$ is a polynomial. Direct use of the formula for integer $x < n$ requires care in its interpretation.

EXAMPLE 1. The analog of De Moivre's identity here is

$$\sum_{k=0}^{\infty} |k - \lambda| \frac{e^{-\lambda}\lambda^k}{k!} = 2\lambda \frac{e^{-\lambda}\lambda^{[\lambda]}}{[\lambda]!},$$

with $[\lambda]$ the greatest integer less than or equal to λ.

Billingsley (1986, pages 381–382) bases a proof of Stirling's formula on this identity. Similar proofs can be based on the other identities of this section.

EXAMPLE 2 (Stein's Identity). The Poisson distribution is characterized by the identity

$$E[\lambda f(X+1)] = E[Xf(X)]$$

for every bounded function f on the integers. Solving for f in terms of g in

$$\lambda f(x+1) = xf(x) = g(x)$$

leads to

$$(6.2) \qquad f(x) = \frac{1}{\lambda q_\lambda(x-1)} \sum_{j=0}^{x-1} g(j) q_\lambda(j),$$

$x \geq 1$; $f(0) = 0$. As usual, $\sum_{x=0}^{\infty} g(x)q_\lambda(x) = 0$ is assumed. Stein (1986, Chapter 9) gives background and motivation.

Comparison with Lemma 1 shows that the Charlier polynomials give a singular value decomposition for U. To state this explicitly, we use

$$\sum_{l=0}^{\infty} C_i(l)C_j(l)q_\lambda(l) = \lambda^i i! \delta_{ij}$$

to form orthonormal polynomials $\overline{C}_n = C_n / \sqrt{\lambda^n n!}$. Let $\{\overline{C}_j\}_{j=1}^{\infty}$ be an orthonormal basis for $L_0^2(q_\lambda)$. Define $p_\lambda(j) = q_\lambda(j-1)$. Let

$$L^2(p_\lambda) = \left\{ f: \{1, 2, \dots\} \to \mathbb{R}: \sum_{j=1}^{\infty} f^2(j) p_\lambda(j) < \infty \right\}.$$

Thus $\overline{D}_n(j) = \overline{C}_n(j-1)$ form an orthonormal basis for $L^2(p_\lambda)$, $0 \leq n \leq \infty$.

COROLLARY 1. *The operator U defined by* (6.2) *is a bounded linear operator from $L_0^2(q_\lambda)$ to $L^2(p_\lambda)$. If these spaces are given bases $\{\overline{C}_n\}_{n=1}^{\infty}$ and $\{\overline{D}_n\}_{n=0}^{\infty}$, then U has singular values given by*

$$U(\overline{C}_n) = \frac{-1}{\sqrt{\lambda n}} \overline{D}_{n-1}.$$

6.2 The Binomial Distribution and Krawtchouk Polynomials

As before, let $b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$ denote the binomial density on $\{0, 1, \dots, n\}$. The Krawtchouck polynomials are orthogonal for $b(k; n, p)$. They are given by the explicit formula

$$(6.3) \qquad P_k^n(x) = \sum_{j=0}^{k} (-1)^j \gamma^{k-j} \binom{x}{j} \binom{n-x}{k-j};$$

$0 \leq k \leq n$, $\gamma = p/(1-p)$. Thus $P_0 = 1$, $P_1 = \gamma(n-x) - x$ and P_2 is

$$P\gamma^2 \frac{(n-x)(n-x-1)}{2} - \gamma x(n-x) + \frac{x(x-1)}{2}.$$

The orthogonality relation is

$$\sum_{i=0}^{n} b(i; n, p) P_r(i) P_s(i) = \left(\frac{p}{1-p}\right)^r \binom{n}{r} \delta_{rs}.$$

Basic properties of the Krawtchouck polynomials with extensive references are given by Macwilliams and Sloane (1977, pages 150–153).

