The Annals of Mathematical Statistics
1972, Vol. 43, No. 1, 58-64

THE CALCULATION OF DISTRIBUTIONS OF TWO-SIDED KOLMOGOROV-SMIRNOV TYPE STATISTICS

BY MARC NOÉ

M.B.L.E. Research Laboratory, Brussels

Let $X_1^n \leqq X_2^n \leqq \cdots \leqq X_n^n$ be the order statistics of a size n sample from any distribution function F not necessarily continuous. Let α_j, β_j, $(j = 1, 2, \cdots, n)$ be any numbers. Let $P_n = P(\alpha_j < X_j^n \leqq \beta_j,\ j = 1, 2, \cdots, n)$. A recursion is given which calculates P_n for any F and any α_j, β_j. Suppose now that F is continuous. A two-sided statistic of Kolmogorov-Smirnov type has the distribution function

$$P_{KS} = P[\sup n^{\frac{1}{2}}\psi(F) \cdot |F^n - F| \leqq \lambda],$$

where F^n is the empirical distribution function of the sample and $\psi(x)$ is any nonnegative weight function. As P_{KS} has the form P_n, its calculation as a function of λ can be carried out by means of the recursion. This has been done for the case $\psi(x) = [x(1-x)]^{-\frac{1}{2}}$. Curves are given which represent λ versus $1 - P_{KS}$ for $n = 1, 2, 10, 100$. From additional computations, the precision of a truncated development of $1 - P_{KS}$ in powers of λ^{-2} has been determined.

1. Introduction. Let Z be any random variable with distribution function $P(Z \leqq z) = F(z)$, not necessarily continuous. Let $X_1^n \leqq X_2^n \leqq \cdots \leqq X_n^n$ be the order statistics of a size n sample from $F(z)$. Let $\{\alpha_j, \beta_j; j = 1, 2, \cdots, n\}$ be any numbers which we call α-boundaries and β-boundaries respectively. We are interested in the probability

$$P_n = P(\alpha_j < X_j^n \leqq \beta_j; j = 1, 2, \cdots, n). \tag{1}$$

In the two special cases $\alpha_j = -\infty$, and $\beta_j = +\infty$, $(j = 1, 2, \cdots, n)$, this probability will be written $\underline{P}_n$ and $\bar{P}_n$ respectively.

Such probabilities are related to the statistics of the Kolmogorov-Smirnov type. Suppose that $F(z)$ is continuous and let

$$L = \sup_z n^{\frac{1}{2}}[F^n(z) - F(z)]\psi_L[F(z)], \tag{2}$$

$$M = \sup_z n^{\frac{1}{2}}[F(z) - F^n(z)]\psi_M[F(z)], \tag{3}$$

where $F^n(z)$ is the empirical distribution function of the sample and $\psi_L(x)$, $\psi_M(x)$ are some nonnegative weight functions. A two-sided statistic of the Kolmogorov-Smirnov type has the distribution function $P(L \leqq \lambda, M \leqq \lambda)$ which is of the form (1) for well-chosen boundaries. The distribution of a one-sided statistic is $P(L \leqq \lambda)$ or $P(M \leqq \lambda)$ and has the form of $\bar{P}_n$ respectively $\underline{P}_n$.

In Section 2, a recursion formula for P_n is given which is valid for any $F(z)$ and any boundaries (8). As can be expected, the function $F(z)$ is involved only

Received December 15, 1970.

as $F(\alpha_j)$ and $F(\beta_j)$, $(j = 1, 2, \cdots, n)$. An elementary proof of the formula is given in Section 3, where it is incidentally shown that this formula is only one of the simplest recursions among a set of similar recursions implying the same amount of computation.

