The Annals of Mathematical Statistics
1968, Vol. 39, No. 1, 233–241

THE CALCULATION OF DISTRIBUTIONS OF KOLMOGOROV-SMIRNOV TYPE STATISTICS INCLUDING A TABLE OF SIGNIFICANCE POINTS FOR A PARTICULAR CASE

BY MARC NOÉ AND GEORGES VANDEWIELE

M.B.L.E. Research Laboratory, Brussels and Free University of Brussels

1. Introduction. Let $X_1 \leqq X_2 \leqq \cdots \leqq X_n$ be the order statistics of a sample of size n from a continuous distribution function $F(x)$ and let $F_n(x)$ be the corresponding empirical distribution function. Let $G(x)$ and $H(x)$ be two functions of x. We will consider the probabilities

$$(1) \qquad \bar{P}_n = P\{\inf_x [G(x) - F_n(x)] \geqq 0\},$$

$$\underline{P}_n = P\{\inf_x [F_n(x) - H(x)] \geqq 0\};$$

$$(2) \quad P_n = P\{\inf_x [G(x) - F_n(x)] \geqq 0, \inf_x [F_n(x) - H(x)] \geqq 0\}.$$

These probabilities are related to the statistics of the Kolmogorov-Smirnov type in the following way: The corresponding one-sided statistic has the distribution function

$$(3) \qquad P\{\sup_x m^{\frac{1}{2}}[F_n(x) - F(x)]\psi[F(x)] \leqq \lambda\}$$

which is a probability of the form (1). The two-sided statistic has the distribution

$$(4) \qquad P\{\sup_x n^{\frac{1}{2}}|F_n(x) - F(x)|\psi[F(x)] \leqq \lambda\}$$

which is a special case of (2). In these expressions $\psi(x)$ is a (non-negative) weight function. A discussion of these statistics can be found e.g. in Kendall and Stuart [4].

Wald and Wolfowitz [8] [9] have given recursion formulas for computing $\bar{P}_n$, $\underline{P}_n$ and P_n. Daniels [2] was led to a probability of the same form as $\bar{P}_n$ or $\underline{P}_n$ in connection with a study of the strength of bundles of threads. He found recursions slightly more general than in [8] by a very similar method. In Section 2 we give a simple derivation of still more general formulas for $\bar{P}_n$ and $\underline{P}_n$, which contain a wider choice of recursions, so that the numerical computability can be taken into account. Note that two non-recursive formulas for $\bar{P}_n$ or $\underline{P}_n$ are given by Daniels [2], but unfortunately they are not easily tractable.

In Section 3 a formula for P_n is derived which is simpler than the corresponding formula of [8] but is valid only under certain conditions. Furthermore in Sections 2 and 3 bounds of $\bar{P}_n$, $\underline{P}_n$ and P_n are given with a view to approximate calculations.

Received 1 June 1967; revised 11, July 1967.

Section 4 is devoted to the particular case where the weight function in (3) and (4) is

$$\psi(x) = [x(1 - x)]^{-\frac{1}{2}}. \tag{5}$$

We give a table of percentage points and a truncated power series which approximates $\bar{P}_n$ in this case. The weight function (5) seems to have been proposed first by L. J. Savage. As Anderson and Darling [1] pointed out, it assigns in a certain sense the same weight to each point of $F(x)$. Some arguments in favour of it and a different and less complete treatment have been given by Vandewiele and de Witte [6]. Anderson and Darling [1] give information on the asymptotic behaviour of the distribution (4). However the question whether the limit distribution is non-degenerate in the case of the weight function (5) is open.

2. The one-sided statistic.

2.1. *The fundamental formula.* The event

$$\inf_x [G(x) - F_n(x)] \geqq 0$$

is equivalent to the event that every order statistic X_j is not smaller than some well defined number α_j, i.e. $X_j \geqq \alpha_j$, $j = 1, \cdots, n$, with $\alpha_1 \leqq \alpha_2 \leqq \cdots \leqq \alpha_n$, or equivalently, since $F(x)$ is monotonous, $F(X_j) \geqq a_j$, $j = 1, \cdots, n$, with $a_j = F(\alpha_j)$. Thus $0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_n \leqq 1$.

