ON A CLASS OF DISTRIBUTIONS THAT APPROACH THE NORMAL DISTRIBUTION FUNCTION[1]

By George B. Dantzig

1. **Formulation of the Problem.** An important property of a sequence of binomial coefficients is that, when suitably normalized and transformed, it converges to the normal distribution.[2] The object of this paper is to exhibit a large class of other sequences which also possess this property.

The Pascal recurrence formula may be taken as the defining property of the binomial coefficients. Let the combination of n things taken x at a time be denoted by $\binom{n}{x}$. If we set $f_n(x) = (\frac{1}{2})^n \cdot \binom{n}{x}$ for $0 \leq x \leq n$ and $f_n(x) = 0$ for $x < 0$ or $x > n$, then $f_n(x)$ is defined for all integers x. With this notation Pascal's recurrence formula, $\binom{n}{x} = \binom{n-1}{x} + \binom{n-1}{x-1}$, may be written

$$f_n(x) = \tfrac{1}{2}\,[f_{n-1}(x) + f_{n-1}(x-1)]\,, \tag{1}$$

where this new form is valid for all integers x extending from $-\infty$ to $+\infty$.

In order to generalize, we may consider a sequence of distributions $f_1(x)$, $f_2(x)$, $\cdots$, $f_n(x)$, $\cdots$ each defined in terms of the preceding one by means of the recurrence formula

$$f_n(x) = \frac{1}{a_n + 1}\,[f_{n-1}(x-0) + f_{n-1}(x-1) + f_{n-1}(x-2) + \cdots + f_{n-1}(x-a_n)], \tag{2}$$

where the x are integers, and a_n is a positive integer which may change in value from one distribution to the next. The problem is to find conditions under which $f_n(x)$, in normalized form, approaches the normal distribution. The normalization of $f_n(x)$ is effected by the affine transformation

$$u = \frac{x - \bar{x}_n}{\sigma_n}\,; \qquad \varphi_n(u) = f_n(x)\,, \tag{3}$$

---

[1] Presented November 21, 1938 before a joint meeting of the Columbia Mathematics Club and the Statistical Seminar of the Graduate School of the Department of Agriculture; also December 10, 1938 before a meeting of the American Mathematical Association at the University of Maryland.

[2] Due to DeMoivre, 1731. By a variable distribution approaching the normal distribution, we mean that the integral under the variable distribution between any two limits approaches the corresponding integral under the normal curve.

where $\bar{x}_n$ and σ_n are the mean and standard deviation of the distribution $f_n(x)$. The normal (cumulative) distribution function is taken in the standard form

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{1}{2}x^2}\, dx\,. \tag{4}$$

The theorem whose proof forms the theme of this paper may be stated as follows:

THEOREM: *A necessary and sufficient condition that* $\varphi_n(u) \to \varphi(u)$ *as* $n \to \infty$ *is that* $\Gamma = 0$, *where*

$$\Gamma = \operatorname*{Lim}_{n\to\infty} \sum_{i=2}^{n} \gamma_i^2 \Big/ \left(\sum_{i=2}^{n} \gamma_i\right)^2; \qquad 4\gamma_i = a_i^2 + 2a_i\,. \tag{5}$$

2. Liapounoff Condition; the general case. The recurrence formula (2) is a special case of the most general linear recurrence formula

$$f_n(x) = \sum_{i=-\infty}^{+\infty} g_n(i) f_{n-1}(x - i), \tag{6}$$

where $g_n(i)$ are a given set of weight functions generating the sequence $f_1(x)$, $f_2(x), \cdots, f_n(x), \cdots$. We may form the recurrence formula (2) by setting

$$\begin{aligned} g_n(i) &= \frac{1}{a_n + 1} \qquad \text{if} \qquad 0 \leq i \leq a_n\,, \\ g_n(i) &= 0 \qquad\quad\ \ \text{if} \qquad i < 0 \text{ or } i > a_n\,. \end{aligned} \tag{7}$$

