ON THE APPROXIMATE FORMULA TO THE DISTRIBUTION OF THE TWO SAMPLE SMIRNOV TEST

BY

RYUZO KANNO

1. Introduction. Let $x_1^{(1)}, \ldots, x_n^{(1)}$ and $x_1^{(2)}, \ldots, x_m^{(2)}$ be the two random samples from populations having continuous cumulative distribution functions $F_1(x)$ and $F_2(x)$ respectively, and let $F_{1n}(x)$, $F_{2m}(x)$ denote the corresponding empirical distributions. Further without loss of generality let us suppose that $n \leqq m$. The Smirnov statistic for testing the hypothesis $F_1(x) = F_2(x)$ is

$$D_{nm} = \sup_x |F_{1n}(x) - F_{2m}(x)|.$$

The exact distribution of D_{nm} for equal-sized samples, *i.e.* $n = m$, has been found explicitly by Gnedenko-Korolyuk [5] and independently by Drion [4], and the table for $1 \leqq n = m \leqq 40$ has been given by Massey [7]. Korolyuk [6] and Blackman [1], [2] studied for the case where one sample size is an integer multiple of the other and Depaix [3] for the general case. However in the case of unequal-sized samples the expressions for the distribution are extremely complicated and poorly suited for computation. In practice, the small sample distribution of D_{nm} may be computed numerically with the aid of a high speed digital computer based on the recursion relation given by Massey [8]. He has also given a small table for $1 \leqq n \leqq m \leqq 10$ and certain other selected values of n, $m \leqq 20$. The problem for the distribution of D_{nm} has been studied by many other authors. For the further investigations of this problem, for example, refer to the references in Steck [9].

The purpose of this paper is to give an approximate formula of the distribution of D_{nm} for two samples with slightly different sizes. Though we make an experimental design for equal-sized samples, yet obtaining the samples with missing values we must deal with two samples having slightly different sizes in practice. In this paper, at first we shall find the exact distribution of D_{nm} in some restricted range and at the next step the approximate formula which is used in general will be constructed with a linear combination of two equal-sized sample distributions.

2. Exact Distribution of D_{nm} in some Restricted Range. To find the distribution of D_{nm} we make use of the graphical representation as mentioned in Gnedenko-Korolyuk

*Received June 20, 1975

[5] (or in Wilks [10], P.455–P.459). Let the order statistics of two samples combined be $z_{(1)} < z_{(2)} < \ldots < z_{(n+m)}$, and let ζ_t be a random variable defined as follows:

$$\zeta_t = \begin{cases} 1/n \text{ if } z_{(t)} \text{ belongs to the 1st sample} \\ -1/m \text{ if } z_{(t)} \text{ belongs to the 2nd sample} \end{cases} \quad (t = 1, 2, \ldots, n+m)$$

We put $s_t = \zeta_1 + \ldots + \zeta_t$ $(s_0 = 0)$ and consider the graph of the points (t, s_t), $t = 0, 1, \ldots, n+m$, in the (t, s)-plane; that is, connecting the sequence of points by line segments, we have a path which begins at $O(0, 0)$ and ends at $P(n+m, 0)$. Then all possible sequences of $x^{(1)}$'s and $x^{(2)}$'s among the order statistics $z_{(1)} < z_{(2)} < \ldots < z_{(n+m)}$ will be represented by all possible paths joining the diagonal corners of a parallelogram's lattice of sides n and m. The number of all possible paths from O to P is $\binom{n+m}{n}$ and under the null hypothesis $F_1(x) = F_2(x)$ all of these paths are equally probable. The problem of finding the value of $P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right)$ is equivalent to determining the number of paths which do not lie entirely between the lines $s = \pm\left(1 - \frac{a}{n} - \frac{b}{m}\right)$ and dividing this number by $\binom{n+m}{n}$. We here denote these lines by L^+_{ab} and L^-_{ab}, respectively.

Now we further consider the following two lines:

l^+_i: line through two points $\left(n-i, 1-\frac{i}{n}\right)$ and $\left(n+i, 1-\frac{i}{m}\right)$.

l^-_i: line through two points $\left(m-i, -1+\frac{i}{m}\right)$ and $\left(m+i, -1+\frac{i}{n}\right)$

and let P^+_{ij} denote the $(j+1)$th lattice-point from the left-hand side on the line l^+_{i+j}, similarly, let P^-_{ij} be defined to l^-_{i+j}. There are $i+1$ lattice-points on the line l^+_i and these are represented by $\{P^+_{i0}, P^+_{i-1,1}, \ldots, P^+_{0i}\}$. Furthermore we introduce the following notations for the set of lattice-points.

