

A THREE-SAMPLE KOLMOGOROV-SMIRNOV TEST

BY HERBERT T. DAVID

Iowa State College

1. Introduction. In 1951, Gnedenko and Korolyuk published an elegant derivation ([6])¹ of the null distribution of the Kolmogorov-Smirnov statistic $D_{2,n}$ for two samples of equal size n . The statistic $D_{2,n}$ is given by

$$(1) \quad D_{2,n} = \sup_t | F_{2,n}(t) - F_{1,n}(t) |,$$

where $F_{i,n}(t)$ is the sample cumulative distribution function for the i th sample. The distribution derived by Gnedenko and Korolyuk is

$$(2) \quad \Pr \left\{ D_{2,n} \geq \frac{l}{n} \right\} = 2 \binom{2n}{n}^{-1} \sum_{i=1}^{\lfloor n/l \rfloor} (-1)^{i+1} \binom{2n}{n-il}.$$

Since

$$(3) \quad \lim_{n \rightarrow \infty} \frac{\binom{2n}{n-k\sqrt{n}}}{\binom{2n}{n}} = e^{-k^2},$$

(2) easily leads to the familiar asymptotic result

$$(4) \quad \lim_{n \rightarrow \infty} \Pr \left\{ n^{1/2} D_{2,n} \geq \lambda \right\} = 2 \sum_{i=1}^{\infty} (-1)^{i+1} e^{-(i\lambda)^2}.$$

Gnedenko and Korolyuk's proof hinges on the fact that, in the null case (for two samples drawn from the same continuous distribution), $\Pr\{D_{2,n} \geq l/n\}$ equals the probability that the maximum deviation from the origin of a certain random walk in the line is at least l . The random paths involved in this random walk start at the origin, and consist of $2n$ unit steps, n to the left and n to the right, with all possible permutations of left and right steps equally likely. The probability $\Pr\{D_{2,n} \geq l/n\}$ is thus equal to, say, $M / \binom{2n}{n}$, where $\binom{2n}{n}$ is the total number of equally likely paths, and M is the number of these paths with maximum deviation from the origin at least l . M can be computed by the reflection principle in the line ([2], [1]), leading to (2).

In this paper I show that the null distribution of the three-sample extension $D_{3,n}$ (see (6) below) of $D_{2,n}$ can be derived by extending the geometric approach of [6] from the line to the plane.

Received October 21, 1957; revised February 25, 1958.

¹ The review of this paper in *Mathematical Reviews* [3] was brought to my attention by Murray Rosenblatt.

$D_{3,n}$ is but one of several “distance” criteria that have recently appeared in the literature. Fisz and Kiefer [4], [7]² have shown that the criterion

$$R_n = \max \left\{ \sup_t |F_{3,n}(t) - F_{2,n}(t)|, \sup_t |2F_{1,n}(t) - F_{2,n}(t) - F_{3,n}(t)| \right\},$$

and extensions of R_n to k samples and unequal sample sizes, can be used with existing Kolmogorov-Smirnov tables because the events

$$A: \left[\sup_t |F_{3,n}(t) - F_{2,n}(t)| \leq \lambda_1 \right]$$

and

$$B: \left[\sup_t |2F_{1,n}(t) - F_{2,n}(t) - F_{3,n}(t)| \leq \lambda_2 \right]$$

are independent. It may be of interest to note that the criterion R_n corresponds to using a rectangular boundary on the hexagonal grid of Figure 1, and that the independence of the events A and B , and distribution of R_n , follow easily from this representation.

Ozols’ [8]² treatment of the criterion

$$S_n = \max \left\{ \sup_t (F_{3,n}(t) - F_{2,n}(t)), \sup_t (F_{2,n}(t) - F_{1,n}(t)) \right\},$$

is similar to my treatment of the criterion $D_{3,n}$. The boundary corresponding to S_n is an infinite 60° wedge on the hexagonal grid of Figure 1.