In the one-sided case, recursion (8) evidently can be simplified. The simplified formula for $\bar{P}_n$ is (9) and was already obtained by Noé and Vandewiele (1968) for a continuous $F(z)$. Faster recursions than (9) are known for $\bar{P}_n$ and $\underline{P}_n$, but they are not suited to numerical computation because they involve small differences of large numbers (Wald and Wolfowitz (1939), Daniels (1945), Noé and Vandewiele (1968)). On the contrary, formulas (8) and (9) involve sums of nonnegative terms only. As regards the two-sided case with a continuous $F(z)$, Wald and Wolfowitz give a general recursion for P_n but involving integrals such that it is not directly usable. An usable recursion is given by Noé and Vandewiele but only for the special case $\alpha_n \leqq \beta_1$. Steck (1971) expresses P_n by means of certain determinants.

Recursion (8) also calculates the power of a test $\{\alpha_j < X_j^n \leqq \beta_j\}$ with respect to an alternative hypothesis $F'(z)$, because this power is expressed by $(1 - P_n')$ where P_n' is defined as in (1) replacing the probabilisation $P(Z \leqq z) = F(z)$ by $P'(Z \leqq z) = F'(z)$.

For Kolmogorov-Smirnov type statistics, the particular weight function

$$\psi_L(x) = \psi_M(x) = [x(1 - x)]^{-\frac{1}{2}} \tag{4}$$

is sometimes chosen because it assigns, in a certain sense, the same weight to each point of $F(z)$ and because it has certain asymptotical properties of optimality as for the power of the test. See Anderson and Darling (1952), Vandewiele and de Witte (1966), Borokov and Sycheva (1968). Tables of significance points for the corresponding one-sided test were calculated by Noé and Vandewiele by means of the general formula (9), and by Borokov and Sycheva by means of a formula restricted to this particular weight function. Section 4 of the present paper is devoted to the corresponding two-sided test. Several curves represent P_n as a function of λ. The precision of a truncated power series for P_n is given too.

2. The recursion. Without loss of generality we assume

$$\alpha_1 \leqq \alpha_2 \leqq \cdots \leqq \alpha_n\,, \qquad \beta_1 \leqq \beta_2 \leqq \cdots \leqq \beta_n\,. \tag{5}$$

Furthermore we exclude trivial cases assuming

$$\alpha_j < \beta_j\,; \qquad j = 1, 2, \cdots, n\,. \tag{6}$$

Let now $\{\gamma_1, \gamma_2, \cdots, \gamma_{2n-1}, \gamma_{2n}\}$ be the $2n$ boundaries arranged in non-decreasing order. Let $\alpha_0 = \beta_0 = \gamma_0 = -\infty$ and $\alpha_{n+1} = \beta_{n+1} = \gamma_{2n+1} = +\infty$. We thus have $\gamma_0 \leqq \gamma_1 \leqq \cdots \leqq \gamma_{2n} \leqq \gamma_{2n+1}$. Let $g(m)$, $(m = 0, 1, \cdots, 2n)$, be the number

of α-boundaries in $\{\gamma_0, \gamma_1, \cdots, \gamma_m\}$. Let $h(m) - 1$, $(m = 1, 2, \cdots, 2n + 1)$, be the number of β-boundaries in $\{\gamma_0, \gamma_1, \cdots, \gamma_{m-1}\}$. In particular $g(0) = 0$, $g(2n) = n$, $h(1) = 1$, $h(2n + 1) = n + 1$. Clearly one has for $m = 1, 2, \cdots, 2n + 1$

$$\begin{aligned} &\alpha_{g(m-1)} \leqq \gamma_{m-1} \leqq \gamma_m \leqq \alpha_{g(m-1)+1}\,, \\ &\beta_{h(m)-1} \leqq \gamma_{m-1} \leqq \gamma_m \leqq \beta_{h(m)}\,. \end{aligned} \tag{7}$$

Let $p_m = F(\gamma_m) - F(\gamma_{m-1})$, $(m = 1, 2, \cdots, 2n + 1)$. The probability P_n defined by (1) can be calculated by the following recursion which is proved in Section 3 and which involves $F(z)$ in the form of the probabilities p_m. It is understood that ${p_m}^0 = 1$ even if $p_m = 0$.

