The Annals of Mathematical Statistics
1968, Vol. 39, No. 2, 398–411

THE PROBABILITY THAT THE SAMPLE DISTRIBUTION FUNCTION LIES BETWEEN TWO PARALLEL STRAIGHT LINES[1]

BY J. DURBIN

London School of Economics and Political Science

1. Introduction. Suppose that $0 \leqq x_1 \leqq \cdots \leqq x_n \leqq 1$ is an ordered sample of n independent observations from the uniform $(0, 1)$ distribution. The sample distribution function is

$$\begin{aligned} F_n(x) &= r/n, \qquad & x_r \leqq x < x_{r+1}, \\ &= 0, & x < x_1, \\ &= 1, & x \geqq x_n. \end{aligned}$$

Let S denote the sample path of $F_n(x)$ as x moves from 0 to 1. In this paper we consider the probability $p_n(a, b, c)$ that S lies entirely in the region R between the lines $ny = a + (n + c)x$ and $ny = -b + (n + c)x$, $(a, b, a + c, b - c > 0)$. A knowledge of this probability is important for problems arising in tests of goodness of fit, tests for the Poisson process and tests of serial independence.

For the case $a = b =$ an integer and $c = 0$, a set of simultaneous recurrence relations were obtained by Kolmogorov (1933) as a preliminary step in the development of the asymptotic form of $p_n(a, a, 0)$. These were solved by Massey (1950) to give a linear difference equation of order $2a - 1$ in the quantity $p_n(a, a, 0)n^n/n!$. In Section 2 we obtain the general form of Kolmogorov's relations and from them deduce a generalization of Massey's difference equation expressed in terms of the quantities $q_n(a, b, c) = p_n(a, b, c)(n + c)^n/n!$. Surprisingly, the coefficients of this difference equation do not depend on c or n and in fact depend only on $a + b$. Initial conditions are given whence values of $p_n(a, b, c)$ can be obtained by repeated applications of the difference equation.

For the case where c is a positive integer an explicit generating function for $q_n(a, b, c)$ is given in Section 3, generalizing results of Kemperman (1961) for the case $c = 0$. This is applied to the study of the distribution of the two-sided Kolmogorov statistic $C_n = \max_j |x_j - j/(n + 1)|$ derived from Pyke's (1959) modified sample distribution function. The methods used are different from Kemperman's and are more elementary.

The asymptotic form of $p_n(a, b, 0)$ was obtained by Doob (1949) by methods based on the reflection principle. In Section 5 we consider the application of a variant of this principle to the finite-sample case. It turns out that while the technique does not give exact results in a simple form, some sharp inequalities

Received September 7, 1967.

[1] This report was supported by the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

which should be adequate for many practical applications can be obtained. The simplicity of these results may be contrasted with the complicated character of some exact formulae obtained by Blackman (1958) by a reflection method. The limiting form of $p_n(a, b, c)$ for $a = \alpha n^{\frac{1}{2}}$, $b = \beta n^{\frac{1}{2}}$, $c = \gamma n^{\frac{1}{2}}$ is given in Section 6.

Finally, we mention that the Appendix to the paper by Barton and Mallows (1965) contains a useful bibliography of recent work on Kolmogorov-Smirnov statistics.

2. A generating function for $q_n(a, b, c)$. Let $(X_j, j/n)$ denote the points at which the line with slope $(n + c)/n$ passing through the point $(1, 1)$ meets the horizontal lines $y = j/n, j = n, n - 1, \cdots, -[c] + 1, -[c]$. Clearly $X_n = 1$ and $X_{-[c]}$ is the last value $\geqq 0$ in the sequence $X_n, X_{n-1}, \cdots$. $[c]$ denotes the largest integer $\leqq c$, regardless of the sign of c.

Consider a Poisson process with occurrence rate $n + c$ and sample path S' starting at the point $(0, 0)$. The probability that the path S of $F_n(x)$ remains inside R as x moves from 0 to 1 is the same as the conditional probability that S' remains inside R given that it passes through $(1, 1)$. Let us therefore consider the latter probability. For convenience assume that $a + c$ and $b - c$ are not integers; results for the integral case then follow immediately as limiting values.

The possible points at which S' can cross the line $x = X_j$ and remain inside R have y-coordinates $[j - b + c + 1]/n, [j - b + c + 2]/n, \cdots, [j + a + c]/n$. Denote these points by $A_{j1}, \cdots, A_{jp}$ where $p = [b - c] + [a + c] + 1$. Given that S' passes through A_{ji} the probability that it passes through $A_{j+1,k}$ is the probability of exactly $k - i + 1$ observations in the interval (X_j, X_{j+1}), i.e. $e^{-1}/(k - i + 1)!\,(i, k = 2, 3, \cdots, p - 1; k \geqq i - 1)$.

For $i, k = 1, p$ we have to allow for the fact that $b - c$ and $a + c$ are not integers. Let $1 - \delta = b - c - [b - c]$ and $1 - \epsilon = a + c - [a + c]$. In moving from A_{j1} to $A_{j+1,k}$ $(k = 1, \cdots, p - 1)$ the path S' will remain in R only if at least one observation occurs in the interval $(X_j, X_j + (1 - \delta)/(n + c))$. The probability of k observations in (X_j, X_{j+1}), at least one of which is in $(X_j, X_j + (1 - \delta)/(n + c))$ is $e^{-1}(1 - \delta^k)/k!$. This is therefore the probability of S moving from A_{j1} to $A_{j+1,k}$ and remaining inside R. Similarly, the probability of moving from A_{ji} to $A_{j+1,p}$ inside R is $e^{-1}(1 - \epsilon^{p-i+1})/(p - i + 1)!$, $i = 2, \cdots; p$. Finally, the probability of moving from A_{j1} to $A_{i+1,p}$ inside R is $e^{-1}(1 - \delta^p - \epsilon^p)/p!$ for $\delta + \epsilon \leqq 1$ and is $e^{-1}\{1 - \delta^p - \epsilon^p + (\delta + \epsilon - 1)^p\}/p!$ for $\delta + \epsilon > 1$.