Finally, Kiefer [7] and Gihman [5]³ consider a criterion T_n (or D_k^2) of form

$$\sup_t \left(\sum_{i=1}^k (F_{i,n}(t) - \overline{F_n(t)})^2 \right), \quad \overline{F_n(t)} = \sum_{i=1}^k F_{i,n}(t)/k,$$

and extensions of this criterion to unequal sample sizes; Kiefer [7] also considers the k -sample extension V_n of the statistic (5) given below in section 2.

Kiefer has shown in [7] that “distance” criteria of the type discussed above have good power properties. Among such criteria, one might suspect on heuristic grounds that $D_{3,n}$ has especially good power characteristics against the “one-sided” alternative $H_A: [(X < Y < Z) \text{ or } (Y < Z < X) \text{ or } (Z < X < Y)]$. This is because H_A tends to generate paths, on the grid of Figure 1, in the directions $\pi/6, \pi/6, +2\pi/3$, or $\pi/6 + 4\pi/3$.

2. A three-sample Kolmogorov-Smirnov statistic and its small-sample null distribution. A natural three-sample extension of (1) would be

$$(5) \quad \text{Max} \left\{ \sup_t |F_{2,n}(t) - F_{1,n}(t)|, \sup_t |F_{3,n}(t) - F_{2,n}(t)|, \right. \\ \left. \sup_t |F_{1,n}(t) - F_{3,n}(t)| \right\}.$$

² I owe these references to an associate editor.

³ I owe this reference to Milton Sobel.

But (5) does not lend itself easily to an extension of Gnedenko and Korolyuk's geometric method; a statistic that does so lend itself is that obtained from (5) by deleting the absolute value signs:

$$(6) \quad D_{3,n} = \text{Max} \left\{ \sup_t (F_{2,n}(t) - F_{1,n}(t)), \sup_t (F_{3,n}(t) - F_{2,n}(t)), \sup_t (F_{1,n}(t) - F_{3,n}(t)) \right\}.$$

The null distribution of $D_{3,n}$ is its distribution when the three samples are drawn from the same continuous population. This null distribution is derived as follows.

A step of type *A* in the plane is defined to be a unit step to the right (direction 0); a step of type *B* is a unit step in the direction $2\pi/3$, and a step of type *C* is a unit step in the direction $4\pi/3$.

In the null case considered, ties occur with probability zero; hence (almost) every set of three samples of n leads to a ranking of the $3n$ sample values making up the three samples. Corresponding to each set of three samples, consider a path $p_{3,n}$ from the origin, composed of $3n$ unit steps, with the k th step of $p_{3,n}$ of type *A* if the rank k belongs to the first sample, etc. Clearly every $p_{3,n}$ contains n steps of each of the three types *A*, *B* and *C*.

Next, consider the equilateral triangle in the plane that is centered at the origin, has sides of length $3l$, and is oriented such that one of its sides is horizontal. Call this equilateral triangle Γ_l . Clearly

$$(7) \quad \left\{ D_{3,n} \geq \frac{l}{n} \right\} \Leftrightarrow \{ (p_{3,n} \cap \Gamma_l) \text{ is not empty} \}.$$

But in the null case every path $p_{3,n}$ (permutation of $3n$ steps, n each of type *A*, *B* and *C*) is possible, and each of the $(3n)!/(n!)^3$ such paths is equally likely. Hence (7) implies

$$(8) \quad \Pr \left\{ D_{3,n} \geq \frac{l}{n} \right\} = \Pr \{ (P_{3,n} \cap \Gamma_l) \text{ is not empty} \} = N/(3n)!/(n!)^3,$$

where N is the number of paths $p_{3,n}$ touching or piercing Γ_l . The small-sample problem is therefore solved if N can be evaluated.

N is evaluated by extending to the plane the principle of reflection that yielded M . Consider a hexagonal grid in the plane, consisting of " \oplus " points and " \ominus " points, as indicated in figure 1 for the case ($n = 7, l = 2$). The extent of the grid is fixed by the fact that the distance between the origin 0 and each of the three "vertices" (indicated by the letters V_1, V_2, V_3 in figure 1) is $(3l)(\lfloor n/l \rfloor)$. This distance is of course $(3 \cdot 2)(\lfloor 7/2 \rfloor) = 18$ for the case illustrated by figure 1. The central triangle indicated by the heavy line in figure 1 represents Γ_l .