Let $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_i$ be the order statistics of a sample of size i from $F(x)$ with $i \leqq n$. Let

$$Q_i(t) = P\{\textstyle\bigcap_{j=1}^{i} [F(Y_j) \geqq a_j,\ F(Y_j) \leqq t]\}, \qquad i = 1, 2, \cdots, n, \tag{6}$$

$$Q_0(t) = 1.$$

In particular $Q_n(1) = \bar{P}_n$. Clearly

$$Q_i(t) = 0 \quad \text{if} \quad t \leqq a_i. \tag{7}$$

If now $0 \leqq a_i \leqq t' \leqq t \leqq 1$, one has

$$P\{[\textstyle\bigcap_{j=1}^{k} a_j \leqq F(Y_j) \leqq t'] \cap [\bigcap_{j=k+1}^{i} t' \leqq F(Y_j) \leqq t]\} = C_i^k Q_k(t')(t - t')^{i-k}$$

and consequently

$$Q_i(t) = \textstyle\sum_{k=0}^{i} C_i^k Q_k(t')(t - t')^{i-k}. \tag{8}$$

Thus $Q_i(t)$ is a polynomial in t of degree i when $a_i \leqq t \leqq 1$. Let $q_i(t)$ be a polynomial in t of degree i defined for all values of t and such that $q_i(t) = Q_i(t)$ everywhere in $[a_i, 1]$. It is clear that

$$q_i(t) = \textstyle\sum_{k=0}^{i} C_i^k q_k(t')(t - t')^{i-k} \tag{9}$$

for all t and t'. By setting $t' = 0$ one of the formulas of Wald and Wolfowitz [8] is obtained.

2.2. *Computation of* $\bar{P}_n$. The formula (9) can be used in several ways for computing $\bar{P}_n$. Setting in (9) successively $t = a_i$ and $t = 1, i = n$, one obtains the following recursion formulas:

$$q_0(t') = 1,$$

$$\sum_{k=0}^{i} C_i^k q_k(t')(a_i - t')^{i-k} = 0, \qquad i = 1, 2, \cdots, n; \tag{10}$$

$$\bar{P}_n = Q_n(1) = \sum_{i=0}^{n} C_n^i q_i(t')(1 - t')^{n-i}. \tag{11}$$

Formula (10) is given by Daniels [2]. By setting $t' = 0$ in (10) and (11) the recursion formulas proposed by Wald and Wolfowitz [8] are obtained. The choice $t' = 1$ leads to the more direct recursion considered by Daniels [2]:

$$Q_0(1) = 1,$$

$$\sum_{k=0}^{i} C_i^k Q_k(1)(a_i - 1)^{i-k} = 0, \qquad i = 1, 2, \cdots, n.$$

These recursions are however not suited to numerical computation as they involve small differences of large numbers. This feature can be avoided at the cost of a greater amount of calculation by using the following recursion: Setting in (9) successively $t = a_j$, $t' = a_{j-1}$ and $t = 1$, $t' = a_n$, $i = n$, one obtains

$$Q_0(a_j) = 1, \qquad j = 1, 2, \cdots, n;$$

$$Q_i(a_j) = \sum_{k=0}^{i-1} C_i^k Q_k(a_{j-1})(a_j - a_{j-1})^{i-k} \tag{12}$$

$$i = 1, 2, \cdots, j - 1; j = 1, 2, \cdots, n;$$

$$\bar{P}_n = Q_n(1) = \sum_{i=0}^{n-1} C_n^i Q_i(a_n)(1 - a_n)^{n-i}. \tag{13}$$

These formulas involve sums of non-negative terms only.