Let $e^{-(j+c)}u_{ji}$ be the probability that S' passes through A_{ji} while remaining inside R and let $u_j' = [u_{j1}, \cdots, u_{jp}]$, $j = c', c' + 1, \cdots, n$, where for notational simplicity we write $c' = -[c]$. (The term in the exponent comes from the fact that $(j + c)/(n + c)$ is the y-coordinate of A_{ji}). The transition from u_j to u_{j+1} is then given by the relation

$$u_{j+1} = Hu_j, \qquad j = c', \cdots, n - 1, \tag{1}$$

where the transition matrix H is given by

$$(2)\quad H = \begin{bmatrix} 1-\delta & 1 & 0 & 0 & \cdots & 0 \\ \frac{1-\delta^2}{2!} & 1 & 1 & 0 & & \vdots \\ \frac{1-\delta^3}{3!} & \frac{1}{2!} & 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & & \ddots & 1 \\ \frac{1-\delta^p-\epsilon^p+h}{p!} & \frac{1-\epsilon^{p-1}}{(p-1)!} & \cdots & \frac{1-\epsilon^2}{2!} & \cdot & 1-\epsilon \end{bmatrix},$$

where $h = 0$ if $\delta + \epsilon \leqq 1$ and $h = (\delta + \epsilon - 1)^p$ if $\delta + \epsilon > 1$.

To obtain the probability of S' arriving at $(1, 1)$ we require the element u_{ni} of u_n corresponding to the point $(1, 1)$; this has $i = [b - c] + 1$. Denote the vector with one in the $([b - c] + 1)$th position and zeros elsewhere by w. Then the required element is $w'H^{[n+c]}u_{c'}$. The probability that S' passes through $(1, 1)$ after remaining inside R is therefore $e^{-(n+c)}w'H^{[n+c]}u_{c'}$. Since the unconditional probability of S' passing through $(1, 1)$ is $e^{-(n+c)}(n + c)^n/n!$, the conditional probability $p_n(a, b, c)$ that S' reaches $(1, 1)$ after remaining in R, given that it reaches $(1, 1)$, is $n!\, w'H^{[n+c]}u_{c'}/(n + c)^n$.

In Section 1 we defined $q_n(a, b, c)$ as $(n + c)^n p_n(a, b, c)/n!$. Consequently $q_n(a, b, c) = w'H^{[n+c]}u_{c'}$ which is the coefficient of z^n in the generating function

$$\begin{aligned} f(z) &= \textstyle\sum_{r=0}^{\infty} w'H^r u_{c'} z^{r+c'} \\ &= z^{c'}w'[I - zH]^{-1}u_{c'}, \end{aligned}$$

the series expansion of $[I - zH]^{-1}$ being valid for $|z\lambda| < 1$ where λ is the largest eigenvalue of H in modulus. Let Γ be the adjoint matrix of $I - zH$. Then

$$(3)\qquad f(z) = z^{c'}w'\Gamma u_{c'}/|I - zH|,$$

where $w'\Gamma u_{c'}$ and $|I - zH|$ are polynomials in z of orders $p - 1$ and p at most respectively.

Putting $z = -y^{-1}$ we have $|I - zH| = y^{-p}\,|H + yI|$. We prove by induction that

$$(4)\qquad |H + yI| = \textstyle\sum_{j=0}^{[a+b]} [(a + b - j)^j/j!]y^{p-j}.$$

Recalling that $1 - \delta = b - c - [b - c]$, $1 - \epsilon = a + c - [a + c]$ and $p = [b - c] + [a + c] + 1$ we have $p = a + b + \delta + \epsilon - 1$. Thus the upper limit of summation in (4) is p if $\delta + \epsilon \leqq 1$ and $p - 1$ if $\delta + \epsilon > 1$.

Denoting by $D_r(\delta, \epsilon)$ the determinant

$$\begin{vmatrix} 1-\delta+y & 1 & 0 & 0 \cdots\cdots 0 \\ (1-\delta^2)/2! & 1+y & 1 & 0 \qquad \cdot \\ (1-\delta^3)/3! & 1/2! & 1+y & 1 \;\cdot \qquad \cdot \\ \cdot & \cdot & \cdot & \ddots \quad 0 \\ \cdot & \cdot & & \ddots \quad 1 \\ \cdot & \cdot & & \\ (1-\delta^r-\epsilon^r+h)/r! & (1-\epsilon^{r-1})/(r-1)! & \cdots\cdots & 1-\epsilon+y \end{vmatrix}$$

we prove (4) by showing that

$$D_r(\delta, \epsilon) = \textstyle\sum_{j=0}^{r'} [(r + 1 - \delta - \epsilon - j)^j/j!]y^{r-j}, \tag{5}$$

where when $\delta + \epsilon \leqq 1$, $r' = r$ and $h = 0$, while when $\delta + \epsilon > 1$, $r' = r - 1$ and $h = (\delta + \epsilon - 1)^r$. Expanding by the first column we have

$$\begin{aligned} D_r(\delta, \epsilon) &= (1 - \delta + y)D_{r-1}(0, \epsilon) - ((1 - \delta^2)/2!)D_{r-2}(0, \epsilon) \\ &\quad + \cdots + (-1)^{r-2}((1 - \delta^{r-1})/(r - 1)!)(1 - \epsilon + y) \\ &\quad + (-1)^{r-1}(1 - \delta^r - \epsilon^r + h)/r!, \quad r = 3, 4, \cdots . \end{aligned} \tag{6}$$