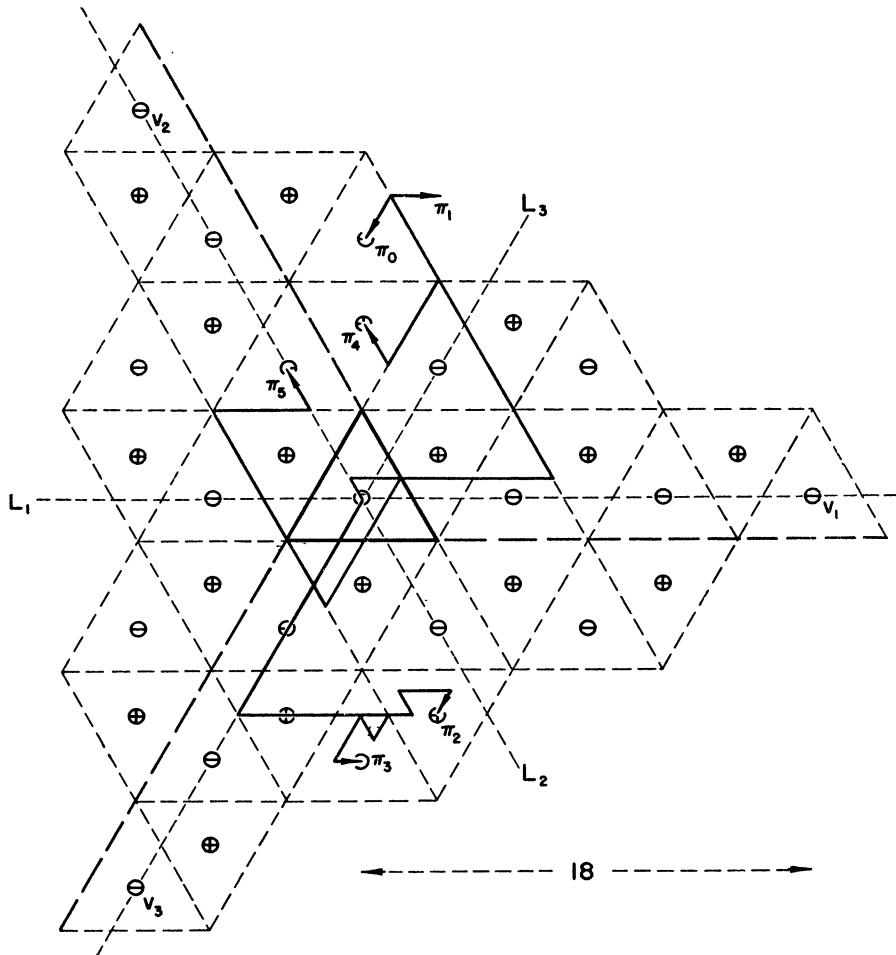


FIG. 1

Any path from the origin 0 to a \oplus point, that consists of $3n$ steps of type A , B or C , is called a path of type \oplus . A path of type \ominus is defined similarly. A path of type \oplus or \ominus is called an auxiliary path π . Again, any path from the origin to the origin, that consists of $3n$ steps of type A , B or C , and that touches the boundary Γ_i , is called a boundary path β . Finally, N_{\oplus} , N_{\ominus} and N_0 are, respectively, the total number of paths of type \oplus , the total number of paths of type \ominus , and the total number of boundary paths.

The argument now is as follows.

- (9) For any particular endpoint, whether it be a \oplus point, a \ominus point, or the origin 0, the specification that there be $3n$ steps in a boundary path or auxiliary path from the origin to that endpoint actually determines the numbers m_A , m_B and m_C of steps of types A , B and C involved in the path.