$A(l)$: set of lattice-points lying above or just on a line l.

$B(l)$: set of lattice-points lying below or just on a line l

S^+_{ij}: set of j lattice-points counted from the right-hand side on l^+_i, *i.e.* $\{P^+_{j-1,i-j+1}, \ldots, P^+_{1,i-1}, P^+_{0i}\}$

S^-_{ij}: set of j lattice-points counted from the left-hand side on l^-_i, *i.e.* $\{P^-_{i0}, P^-_{i-1,1}, \ldots, P^-_{i-j+1,j-1}\}$

Let M denote the largest integer of k satisfying $1 \leqq \frac{m}{n} < \frac{k+1}{k}$, then we see that

[i] if $a = 0, 1, 2, \ldots, M$,

$$\text{(2.1)} \qquad \begin{cases} A(L^+_{a0}) = A(l^+_a) = \bigcup_{\xi=0}^{a} S^+_{\xi,\xi+1} \\ B(L^-_{a0}) = B(l^-_a) = \bigcup_{\xi=0}^{a} S^-_{\xi,\xi+1} \end{cases}$$

[ii] if $a = 0, 1, 2, \ldots, M$ and b such that $1 \leqq a+b \leqq M+1$,

$$\text{(2.2)} \qquad \begin{cases} A(L^+_{ab}) = A(L^+_{a+b-1,0}) \cup S^+_{a+b,a+1} \\ B(L^-_{ab}) = B(L^-_{a+b-1,0}) \cup S^-_{a+b,a+1} \end{cases}$$

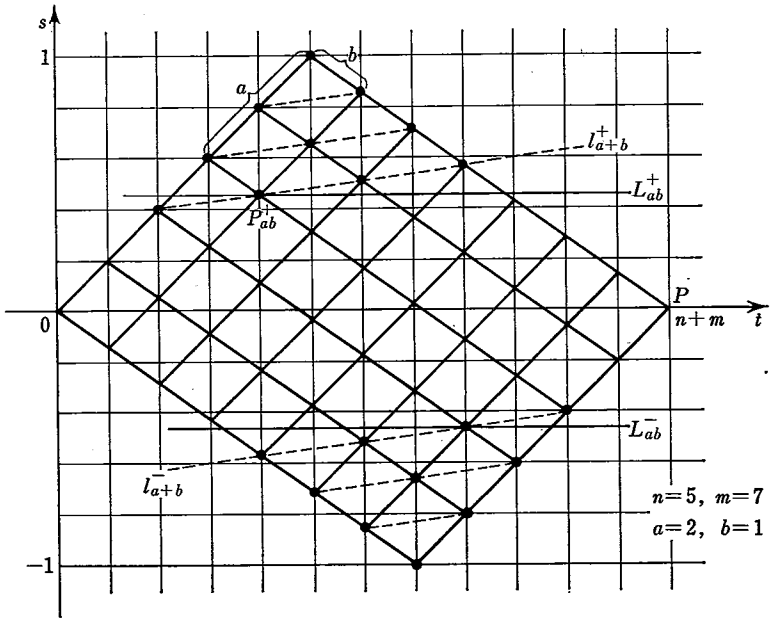

Fig. 1

Table 1 shows the values of M for $1 \leqq n \leqq 20$ and $n \leqq m \leqq n + 6$. It should be noted that for equal-sized samples we have $M = n = m$ especially.

From (2.1) and (2.2), it may be shown that for $a = 0, 1, 2, \ldots, M$ and b such as $1 \leqq a + b \leqq M + 1$,

$$A(L^+_{ab}) \cup B(L^-_{ab}) = \{A(l^+_{a+b-1}) \cup B(l^-_{a+b-1})\} \cup \{S^+_{a+b,a+1} \cup S^-_{a+b,a+1}\} \tag{2.3}$$