Assuming that (5) is true when r is replaced by $2, 3, \cdots, r - 1$ and taking first the case $\delta + \epsilon \leqq 1$ we find that the coefficient of y^k on the right-hand side of (6) $= 1$, $k = r$; $= (k + 1 - \delta - \epsilon)^{r-k}/(r - k)!$, $k = r - 1, \cdots, 1$; and $= (1 - \delta - \epsilon)^r/r!$, $k = 0$. Thus (5) is true for $\delta + \epsilon \leqq 1$. When $\delta + \epsilon > 1$ the term $(-1)^{r-1}h/r!$ in (6) contributes an amount $(-1)^{r-1}(\delta + \epsilon - 1)^r/r!$ which cancels out the constant term $(1 - \delta - \epsilon)^r/r!$. Thus (5) is true for $\delta + \epsilon > 1$ with $r' = r - 1$. That (5) is true for $r = 2$ is easily verified. Thus it is true for $r = 3, 4, \cdots$; in particular for $r = p$, and for all δ, ϵ.

Substituting $y = -z^{-1}$ in (4) we obtain

$$\begin{aligned} |I - zH| &= \textstyle\sum_{j=0}^{[a+b]} (-1)^j((a + b - j)^j/j!)z^j \\ &= g(z, a + b) \quad \text{say.} \end{aligned} \tag{7}$$

Cross-multiplying in (3) and writing

$$f(z) = \textstyle\sum_{r=c'}^{\infty} q_r(a, b, c)z^r \tag{8}$$

we have

$$g(z, a + b) \textstyle\sum_{r=c'}^{\infty} q_r(a, b, c)z^r = z^{c'}w'\Gamma u_{c'} . \tag{9}$$

Since the right-hand side of (9) is a polynomial of degree at most $c' + p - 1 = -[c] + [a + c] + [b - c]$, on equating the coefficients of z^r on both sides of (9) we have

$$(10)\quad \textstyle\sum_{j=0}^{[a+b]} (-1)^j((a+b-j)^j/j!)q_{r-j}(a, b, c) = 0,$$

$$r = -[c] + [a+c] + [b-c] + 1, -[c] + [a+c] + [b-c] + 2, \cdots$$

where $q_s(a, b, c) = 0$ for $s < 0$.

This is the required generalisation of Massey's difference equation ((1950), p. 118; note, however, the misprint $h = 1$ for $h = 0$ in Massey's formula). For the case $c = 0$, (10) is implied by Kemperman's generating function [(1961), 5.34 and 5.40] It is rather surprising that as a function of a, b, c and n (10) depends only on $a + b$, apart from the range of r. The required probability of S remaining in R is then obtained from the relation $p_n(a, b, c) = q_n(a, b, c)n!(n + c)^{-n}$.

To use (10) in practice we need suitable initial conditions. The most important case is that where c is an integer. $u_{c'}$ then gives the probabilities of S' passing through points on the line $x = 0$, i.e. unity for the point $(0, 0)$ and zero otherwise. Thus $u_{c'}$ is the vector with unity in the $[b + 1]$th position and zeros elsewhere. But $q_n(a, b, c) = w'H^{[n+c]}u_{c'}$ where w is the vector with unity in the $[b - c + 1]$th position and zero elsewhere. Thus $q_r(a, b, c)$ is the $([b - c + 1], [b + 1])$th element of the matrix H^{r+c}, $r = -c, -c + 1, \cdots, [a] + [b] - c$. The initial conditions can therefore be obtained simply by calculating powers of the matrix H. For the case where c is $\geqq 0$ as well as being integral an alternative is to recognise that for $r = -c, -c + 1, \cdots, -c + [a] + [b]$, the probability of leaving R given a sample size of r is the sum of the probabilities of crossing the two lines bounding R, since for $r \leqq [a] + [b] - c$ it is not possible for a single path to cross both lines. The probabilities of crossing the separate lines are, for this purpose, most easily obtained from Dempster's formula (1959, (5')). The required $q_r(a, b, c)$ are then derived from the values so obtained. The method is exemplified in Section 4. A further method is given in the next section immediately following equation (11).

For the general case a method which is straightforward in principle is to calculate $u_{c'}$ directly and to obtain $q_r(a, b, c)$ as the $[b - c + 1]$th element of $H^{r+c}u_{c'}$.

3. Explicit forms. When c is an integer $\geqq 0$ the generating function (3) can be simplified since $w'\Gamma u_{c'}$ is then the $([b - c + 1], [b + 1])$th element of Γ, which is the $([b + 1], [b - c + 1])$th co-factor of $I - zH$. By deleting the $[b + 1]$th row and the $[b - c + 1]$th column of $I - zH$ we see that this co-factor is $(-1)^c$ times $(-z)^c$ times the product of the following two determinants of orders $[b - c]$ and $p - 1 - [b] = [a]$ respectively:

$$\begin{vmatrix} 1 - z(1 - \delta) & -z & 0 \cdots\cdots 0 \\ -z((1 - \delta^2)/2!) & 1 - z & -z \cdots\cdots \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ -z((1 - \delta^{[b-c]})/[b - c]!) & & 1 - z \end{vmatrix}$$

and

$$\begin{vmatrix} 1 - z & -z & 0 & \cdots\cdots\cdots & 0 \\ -z/2! & 1 - z & -z & \cdots\cdots\cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ & & & & -z \\ -z((1 - \epsilon^{[a]})/[a]!) & \cdot\ \cdot & & & 1 - z(1 - \epsilon) \end{vmatrix}$$

But these have the form $|I - zH|$ with $p = [b - c]$, $\epsilon = 0$ and with $p = [a]$, $\delta = 0$ respectively. Their product is therefore $g(z, b - c)g(z, a)$ by (7). Substituting in (3) we have

$$(11) \qquad f(z) = g(z, a)g(z, b - c)/g(z, a + b).$$

Equating the coefficients of z^r on both sides of the relation $g(z, a + b)f(z) = g(z, a)g(z, b - c)$ for $r = 0, 1, \cdots, -c + [a] + [b]$ gives an alternative way of getting the initial conditions needed for the application of the difference equation (10).