(9) follows from the fact that the location of the endpoint provides three equations in m_A , m_B and m_C , which, together with

$$(10a) \quad m_A + m_B + m_C = 3n,$$

yield m_A , m_B and m_C . These three equations are

$$(10b) \quad m_C - m_B = K_1$$

$$(10c) \quad m_A - m_C = K_2$$

$$(10d) \quad m_B - m_A = K_3.$$

K_1 , K_2 and K_3 are determined by the signed perpendicular distances d_1 , d_2 and d_3 of the endpoint from the lines L_1 , L_2 and L_3 (see figure 1). For example, $K_1 = 2d_1/3^\dagger$ if the endpoint is d_1 units below L_1 , and $K_1 = -2d_1/3^\ddagger$ if the endpoint is d_1 units above L_1 .

In particular, for every path from the origin to the origin, (10b), (10c) and (10d) become $m_C - m_B = m_A - m_C = m_B - m_A = 0$, which, together with (10a), yield $m_A = m_B = m_C = n$. This last implies that the boundary paths are exactly the paths enumerated by N , or

$$(11) \quad N_0 = N.$$

Next, we introduce the operation of reflection. Reflection is an operation performed on an auxiliary path π that yields a path $p(\pi)$ which can be either an auxiliary path or a boundary path. Reflection is defined as follows.

Let π be an auxiliary path whose last point of contact (proceeding along π from the origin) with Γ_l is the point u .

1) Suppose first that u is not a vertex of Γ_l . Suppose for example that u lies within the horizontal side of Γ_l (i.e. the side oriented in the direction of a step of type A). Then $p(\pi)$ is obtained from π by replacing every step of type B occurring after u by a step of type C , and every step of type C occurring after u by a step of type B . Analogously, if u lies within the side of Γ_l oriented in the direction of a step of type B , then $p(\pi)$ is obtained from π by replacing steps of type A occurring after u by steps of type C , and vice-versa; if u lies within the side of Γ_l oriented in the direction of a step of type C , the transposition of step types involves types A and B .

For example, reflection of the path π_4 (see figure 1) leads to the path π_5 .

2) If u is a vertex of Γ_l , then reflection consists, as in 1), of a transposition of two step types. Which two step types are involved is determined by the requirement that the step occurring immediately after u be converted into a step lying in Γ_l . Thus, for example, if u is the vertex of Γ_l lying on both of the two non-horizontal sides of Γ_l , and if the step occurring immediately after u is a step of type A , then the two step types involved in the transposition are types A and C ; in other words, $p(\pi)$ is obtained from π by replacing every step of type A occurring after u by a step of type C , and every step of type C occurring after u by a step of type A .

The operations of reflection, performed on an auxiliary path π , yields a path $p(\pi)$ which 1) contains $3n$ steps, 2) contains no steps of types other than A , B and C ; and 3) begins at the origin and ends at a \oplus point, at a \ominus point, or at the origin. (Endpoints exterior to the grid of figure 1, such as the endpoint of the path π_1 for example, cannot result from reflection, because, for such endpoints, equations (10) have at least one negative solution). Finally, it is clear that: 4) the number of steps of π exterior to Γ_l , from the point of last contact of π with Γ_l to the endpoint of π , is greater by at least one than the number of steps of $p(\pi)$ exterior to Γ_l , from the point of last contact of $p(\pi)$ with Γ_l to the endpoint of $p(\pi)$.

By 1), 2) and 3), $p(\pi)$ is either an auxiliary path or a boundary path, and, by 4), successive reflection $p_1(\pi)$, $p_2(p_1(\pi))$, $p_3(p_2(p_1(\pi)))$, \dots eventually lead to a boundary path, say $p_k(p_{k-1}(\dots p_1(\pi)\dots))$; this boundary path is called the image $\beta(\pi)$ of π .

Our discussion of reflection can be summarized by:

- (12) To every auxiliary path π there corresponds a unique image path $\beta(\pi)$, which is a boundary path obtained from π by successive reflections.