2.3. *Bounds of* $\bar{P}_n$. Consider two sets of numbers $\{a_j', j = 1, \cdots, n\}$ and $\{a_j'', j = 1, \cdots, n\}$ and let $Q_i'(t)$ and $Q_i''(t)$ be defined by (6) where a_j has been replaced by a_j' resp. a_j''. Let $\bar{P}_n' = Q_n'(1)$ and $\bar{P}_n'' = Q_n''(1)$. When $a_j' \leqq a_j \leqq a_j''$, $j = 1, \cdots, n$, it is clear that

$$\bar{P}_n' \geqq \bar{P}_n \geqq \bar{P}_n''. \tag{14}$$

The following particular choice of a_j' and a_j'', where l is any integer in $[1, n]$, will provide useful upper and lower bounds of $\bar{P}_n$:

$$a_j' = a_j'' = a_j, \qquad j = 1, \cdots, l - 1,$$

$$a_j' = a_l, \qquad j = l, \cdots, n,$$

$$a_j'' = a_n, \qquad j = l, \cdots, n.$$

It is clear that $Q_i'(t) = Q_i''(T) = Q_i(t)$ for $i < l$. According to (7),

$$Q_i'(a_l) = Q_i'(a_i') = 0 \quad \text{for } i \geqq l,$$

$$Q_i''(a_n) = Q_i''(a_i'') = 0 \quad \text{for } i \geqq l.$$

According to (8), with $i = n$, $t' = a_l$, $t = 1$, one thus has

$$(15) \quad \bar{P}_n' = Q_n'(1) = \sum_{i=0}^{n} C_n^i Q_i'(a_l)(1 - a_l)^{n-i} = \sum_{i=0}^{l-1} C_n^i Q_i(a_l)(1 - a_l)^{n-i}.$$

On the other hand, setting in (8) $i = n$, $t' = a_n$, $t = 1$, one obtains

$$(16) \quad \bar{P}_n'' = Q_n''(1) = \sum_{i=0}^{n} C_n^i Q_i''(a_n)(1 - a_n)^{n-i} = \sum_{i=0}^{l-1} C_n^i Q_i(a_n)(1 - a_n)^{n-i}.$$

The terms of (15) can be computed using recursion (12) for $j = 1, \cdots, l$. The terms of (16) then follow from

$$Q_i(a_n) = \sum_{k=0}^{i} C_i^k Q_k(a_l)(a_n - a_l)^{i-k}.$$

Again sums of non-negative numbers only are involved.

Note that (16) simply corresponds to a truncation of (13). According to (14) the expressions (15) and (16) are upper resp. lower bounds of $\bar{P}_n$ and can be used for approximating it. The larger l is chosen, the nearer to $\bar{P}_n$ the bounds will be. Unfortunately actual computations need large values of l because (16), unlike (15), converges slowly. For the computations described in section 4.3, the use of an electronic computer was necessary.

3. The two-sided statistic.

3.1. *A particular case.* The event

$$\inf_x [F_n(x) - H(x)] \geqq 0$$

is equivalent to the event that every order statistic X_j is not larger than some well defined number β_j, i.e.

$$X_j \leqq \beta_j, \qquad j = 1, \cdots, n,$$

with $\beta_1 \leqq \beta_2 \leqq \cdots \leqq \beta_n$, or equivalently, since $F(x)$ is monotonous,

$$F(X_j) \leqq 1 - b_{n-j+1}, \qquad j = 1, \cdots, n,$$

with $1 - b_{n-j+1} = F(\beta_j)$. Thus $0 \leqq b_1 \leqq b_2 \leqq \cdots \leqq b_n \leqq 1$.

We then have

$$P_n = P[\textstyle\bigcap_{j=1}^{n} a_j \leqq F(X_j) \leqq 1 - b_{n-j+1}].$$

We will derive an expression of P_n in the particular case where $\alpha_n \leqq \beta_1$ or equivalently

$$(17) \qquad a_n \leqq 1 - b_n.$$

Let t and t' be such that $a_n \leqq t \leqq 1 - t' \leqq 1 - b_n$. The probability that there are i observations in $[-\infty, F^{-1}(t)]$, j observations in $[F^{-1}(1 - t'), \infty]$ and $n - i - j$ observations in $[F^{-1}(t), F^{-1}(1 - t')]$ and that at the same time $H(x) \leqq F_n(x) \leqq G(x)$ for all x is

$$[n!/i!j!(n - i - j)!]Q_i(t) \cdot Q_j'(t') \cdot (1 - t - t')^{n-i-j}.$$

In this expression $Q_i'(t)$ is defined by (6) where a_j has been replaced by b_j. Hence

$$P_n = \sum_{i+j\leqq n} [n!/i!j!(n-i-j)!] Q_i(t)\cdot Q_j'(t')\cdot(1-t-t')^{n-i-j}.$$