Picking out the coefficient of z^n in (11) and multiplying by $n!/(n + c)^n$ we have the exact form

$$(12) \quad p_n(a, b, c) \\ = (n!/(n + c)^n) \textstyle\sum_{r=0}^{[a]} \sum_{s=0}^{[b-c]} (-1)^{r+s}\gamma_{n-r-s}(a - r)^r(b - c - s)^s/r!\,s!$$

for $n \geqq [a] + [b - c]$, where γ_j is the coefficient of z^j in the expansion of $g(z, a + b)^{-1}$, i.e.

$$(13) \quad \gamma_j = (-1)^j \textstyle\sum' \{(i_1 + \cdots + i_m)!/i_1! \cdots i_m!\} \prod_{h=1}^{m} (-1)^{i_h} \\ \cdot\{(a + b - h)^h/h!\}^{i_h},$$

where $\sum'$ indicates summation over all sets $(i_1, \cdots, i_m)$ of non-negative integers satisfying $i_1 + 2i_2 + \cdots + mi_m = j$ and where $m = [a + b]$. This generalises Kemperman's [(1961), 5.40] result for $c = 0$.

4. Application to Pyke's modified sample distribution function. Pyke (1959) has suggested a form of the sample distribution function based on plotting the points $(x_i, i/(n + 1))$ instead of $(x_i, i/n)$ and has given reasons of symmetry supporting the suggestion. There are other reasons for considering tests based on this form arising from the fact that these tests occur more naturally in problems of testing serial independence in time-series analysis and in testing the hypothesis of a Poisson process than do tests based on the usual form $F_n(x)$ [Brunk (1962); Durbin (1967)].

A two-sided statistic of Kolmogorov type based on this approach is

$$(14) \qquad C_n = \max_j |x_j - j/(n + 1)|.$$

Its distribution is given by

$$(15)\quad \Pr(C_n \leqq a/(n+1)) = \Pr(-a/(n+1) + j/(n+1) \leqq x_j \leqq a/(n+1) + j/(n+1) \quad \text{for all} \quad j = 1, \cdots, n).$$

Now the event $F_n(x) \leqq a/n + (n+c)x/n$ for all x in $(0, 1)$ is the same as the event $j/n \leqq a/n + (n+c)x_j/n$, i.e. $x_j \geqq -a/(n+c) + j/(n+c)$, for all $j = 1, \cdots, n$. Similarly the event $F_n(x) \geqq -b/n + (n+c)x/n$ for all x in $(0, 1)$ is the same as the event $j/n \geqq -b/n + 1/n + (n+c)x/n$, i.e. $x_j \leqq (b-1)/(n+c) + j/(n+c)$ for all $j = 1, \cdots, n$. Since $p_n(a, b, c) = \Pr\{-b/n + (n+c)x/n \leqq F_n(x) \leqq a/n + (n+c)x/n$ for all x in $(0, 1)\}$ it follows that

$$(16)\quad p_n(a, b, c) = \Pr(-a/(n+c) + j/(n+c) \leqq x_j \leqq (b-1)/(n+c) + j/(n+c) \quad \text{for all} \quad j = 1, \cdots, n).$$

Comparing (15) and (16) we deduce, by taking $b = a + 1$ and $c = 1$,

$$(17)\qquad \Pr(C_n \leqq a/(n+1)) = p_n(a, a+1, 1).$$

The generating function (11) reduces to

$$(18)\qquad f(z) = \{g(z, a)\}^2/g(z, 2a+1).$$

Writing $q_n(a, a+1, 1) = q_n(a)$ the difference equation (10) gives

$$(19)\quad \textstyle\sum_{j=0}^{[2a+1]} (-1)^j((2a+1-j)^j/j!)q_{r-j}(a) = 0, r = 2[a]+1, 2[a]+2, \cdots.$$

To obtain initial conditions we first note that $q_{-1}(a) = 0$. For $r = 0, 1, \cdots, [a]$ it is clear that if S' reaches the point $(X_r, r/n)$ it cannot previously have left R. Thus the conditional probability that S' remains in R and reaches $(X_r, r/n)$ is the same as the unconditional probability, namely $e^{-(r+1)}(r+1)^r/r!$. It follows that $q_r(a) = (r+1)^r/r!$ for $r = 0, \cdots, [a]$.

For $r = [a+1], \cdots, 2[a]$, S' can cross either the upper line or the lower line but not both on its way from $(0, 0)$ to $(X_r, r/n)$. Thus the probability that it leaves R is the sum of the probabilities that it crosses either line. By symmetry, these two probabilities are equal. The conditional probability of crossing either line given that S' passes through $(X_r, r/n)$ is easily obtained from a formula of Dempster (1959), (5′). Dempster showed that the probability that the path of $F_n(x)$ crosses the line through the points $(0, \delta)$, $(1-\epsilon, 1)(\delta, \epsilon > 0)$ is

$$(20)\quad Q_n(n, \delta, \epsilon) = \epsilon \textstyle\sum_{j=0}^{[n(1-\delta)]} \binom{n}{j}(\epsilon + ((1-\epsilon)/(1-\delta))j/n)^{j-1} \cdot(1 - \epsilon - ((1-\epsilon)/(1-\delta))j/n)^{n-j}.$$