Further,

- (13) among all the auxiliary paths with the same image path, the number of paths of type \oplus exceeds the number of paths of type \ominus by one.

(13) follows from the fact that the auxiliary paths with the same image path β come in pairs of type (\oplus, \ominus) , as illustrated in figure 1 by paths π_2 and π_3 , except for a single "bachelor" path of type \oplus from the origin to one of the three \oplus points immediately next to Γ_l .

The bachelor path of type \oplus is the auxiliary path yielding β after only one reflection; it is uniquely defined for any boundary path β , and is constructed from β as follows. Let v be the last point of contact of β with Γ_l , proceeding along β from the origin in accordance with the directions associated with each of the three step types. (Note that β has at least one point of contact with Γ_l , since β is a boundary path). The bachelor auxiliary path is constructed from β by "reflecting" the portion of β following v . (The word "reflection" is put in quotes because, up to now, reflection has been defined only as an operation on auxiliary paths. But the construction involved here is entirely analogous to the earlier operation.) For example, if v lies in the horizontal side of Γ_l , then "reflection" of the portion of β following v consists of replacing every step of type B by a step of type C , and every step of type C by a step of type B ; the procedure is analogous if v lies in one of the other two sides of Γ_l . (Note that v is never a vertex of Γ_l).

The pairing of the other auxiliary paths with image β is accomplished by "reflection" about the last point of contact with the triangular grid lines indicated by the dashed lines in figure 1. (The word "reflection" again is put in quotes, because the usage here does not correspond exactly to the operation

yielding $p(\pi)$ from π). For example, consider an auxiliary path π_2 with image β , and let the last point of contact of π_2 with the triangular grid lines be w ; suppose for example that w lies on a grid line oriented in the direction of a step of type B (as illustrated in figure 1). Then, as indicated in figure 1, the mate π_3 of π_2 is obtained from π_2 by replacing every step of type C occurring after w by a step of type A , and every step of type A by a step of type C . The same "reflection" operation, applied to π_3 , yields π_2 , which establishes the pairing.

That π_2 and its mate π_3 have the same image β is best verified by imagining π_2 and π_3 as undergoing reflection simultaneously.

Except for the single bachelor path, auxiliary paths with the same image thus come in pairs of type (\oplus, \ominus) , except possibly in the case of an auxiliary path, such as that indicated by π_0 in figure 1, whose potential mate π_1 is not one of the auxiliary paths. However, auxiliary paths such as π_0 do not exist, and this is shown as follows.

Suppose there were an auxiliary path, such as π_0 , to an endpoint at the outer edge of the hexagonal grid of \oplus points and \ominus points, which entered the triangular cell containing this endpoint from an "exterior" side of the cell. The four equations (10a), (10b), (10c) and (10d) yield $m_C = n - l\lfloor n/l \rfloor$ for any auxiliary path to any endpoint between the two vertices V_1 and V_2 . (Correspondingly $m_B = n - l\lfloor n/l \rfloor$ and $m_A = n - l\lfloor n/l \rfloor$ for the other two sets of "outer" endpoints). Hence, if π_0 existed, it would contain $n - l\lfloor n/l \rfloor$ steps of type C . But then π_1 would contain $n - l\lfloor n/l \rfloor - l$ steps of type C , which could not be because $n - l\lfloor n/l \rfloor - l$ is negative.

Finally,

(14) Every boundary path is the image of at least one auxiliary path,

because every boundary path is the image at least of its corresponding "bachelor" path.