Let $F_k(t) = \sum_{x<t} f_k(x)$ express[3] the probability that a variable $x_k < t$, where the distribution function of x_k is defined as $f_k(x)$; and in a similar manner let the probability that a variable $s_k < t$ be given by $G_k(t) = \sum_{x<t} g_k(x)$. By summing $f_n(x)$ for all x less than t, we obtain

$$F_n(t) = \sum_{i=-\infty}^{+\infty} F_{n-1}(t - i) g_n(i) = \int_{-\infty}^{+\infty} F_{n-1}(t - i)\, dG_n(i)\,, \tag{8}$$

where we have replaced the summation by a Stieltjes Integral. In the latter form the integral gives, in general, the probability that the *sum* of two independent variables x_{n-1} and s_n is less than t. From the above equation we see that the probability that $x_{n-1} + s_n < t$ is the same as that of $x_n < t$, so that we may set $x_n = x_{n-1} + s_n$. By iteration one obtains

$$x_n = s_1 + s_2 + \cdots + s_n \tag{9}$$

for all n. Thus we have established that *if a distribution function of a variable* s_k *is defined as* $g_k(x)$, *then the distribution function of the sum* $s_1 + s_2 + \cdots + s_n = x_n$ *is* $f_n(x)$.

---

[3] The summation extends over all values x less than t.

The limit of the distribution function of the sum of n independent variables as $n \to \infty$ has been considered by Laplace, Liapounoff, Lindeberg, and others. We shall make use of a sufficient condition given by Liapounoff that the normalized distribution function of x_n approaches $\varphi(u)$.

LAPLACE-LIAPOUNOFF THEOREM:[4] *A sufficient condition for the normalized distribution function of the sum of n independent variables $s_1, s_2, \cdots, s_n$ to approach the normal distribution function with increasing n is $\Gamma' = 0$, where*

$$\Gamma' = \lim_{n\to\infty} \frac{M_4(1) + M_4(2) + \cdots + M_4(n)}{[M_2(1) + M_2(2) + \cdots + M_2(n)]^2}, \tag{10}$$

and where $M_2(k)$ and $M_4(k)$ are defined as the second and fourth moments of s_k whose distribution is $g_k(x)$.

Thus we have shown that if a sequence of distributions $f_n(x)$ is defined by the general linear recurrence formula (6),

$$f_n(x) = \sum_{i=-\infty}^{+\infty} g_n(i) \cdot f_{n-1}(x - i),$$

then a sufficient condition that $\varphi_n(u) \to \varphi(u)$ as $n \to \infty$ is given by $\Gamma' = 0$, where $\varphi_n(u)$ is the normalized form of $f_n(u)$.

3. **Sufficiency of the Condition $\Gamma = 0$.** We may simplify the condition $\Gamma' = 0$ for the more restricted case of a sequence of distributions defined by the recurrence formula (2). In general, the second and fourth moments of $g_n(x)$ are given by

$$\begin{aligned} M_2(k) &= \sum_{x=-\infty}^{+\infty} g_k(x)(x - \bar{s}_k)^2, \\ M_4(k) &= \sum_{x=-\infty}^{+\infty} g_k(x)(x - \bar{s}_k)^4, \end{aligned} \tag{11}$$

where $\bar{s}_k$ is the mean value of the distribution. Equations (7) give the special values of $f_k(x)$; substituting these values in (11), and remembering the Bernoulli summation by which $1^p + 2^p + 3^p + \cdots + n^p$ may be expressed as a polynomial in n of degree $p + 1$, we obtain

$$\begin{aligned} M_2(k) &= \sum_{x=0}^{a_k} \frac{1}{a_k + 1}\left(x - \frac{1}{2}a_k\right)^2 = \frac{1}{3}\left[\frac{a_k^2 + 2a_k}{4}\right] = \frac{1}{3}\gamma_k, \\ M_4(k) &= \sum_{x=0}^{a_k} \frac{1}{a_k + 1}\left(x - \frac{1}{2}a_k\right)^4 \\ &= \frac{1}{5}\left[\frac{a_k^2 + 2a_k}{4}\right]^2 - \frac{1}{15}\left[\frac{a_k^2 + 2a_k}{4}\right] = \frac{1}{5}\gamma_k^2 - \frac{1}{15}\gamma_k; \end{aligned} \tag{12}$$

