SOME EXACT RESULTS FOR ONE-SIDED DISTRIBUTION TESTS OF THE KOLMOGOROV-SMIRNOV TYPE[1]

By P. Whittle

Statistical Laboratory, University of Cambridge

0. Summary. I consider the calculation of the probability P_n that the graph of a sample distribution function lie wholly to one side of a given arbitrary contour. A generating function approach is described in Section 2, and P_n calculated exactly for some simple types of contour. Upper and lower bounds of the correct asymptotic form (relations (14), (15)) are obtained for P_n in the case of a straight line contour.

1. Introduction. Assume, as usual, that the observations are distributed rectangularly in (0, 1), and so have distribution function

$$F(x) = x \qquad (0 \leqq x \leqq 1).$$

Let the sample consist of n ordered observations

$$0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq 1,$$

and let $F_n(x)$ denote the sample distribution function

$$F_n(x) = \sum_{x_j < x} \frac{1}{n}.$$

My aim is to use the methods developed by Wald and Wolfowitz [7] and Daniels [3], and previously applied by Birnbaum and Tingey [1], to obtain an exact calculation of the probability $P_n = \Pr[F_n(x) \leqq G(x);\ 0 \leqq x \leqq 1]$ for certain functions $G(x)$.

2. General formulae. Suppose that the function $G(x)$ is monotone non-decreasing. Then we can uniquely define a non-decreasing sequence of constants α_j by

$$\alpha_j = G^{-1}(j/n), \qquad (j = 1, 2 \cdots n) \tag{1}$$

and we can rewrite $\Pr[F_n(x) \leqq G(x)]$ as

$$P_n = \Pr(x_1 \geqq \alpha_1, x_2 \geqq \alpha_2 \cdots, x_n \geqq \alpha_n).$$

Following Wald and Wolfowitz [7], let us introduce the polynomials,

$$P_0(x) = 1,$$

$$P_j(x) = \int_{\alpha_j}^{x} dx_j \int_{\alpha_{j-1}}^{x_j} dx_{j-1} \cdots \int_{\alpha_1}^{x_2} dx_1, \qquad (j \geqq 1)$$

Received June 1, 1959; revised July 1, 1960.

[1] This paper was written while the author was a member of the Applied Mathematics Laboratory, Department of Scientific and Industrial Research, Wellington, New Zealand.

which are related to one another and to the probability P_n by the equations,

$$P_j(x) = \int_{\alpha_j}^{x} P_{j-1}(u)\, du, \tag{2}$$

$$P_n = n!\, P_n(1). \tag{3}$$

Since $P_j(x)$ is a polynomial in x of order j exactly, and

$$P_j'(x) = P_{j-1}(x), \qquad (j = 1, 2, \cdots)$$

the $P_j(x)$'s constitute an *Appell set of polynomials* ([4] p. 235) and their formal generating function,

$$T(\theta) = \sum_0^\infty P_j(x)\theta^j, \tag{4}$$

is of the form

$$T(\theta) = A(\theta)e^{x\theta}, \tag{5}$$

where $A(\theta)$ is the formal generating function of the constant terms in the polynomials. ("Formal" in the sense that we are concerned only with relations between the first few terms in the expansions of $A(\theta)$, $T(\theta)$, and not in the convergence of these expansions.)

$A(\theta)$ (or at least its first $n + 1$ terms) can be regarded as being determined (for a given $G(x)$) by the conditions that $P_0 = 1$ and that α_j be the greatest real zero of $P_j(x)$, $(j = 1, 2 \cdots n)$. (Equation (2) shows that α_j is a zero of $P_j(x)$, and it must be the greatest zero, since $P_j(x)$ is intrinsically positive for $x > \alpha_j$.)

It follows from equations (1), (2) and (3) that

$$\frac{P_n}{n!} = \text{coefficient of } \theta^n \text{ in } A(\theta)e^{\theta}. \tag{6}$$

Rather than regarding $A(\theta)$ as being determined by $G(x)$, it may be simpler to prescribe $A(\theta)$, calculate the α_j's from the relation

$$P_j(\alpha_j) = 0, \qquad (j = 1, 2 \cdots n)$$

and so effectively determine $G(x)$. This we proceed to do for some simple cases in the next section.