(12), (13), and (14) imply

$$(15) \quad N_0 = N_{\oplus} - N_{\ominus}.$$

(15) is shown as follows. Let π denote an auxiliary path, let β denote a boundary path, and define the function $f(\pi, \beta)$ as follows.

$$f(\pi, \beta) = 1 \quad \text{if } \beta \text{ is the image of } \pi, \text{ and } \pi \text{ is a path of type } \oplus.$$

$$f(\pi, \beta) = -1 \quad \text{if } \beta \text{ is the image of } \pi, \text{ and } \pi \text{ is a path of type } \ominus.$$

$$f(\pi, \beta) = 0 \quad \text{if } \beta \text{ is not the image of } \pi.$$

Now, for any fixed β ,

$$\sum_{\pi} f(\pi, \beta) = 1$$

by (13) and (14), so that

$$(16) \quad \sum_{\beta} [\sum_{\pi} f(\pi, \beta)] = N_0$$

Again, by (12), it is true for every fixed π that

$$\begin{aligned} \sum_{\beta} f(\pi, \beta) &= +1 \text{ for } \pi \text{ of type } \oplus \\ &= -1 \text{ for } \pi \text{ of type } \ominus, \end{aligned}$$

so that

$$(17) \quad \sum_{\pi} [\sum_{\beta} f(\pi, \beta)] = N_{\oplus} - N_{\ominus},$$

and (15) follows from (16) and (17).

(11) and (15) yield

$$(18) \quad N = N_{\oplus} - N_{\ominus}$$

In view of (8), equation (18) represents the solution of the small-sample problem, because the computation of N_{\oplus} and of N_{\ominus} is straightforward. For example, N_{\oplus} is the total number of paths of type \oplus , which is easily computed because the number of paths to any particular \oplus point is given by the usual trinomial coefficient, the count being entirely unrestricted. The three arguments of this trinomial coefficient are the numbers of steps of types A , B and C involved in any auxiliary path to this \oplus point; these numbers are of course fixed by the location of the \oplus point, in view of equations (10). There remains only the problem of efficient enumeration of \oplus points and θ points; one such enumeration gives for $\Pr \{D_{3,n} \geq l/n\}$ the expression

$$(19) \quad 3 \sum_{i=1}^{\lfloor n/l \rfloor} \sum_{j \in J(i)} (\pm)(n!)^3 / (n - il)!(n + jl)!(n + (i - j)l)!,$$

where the set $J(i)$ consists of the integers $(2 - i, 3 - i, 5 - i, 6 - i, 8 - i, 9 - i, 11 - i, 12 - i, \dots, 2i)$, and where the (\pm) sign indicates that, for fixed i , successive terms in the finite series indexed by j have alternating signs, beginning with $+$ for $j = 2 - i$, $-$ for $j = 3 - i$, $+$ for $j = 5 - i$, etc.

3. Large-sample distribution. The asymptotic distribution of $D_{3,n}$ is given by the following theorem.

THEOREM. For $\lambda n^{\frac{1}{3}}$ integral

$$\lim_{n \rightarrow \infty} \Pr \{n^{\frac{1}{3}} D_{3,n} \geq \lambda\} = 3 \sum_{i=1}^{\infty} \sum_{j \in J(i)} (\pm) e^{-\lambda^2(i^2 + j^2 - ij)}$$

where the set $J(i)$ and the sign (\pm) are as defined in (19).

PROOF. Put $l = \lambda n^{\frac{1}{3}}$ in (19). Since, for fixed k_1, k_2, k_3 with $k_1 + k_2 + k_3 = 0$,

$$(20) \quad \lim_{n \rightarrow \infty} \frac{(n!)^3}{(n + k_1 n^{\frac{1}{3}})!(n + k_2 n^{\frac{1}{3}})!(n + k_3 n^{\frac{1}{3}})!} = e^{-\frac{1}{3}(k_1^2 + k_2^2 + k_3^2)},$$

it suffices to show that,

$$(21) \quad \left\{ \begin{array}{l} \text{for } k \text{ large enough,} \\ R(k, n, \lambda) = \left| \sum_{i=k}^{\lfloor n^{1/2}/\lambda \rfloor} \sum_{j \in J(i)} (\pm)(n!)^3 / (n - i\lambda n^{1/2})! \right. \\ \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \cdot (n + j\lambda n^{1/2})!(n + (i - j)\lambda n^{1/2})! \\ \left. \right| \\ \text{is arbitrarily small, uniformly in } n \text{ for large } n. \end{array} \right.$$