Let $q_i'(t)$ be a polynomial in t of degree i defined for all values of t and such that $q_i'(t) = Q_i'(t)$ everywhere in $[b_i, 1]$. It is clear that

$$P_n = \sum_{i+j\leqq n} [n!/i!j!(n-i-j)!] q_i(t)\cdot q_j'(t')(1-t-t')^{n-i-j}$$

for all t and t'. This can be written

$$(18) \qquad P_n = \sum_{k=0}^{n} C_n^k (1-t-t')^{n-k} \sum_{i=0}^{k} C_k^i q_i(t)\cdot q_{k-i}'(t').$$

By suitable choices of t and t' convenient methods of computation of P_n can be devised, as was the case in Section 2.2.

3.2. *Bounds of P_n*. Define the events

$$A = \{\inf_x [G(x) - F_n(x)] \geqq 0\}, \qquad B = \{\inf_x [F_n(x) - H(x)] \geqq 0\};$$

and let $\bar{A}$ and $\bar{B}$ be the complementary events of A and B. Then

$$\bar{P}_n = P(A \cap B) + P(A \cap \bar{B}),$$

$$\underline{P}_n = P(A \cap B) + P(\bar{A} \cap B),$$

$$P_n = P(A \cap B).$$

Hence

$$(19) \qquad P(\bar{A} \cap \bar{B}) + \bar{P}_n + \underline{P}_n = P_n + 1.$$

Let now $G'(x)$ and $H'(x)$ be functions of x such that $G'(x) \geqq G(x)$ and $H'(x) \leqq H(x)$. Define A', B', $\bar{A}'$, $\bar{B}'$, $\bar{P}_n'$, $\underline{P}_n'$, P_n' for $G'(x)$ and $H'(x)$ in a similar manner as A, B, $\bar{A}$, $\bar{B}$, $\bar{P}_n$, $\underline{P}_n$, P_n were defined for $G(x)$ and $H(x)$. Since clearly

$$P(\bar{A}' \cap \bar{B}') \leqq P(\bar{A} \cap \bar{B}),$$

relation (19) provides a lower bound for P_n:

$$P_n \geqq P_n' + (\bar{P}_n - \bar{P}_n') + (\underline{P}_n - \underline{P}_n').$$

An upper bound of P_n is given by the following inequality which has been proved by the authors [7]:

$$P_n \leqq P_n'\cdot\bar{P}_n\cdot\underline{P}_n\cdot[\bar{P}_n'\cdot\underline{P}_n']^{-1}.$$

These bounds make possible an approximate calculation of P_n if one knows functions $G'(x)$ and $H'(x)$ such that P_n' is more easily calculated than P_n. In particular, choosing $G'(x) \geqq 1$ and $H'(x) \leqq 0$ for all x, one has $P_n' = \bar{P}_n' = \underline{P}_n' = 1$, and the bounds are written as functions of the corresponding one-sided probabilities exclusively:

$$P_n \geqq \bar{P}_n + \underline{P}_n - 1, \tag{20}$$

$$P_n \leqq \bar{P}_n \cdot \underline{P}_n . \tag{21}$$

On the other hand any alternative choice of $G'(x)$ and $H'(x)$ provides closer bounds than (20) and (21). Wald and Wolfowitz [8] proved (20) and conjectured (21). Professor W. Hoeffding pointed out to the authors that (21) also can be deduced from a theorem of Lehmann [5]. This is explained in [7].

When $G(x)$ and $H(x)$ are such that (17) is not satisfied, formula (18) can nevertheless be used for finding bounds of P_n. Let u be any number and let

$$\begin{aligned} G'(x) &= G(x) && \text{for} \quad x \leqq u \\ &= 1 && \text{for} \quad x > u; \\ H'(x) &= 0 && \text{for} \quad x < u \\ &= H(x) && \text{for} \quad x \geqq u. \end{aligned}$$

Then P_n' can be calculated by (18), since clearly $a_n' \leqq 1 - b_n'$.