3. Some special cases. If $A(\theta)$ is of the form

$$A(\theta) = \sum_0^m A_k\theta^k,$$

then α_j will be the greatest root of the equation

$$P_j(\alpha) = \sum_{k=0}^{\min(j,m)} \frac{A_k\, \alpha^{j-k}}{(j-k)!} = 0. \tag{7}$$

As a special case, consider

$$A(\theta) = 1 - B\theta^m \qquad (m \text{ integral}, B > 0).$$

for which it follows from (6) and (7) that we shall have

$$\alpha_j = [Bj(j-1)(j-2)\cdots(j-m+1)]^{1/m}, \qquad (j = 1, 2\cdots n), \tag{8}$$

$$P_n = 1 - \alpha_n^m. \tag{9}$$

B and m are, of course, so chosen that $\alpha_n \leqq 1$. In the figure we have taken coordinates x, $y = G(x)$, and the upper heavy contour is a curve drawn through the points

$$x = \alpha_j,$$
$$y = G(\alpha_j) = j/n,$$

where α_j is given by formula (8). The contour initially rises vertically (to the point corresponding to $j-1$) after which it rises convexly to the x-axis and is quickly asymptotic to the straight line $\alpha_j = B^{1/m}(j - \frac{1}{2}(m-1))$. Seeing that one expects the greatest deviations midway in the range $x = [0, 1]$, it is reasonable

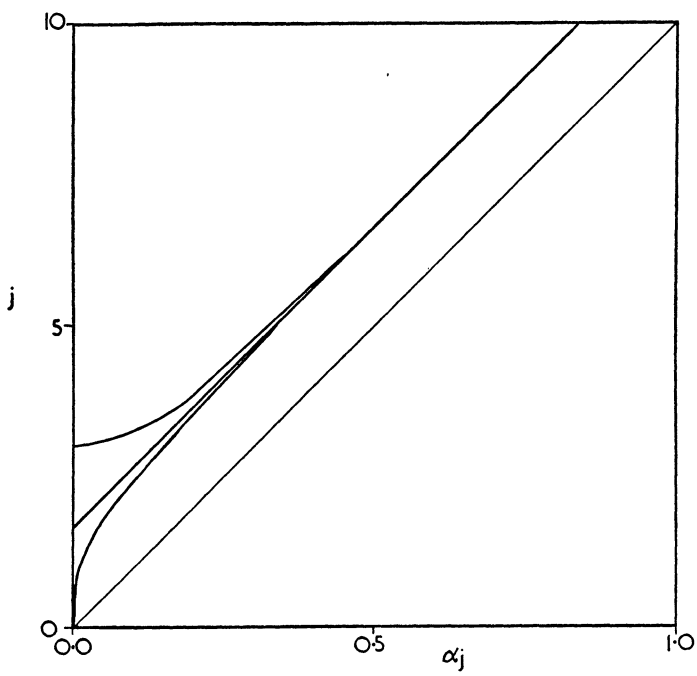

FIG. 1. The upper and lower curves are those described by equations (8) and (10) respectively, with the following numerical values of the parameters: $n = 10$; $m = 4$, $B = 0.0^49651$; $\beta = 0.3767$, $\gamma = 3.723$. It will be seen that both curves are quickly convergent to a common straight line, and that for both curves $s = 1 - \alpha_{10} = 0.1649$. The probabilities, P_{10}, that an empirical distribution curve should lie completely below the upper or the lower curve are 0.5136 and 0.4192 respectively, by equations (9) and (11). The Kolmogorov probability for the region bounded by the straight line asymptote is $1 - e^{-20s^2} = 0.4195$.

to choose a function which is concave, and it is unfortunate that the present calculations have led to a $G(x)$ which is convex. However, we shall find a use for formulae (8), (9) in the next section.