[4] J. V. Uspensky, *Introduction to Mathematical Probability* (McGraw-Hill, 1937), pages 284–292; the theorem is proved there by the method of characteristic functions.

whence by substitution in (10), Γ' becomes

$$\Gamma' = \operatorname*{Lim}_{n\to\infty} \frac{\frac{1}{5}\sum_{i=2}^{n}\gamma_i^2 - \frac{1}{15}\sum_{i=2}^{n}\gamma_i + M_4(1)}{\left[\frac{1}{3}\sum_{i=2}^{n}\gamma_i + M_2(1)\right]^2}. \tag{13}$$

Since $a_i \geq 1$, $\gamma_i \geq 3/4$, and thus $\sum_{i=2}^{n}\gamma_i \to \infty$ as $n \to \infty$, we may reduce Γ' in the limit to

$$\Gamma' = \frac{3}{5}\operatorname*{Lim}_{n\to\infty} \sum_{i=2}^{n}\gamma_i^2 \Big/ \left[\sum_{i=2}^{n}\gamma_i\right]^2. \tag{14}$$

Since $\Gamma' = \frac{3}{5}\Gamma$, the Liapounoff condition $\Gamma' = 0$ for normality becomes by (5), $\Gamma = 0$.

4. **Necessity of the Condition $\Gamma = 0$.** A necessary condition for normality can be found by noting that if $\varphi_n(u)$ approaches $\varphi(u)$, then the moments of $\varphi_n(u)$ must approach the corresponding moments of $\varphi(u)$.[5] Letting $\mu_4(n)$ be the 4th moment of $\varphi_n(u)$ and μ_4 the corresponding moment of the normal curve, a necessary condition is that $\mu_4(n) \to \mu_4$ as $n \to \infty$, and $\mu_4 = 3$. The 4th moment of $\varphi_n(u)$ may be expressed simply in terms of the moment of $f_n(x)$. If the symbol E stands for expected value, the second and fourth moments of $f_n(x)$ are $E(x_n - \bar{x}_n)^2$ and $E(x_n - \bar{x}_n)^4$ respectively, and the relationship is then

$$\mu_4(n) = \frac{E(x_n - \bar{x}_n)^4}{[E(x_n - \bar{x}_n)^2]^2} = \frac{E\left[\sum_{i=1}^{n}(s_n - \bar{s}_n)\right]^4}{\left\{E\left[\sum_{i=1}^{n}(s_n - \bar{s}_n)\right]^2\right\}^2}. \tag{15}$$

Expanding the sums by the multinomial theorem and taking the expected value of each term we obtain

$$E(x_n - \bar{x}_n)^2 = \sum_{i=1}^{n} E(s_i - \bar{s}_i)^2 + 2\sum_{i<j=1}^{n} E(s_i - \bar{s}_i)E(s_j - \bar{s}_j) = \sum_{i=1}^{n} M_2(i), \tag{16}$$

where $M_2(i)$ is the second moment of $g_i(x)$. In a similar manner we have

$$\begin{aligned} E(x_n - \bar{x}_n)^4 &= \sum_{i=1}^{n} M_4(i) + 6\sum_{i<j=1}^{n} M_2(i)M_2(j) \\ &= \sum_{i=1}^{n} M_4(i) + 3\left[\sum_{i=1}^{n} M_2(i)\right]^2 - 3\sum_{i=1}^{n} M_2^2(i); \end{aligned} \tag{17}$$