$$\begin{aligned} Q_0(0) &= 1 \\ Q_i(m) &= \textstyle\sum_{k=h(m)-1}^{i} C_i^k \cdot Q_k(m-1) \cdot {p_m}^{i-k} \\ &\qquad\qquad i = h(m+1) - 1, h(m+1), \cdots, g(m-1)\,, \\ Q_{g(m-1)+1}(m) &= 0 \qquad\qquad m = 1, 2, \cdots, 2n \\ P_n &= Q_n(2n)\,. \end{aligned} \tag{8}$$

At the beginning of a step m of this recursion, including step 1, the numbers $Q_{h(m)-1}(m-1)$, $Q_{h(m)}(m-1)$, $\cdots$, $Q_{g(m-1)}(m-1)$ are available. Since clearly $g(m) \leqq g(m-1) + 1$, step m provides the analogous numbers for $m + 1$ instead of m. From (6) one verifies that $h(m + 1) - 1 > g(m - 1)$ never holds and hence that the summation involved in (8) never is void.

In the one-sided case $\beta_j = +\infty$, $(j = 1, 2, \cdots, n)$, one has $\gamma_m = \alpha_m$, $(m = 0, 1, \cdots, n)$, $\gamma_m = +\infty$, $(m = n + 1, n + 2, \cdots, 2n + 1)$, and hence $p_m = 0$, $(m = n + 2, n + 3, \cdots, 2n + 1)$, $g(m) = m$, $(m = 0, 1, \cdots, n)$, $h(m) = 1$, $(m = 1, 2, \cdots, n + 1)$. Recursion (8) then becomes

$$\begin{aligned} Q_0(0) &= 1 \\ Q_i(m) &= \textstyle\sum_{k=0}^{i} C_i^k \cdot Q_k(m-1) \cdot {p_m}^{i-k} \\ &\qquad\qquad i = 0, 1, \cdots, m-1\,, \\ Q_m(m) &= 0 \qquad\qquad m = 1, 2, \cdots, n+1 \\ \bar{P}_n &= Q_n(n+1)\,, \end{aligned} \tag{9}$$

where $p_m = F(\alpha_m) - F(\alpha_{m-1})$, $(m = 1, 2, \cdots, n + 1)$. This is recursion (12) (13) of [4]. (Corrigenda in [4]: In formula (12), $\sum_{k=0}^{i-1}$ should be read $\sum_{k=0}^{i}$. Formula $Q_j(a_j) = 0$ should be added).

For a continuous $F(z)$, it is pointed out in [4] that, in the practically important region, P_n can be closely approximated from $\underline{P}_n$ and $\bar{P}_n$ by means of the inequalities $\underline{P}_n + \bar{P}_n - 1 \leqq P_n \leqq \underline{P}_n \cdot P_n$. In a symmetrical case, i.e., when $\underline{P}_n = \bar{P}_n$, one can approximate $\underline{P}_n = \bar{P}_n$ from P_n by these inequalities rewritten $P_n^{\frac{1}{2}} \leqq \underline{P}_n = \bar{P}_n \leqq (1 + P_n)/2$. As the sum involved in (8) has fewer terms than the sum involved

in (9), it would be interesting to investigate to what extent the latter approximation saves computation with respect to an exact calculation of $\underline{P}_n = \bar{P}_n$.

3. Proof of the recursion. Let r and t be any numbers such that $r \leqq t$. Let $B_{l,i}(r, t)$ be the event $\{\alpha_{l+j} < X_j^{i-l} \leqq \beta_{l+j},\ r < X_j^{i-l} \leqq t;\ j = 1, 2, \cdots, i - l\}$ when $i - l > 0$ and let it be the certain event when $i - l = 0$. Let $R_{l,i}(r, t) = P[B_{l,i}(r, t)]$. In particular $R_{0,n}(-\infty, +\infty) = P_n$.