where the sets $A(l^+_{a+b-1}) \cup B(l^-_{a+b-1})$ and $S^+_{a+b,a+1} \cup S^-_{a+b,a+1}$ are disjoint. Therefore when we wish to find the number of paths which do not lie entirely between the lines L^+_{ab} and L^-_{ab}, it is sufficient to discuss the number of paths in the following two cases: one is to pass through at least one point in $A(l^+_i) \cup B(l^-_i)$, and the other is to lie between the two lines l^+_i, l^-_i and pass through at least one point in $S^+_{ij} \cup S^-_{ij}$. We here denote these number of paths by $Q_{nm}(i)$ and $R_{nm}(i, j)$, respectively. Namely the number of paths which do not lie entirely between the lines L^+_{ab} and L^-_{ab} is represented by $Q_{nm}(a + b - 1) + R_{nm}(a + b, a + 1)$. Therefore we have, for $a = 0, 1, 2, \ldots, M$ and b such as $1 \leqq a + b \leqq M + 1$,

$$P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right) = [Q_{nm}(a + b - 1) + R_{nm}(a + b, a + 1)] \Big/ \binom{n+m}{n}, \tag{2.4}$$

where it should be noticed that $Q_{nm}(i) + R_{nm}(i + 1, i + 2) = Q_{nm}(i + 1)$.

We next consider to find the values of $Q_{nm}(i)$ and $R_{nm}(i, j)$. Fortunately, for the computation of $Q_{nm}(i)$ we may apply the similar way used in Gnedenko-Korolyuk [5]. The result is

Table 1. Values of M which are defined to the largest integer of k satisfying $1 \leqq \frac{m}{n} < \frac{k+1}{k}$, for $1 \leqq n \leqq 20$ and $n \leqq m \leqq n+6$

n	$m=$ n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$
1	1						
2	2	1					
3	3	2	1				
4	4	3	1	1			
5	5	4	2	1	1		
6	6	5	2	1	1	1	
7	7	6	3	2	1	1	1
8	8	7	3	2	1	1	1
9	9	8	4	2	2	1	1
10	10	9	4	3	2	1	1
11	11	10	5	3	2	2	1
12	12	11	5	3	2	2	1
13	13	12	6	4	3	2	2
14	14	13	6	4	3	2	2
15	15	14	7	4	3	2	2
16	16	15	7	5	3	3	2
17	17	16	8	5	4	3	2
18	18	17	8	5	4	3	2
19	19	18	9	6	4	3	3
20	20	19	9	6	4	3	3

$$(2.5)\qquad Q_{nm}(i) = 2\sum_{\alpha=1}^{p}\binom{n+m}{\alpha n+\alpha m-(2\alpha-1)i} - \sum_{\beta=1}^{q}\binom{n+m}{(\beta+1)n+\beta m-2\beta i} - \sum_{\gamma=1}^{r}\binom{n+m}{\gamma n+(\gamma+1)m-2\gamma i},$$

where $p=\left[\frac{n+m-i}{n+m-2i}\right]$, $q=\left[\frac{m}{n+m-2i}\right]$, $r=\left[\frac{n}{n+m-2i}\right]$. $[x]$ denotes the largest integer less than or equal to x. On the the other hand, the computation of $R_{nm}(i,j)$ is very complicated in general. Hence we consider to find the values of $R_{nm}(i,j)$ in the special case: that is, i and j are subjected to the condition that $j \leqq i+1 \leqq 1+\min\left(M, \left[\frac{n-1}{2}\right]\right)$ or $j \leqq i = 1+\min\left(M, \left[\frac{n-1}{2}\right]\right)$. Since under these conditions there is no path joining each other's point in S_{ij}^{+} and S_{ij}^{-}, $R_{nm}(i,j)$ may be easily found by using only the following result: Let $U(a,b)$ be the number of paths from O to A in Fig. 2. Then we have $U(a,b) = \binom{a+b}{a} - \binom{a+b}{a-2}$, where $U(0,b)=1$ and $U(1,b)=b+1$.

By using this result, for example, in the case of Fig. 1, that is, when $n=5$, $m=7$, $a=2$ and $b=1$, the number of paths which lies between the two linens l_3^{+}, l_3^{-} and passes through the point P_{21}^{+} is given by the product of $U(1,2)$ and $U(2,5)$.