[5] Uspensky, loc. cit., pages 383–388.

whence

$$\mu_4(n) = 3 + \frac{\sum_{i=1}^{n} M_4(i) - 3\sum_{i=1}^{n} M_2^2(i)}{\left[\sum_{i=1}^{n} M_2(i)\right]^2}. \tag{18}$$

Since a necessary condition for normality is that Lim $\mu_4(n) \to \mu_4 = 3$, the fraction in the above equation must in the limit approach zero. Substituting $M_2(i) = \frac{1}{3}\gamma_i$ and $M_4(i) = \frac{1}{5}\gamma_i^2 - \frac{1}{15}\gamma_i$, we find that this ratio reduces immediately in the limit to the condition $\Gamma = 0$.

5. **Application to the Distribution of Inversions.** A frequency table may be set up for the number of permutations of n objects that give rise to a fixed number of inversions. Three objects marked 1, 2, 3 may be permuted in 6 ways:

(123), (132), (213), (231), (312), (321).

If (123) is taken as standard position, the number of inversions associated with the above set to bring each one into standard position are respectively 0, 1, 1, 2, 2, 3. Thus we pass from (321) to (123) by the following three inversions or adjacent interchanges: (312), (132), (123). Among the six permutations there is one giving rise to 0 inversions, two having 1 inversion, two having 2 inversions, and one having 3 inversions.

The distribution of inversions finds its application in a test of significance. The standard position is taken as a *hypothesis* of rank order, and the difference between an observed set of ranks and the hypothetical one is measured by the number of inversions. The distribution may then be used for finding the probability of obtaining by chance the number of inversions found, or less. For a moderate number of ranks (six or more), the distribution of inversions may be approximated by a normal curve. We shall show that as the number of ranks is increased, the normalized distribution of inversions approaches the normal distribution. The distribution of inversions of 1, 2, 3, 4, objects will be found in the table below.

Inversions: x	0	1	2	3	4	5	6
$1 \cdot f_1(x)$	1						
$1 \cdot 2 \cdot f_2(x)$	1	1					
$1 \cdot 2 \cdot 3 \cdot f_3(x)$	1	2	2	1			
$1 \cdot 2 \cdot 3 \cdot 4 \cdot f_4(x)$	1	3	5	6	5	3	1

By induction one may show that the following relationships hold between successive distributions:

$$
\begin{aligned}
f_2(x) &= \tfrac{1}{2}[f_1(x-0) + f_1(x-1)], \\
f_3(x) &= \tfrac{1}{3}[f_2(x-0) + f_2(x-1) + f_2(x-2)], \\
&\vdots \\
f_n(x) &= \frac{1}{n}[f_{n-1}(x-0) + f_{n-1}(x-1) \\
&\qquad + f_{n-2}(x-2) + \cdots + f_{n-2}(x-n+1)].
\end{aligned} \tag{19}
$$

Since this satisfies the basic recurrence formula (2), where $a_n = n - 1$, we may find out whether the normalized distributions of inversions approaches $\varphi(u)$. With $\gamma_n = n^2 - 1$ the condition $\Gamma = 0$ becomes $\lim_{n\to\infty} \sum_{i=2}^{n} (i^2-1)^2 \Big/ \Big[\sum_{i=2}^{n} (i^2-1)\Big]^2$. The numerator sums to a polynomial of the 5th degree in n, while the brackets of the denominator sums to a 3d degree polynomial, which after squaring is of the 6th degree; so that as $n \to \infty$ we have in the limit $\Gamma = 0$. Thus the normalized distribution function of the inversions of n objects approaches $\varphi(u)$ as $n \to \infty$.