4. The particular weight function (5).

4.1. *The quantities a_i and b_i.* When the particular weight function (5) is chosen, the function $G(x)$ is defined by

$$G(x) = F(x) + \lambda n^{-\frac{1}{2}}\{F(x)[1 - F(x)]\}^{\frac{1}{2}}$$

and consequently

$$i/n = a_i + \lambda n^{-\frac{1}{2}}[a_i(1 - a_i)]^{\frac{1}{2}}.$$

Unfortunately this expression of a_i, when put in formula (10), does not lead to a simple expression for $\bar{P}_n$. However a development of a_i in powers of λ^{-2} will lead in Section 4.2 to a useful series for $\bar{P}_n$.

Setting

$$\varphi(a_i) = (i - na_i)^2 n^{-1}(1 - a_i)^{-1}$$

one obtains $a_i = \lambda^{-2}\varphi(a_i)$ and, since $a_i = 0$ when $\lambda^{-2} = 0$, the development of a_i is given by Lagrange's formula (see e.g. Goursat [3]) in the form

$$a_i = \sum_{k=1}^{\infty} k!^{-1}\lambda^{-2k}[d^{k-1}[\varphi(a_i)]^k/da_i^{k-1}]_{a_1=0} .$$

Using complete induction for finding the derivative one obtains finally

$$a_i = \sum_{k=1}^{\infty} (-1)^{k+1}k^{-1}i^{2k}n^{-k}\lambda^{-2k} \cdot \sum_{l=0}^{k-1} C_{2k}^l n^l i^{-l} \sum_{m=0}^{k-l-1} (-1)^m C_{2k-l}^m (k - m - l). \tag{22}$$

On the other hand one verifies that $b_i = a_i$ for $i = 1, 2, \cdots, n$, and $\bar{P}_n = \underline{P}_n$. Hence by (20) and (21)

$$1 - \bar{P}_n^2 \leqq 1 - P_n \leqq 2 - 2\bar{P}_n . \tag{23}$$

These inequalities will be used in Section 4.3.

4.2. *Development of $\bar{P}_n$ and its inversion.* When in formula (10) a_i is replaced by its expression (22), one obtains by (11), setting $t' = 0$

$$
\begin{aligned}
\bar{P}_n = 1 &- \lambda^{-2} - (2 - 3n^{-1})\lambda^{-4} - (10 - 57n^{-1} + 48n^{-2})\lambda^{-6} \\
&- (74 - 1021n^{-1} + 2743n^{-2} - 1797n^{-3})\lambda^{-8} \\
&- (706 - 19123n^{-1} + 111905n^{-2} - 213619n^{-3} + 120132n^{-4})\lambda^{-10} \\
&- \cdots .
\end{aligned} \tag{24}
$$

By inversion of this series one obtains

$$
\begin{aligned}
\lambda^{-2} = (1 - \bar{P}_n) &- (2 - 3n^{-1})(1 - \bar{P}_n)^2 - (2 - 33n^{-1} + 30n^{-2})(1 - \bar{P}_n)^3 \\
&- (14 - 481n^{-1} + 1678n^{-2} - 1212n^{-3})(1 - \bar{P}_n)^4 \\
&- (134 - 8083n^{-1} + 65246n^{-2} - 146110n^{-3} + 88812n^{-4})(1 - \bar{P}_n)^5 \\
&- \cdots .
\end{aligned} \tag{25}
$$

Unfortunately we did not succeed in finding the general term and a truncation error for these series. Their comparison with exact calculations carried out on a digital computer is discussed in the next section.

4.3. *Numerical results.* Let

$$
1 - \bar{P}_n = P\{\sup_x n^{\frac{1}{2}}\{F(x)[1 - F(x)]\}^{-\frac{1}{2}}[F_n(x) - F(x)] \leqq \lambda\}. \tag{26}
$$

Table 1 gives the values of λ which are necessary for carrying out the one-sided test using standard levels of significance. (As will be seen the unusual levels 0.025 and 0.005 are necessary for the two-sided case). The columns with headings 0.1 and 0.05 were computed by successive interpolations with the help of formulas (15) and (16) where l was chosen so as to ensure the precision mentioned in the table. The other columns were computed by series (25) taking all five terms fully written and after checking its precision in the proper region.

The curves of Figure 1 are derived from additional computations by formulas (15) and (16). All computations were carried out on a digital computer.