For the present application we find we need to take $\delta = a/r$, $\epsilon = (a+1)/(r+1)$ and $n = r$. On substitution, (20) gives for the probability that S' lies between the two lines,

$$(a+1)(r+1)^{-r} \textstyle\sum_{j=0}^{[r-a]} \binom{r}{j}(a+1+j)^{j-1}(r-a-j)^{r-j}.$$

To obtain $q_r(a)$ we must multiply this by the unconditional probability of S' passing through $(X_r, r/n)$, namely $e^{-(r+1)}(r+1)^r/r!$, and divide by e^{-r}. Performing this calculation we obtain for the set of initial conditions needed for the full solution of (19),

$$\begin{aligned}
q_r(a) &= 0, && r = -1,\\
&= (r+1)^r/r!, && r = 0, \cdots, [a], \qquad (21)\\
&= (r+1)^r/r! - 2(a+1) \textstyle\sum_{j=0}^{[r-a]} ((a+1+j)^{j-1}/j!)\\
&\qquad \cdot (r-a-j)^{r-j}/(r-j)!, && r = [a+1], \cdots, 2[a].
\end{aligned}$$

It is interesting to compare these results with the corresponding ones for Kolmogorov's statistic $D_n = \max_x |F_n(x) - x|$. The generating function for $\Pr(D_n \leqq a/n)$ corresponding to (19) is

$$f(z) = \{g(z, a)\}^2/g(z, 2a) \tag{22}$$

with consequential difference equation

$$\textstyle\sum_{j=0}^{[2a]} (-1)^j((2a-j)^j/j!)q_{r-j}(a) = 0, \quad r = 2[a]+1, 2[a]+2, \cdots \tag{23}$$

and initial conditions

$$\begin{aligned}
q_r(a) &= 1, && r = 0\\
&= r^r/r!, && r = 1, \cdots, [a], \qquad (24)\\
&= r^r/r! - 2a \textstyle\sum_{j=0}^{[r-a]} ((a+j)^{j-1}/j!)\\
&\qquad \cdot (r-a-j)^{r-j}/(r-j)!, && r = [a+1], \cdots, 2[a],
\end{aligned}$$

where now $qn\ (a) = qn\ (a, a, 0)$.

5. Inequalities for $p_n(a, b, c)$. Denote the lines $ny = a + (n+c)x$ and $ny = -b + (n+c)x$ by A, B. As x moves from 0 to 1 let A_1 be the event S crosses A at least once, A_2 the event S crosses A then B at least once, A_3 the event S crosses A then B then A at least once, and so on. Similarly, let B_1 be the event S crosses B at least once, B_2 the event S crosses B then A at least once and so on. Let $\mathcal{A}_1$ be the event S crosses A but not B, $\mathcal{A}_2$ the event S crosses A before crossing B, then crosses B but does not subsequently cross A, and so on. Let E be the event S crosses A or B at least once. Then

$$\begin{aligned}
E &= B_1 + \mathcal{A}_1\\
&= A_1 + B_1 - B_2 - \mathcal{A}_2\\
&\ \ \vdots\\
&= \textstyle\sum_{j=1}^{2k-1} (-1)^{j-1}(A_j + B_j) - B_{2k} - \mathcal{A}_{2k} && (25)\\
&= \textstyle\sum_{j=1}^{2l} (-1)^{j-1}(A_j + B_j) + B_{2l+1} + \mathcal{A}_{2l+1}\,. && (26)
\end{aligned}$$

Let $\alpha_j = \Pr(A_j)$ and $\beta_j = \Pr(B_j)$. Since $p_n(a, b, c) = 1 - \Pr(E)$ we have from (25) and (26).

$$(27)\quad 1 + \textstyle\sum_{j=1}^{2k-1}(-1)^j(\alpha_j + \beta_j) + \beta_{2k} \leqq p_n(a, b, c) \leqq 1 + \sum_{j=1}^{2l}(-1)^j(\alpha_j + \beta_j) - \beta_{2l+1}, \quad k, l = 1, 2, \cdots.$$

A similar pair of inequalities obtained by replacing β_{2k}, β_{2l+1} in (27) by α_{2k}, α_{2l+1} can be determined by an identical argument. Which pair is preferred depends in part on the relative magnitudes of the α's and β's, but generally speaking the form given is to be preferred since sharper inequalities can be given for β_j than for α_j when j is even (there is no advantage when j is odd).

(27) implies the further sets of inequalities

$$(28)\quad 1 + \textstyle\sum_{j=1}^{2k-1}(-1)^j(\alpha_j + \beta_j) + \beta_{2k} \leqq p_n(a, b, c) \leqq 1 + \sum_{j=1}^{2k}(-1)^j(\alpha_j + \beta_j);$$

$$(29)\quad 1 + \textstyle\sum_{j=1}^{2k+1}(-1)^j(\alpha_j + \beta_j) \leqq p_n(a, b, c) \leqq 1 + \sum_{j=1}^{2k}(-1)^j(\alpha_j + \beta_j) - \beta_{2k+1}, \quad k = 1, 2, \cdots.$$

For practical work these are preferable to (27) since the maximum value of r in the α_r, β_r required is the same on both sides.