Thus under the conditions that $j \leqq i+1 \leqq 1+\min\left(M, \left[\frac{n-1}{2}\right]\right)$ or $j \leqq i = 1+\min$

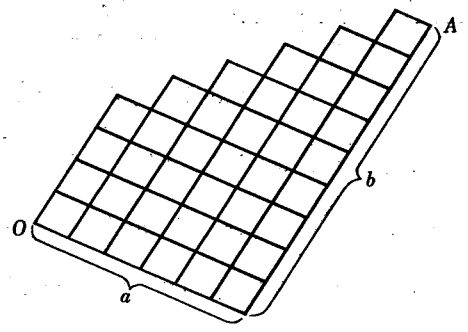

Fig. 2

$\left(M, \left[\frac{n-1}{2}\right]\right)$, $R_{nm}(i, j)$ may be found as follows:

$$(2.6)\quad R_{nm}(i, j) = 2\sum_{\xi=0}^{j-1} U(\xi, m-i-2+\xi)\, U(i-\xi, n-1-\xi)$$
$$= 2\sum_{\xi=0}^{j-1}\left[\binom{m-i-2+2\xi}{\xi} - \binom{m-j-2+2\xi}{\xi-2}\right]\left[\binom{n+i-1-2\xi}{i-\xi} - \binom{n+i-1-2\xi}{i-\xi-2}\right].$$

NOTE: It should be noticed that if we put $n = m$, $a = i$, $b = 0$, then (2.4) becomes as follows:

$$(2.7)\quad P\left(D_{nn} \geqq 1 - \frac{i}{n}\right) = \frac{1}{\binom{2n}{n}}\left[Q_{nn}(i-1) + R_{nn}(i, i+1)\right]$$
$$= \frac{1}{\binom{2n}{n}} Q_{nn}(i)$$

and (2.5) becomes

$$(2.8)\quad Q_{nn}(i) = 2\sum_{\alpha=1}^{p}\binom{2n}{2\alpha n - (2\alpha-1)i} - 2\sum_{\beta=1}^{q}\binom{2n}{(2\beta+1)n - 2\beta i}$$
$$= 2\left[\binom{2n}{2n-i} - \binom{2n}{3n-2i} + \binom{2n}{4n-3i} - \binom{2n}{5n-4i} + \cdots\right]$$
$$= 2\left[\binom{2n}{n+(n-i)} - \binom{2n}{n+2(n-i)} + \binom{2n}{n+3(n-i)} - \binom{2n}{n+4(n-i)} + \cdots\right]$$
$$= 2\sum_{\xi=1}^{r}(-1)^{\xi+1}\binom{2n}{n+\xi(n-i)},$$

where $r = \left[\frac{n}{n-i}\right]$.

Hence we have the following result:

$$(2.9)\qquad P\left(D_{nn} \geqq 1-\frac{i}{n}\right)=\frac{2}{\binom{2n}{n}}\sum_{\xi=1}^{r}(-1)^{\xi+1}\binom{2n}{n+\xi(n-i)}. \quad i=0,\ 1,\ 2,\ \ldots,n$$

This result coincides with that of Gnedenko-Korolyuk [5] and Drion [4] for equalsized samples.

From (2.4), (2.5) and (2.6), we may obtain immediately the following theorem.

THEOREM. *Let M be the largest integer of k satisfying $1 \leqq \frac{m}{n} < \frac{k+1}{k}$. Then, for $a=0,1,2,\ldots,\min\left(M,\left[\frac{n-1}{2}\right]\right)$ and b such that $0 \leqq a+b \leqq 1+\min\left(M,\left[\frac{n-1}{2}\right]\right)$, we have*

$$(2.10)\qquad P\left(D_{nm} \geqq 1-\frac{a}{n}-\frac{b}{m}\right)=2\left[\binom{n+m}{a+b-1}+\sum_{\xi=0}^{a}\left\{\binom{m-a-b-2+2\xi}{\xi}-\binom{m-a-b-2+2\xi}{\xi-2}\right\}\cdot\left\{\binom{n+a+b-1-2\xi}{a+b-\xi}-\binom{n+a+b-1-2\xi}{a+b-\xi-2}\right\}\right]\Big/\binom{n+m}{n}$$

and in particular when $a=i$ and $b=0$, for $i=0,1,2,\ldots,\min\left(M,\left[\frac{n-1}{2}\right]\right)$,

$$(2.11)\qquad P\left(D_{nm} \geqq 1-\frac{i}{n}\right)=2\binom{n+m}{i}\Big/\binom{n+m}{n}$$
$$=\frac{(n+1)(n+2)\ldots(n+r)}{(2n-i+1)(2n-i+2)\ldots(2n-i+r)}\cdot P\left(D_{nn} \geqq 1-\frac{i}{n}\right)$$