The basic identity for the binomial density is:

LEMMA 1. *For P_k^n defined by* (6.3), *with $k \geq 1$ and a an integer, $0 \leq a \leq n$,*

$$\sum_{i=0}^{a} P_k^n(i) b(i; n, p) = \frac{p}{1-p} \frac{n-a}{k} b(a; n, p) P_{k-1}^{n-1}(a).$$

PROOF. Macwilliams and Sloane (1977, page 152) give the identity

$$(6.4) \qquad P_0^n(x) + P_1^n(x) + \cdots + P_a^n(x) = P_a^{n-1}(x-1)$$

and the duality relation

$$(6.5) \qquad P_k(i) = \frac{\gamma^{k-1}\binom{n}{k}}{\binom{n}{i}} P_i(k).$$

Substituting (6.5) into (6.4) and simplifying gives the desired result. □

REMARK. The Krawtchouck polynomials have a Rodrigues-type representation, which can be used to give an alternative proof for Lemma 1. The Christoffel-Darboux formula and duality can also be used as in our treatment of the Poisson distribution. Finally, as was the case for us originally, the correct formula can be guessed at from small cases and proved directly from (6.3).

EXAMPLE 1 (Stein's Identity). Binomial random variables are characterized by the identity

$$E\big(p(n-X)f(X)\big) = E\big(qXf(X-1)\big)$$

for every $f:\{-1, 0, 1, \ldots, n\}$ into $\mathbb{R}$, where $q = 1 - p$. The study of this identity involves solving for f given g in the following equation:

$$p(n-x)f(x) - qxf(x-1) = g(x) \qquad \text{where} \quad E_{n,p}\big(g(X)\big) = 0.$$

This can be solved explicitly as

$$f(x) = (Ug)(x)$$

$$(6.6) \quad = \frac{1}{p(n-x)b(x;n,p)} \sum_{i=0}^{x} b(i;n,p)g(i), \qquad 0 \le x < n-1.$$

The value of $f(-1)$ and $f(n)$ can be chosen arbitrarily.

Lemma 1 translates into a singular value decomposition for U after introducing orthonormal bases

$$\bar{P}_r^n(x) = P_r^n \Big/ \sqrt{\binom{n}{r}\left(\frac{p}{q}\right)^r}.$$

COROLLARY 1. *For U defined by (6.6), let $L_0^2(b(k;n,p))$ and $L^2(b(k;n-1,p))$ have $\{\bar{P}_i^n\}_{i=1}^n$ and $\{\bar{P}_i^{n-1}\}_{i=0}^{n-1}$ as orthonormal bases. Then*

$$U\big(\bar{P}_i^n\big) = P_{i-1}^{n-1} \Big/ \sqrt{\frac{inp}{q}},$$

so that U is a 1-1, onto, linear map with singular values $1/\sqrt{inp/q}$, $1 \le i \le n$.

7. OTHER DENSITIES

Very similar results can be derived for other densities. What is needed is either a Rodrigues-type formula or a duality result along with the Christoffel-Darboux identity as outlined in Sections 6.1 and 6.2 above. For example, the geometric and negative binomial distributions give rise to Meixner polynomials, the hypergeometric distribution to Hahn polynomials. There is, in fact, a discrete analog of Theorem 1 of Section 5.3 characterizing all discrete measures having Rodrigues-type formulas; see Weber and Erdélyi (1952). Along these lines, see Chihara (1978, Chapter 5, Section 3). Eagleson (1968) characterizes discrete orthogonal polynomials which admit a duality relation.

The "sixth family" of Morris (1982) is related to Pollaczek polynomials. It is not covered by the results above (it is not in the Pearson family). It would be interesting to see that formulas are available for the cubic exponential families of Letac and Mora (1990).

8. CODA

This article had its origin in the simple observation that buried in Problem 72 of De Moivre's *Doctrine of Chances* was the L_1 law of large numbers for Bernoulli trials. Somewhat to our surprise, however, what was initially regarded as a fairly straightforward (and short!) historical note soon began to acquire a life of its own: no sooner did we think that we had tracked down the earliest rediscovery of the result, then another cropped up; a routine intellectual credit check on Sir Alexander Cuming ended up leading us down the path of an 18th century con artist (see the concluding postscript below); and an attempt to understand Todhunter's proof of De Moivre's formula ultimately resulted in the discovery of a much more general phenomenon, valid for many of the classical distributions.

Most of us have probably had this experience at one time or another. But (for us, at least) it seems to happen with uncanny frequency when trying to read and understand the past masters of our subject, which is one reason why we enjoy it so much.