Let s be any number such that $r \leqq s \leqq t$. The event $B_{l,i}(r, t)$ can be realized in a number of mutually exclusive ways: exactly $(k - l)$ components $(k - l = 0, 1, \cdots, i - l)$ out of the sample considered are located in $(r, s]$, the remaining $(i - k)$ components then being located in $(s, t]$. It follows that

$$R_{l,i}(r, t) = \sum_{k=l}^{i} C_{i-l}^{k-l} \cdot R_{l,k}(r, s) \cdot R_{k,i}(s, t)\ . \tag{10}$$

If $\alpha_i \leqq r \leqq t \leqq \beta_{l+1}$, one has by (5) $\alpha_{l+j} \leqq r \leqq t \leqq \beta_{l+j}$ $(j = 1, 2, \cdots, i - l)$. Then $B_{l,i}(r, t) = \{r < X_j^{i-l} \leqq t;\ j = 1, 2, \cdots, i - l\}$. Hence

$$\text{if}\quad \alpha_i \leqq r \leqq t \leqq \beta_{l+1}\quad \text{then}\quad R_{l,i}(r, t) = [F(t) - F(r)]^{i-l}\ . \tag{11}$$

Furthermore it is clear that

$$\text{if}\quad \alpha_i \geqq t\quad \text{then}\quad R_{l,i}(r, t) = 0\ , \tag{12}$$

$$\text{if}\quad \beta_{l+1} \leqq r\quad \text{then}\quad R_{l,i}(r, t) = 0\ . \tag{13}$$

Let now $1 \leqq m \leqq 2n + 1$ and $r \leqq \gamma_{m-1} \leqq \gamma_m \leqq t$. If $i \leqq g(m - 1)$ and $h(m) \leqq l + 1$, one has by (5) and (7) $\alpha_i \leqq \alpha_{g(m-1)} \leqq \gamma_{m-1} \leqq \gamma_m \leqq \beta_{h(m)} \leqq \beta_{l+1}$. Hence by (11), since $F(\gamma_m) - F(\gamma_{m-1}) = p_m$,

$$\text{if}\quad h(m) - 1 \leqq l \leqq i \leqq g(m - 1)\quad \text{then}\quad R_{l,i}(\gamma_{m-1}, \gamma_m) = {p_m}^{i-l}\ . \tag{14}$$

If $g(m - 1) + 1 \leqq i$, one has by (7) and (5) $\gamma_m \leqq \alpha_{g(m-1)+1} \leqq \alpha_i$. Hence by (12)

$$\text{if}\quad g(m - 1) < i\quad \text{then}\quad R_{l,i}(r, \gamma_m) = 0\ . \tag{15}$$

If $l + 1 \leqq h(m) - 1$, one has by (5) and (7) $\beta_{l+1} \leqq \beta_{h(m)-1} \leqq \gamma_{m-1}$. Hence by (13)

$$\text{if}\quad l < h(m) - 1\quad \text{then}\quad R_{l,i}(\gamma_{m-1}, t) = 0\ . \tag{16}$$

Setting $s = \gamma_{m-1}$ and $t = \gamma_m$, one has by (10), (14), (16),

$$\text{if}\quad \max[l, h(m) - 1] \leqq i \leqq g(m - 1)$$
$$\text{then}\quad R_{l,i}(r, \gamma_m) = \sum_k C_{i-l}^{k-l} \cdot R_{l,k}(r, \gamma_{m-1}) \cdot {p_m}^{i-k}\ , \tag{17}$$

where the summation is over $\{\max[l, h(m) - 1] \leqq k \leqq i\}$.

From relations (17) and (15), quite a set of recursions can be devised for the calculation of $R_{0,n}(-\infty, +\infty)$, all implying the same amount of computation. We choose one which formally is as simple as possible. Let us set $r = -\infty$ and $l = 0$. Clearly

$$R_{0,0}(-\infty, \gamma_0) = 1\ ,\qquad R_{0,n}(-\infty, \gamma_{2n}) = P_n\ . \tag{18}$$

Let us write $Q_i(m)$ for $R_{0,i}(-\infty, \gamma_m)$. Relations (18), (17), (15) then prove formula (8).