and when $a=0$ and $b=i$, for $i=0,1,2,\ldots,1+\min\left(M,\left[\frac{n-1}{2}\right]\right)$,

$$(2.12)\qquad P\left(D_{nm} \geqq 1-\frac{i}{m}\right)=2\left[\binom{n+m}{i-1}+\binom{n+i-1}{i}-\binom{n+i-1}{i-2}\right]\Big/\binom{n+m}{n}$$
$$=\frac{(2n-i+r+2)(2n-i+r+3)\ldots(2n-i+2r+1)}{(n+1)(n+2)\ldots(n+r)}\cdot P\left(D_{mm} \geqq 1-\frac{i-1}{m}\right)+2\left[\binom{n+i-1}{i}-\binom{n+i-1}{i-2}\right]\Big/\binom{n+m}{n},$$

where $r=m-n \geqq 0$.

3. Approximation to the Distribution of D_{nm}. We here consider the approximate formula which is computed by using equal-sized distribution, and that works well when n and m are slightly different.

Now, $P\left(D_{nm} \geqq 1-\frac{a}{n}-\frac{b}{m}\right)$ has the following relation:

$$P\left(D_{nm} \geqq 1-\frac{a+b}{n}\right)=P\left\{D_{nm} \geqq \left(1-\frac{a}{n}-\frac{b}{m}\right)-b\left(\frac{1}{n}-\frac{1}{m}\right)\right\}<\ldots$$

$$
\begin{aligned}
&< P\left\{D_{nm} \geqq \left(1 - \frac{a}{n} - \frac{b}{m}\right) - 2\left(\frac{1}{n} - \frac{1}{m}\right)\right\} \\
&< P\left\{D_{nm} \geqq \left(1 - \frac{a}{n} - \frac{b}{m}\right) - \left(\frac{1}{n} - \frac{1}{m}\right)\right\} \\
&< P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right) \\
&< P\left\{D_{nm} \geqq \left(1 - \frac{a}{n} - \frac{b}{m}\right) + \left(\frac{1}{n} - \frac{1}{m}\right)\right\} \\
&< \cdots \quad < P\left\{D_{nm} \geqq \left(1 - \frac{a}{n} - \frac{b}{m}\right) + a\left(\frac{1}{n} - \frac{1}{m}\right)\right\} \\
&= P\left(D_{nm} \geqq 1 - \frac{a+b}{m}\right).
\end{aligned}
$$

Hence, when we construct the approximation of $P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right)$ based on linear interpolation, it follows that

$$
\begin{aligned}
(3.2) \qquad P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right) \fallingdotseq{}& \frac{a+1}{a+b+1} P\left(D_{nm} \geqq 1 - \frac{a+b}{n}\right) \\
&+ \frac{b}{a+b+1} \cdot P\left(D_{nm} \geqq 1 - \frac{a+b}{m}\right).
\end{aligned}
$$

Thus, from (2.11), (2.12) and (3.2), we can obtain the approximate formula which is computed by using the values of $P\left(D_{nn} \geqq 1 - \frac{a+b}{n}\right)$ and $P\left(D_{mm} \geqq 1 - \frac{a+b-1}{m}\right)$. However, since it is rather complicated, we here propose experimentally the following approximate formula:

$$
\begin{aligned}
(3.3) \qquad P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right) \fallingdotseq{}& \frac{ar+1}{(a+b)r+1}\left(\frac{m}{n+m-a-b}\right)^r P\left(D_{nn} \geqq 1\right. \\
&\left. - \frac{a+b}{n}\right) + \frac{br}{(a+b)r+1}\left(\frac{2m-a-b+1}{m}\right)^r \cdot \\
&P\left(D_{mm} \geqq 1 - \frac{a+b-1}{m}\right),
\end{aligned}
$$

where $r = m - n$. When there are multiple sets of (a, b) which give the same value to $1 - \frac{a}{n} - \frac{b}{m}$, let $P\left(D_{nm} \geqq 1 - \frac{a}{n} - \frac{b}{m}\right)$ be assigned to the average of the values which are calculated from each set of (a, b).

To examine the adequateness of this approximation, and to comparison with the other approximation which results in one sample case (*i.e.* by putting $l = nm/(n+m)$, it is computed from the distribution of Kolmogorov statistic $d_l = \max_x |F(x) - F_l(x)|$), numerical examination was made for several values of n and m. The results are shown in Table 2. In many numerical examples, it appears that when $r = m - n$ is small, *i.e.* less than 5 or so, our approximation is reasonable and better than the approximation which results in one sample case.