The precision of series (24) has been examined for $n \leqq 100$. Taking the five first terms the relative error on $1 - \bar{P}_n$ 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 two first terms only this error is smaller than 10^{-2} when $\lambda \geqq 6$. For smaller n the convergence is somewhat more rapid.

Using formula (23) the two-sided test can be carried out with a good approximation by means of Table 1. As an example suppose $n = 20$ and take $\lambda = 6.477$; then $\bar{P}_n = 0.975$ and $0.049{,}375 \leqq 1 - P_n \leqq 0.05$. The corresponding significance level thus certainly is smaller than 5% with a possible deviation of 0.000,625 or a possible relative deviation of 1.25%.

TABLE 1

The variable λ as a function of n and $(1 - \bar{P}_n)$ according to equation (26). Values of λ either read directly or interpolated linearly for additional values of n are guaranteed to entail a relative error on $(1 - \bar{P}_n)$ smaller than 0.1 %.

n	$1 - \bar{P}_n$				
	0.1	0.05	0.025	0.01	0.005
1	3.000	4.359	6.245	9.950	14.107
2	3.169	4.498	6.352	10.022	14.159
3	3.247	4.556	6.394	10.047	14.177
4	3.294	4.589	6.416	10.061	14.186
5	3.327	4.611	6.430	10.069	14.191
6	3.351	4.627	6.440	10.074	14.195
7	3.370	4.639	6.447	10.078	14.198
8	3.386	4.648	6.452	10.081	14.200
9	3.399	4.655	6.457	10.084	14.201
10	3.409	4.662	6.460	10.085	14.202
11	3.419	4.667	6.463	10.087	14.203
12	3.427	4.671	6.465	10.088	14.204
13	3.434	4.675	6.468	10.089	14.205
14	3.441	4.678	6.469	10.090	14.206
15	3.446,9	4.681	6.471	10.091	14.206
20	3.469,6	4.692	6.477	10.094	14.208
25	3.485,2	4.699	6.480	10.096	14.209
30	3.497,0	4.704	6.482	10.097	14.210
40	3.513,7	4.711	6.485	10.098	14.211
50	3.525,5	4.715	6.487	10.099	14.211
70	3.541,3	4.720	6.489	10.100	14.212
100	3.556,0	4.724	6.490	10.101	14.213

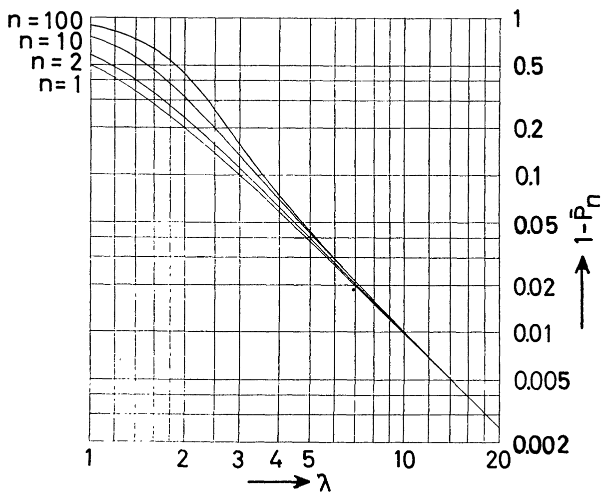

FIG. 1. The variable λ defined by equation (26) as a function of $(1 - \bar{P}_n)$ for n = 1, 2, 10 and 100.

Acknowledgment. We are indebted to M. F. Heymans for programming the numerical computations. We wish to thank Professor W. Hoeffding for his criticisms.

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

[3] GOURSAT, E. (1959). *A Course in Mathematical Analysis* **1**. Dover, New York.

[4] KENDALL, M. G. and STUART, A. (1961). *The Advanced Theory of Statistics* **2**. Griffin, London.

[5] LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.

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

[7] VANDEWIELE, G. and NOÉ, M. (1967). An inequality concerning tests of fit of the Kolmogorov-Smirnov type. *Ann. Math. Statist.* **38** 1240–1244.

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

[9] WALD, A. and WOLFOWITZ, J. (1941). Note on confidence limits for continuous distribution functions. *Ann. Math. Statist.* **12** 118–119.