4. The particular weight function (4). In this section, we suppose that $F(z)$ is continuous. We consider the distribution of the two-sided Kolmogorov-Smirnov type statistic, i.e., $P(L \leqq \lambda, M \leqq \lambda)$ where L and M are defined by (2), (3). We choose the weight functions (4). It is well known that the corresponding distribution function does not depend on $F(z)$. One verifies that it has the form P_n provided that $F(\alpha_j)$, $F(\beta_j)$, $(j = 1, 2, \cdots, n)$ satisfy

$$j/n = a_j + \lambda n^{-\frac{1}{2}}[a_j(1 - a_j)]^{\frac{1}{2}} \tag{19}$$
$$(j - 1)/n = b_j - \lambda n^{-\frac{1}{2}}[b_j(1 - b_j)]^{\frac{1}{2}},$$

where a_j and b_j stand for $F(\alpha_j)$ and $F(\beta_j)$. We thus write

$$1 - P_n = 1 - P\{\sup_z n^{\frac{1}{2}}\{F(z)[1 - F(z)]\}^{-\frac{1}{2}}|F^n(z) - F(z)| \leqq \lambda\}. \tag{20}$$

The curves of Figure 1 represent $(1 - P_n)$ as a function of λ for $n = 1, 2, 10, 100$. They were calculated on a computer by means of recursion (8) applied to the values of $F(\alpha_j)$, $F(\beta_j)$ drawn from (19), They allow to carry out, with moderate accuracy, the corresponding test for any level of significance. The same curves were obtained approximately by Vandewiele and de Witte by means of a Monte-Carlo method.

Vandewiele and de Witte also gave the first five terms of the development of $(1 - P_n)$ in powers of λ^{-2} in the case $a_n \leqq b_1$, i.e., in the case $\lambda^2 \geqq n$. (Corrigenda in [6]: the last term is erroneous and should be corrected as follows):

$$\begin{aligned} 1 - P_n = {} & 2\lambda^{-2} + (3 - 5n^{-1})\lambda^{-4} + (16 - 96n^{-1} + 82n^{-2})\lambda^{-6} \\ & + (124 - 1760n^{-1} + 4779n^{-2} - 3145n^{-3})\lambda^{-8} \\ & + (1224 - 33696n^{-1} + 198636n^{-2} - 380616n^{-3} \\ & \qquad + 214454n^{-4})\lambda^{-10} \\ & + \cdots . \end{aligned} \tag{21}$$

Neither the general term nor a general truncation error bound are known. By reversion of this series one obtains

$$\begin{aligned} \lambda^{-2} = {} & 2^{-1}(1 - P_n) - 2^{-3}(3 - 5n^{-1})(1 - P_n)^2 \\ & - 2^{-5}(14 - 132n^{-1} + 114n^{-2})(1 - P_n)^3 \\ & - 2^{-7}(151 - 4035n^{-1} + 12981n^{-2} - 9105n^{-3}(1 - P_n)^4 \\ & - 2^{-9}(2706 - 139992n^{-1} + 1041576n^{-2} - 223900n^{-3} \\ & \qquad + 1334894n^{-4})(1 - P_n)^5 \\ & - \cdots . \end{aligned}$$

These truncated series provide excellent approximations of (20) when λ is sufficiently large, even if $\lambda^2 \geqq n$ does not hold. A numerical comparison of