The probability that S crosses the line $A \equiv ny = a + (n + c)x$ for $a > 0$, $a + c > 0$ is given by the known formulae

$$(30)\quad h_n(a, c) = 1 - (a + c)(n + c)^{-n}\textstyle\sum_{j=0}^{[a]}\binom{n}{j}(j - a)^j(a + c + n - j)^{n-j-1}$$

$$(30')\quad = (a + c)(n + c)^{-n}\textstyle\sum_{j=[a]+1}^{n}\binom{n}{j}(j - a)^j(a + c + n - j)^{n-j-1}$$

[Pyke (1959), Dempster, (1959)]. Since throughout this section n and c remain constant we will write $h(a)$ for $h_n(a, c)$. By reversing the order of the sample it follows that the probability of S crossing B is $h(b - c)$. Thus $\alpha_1 = h(a)$ and $\beta_1 = h(b - c)$.

To evaluate β_2, consider a path in the class B_2, i.e. one that crosses B then A at least once. Let P_B be the first point at which S crosses B and let P_A be the first point after P_B at which S crosses A *from above*. Let p_B, p_A be the x-coordinates of P_B, P_A.. Construct a new sample $y_1, \cdots, y_n$ by interchanging the intervals $(0, p_B)$, (p_B, p_A), the relative positions of sample points within intervals being kept unchanged.

Let the sample path for the new sample be denoted by S'. As x moves from 0 to $p_A - p_B$, S' rises through a vertical distance $(p_A - p_B)(n + c)/n + a/n + b/n$ since over the interval (p_B, p_A) S rose from line B to line A. Let $L(d)$ denote the line $ny = d + (n + c)x$. It follows that S' crosses the line $L(a + b)$ at the point with x-coordinate $p_A - p_B$, say P_B'. Moreover this is the first point at which S' crosses $L(a + b)$ from above.

Conversely, let S' be a sample path which first crosses the line $L(a + b)$ from above at a point P_B' and subsequently first crosses A at P_A. Let p_B', p_A be the x-coordinates of P_B', P_A. Interchange the intervals $(0, p_B')$, (p_B', p_A)

without altering the relative positions of points within the intervals. Let S be the sample path of the new sample. By a similar argument to the above it follows that S crosses B then A. Moreover the point P_B at which S crosses B is the first such point and P_A is the first point at which A is crossed from above after P_B.

We have therefore shown that there is a one-to-one correspondence between sample paths crossing B then A and those crossing $L(a + b)$. It remains to show that corresponding paths are equi-probable. This can either be taken as immediately obvious because of the uniformity of the distribution or can be shown formally as follows.

LEMMA 1. *Let x_r, x_s be the largest of the observations $x_1, \cdots, x_n$ less than p_B, p_A respectively. Let*

$$\begin{aligned} y_i &= x_{i+r} - p_B, & i &= 1, \cdots, s - r, \\ &= x_{i-s+r} + p_A - p_B, & i &= s - r + 1, \cdots, s, \\ &= x_i, & i &= s + 1, \cdots, n. \end{aligned}$$

Then the probability density at $(y_1, \cdots, y_n)$ is the same as that at $(x_1, \cdots, x_n)$.

PROOF. The probability of $s - r$ observations in the interval $(0, p_A - p_B)$, r in $(p_A - p_B, p_A)$, $n - s$ in $(p_A, 1)$ is

$$n![(s - r)!r!(n - s)!]^{-1}(p_A - p_B)^{s-r}p_B^{r}(1 - p_A)^{n-s},$$

which is independent of the order of the intervals. Multiplying by the conditional density at $y_1, \cdots, y_{s-r}$ given that there are $s - r$ observations in $(0, p_A - p_B)$, which is the same as that of $x_{r+1}, \cdots, x_s$ given that there are $s - r$ observations in (p_B, p_A), i.e. $(s - r)!/(p_A - p_B)^{s-r}$, and similarly for the other two intervals we have for the density at $(y_1, \cdots, y_n)$ $n!$, which is the same as the density at $(x_1, \cdots, x_n)$.

It follows that the probability of S crossing B then A is the same as the probability of S crossing $L(a + b)$, i.e. $\beta_2 = h(a + b)$.

Unfortunately the same argument cannot be applied to α_2 since a one-to-one correspondence does not hold between paths crossing $L(a + b)$ or $L(-a - b)$ and paths crossing A then B. However, we are able to obtain some useful inequalities. Suppose S is a path crossing A then B at least once. Let P_A be the point at which it first crosses A *from above* and let P_B be the first point after P_A at which it crosses B. Letting p_A, p_B be the x-coordinates of P_A, P_B construct a new sample by interchanging the intervals $(0, p_A)$, (p_A, p_B). Arguing as before we find that the new sample path, S' say, crosses the line $L(-a - b)$. By Lemma 1 paths S and S' are equi-probable, so that to every path crossing A then B corresponds an equi-probable path crossing $L(-a - b)$. Thus $\alpha_2 \leqq$ the probability of S crossing $L(-a - b)$, i.e. $\alpha_2 \leqq h(a + b - c)$. We can only obtain an inequality because the correspondence is not one-to-one, i.e. not every path which crosses $L(-a - b)$ subsequently crosses B *from above* on its way to $(1, 1)$.

We now seek a correspondence the other way round and hence deduce a lower bound for α_2. Let S' be a sample path crossing $L(-a - b - 1)$ and let

P_A' be the first point at which it crosses $L(-a - b - 1)$ as x moves from 0 to 1. Let P_B' be the first point of S' after P_A' of the form $(x_j, j/n)$ which lies on or above the line $L(-b - 1)$. Let d be the vertical distance of P_B' above $L(-b - 1)$. Then $0 \leqq d < 1/n$ so P_B' lies below the line $L(-b)$, i.e. below B. Letting p_A', p_B' be the x-coordinates of P_A', P_B' construct a new sample by interchanging the intervals $(0, p_A')$, (p_A', p_B').