Table 2. Comparison of the approximate values and the exact distribution of D_{nm}

Example 1. $n = 8, m = 9, l = nm/(n + m) \doteqdot 4.23$

h	Exact values of $P(D_{nm} \geqq h/72)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/72)$
55	.00831	.00793	.00705
54	.01119	.01119	.00786
48	.02024	.02237	.05665
47	.03357	.03356	.06469
46	.04689	.04475	.07287
45	.05594	.05594	.08091

Example 2. $n = 10, m = 12, l \doteqdot 5.45$

h	Exact values of $P(D_{nm} \geqq h/60)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/60)$
40	.00673	.00667	.01284
39	.01054	.01083	.01698
38	.01531	.01499	.02111
37	.01981	.01915	.02524
36	.02262	.02331	.02937
35	.02769	.02745	.04492
34	.03698	.03868	.06048
33	.04889	.04992	.07604
32	.06175	.06115	.09159

Example 3. $n = 12, m = 15, l \doteqdot 6.66$

h	Exact values of $P(D_{nm} \geqq h/60)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/60)$
36	.00955	.00920	.01537
35	.01308	.01216	.01825
34	.01703	.01873	.02316
33	.02187	.02312	.03293
32	.02967	.03117	.04276
31	.03980	.03858	.05260
30	.05072	.04600	.06237

Example 4. $n = 16, m = 20, l \doteqdot 8.88$

h	Exact values of $P(D_{nm} \geqq h/80)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/80)$
41	.01210	.01143	.01737
40	.01542	.01672	.02132
39	.01968	.02419	.02523
38	.02511	.02786	.02918
37	.03136	.03153	.03309
36	.03931	.04504	.03704
35	.04889	.05082	.04975
34	.05974	.06972	.06973

Example 5. $n = 10, m = 15, l = 6$

h	Exact values of $P(D_{nm} \geqq h/30)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/30)$
20	.00551	.00621	.00377
19	.01003	.01055	.01614
18	.01813	.01893	.02850
17	.02958	.03523	.04086
16	.04983	.06025	.05322
15	.07740	.07898	.06559

Example 6. $n = 9, m = 15, I \doteqdot 5.62$

h	Exact values of $P(D_{nm} \geqq h/45)$	Approximation by formula (3.3)	Approximation by $P(d_l \geqq h/45)$
30	.00728	.00824	.00996
29	.01038	.01250	.01636
28	01485	.02664	.02273
27	.02231	.03118	.02909
26	.02973	.04127	.04598
25	.04180	.05029	.06271

Acknowledgment The author would like to express his sincere appreciation to Prof. Y. Tumura for his guidance and his effective advice given to the author through this work.

REFERENCES

[1] Blackman, J. (1956): An extension of the Kolmogorov distribution. *Ann. Math. Statist., 27,* 513–520.

[2] Blackman, J. (1958): Correction to "An extension of the Kolmogorov distribution." *Ann. Math. Statist., 29,* 318–324.

[3] Depaix, M. (1962): Distributions de déviations maximales bilatérales entre deux échantillons indépendents de même loi continue. *Comptes Rendues Acad. Sci. Paris, 255,* 2900–2902.

[4] Drion, E. F. (1952): Some distribution-free tests for the difference between two empirical cumulative distribution functions. *Ann. Math. Statist., 23,* 563–574.

[5] Gnedenko, B. V. and Korolyuk, V. S. (1951): On the maximum discrepancy between two empirical distributions. (in Russian). *Doklady Akad. Nauk SSSR, 80,* 525–528.

[6] Korolyuk, V. S. (1955): On the deviation of empirical distributions for the case of two independent samples. (in Russian). *Izv. Akad. Nauk. SSSR Ser. Mat., 19,* 81–96.

[7] Massey, F. J., JR. (1951): The distribution of the maximum deviation between two sample cumulative step functions. *Ann. Math. Statist., 22,* 125–128.

[8] Massey, F. J., JR. (1952): Distribution table for the deviation between two sample cumulatives. *Ann. Math. Statist., 23,* 435–441.

[9] Steck, G. P. (1969): The smirnov two sample tests as rank tests. *Ann. Math. Statist., 40,* 1449–1466.

[10] Wilks, S. S. (1962): *Mathematical statistics.* New York: John Wily and Sons.