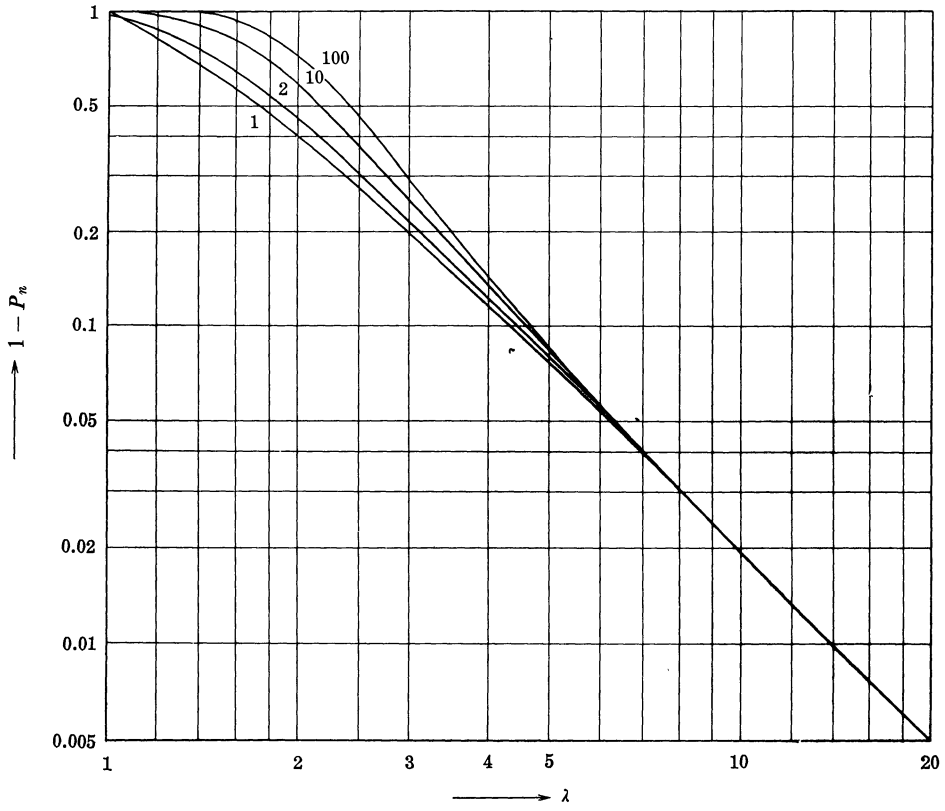

FIG. 1. The variable λ defined by equation (20) as a function of $(1 - P_n)$ for $n = 1, 2, 10$ and 100.

(21) with exact calculations has given the following results for $n \leqq 100$. Taking the five terms, the relative error on $(1 - P_n)$ is smaller than 10^{-2} when $\lambda \geqq 4$, than 10^{-3} when $\lambda \geqq 5$ and than 10^{-4} when $\lambda \geqq 6$. Taking the first two terms only this error is smaller than 10^{-2} when $\lambda \geqq 6$. For smaller n the convergence is somewhat more rapid.

The curves, power series and error bounds corresponding to the one-sided case with the particular weight function (4) are given in [4].

Acknowledgment. The author wishes to thank Professor G. Vandewiele for helpful discussions and for his criticisms on the manuscript.

REFERENCES

[1] ANDERSON, T. W. and DARLING, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann. Math. Statist.* **23** 193-212.

[2] BOROKOV, A. A. and SYCHEVA, N. M. (1968). On asymptotically optimal non-parametric criteria. *Theor. Probability Appl.* **13** 359-393.

[3] DANIELS, H. E. (1945). The statistical theory of the strength of bundles of threads, I, *Proc. Roy. Soc. Ser. A* **183** 405–435.

[4] NOÉ, M. and VANDEWIELE, G. (1968). The calculation of distributions of Kolmogorov-Smirnov type statistics including a table of significance points for a particular case. *Ann. Math. Statist.* **39** 233–241.

[5] STECK, G. P. (1971). Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *Ann. Math. Statist.* **42** 1–11.

[6] VANDEWIELE, G. and de WITTE, P. (1966). A test of goodness of fit. *Statistica Neerlandica*, **20** 87–105.

[7] WALD, A. and WOLFOWITZ, J. (1939). Confidence limits for continuous distribution functions. *Ann. Math. Statist.* **10** 105–118.