Let S be the sample path corresponding to the new sample. As x moves from p_A' to p_B', S' rises through a vertical distance $(p_B' - p_A')(n + c)/n + a/n + d$. Consequently as x moves from 0 to $p_B' - p_A'$, S rises to a point which is above A by an amount d. Since S also passes through P_B', which is below B it follows that to every path S' crossing $L(-a - b - 1)$ corresponds a path S crossing A then B. The equi-probability of S' and S is a consequence of the following.

LEMMA 2. *Let* $0 \leqq y_1 \leqq \cdots \leqq y_n \leqq 1$ *be a sample with density* $n!$, *let* y_r *be the largest of the* y'*s* $\leqq p_A'$ *and let* p_B' *denote* y_s, $s > r$. *Let*

$$\begin{aligned} x_i &= y_{i+r} - p_A', & i &= 1, \cdots, s - r, \\ &= y_{i+r-s} + p_B' - p_A', & i &= s - r + 1, \cdots, s, \\ &= y_i, & i &= s + 1, \cdots, n. \end{aligned}$$

Then the sample density at $x_1, \cdots, x_n$ *is* $n!$

PROOF. The probability of $s - r - 1$ observations in $(0, p_B' - p_A')$, one in $(p_B' - p_A', p_B' - p_A' + dx_{s-r})$, r in $(p_B' - p_A', p_B')$ and $n - s$ in $(p_B', 1)$ is

$$n![(s - r - 1)!r!(n - s)!]^{-1}(p_B' - p_A')^{s-r-1}p_A'^{r}(1 - p_B')^{n-s}\,dx_{s-r}.$$

Multiplying by the conditional densities of $x_1, \cdots, x_{s-r-1}$, of $x_{s-r+1}, \cdots, x_s$ and of $x_{s+1}, \cdots, x_n$ we have the result.

The probability of crossing A then B is therefore $\geqq$ the probability of crossing $L(-a - b - 1)$, i.e. $\alpha_2 \geqq h(a + b - c + 1)$. Putting the results together we have

$$h(a + b - c + 1) \leqq \alpha_2 \leqq h(a + b - c). \tag{31}$$

To deal with the remaining cases we need to state the basic results obtained above in a slightly more general form.

LEMMA 3. *The probability of crossing* $L(-e)$ *then* A *then* B *then* A *then* $B \cdots$ (p *crossings*) *at least once* = *the probability of crossing* $L(a + e)$ *then* B *then* A *then* $B \cdots$ ($p - 1$ *crossings*) *at least once* ($e > 0, p = 2, 3, \cdots$).

LEMMA 4. *The probability of crossing* $L(f)$ *then* B *then* A *then* B *then* $A \cdots$ (p *crossings*) *at least once* $\leqq$ *the probability of crossing* $L(-b - f)$ *then* A *then* B *then* $A \cdots$ ($p - 1$ *crossings*) *at least once and* $\geqq$ *the probability of crossing* $L(-b - f - 1)$ *then* A *then* B *then* $A \cdots$ ($p - 1$ *crossings*) *at least once* ($f > 0$, $p = 2, 3, \cdots$).

The proofs are essentially identical to those for the special cases $e = b, f = a$ and $p = 2$ considered above.

Consider β_{2j} = the probability of crossing B then A then $B \cdots$ ($2j$ crossings).

By Lemma 3 this = the probability of crossing $L(a + b)$ then B then $A \cdots$ $(2j - 1$ crossings). Applying Lemma 4 and Lemma 3 successively $j - 1$ times each we have $\beta_{2j} \leqq$ the probability of crossing $L(ja + jb)$ and $\geqq$ the probability of crossing $L(ja + jb + j - 1)$, i.e.

$$(32) \qquad h(ja + jb + j - 1) \leqq \beta_{2j} \leqq h(ja + jb).$$

Similarly for β_{2j+1} we require j applications each of Lemmas 3 and 4 giving

$$(33) \qquad h(ja + (j + 1)b - c + j) \leqq \beta_{2j+1} \leqq h(ja + (j + 1)b - c).$$

For α_{2j} we need j applications of Lemma 4 and $j - 1$ of Lemma 3 giving

$$(34) \qquad h(ja + jb - c + j) \leqq \alpha_{2j} \leqq h(ja + jb - c),$$

and for α_{2j+1} we have similarly

$$(35) \qquad h((j + 1)a + jb + j) \leqq \alpha_{2j+1} \leqq h((j + 1)a + jb).$$

On substituting in (27), (28) or (29) we obtain the required inequalities on $p_n(a, b, c)$. From a practical point of view the most important of these is the simplest which is obtained by taking $k = 1$ in (28) giving

$$(36) \quad 1 - h(a) - h(b - c) + h(a + b) \leqq p_n(a, b, c) \leqq 1 - h(a) - h(b - c) + h(a + b) + h(a + b - c).$$

Taking $k = 1$ in (29) gives

$$(37) \quad \begin{aligned} &1 - h(a) - h(b - c) + h(a + b) + h(a + b - c + 1) - h(2a + b) \\ &\quad - h(a + 2b - c) \\ &\leqq p_n(a\ b, c) \leqq 1 - h(a) - h(b - c) + h(a + b) + h(a + b - c) \\ &\quad - h(a + 2b - c + 1). \end{aligned}$$

Asymptotically, for p_n moderately near to unity the difference between the bounds is very small. For instance, for $c = 0$ and $a = b = \lambda n^{\frac{1}{2}}$ corresponding to the Kolmogorov statistic, (36) gives $1 - 2h(n^{\frac{1}{2}}\lambda) + h(2n^{\frac{1}{2}}\lambda) \leqq p_n(n^{\frac{1}{2}}\lambda, n^{\frac{1}{2}}\lambda, 0) \leqq 1 - 2h(n^{\frac{1}{2}}\lambda) + 2h(2n^{\frac{1}{2}}\lambda)$ Applying Smirnov's result $h(n^{\frac{1}{2}}\lambda) \rightarrow e^{-2\lambda^2} = \alpha$ say, we have $h(2n^{\frac{1}{2}}\lambda) \rightarrow e^{-8\lambda^2} = \alpha^4$ Taking $\alpha = 0.1$ we see that for $p_n >$ about 0.8 the difference between the upper and lower bounds of (36) is asymptotically less than 0.0001. Similarly, the difference between the bounds (37) is asymptotically $e^{-2(3\lambda)^2} = \alpha^9$.