The scale constant B can be chosen according to several criteria. For instance, we should obtain a fairly symmetric contour if we so chose B that

$$\alpha_n = 1 - m/(n + 1).$$

Varying m, we should then obtain a nested sequence of contours, beginning at $m = 1$ with $\alpha_j = j/(n + 1) = E(x_j)$. Note that formula (9) for the special case $m = 1$ is to be found in the article by Daniels [3] and is also a special case of Lemma 1 in a recent paper by Pyke [6].

Another elementary choice for $A(\theta)$ is

$$A(\theta) = \sum R_k e^{\beta_k \theta}.$$

In this case α_j will be the largest root of

$$\sum R_k(\alpha + \beta_k)^j = 0.$$

Thus, the particular case

$$A(\theta) = (e^\gamma - e^{\beta\theta}) / (e^\gamma - 1)$$

yields

$$\alpha_j = \beta/(e^{\gamma/j} - 1), \tag{10}$$

$$P_n = [e^\gamma - (1 + \beta)^n] / (e^\gamma - 1). \tag{11}$$

Again, β, γ must be so chosen that $\alpha_n \leqq 1$. Formula (10) leads to the lower contour of the figure; which rises rapidly and concavely from zero and is soon asymptotic to a straight line. By varying the constants β and γ one can obtain various families of nested contours whose shape would seem to be an improvement on the straight line contour often adopted.

4. Bounds for the significance points of the Kolmogorov test. Suppose m and B can be chosen in formula (8), and β and γ in formula (10), such that in both cases

$$\alpha_n = 1 - s, \tag{12}$$

$$(\partial\alpha_j/\partial j)_{j=n} = 1/n, \tag{13}$$

where s is prescribed, and the derivative is defined by considering j as a continuous variable in the expressions given for α_j. On account of the respective convexity and concavity of the two curves, the contours (1) will respectively lie completely above and completely below the line $y = x + s$ in the square $0 \leqq x, y \leqq 1$. The situation is presented in the figure. That is, the two values of P_n constitute respectively upper and lower bounds for the probability

$$\Pr[\max[F_n(x) - F(x)] \leqq s] = 1 - Q_n(s).$$

We shall press this calculation somewhat further to obtain the following

THEOREM:

$$Q_n(s) = \Pr[\max [F_n(x) - F(x)] > s]$$
$$> (1-s)^{2ns+2} > (1-s)^2 e^{-2ns^2 - ns^3/(1-s)} \quad (0 \leqq s \leqq 1) \tag{14}$$

$$Q_n(s) < e^{-2ns^2+3.83n[s/(1-s)]^3} \quad (0 \leqq s \leqq 0.31). \tag{15}$$

These two inequalities together are obviously sufficient to prove the known result that, for fixed t,

$$\lim_{n\to\infty} Q(t/n^{\frac{1}{2}}) = e^{-2t^2}.$$

However, our methods are not strong enough to prove the conjecture, made in [2], that $Q_n(s) < e^{-2ns^2}$. For comparison, note should be made of the result, proved in [5], that there exists a constant c, independent of s and n, such that $Q_n(s) < ce^{-2ns^2}$.

To prove (14), we take the α_j sequence (8). If we use condition (12) to determine B, then we find that equations (13) and (9) amount, if m is integral, to

$$\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-m+1} = \frac{m}{n(1-s)}, \tag{16}$$

$$1 - P_n = (1-s)^m. \tag{17}$$

By eliminating m from these two equations, we obtain an expression for $1 - P_n$ which constitutes a lower bound for the probability $Q_n(s)$. If s is such that there is indeed an integral m satisfying (16) then

$$1 - s = m\left[n\sum_{k=0}^{m-1}(n-k)^{-1}\right]^{-1}$$
$$= \left[\sum_{j=0}^{\infty} n^{-j}(m^{-1}\sum_k k^j)\right]^{-1}$$
$$\leqq \left[\sum_{j=0}^{\infty} n^{-j}(m^{-1}\sum k)^j\right]^{-1} = 1 - (m-1)/(2n),$$

or

$$m \leqq 2ns + 1.$$

If (16) cannot be satisfied by some integral m, then if we are to err on the conservative side we must increase s until such a solution can be found. However, since this procedure will not increase m by as much as unity we shall have, under all circumstances,

$$m < 2ns + 2. \tag{18}$$

Substituting (18) into (17) we obtain

$$1 - P_n > (1 - s)^{2ns+2} = (1 - s)^2 e^{2ns\log(1-s)}$$

from which (14) follows immediately.