We have not exhausted the rich collection of ideas connected to De Moivre's identity. David Aldous has shown us a probabilistic proof of (1.1) and connected it to Tanaka's formula of stochastic calculus. A discrete version of this given by (Csörgő and Révész (1985) yields the identity

$$2\left|S_n - \frac{n}{2}\right| = \sum_{k=0}^{n-1} (2X_{k+1} - 1)\, sgn(2S_k - k) + \xi(n),$$

where X_k are symmetric Bernoulli, $S_k = X_1 + \cdots + X_k$ and $\xi(n)$ the number of k, $0 \le k < n$ such that $S_k = k/2$. Taking expectations gives a formula for the left side of (1.1) as the sum of middle binomial coefficients:

$$2E\left|S_n - \frac{n}{2}\right| = \sum_{j=0}^{\left[\frac{n-1}{2}\right]} \frac{1}{2^{2j}}\binom{2j}{j}.$$

Richard Askey and George Gasper have pointed to "q analogs" of some of the formulas. Stein's operator U is a standard tool in working with Hermite polynomials. Our Corrollary 2 of Section 4.1 is the basic "lowering relation" in that theory; see, for example, Cormier and Greenleaf (1990), Lemma A.3.4b, page 244.

Presumably the list goes on.

CONCLUDING (PARTIALLY UNSCIENTIFIC) POSTSCRIPT: SIR ALEXANDER CUMING

1. STIRLING AND CUMING

In the *Miscellanea Analytica*, De Moivre states that Problem 72 in the *Doctrine of Chances* had been originally posed to him in 1721 by Alexander Cuming, whom he describes as an illustrious man *(vir clarissimus)* and a member of the Royal Society *(Cum aliquando labenta Anno 1721, Vir clarissimus Alex. Cuming Eq. Au. Regiae Societatis Socius, quaestionem infra subjectum mihi proposuisset, solutionem prolematis ei postero die tradideram).*

Thus, we have argued, Cuming was responsible for instigating a line of investigation on De Moivre's part that ultimately led to his discovery of the normal approximation to the binomial. But curiously, Cuming was also directly responsible for James Stirling's discovery of the asymptotic series for $\log(n!)$.

At some point prior to the publication of the *Miscellanea Analytica*, De Moivre discovered that Stirling had also made important discoveries concerning the asymptotic behavior of the middle term of the binomial distribution. Stirling and De Moivre were on good terms, and De Moivre, while obviously wishing to establish that he had been the first to make the discovery, was also clearly anxious to avoid an unpleasant priority dispute (at least two of which he had been embroiled in earlier in his career). And thus, as De Moivre tells us in the *Miscellanea Analytica* (1730, page 170),

> As soon as [Stirling] communicated this solution to me, I asked him to prepare a short description of it for publication, to which he kindly assented, and he generously undertook to explain it at some length, which he did in the letter which I now append.

De Moivre then gave the full text (in Latin) of Stirling's letter, dated 19 June 1729. Stirling wrote:

> About four years ago [i.e., 1725], I informed the distinguished Alexander Cuming that the problems of interpolation and summation of series, and other such matters of that type, which did not fall under the ordinary categories of analysis, could be solved by the differential method of Newton; this illustrious man responded that he doubted whether the problem solved by you several years earlier, concerning the behavior of the middle term of any power of the binomial, could be solved by differentials. I then, prompted by curiousity and feeling confident that I would do something that would please a mathematician of very great merit [i.e., De Moivre], took on the same problem; and I confess that difficulties arose which prevented me from quickly arriving at an answer, but I do not regret the labor if I shall nonetheless have achieved a solution so approved by you that you would see fit to insert it in your own writings. Now this is how I did it.

Stirling then went on to give, at considerable length, an illustration of his solution, but did not derive it, because "it will be described in a tract shortly to appear, concerning the interpolation and summation of series, that I am writing".