By judicious use of the information in Birnbaum's (1952) Table we are able to obtain the following results for finite samples, again for the case of Kolmogorov's statistic for which $c = 0$ and $a = b$.

These results suggest that (36) is a very useful inequality for values of p_n not too far from unity but that the improvement from (36) to (37) is only likely to be important as the sample size becomes fairly large.

6. Asymptotic forms. The limiting form of $p_n(a, b, c)$ is easily obtained by

TABLE

Differences between bounds in (36) *and* (37)

n	a	$p_n(a, a, 0)$	Difference between bounds	
			(36)	(37)
10	3	0.7295	0.0003	0.0003
10	4	.9410	.0000	.0000
20	4	.6473	.0011	.0009
20	5	.8624	.0000	.0000
40	5	.4808	.0054	.0036
40	6	.7016	.0005	.0004
40	7	.8471	.0000	.0000
60	5	.2324	.0312	.0162
60	6	.4478	.0070	.0040
60	7	.6404	.0012	.0007

Doob's heuristic method. Putting $a = \alpha n^{\frac{1}{2}}$, $c = \gamma n^{\frac{1}{2}}$ we have $nF_n(x) < a + (n + c)x$ when $n^{\frac{1}{2}}\{F_n(x) - x\} < \alpha + \gamma x$, $0 \leqq x \leqq 1$. Letting $n \to \infty$, Doob (1949) replaces this by $\zeta(t)(t + 1)^{-1} < \alpha + \gamma t(t + 1)^{-1}$, i.e. $\zeta(t) < \alpha + (\alpha + \gamma)t$, $0 \leqq t \leqq \infty$, where $\zeta(t)$ is a Wiener process. Using Doob's formula (4.2) this gives

$$(38) \qquad \lim_{n\to\infty} h_n(\alpha n^{\frac{1}{2}}, \gamma n^{\frac{1}{2}}) = e^{-2\alpha(\alpha+\gamma)}.$$

A continuous version of the reflection method of Section 5 then gives immediately

$$\begin{aligned} &\lim_{n\to\infty} p_n(\alpha n^{\frac{1}{2}}, \beta n^{\frac{1}{2}}, \gamma n^{\frac{1}{2}}) \\ (39) \quad &= 1 - \textstyle\sum_{j=1}^{\infty} [\exp(-2\{j\alpha + (j-1)\beta\}\{j\alpha + (j-1)\beta + \gamma\}) \\ &\quad + \exp(-2\{(j-1)\alpha + j\beta\}\{(j-1)\alpha + j\beta - \gamma\}) \\ &\quad - \exp(-2j(\alpha + \beta)(j\alpha + j\beta + \gamma)) - \exp(-2j(\alpha + \beta)(j\alpha + j\beta - \gamma))]. \end{aligned}$$

It is worth noting that the methods of this paper permit the development of an elementary derivation of (39) which avoids the mathematical difficulties inherent in the justification of Doob's approach. Following Wilks (1962), p. 339, one puts $a = \alpha n^{\frac{1}{2}}$, $c = \gamma n^{\frac{1}{2}}$ in (30)′ and proceeds to the limit using Stirling's approximation. This gives

$$\begin{aligned} &\lim_{n\to\infty} h_n(\alpha n^{\frac{1}{2}}, \gamma n^{\frac{1}{2}}) \\ &\quad = (\alpha + \gamma)(2\pi)^{-\frac{1}{2}} \textstyle\int_0^1 y^{-\frac{1}{2}}(1 - y)^{-3/2} \exp(-(\alpha + \gamma y)^2/2y(1 - y))\, dy. \end{aligned}$$

Evaluating the integral we obtain (38), Substituting in (32)–(35) and then (27) and proceeding to the limit we obtain (39).

Acknowledgments. This paper is based on research done in part at the University of Washington with the support of the Office of Naval Research. I am grateful to Professors Z. W. Birnbaum and R. Pyke for providing both the stimulus and the facilities which led to the initiation of this work.

REFERENCES

BARTON, D. E. and MALLOWS, C. L. (1965). Some aspects of the random sequence. *Ann. Math. Statist.* **36** 236.

BIRNBAUM, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *J. Amer. Statist. Assoc.* **47** 425.

BLACKMAN, J. (1958). Correction to "An extension of the Kolmogorov distribution". *Ann. Math. Statist.* **29** 318.

BRUNK, H. D. (1962). On the range of the difference between hypothetical distribution function and Pyke's modified empirical distribution function. *Ann. Math. Statist.* **33** 525.

DEMPSTER, A. P. (1959). Generalized D_n^+ statistics. *Ann. Math. Statist.* **30** 593.

DOOB, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20** 277.

DURBIN, J. (1967). Tests of serial independence based on the cumulated periodogram. Paper presented at the 36th Session of the International Statistical Institute, Sydney, 1967.

KEMPERMAN, J. H. B. (1961). *The Passage Problem for a Stationary Markov Chain.* Univ. of Chicago Press.

KOLMOGOROV, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari.* **4** 83.

MASSEY, F. J. (1950). A note on the estimation of a distribution function by confidence limits. *Ann. Math. Statist.* **21** 116.

PYKE, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Statist.* **30** 568.

WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.