Rewriting the terms of (21) and putting the absolute value signs inside the first summation,

$$(22) \quad R(k, n, \lambda) \leq \sum_{i=k}^{\lfloor n^{1/2}/\lambda \rfloor} ((n!)^3 / (n - i\lambda n^{1/2})!(2n + i\lambda n^{1/2})!) \cdot \left(\left| \sum_{j \in J(i)} (\pm) \binom{2n + i\lambda n^{1/2}}{n + j\lambda n^{1/2}} \right| \right).$$

For fixed i , the absolute values of the terms of the alternating series increase monotonically to the maximum

$$\binom{2n + i\lambda n^{1/2}}{n + [i/2] \lambda n^{1/2}},$$

and then decrease monotonically. Hence

$$\left| \sum_{j \in J(i)} (\pm) \binom{2n + i\lambda n^{1/2}}{n + j\lambda n^{1/2}} \right| \leq 2 \binom{2n + i\lambda n^{1/2}}{n + [i/2] \lambda n^{1/2}},$$

and (22) yields

$$(23) \quad R(k, n, \lambda) \leq 2 \left[\sum_{i=k}^{\lfloor n^{1/2}/\lambda \rfloor} \right] b_i,$$

where

$$(24) \quad b_i = (n!)^3 / (n - i\lambda n^{1/2})! \left(n + \left[\frac{i}{2} \right] \lambda n^{1/2} \right)! \left(n + \left(i - \left[\frac{i}{2} \right] \right) \lambda n^{1/2} \right) !.$$

It is easy to show by direct computation that

1) b_i/b_{i+1} is increasing in i ,

2) $b_k/b_{k+1} \geq \left(1 + \left[\frac{k}{2} \right] \lambda n^{-1/2} \right)^{\lambda n^{1/2}}$, which is uniformly close to

$e^{[k/2]\lambda^2}$ for n large.

Hence, by (23), $R(k, n, \lambda)$ is essentially bounded by

$$(25) \quad 2b_k / (1 - e^{-[k/2]\lambda^2})$$

for n large. But, by (20) and (24), (25) is approximated by

$$2e^{-(k^2 + [k/2]^2 - k[k/2])} / (1 - e^{-[k/2]^2})$$

for n large; this establishes (21).

REFERENCES

- [1] S. CHANDRASEKHAR, "Stochastic problems in physics and astronomy", *Selected Papers on Noise and Stochastic Processes* (editor N. Wax), Dover, New York, 1954, pp. 4-9.
- [2] W. FELLER, *Probability theory and its applications*, Wiley, New York, 1950, p. 304 (problem No. 5).
- [3] W. FELLER, Review of B. V. Gnedenko, and V. S. Korolyuk, "On the maximum discrepancy between two empirical distributions", *Mathematical Reviews*, Vol. 13 (1952), pp. 570-571.
- [4] M. FISZ, "A limit theorem for empirical distribution functions", *Bulletin de l'Académie Polonaise des Sciences, classe III*, Vol. 5 (1957), pp. 695-698.
- [5] I. I. GIHMAN, "On a nonparametric criterion for the homogeneity of k samples", *Teoriya Veryotnostei i yeyau primenyeniya*, Vol. 2 (1957), pp. 380-384.
- [6] B. V. GNEDENKO AND V. S. KOROLYUK, "On the maximum discrepancy between two empirical distributions", *Doklady Akad. Nauk. SSSR (N.S.)*, Vol. 80 (1951), pp. 525-528.
- [7] J. KIEFER, "Distance tests with good power for the nonparametric k -sample problem" (abstract), *Ann. Math. Stat.*, Vol. 26 (1955), p. 775.
- [8] V. OZOLS, "Generalization of the theorem of Gnedenko-Korolyuk to three samples in the case of two one-sided boundaries". *Latvijas PSR Zinātņu Akad. Vēstis* 1956, No. 10 (111). pp. 141-152. (Listed in *Mathematical Reviews*, Vol. 18 (1957), p. 833.)