This promised book was Stirling's *Methodus Differentialis* of 1730 (which thus appeared in the same year as De Moivre's *Miscellanea Analytica*), one of the first great works on numerical analysis. In his preface, Stirling again acknowledged the crucial role of Cuming:

> The problem of the discovery of the middle term of a very high power of the binomial had been solved by De Moivre several years before I had accomplished the same thing. *It is improbable that I would have thought about it up to the present day* had it not been suggested by that eminent gentleman, the most learned Alexander Cuming, who indicated that he very much doubted whether it could be solved by Newton's differential method. [Stirling, 1730, Preface; emphasis added.]

Thus Alexander Cuming appears to have played, for De Moivre and Stirling, a role similar to that of the Chevalier de Meré for Pascal and Fermat. Who was he?

2. THE QUEST FOR CUMING

At this remove of time, the question can only be partially answered, but the story that emerges is a strange and curious one, a wholly unexpected coda to an otherwise straightforward episode in the history of mathematics.

The British *Dictionary of National Biography* tells us that Cuming was a Scottish baronet, born about 1690, who briefly served in the Scottish bar (from 1714 to 1718) and then left it, under obscure but possibly disreputable circumstances. Shortly after, Cuming surfaces in London, where he was elected a Fellow of the Royal Society of London on June 30, 1720, the year before that in which De Moivre says Cuming posed his problem. The *DNB* does not indicate the reason for Cuming's election, and there is little if any indication of serious scientific output on his part. (No papers by him appear, for example, in the *Philosophical Transactions of*

the Royal Society of London. This was not unusual, however, at the time; prior to a 19th century reform, members of the aristocracy could become members of the Royal Society simply by paying an annual fee.)

During the next decade, Cuming seems to have taken on the role of intellectual go-between (see Tweedie, 1922, pages 93 and 201). Cuming's chief claim to fame, however, lies in an entirely different direction. In 1729 he undertook an expedition to the Cherokee Mountains in Georgia, several years prior to the time the first settlers went there, led by James Oglethorp, in 1734. Appointed a chief by the Cherokees, Cuming returned with seven of their number to England, presenting them to King George II in an audience at Windsor Castle on June 18, 1730. Before returning, an "Agreement of Peace and Friendship" was drawn up by Cuming and signed by the chiefs, which agreement, as the 19th century *DNB* so charmingly puts it, "was the means of keeping the Cherokees our firm allies in our subsequent wars with the French and American colonists".

This was Sir Alexander's status in 1730, when De Moivre refers to him as an illustrious man and a member of the Royal Society; both conditions, unfortunately, were purely temporary. For the surprising *denouement* to Sir Alexander's career, we quote the narrative of the *DNB*:

> By this time some reports seriously affecting Cuming's character had reached England. In a letter from South Carolina, bearing date 12 June 1730, ... he is directly accused of having defrauded the settlers of large sums of money and other property by means of fictitious promissory notes. He does not seem to have made any answer to these charges, which, if true, would explain his subsequent ill-success and poverty. The government turned a deaf ear to all his proposals, which included schemes for paying off eighty millions of the national debt by settling three million Jewish families in the Cherokee mountains to cultivate the land, and for relieving our American colonies from taxation by establishing numerous banks and a local currency. Being now deeply in debt, he turned to alchemy, and attempted experiments on the transmutation of metals.

Fantastic as Cuming's alleged schemes might seem, they were of a type not new to the governments of his day. A decade earlier, thousands had lost fortunes in England and France with the bursting of the South Sea and Mississippi "bubbles."

For Cuming it was all downhill from here. A few years later, in 1737, the law finally caught up with him, and he was confined to Fleet prison, remaining there perhaps continuously until 1766, when he was moved to the Charterhouse (a hospital for the poor), where he remained until his death on August 23, 1775. He had been expelled from the Royal Society on June 9, 1757 for nonpayment of the annual fee, and when his son, also named Alexander, died some time prior to 1796, the Cuming baronetcy became extinct. By 1738, when the second edition of De Moivre's *Doctrine of Chances* appeared, association with the Cuming name had clearly become an embarrassment, and unlike the corresponding passage in the *Miscellanea Analytica*, no mention of Cuming appears when De Moivre discusses the problem Cuming had posed to him.