To establish (15), we take the α_j sequence (10). Solving for β from (12) and substituting in (13), (11) we obtain

$$s = (e^{-2\phi} - (1 - 2\phi))/(2\phi), \tag{19}$$

$$1 - P_n = [(1 + (2/\phi)\sinh^2\phi)^n - 1]/(e^{2n\phi} - 1), \tag{20}$$

where $\phi = \gamma/(2n)$. Eliminating ϕ from (19), (20) we shall obtain an expression for $1 - P_n$ which majorises $Q_n(s)$.

If two positive quantities c, d satisfy $c/d \leqq 1$, then it is certainly true that $c/d \leqq (c + 1)/(d + 1)$. We thus deduce from (20) that

$$1 - P_n \leqq (1 + (2/\phi)\sinh^2\phi)^n e^{-2n\phi}$$

Let us restrict ourselves to the range

$$0 \leqq \phi \leqq 0.4. \tag{21}$$

Now,

$$\sinh\phi = \phi + R_1$$

where

$$R_1 \leqq \tfrac{1}{6}\phi^3 e^{\phi^2/20}.$$

Thus

$$1 + (2/\phi)\sinh^2\phi = 1 + 2\phi + R_2,$$

where

$$R_2 < (2/\phi)(2R_1\phi + R_1^2) < 0.6811\phi^3, \tag{22}$$

in view of (21). It also follows from (21), (22) that

$$0 \leqq 2\phi + R_2 \leqq 1,$$

so that

$$\begin{aligned}\log(1 + 2\phi + R_2) &= (2\phi + R_2) - \tfrac{1}{2}(2\phi + R_2)^2 + R_3 \\ &= 2\phi - 2\phi^2 + R_4\end{aligned}$$

where

$$\begin{aligned}|R_3| &< \tfrac{1}{3}|2\phi + R_2|^3, \\ |R_4| &< |R_2(1 - 2\phi) - \tfrac{1}{2}R_2^2 + R_3| \\ &< R_2 + \tfrac{1}{2}R_2^2 + |R_3| \\ &< 3.83\phi^3.\end{aligned}$$

Thus

$$1 - P_n \leqq e^{n(-2\phi^2+3.83\phi^3)}. \tag{23}$$

Turning now to the relation between ϕ and s, we note that if the function of ϕ in the right-hand member of (19) is denoted $f(\phi)$, then f is monotone and

$$\phi/(1+\phi) \leqq f(\phi) \leqq \phi.$$

Thus

$$s \leqq \phi \leqq s/(1-s).$$

Relation (15) now follows immediately from (23), (24). It is easily shown that condition (21) ensures that $0 \leqq s \leqq 0.31$.

REFERENCES

[1] Z. W. Birnbaum and Fred H. Tingey, "One sided contours for probability distribution functions," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 592–596.

[2] Z. W. Birnbaum and R. C. McCarty, "A distribution-free upper confidence bound for $\Pr(Y < X)$, based on independent samples of X and Y," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 558–562.

[3] H. E. Daniels, "The statistical theory of the strengths of bundles of threads," *Proc. Roy. Soc. London*, Ser. A, Vol. 183 (1945), pp. 405–435.

[4] A. Erdelyi, et al, *Higher Transcendental Functions*, Vol. 3, McGraw-Hill Book Co., New York, 1953.

[5] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 642–669.

[6] Ronald Pyke, "The supremum and infimum of the Poisson process," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 568–576.

[7] A. Wald and J. Wolfowitz, "Confidence limits for continuous distribution functions," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 105–118.