Equations (12) and (16) permit us to find the mean and standard deviation of the distribution of the inversions of n objects:

$$
\begin{aligned}
\bar{x}_n &= \tfrac{1}{4}n(n-1), \\
\sigma_n^2 &= \frac{1}{72}n(n-1)(2n+5).
\end{aligned} \tag{20}
$$

The sequence of binomial coefficients, and the distribution of inversions are examples of sequences that satisfy recurrence relation (2); it should be noted that their respective values of γ_n, ($\gamma_n = 3/4$ or $\gamma_n = n^2 - 1$), may be considered as *bounded* between two polynomials of the same degree in n. Whenever this is true the condition $\Gamma = 0$ will hold and $\varphi_n(u)$ will approach $\varphi(u)$. On the other hand, if for example, $\gamma_n = r^n$, then $\Gamma \neq 0$ and $\varphi_n(u)$ does not approach $\varphi(u)$.

6. **Smoothing Formulas.** The general recurrence formula (6),

$$
f_n(x) = \sum_{i=-\infty}^{+\infty} g_n(i) f_{n-1}(x-i),
$$

may be considered as a linear smoothing formula. For example, we may obtain the usual three point smoothing formula based on binomial coefficients for smoothing a distribution $f_1(x)$ into $f_2(x)$ by setting in the above equation $n = 2$, $g_2(i) = \frac{1}{4}\binom{2}{i+1}$ for $-1 \leq i \leq +1$, and $g_2(i) = 0$ for $i < -1$ or $i > +1$. Thus

$$
f_2(x) = \tfrac{1}{4}[f_1(x+1) + 2f_1(x) + f_1(x-1)]. \tag{21}
$$

From considerations found in Section 2, we see that if a variable x_1 has for distribution $f_1(x)$ and a variable s_2 has for distribution $g_2(x)$, then their *sum* $s_2 + x_1$ has for distribution function the smoothed distribution $f_2(x)$. From this point of view, the smoothed distribution $f_2(x)$, obtained by applying a linear smoothing formula, is a "cross" between the original unsmoothed distribution $f_1(x)$ and the artificial weight distribution $g_2(x)$.

Often a smoothing formula is used several times; first on the original distribution, then on the smoothed distribution, and then sometimes on the smoothed smoothed distribution. *If a linear smoothing formula is thus iterated* 1, 2, 3, $\cdots$, n, $\cdots$ *times, the sequence of smoothed distributions obtained upon normalization approaches* $\varphi(u)$. This may easily be demonstrated by showing that Liapounoff's condition for normality, $\Gamma' = 0$, is satisfied. Since in this case the weight distribution $g_n(i)$ is the same for all $n \geq 2$, the corresponding moments of these distributions must all be equal; thus we may write $M_4(n) = M_4(2)$ and $M_2(n) = M_2(2)$ where $n \geq 2$. Substituting in (10), we obtain for Γ'

$$\Gamma' = \operatorname*{Lim}_{n\to\infty} \frac{M_4(1) + (n-1)M_4(2)}{[M_2(1) + (n-1)M_2(2)]^2}, \tag{22}$$

where $M_2(1)$ and $M_4(1)$ are the 2d and 4th moments of the unsmoothed distribution $f_1(x)$. The mean value $\bar{x}_n$ and the standard deviation σ_n of the distribution $f_n(x)$ formed by iterating a smoothing formula $n-1$ times are easily shown to be

$$\begin{aligned} \bar{x}_n &= \bar{x}_1 + (n-1)\bar{s}_w, \\ \sigma_n^2 &= \sigma_1^2 + (n-1)\varphi_w^2, \end{aligned} \tag{23}$$

where $\bar{x}_1$ and σ_1 are the mean and standard deviations of the original unsmoothed distribution, and where $\bar{s}_w$ and σ_w are the mean and standard deviation of the weight distribution $g_2(i)$.

The linear smoothing formula is used in practical work to smooth data. Successive application of one or many such linear formulas will usually smooth *any* set of values to the normal curve of error. The above section serves as a warning of what is introduced by the use of such methods.

It is a pleasure to acknowledge the helpful criticisms and advice of Dr. W. E. Deming in the preparation of the manuscript.

WASHINGTON, D C.