Thus Cuming's life in outline. Nevertheless, there remain tantalizing and unanswered questions. The account in the *Dictionary of National Biography* appears largely based on an article by H. Barr Tomkins (1878). Tomkins's article several times quotes a manuscript written by Cuming while in prison (see also Drake, 1872), and this manuscript is presumably the ultimate source for the curious schemes mentioned by the *DNB*. But although they are there presented as serious proposals, at the time that Cuming wrote the manuscript his mind appears to have been substantially deranged for several years, and the evidentiary value of the manuscript is questionable.

ACKNOWLEDGMENTS

We thank Richard Askey, David Bellhouse, Daniel Garrison, George Gasper, Ian Johnstone, Charles Stein, Steve Stigler and Gérard Letac for their comments as our work progressed. Research supported by NSF Grant DMS-89-05874.

REFERENCES

ADAMS, W. J. (1974). *The Life and Times of the Central Limit Theorem*. Kaedmon, New York.

BARDWELL, G. E. (1960). On certain characteristics of some discrete distributions. *Biometrika* **47** 473-475.

BEALE, F. S. (1937). On the polynomials related to Pearson's differential equation. *Ann. Math. Statist.* **8** 206-223.

BEALE, F. S. (1941). On a certain class of orthogonal polynomials. *Ann. Math. Statist.* **12** 97-103.

BERTRAND, J. (1889). *Calcul des probabilités*. Gauthier-Villars, Paris.

BILLINGSLEY, P. (1985). *Probability and Measure*, 2nd ed. Wiley, New York.

BLYTH, C. R. (1980). Expected absolute error of the usual estimator of the binomial parameter. *Amer. Statist.* **34** 155-157.

CHIHARA, T. S. (1978). *An Introduction to Orthogonal Polynomials*. Gordon and Breach, New York.

CLERKE, A. M. (1894). Moivre, Abraham de. *Dictionary of National Biography* **38** 116-117.

CORMIER, L. and GREENLEAF, F. P. (1990). *Representations of Groups and Their Applications. Part 1. Basic Theory and Examples*. Cambridge Univ. Press.

CRYER, C. (1970). Rodrigues' formula and the classical orthogonal polynomials. *Boll. Un. Mat. Ital.* **25** 1-11.

CSÖRGŐ, M. AND RÉVÉSZ, P. (1985). On the stability of the local time of symmetric random walk. *Acta. Sci.* **48** 85-96.

CZUBER, E. (1914). *Wahrscheinlichkeitsrechnung*. Teubner, Leipzig.

DALLAL, S. and HALL, W. (1983). Approximating priors by mixtures of conjugate priors. *J. Roy. Statist. Soc. Ser. B* **45** 278-286.

DASTON, L. (1988). *Classical Probability in the Enlightenment*. Princeton Univ. Press.

DAVID, F. N. (1962). *Games, Gods, and Gambling*. Hafner, New York.

DAW, R. H. and PEARSON, E. S. (1972). Abraham de Moivre's 1733 derivation of the normal curve: A bibliographical note. *Biometrika* **59** 677-680.

DE MOIVRE, A. (1718). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 1st ed. A. Millar, London. (2nd ed. 1738; 3rd ed. 1756.)

DE MOIVRE, A. (1730). *Miscellanea Analytica de Seriebus et Quadraturis*. J. Tonson and J. Watts, London.

DIACONIS, P. and FREEDMAN, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745-764.

DIACONIS, P. and YLVISAKER, D. (1985). Quantifying prior opinion. In *Bayesian Statistics* **2** (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 133-156. North-Holland, Amsterdam.

DRAKE, S. G. (1872). *Early History of Georgia, Embracing the Embassy of Sir Alexander Cuming to the Country of the Cherokees, in the Year 1730*. David Clapp and Son, Boston.

EAGLESON, G. K. (1968). A duality relation for discrete orthogonal systems. *Studia Sci. Math. Hungar.* **3** 127-136.

FELLER, W. (1968). *An Introduction to Probability and Its Applications* **1**, 3rd ed. Wiley, New York.

FELLER, W. (1971). *An Introduction to Probability and Its Applications* **2**, 2nd ed. Wiley, New York.

FRAME, J. S. (1945). Mean deviation of the binomial distribution. *Amer. Math. Monthly* **52** 377-379.

FRISCH, R. (1924). Solution d'un problème du calcul des probabilités. *Skandinavisk Aktuarietidskrift* **7** 153-174.

GRUDER, O. (1930). *9th International Congress of Actuaries* **2** 222.

HALD, A. (1984). Commentary on "De Mensura Sortis." *Internat. Statist. Rev.* **52** 229-236.

HALD, A. (1988). On de Moivre's solutions of the problem of duration of play, 1708-1718. *Arch. Hist. Exact Sci.* **38** 109-134.

HALD, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*. Wiley, New York.

HILDEBRANDT, E. H. (1931). Systems of polynomials connected with the Charlier expansions and the Pearson differential and difference equation. *Ann. Math. Statist.* **2** 379-439.

JOHNSON, N. L. (1957). A note on the mean deviation of the binomial distribution. *Biometrika* **44** 532-533.

JOHNSON, N. L. (1958). The mean deviation with special reference to samples from a Pearson type III population. *Biometrika* **45** 478-483.

KAMAT, A. R. (1965). A property of the mean deviation for a class of continuous distributions. *Biometrika* **52** 288-9.

KAMAT, A. R. (1966a). A property of the mean deviation for the Pearson type distributions. *Biometrika* **53** 287-289.

KAMAT, A. R. (1966b). A generalization of Johnson's property of the mean deviation for a class of discrete distributions. *Biometrika* **53** 285-287.

KRALL, H. L. and FRINK, O. (1949). A new class of orthogonal polynomials: The Bessel polynomials. *Trans. Amer. Math. Soc.* **65** 100-115.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LETAC, G. and MORA, M. (1990). Natural real exponential families with cubic variance functions. *Ann. Statist.* **18** 1-37.

LORENTZ, G. G. (1986). *Bernstein Polynomials*, 2nd ed. Chelsea, New York.

MACWILLIAMS, F. J. and SLOANE, N. J. A. (1977). *The Theory of Error Correcting Codes*. North-Holland, Amsterdam.

MORRIS, C. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65-80.

PEARSON, K. (1924). On the mean error of frequency distributions. *Biometrika* **16** 198-200.

PEARSON, K. (1978). *The History of Statistics in the 17th and 18th Centuries*. Macmillan, New York.

POINCARÉ, H. (1896). *Calcul des Probabilités*, 1st ed. Georges Carré, Paris. (2nd ed. 1912, Gauthier-Villars, Paris.)

RAMASUBBAN, T. A. (1958). The mean difference and the mean deviation of some discontinuous distributions. *Biometrika* **45** 549-556.

ROMANOVSKY, V. (1929). Sur quelques classes nouvelles de polynomes orthogonaux. *C. R. Acad. Sci Paris* **188** 1023-1025.

ROSENLICHT, M. (1976). Integration in finite terms. *Amer. Math. Monthly* **79** 963-972.

SCHNEIDER, I. (1968). Der Mathematiker Abraham de Moivre, 1667-1754. *Arch. Hist. Exact Sci.* **5** 177-317.

SIMMONS, G. F. (1972). *Differential Equations with Applications and Historical Notes*. McGraw-Hill, New York.

STEIN, C. (1986). *Approximate Computation of Expectations*. IMS, Hayward, Calif.

STIGLER, S. (1986). *The History of Statistics*. Harvard Univ. Press.

STIRLING, J. (1730). *Methodus Differentialis*. Gul. Bowyer, London.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability*. Macmillan, London.

TOMKINS, H. B. (1878). Sir Kenneth William Cuming of Culter, Baronet. *The Genealogist* **3** 1-11.

TWEEDIE, C. (1922). *James Stirling*. Clarendon Press, Oxford.

USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.

VON BORTKIEWICZ, L. (1923). Über eine vershiedenen Fehlergesetzen gemeinsame Eigenschaft. *Sitzungsberichte der Berliner Mathematischen Gesellschaft* **22** 21-32.

WALKER, H. M. (1934). Abraham de Moivre. *Scripta Mathematica* **2** 316-333.

WEBER M. and ERDÉLYI, A. (1952). On the finite difference analogue of Rodriques' formula. *Amer. Math. Monthly* **59** 163-168.

ZEILBERGER, D. (1989). A holonomic systems approach to binomial coefficient identities. Technical Report, Drexel Univ.