ON SOME RESULTS AND PROBLEMS IN CONNECTION WITH STATISTICS OF THE KOLMOGOROV-SMIRNOV TYPE

I. VINCZE
MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES

1. Introduction

Since Kolmogorov and Smirnov established their limiting distribution theorems concerning maximal deviations between empirical and theoretical distributions, an increasing amount of scientific work has been done by statisticians in this field. Practical importance and theoretical interest give the motivation. Researchers have worked on distribution laws, power considerations and the limiting process in the last five years with considerable results. The present paper will consider only a few results, those nearest to the author's work and interest of the past few years. The first part, Section 2, concerns the case when the parent distribution is noncontinuous, the third and fourth sections consider the one and two sample problem, in the fifth section an analogous question for a density function due to Révész is discussed, while in the last section the two dimensional problem is considered.

2. The Gnedenko-Korolyuk distribution for discontinuous random variables

In his paper Schmid [12] has given the limiting distribution law of the Kolmogorov and of the Smirnov statistics, that is, of

$$\begin{aligned} D_n &= \sup_{(x)} |F_n(x) - F(x)|, \\ D_n^+ &= \sup_{(x)} [F_n(x) - F(x)] \end{aligned} \tag{2.1}$$

for discontinuous $F(x)$, where $F_n(x)$ denotes the empirical distribution function of a sample of size n from a population distributed according to $F(x)$. Using the ballot lemma Csáki [2] determined the exact distribution of D_n^+ for finite n which corresponds to the well known Smirnov-Birnbaum-Tingey distribution for continuous $F(x)$. His formula has a fairly complicated form.

The two sample case for discontinuous $G(x) \equiv F(x)$ was considered by the author [20], who determined for finite $m = n$ the exact distribution of

$$
\begin{aligned}
D^{+}_{n,n} &= \max_{(x)} [F_n(x) - G_n(x)], \\
D_{n,n} &= \max_{(x)} |F_n(x) - G_n(x)|,
\end{aligned} \tag{2.2}
$$

as well as their limiting forms as $n \to \infty$.

Let $F(x)$ have jumps at $x_1, \cdots, x_r$, with $x_i < x_{i+1}$, and be continuous otherwise. Let $F(x)$ be left continuous satisfying the following relations: with $x_0 = -\infty, x_{r+1} = +\infty$,

$$
\begin{aligned}
F(x_i) - F(x_{i-1} + 0) &= p_i, \qquad & i &= 1, \cdots, r + 1 \\
F(x_i + 0) - F(x_i) &= q_i, \qquad & i &= 1, \cdots, r,
\end{aligned} \tag{2.3}
$$

where $\Sigma_{i=1}^{r+1} p_i + \Sigma_{i=1}^{r} q_i = 1$.

For the limiting distributions we have for $y \geqq 0$,

$$
\lim_{n \to \infty} P\left[\left(\frac{n}{2}\right)^{1/2} D^{+}_{n,n} < y\right] \tag{2.4}
$$

$$
= \frac{1}{(2\pi)^r p_{r+1}} \int \underset{G+}{\cdots} \int \prod_{i=1}^{r+1} \left[1 - \exp\left\{-\frac{2}{p_i}(y - S_{i-1} - T_{i-1})(y - S_i - T_{i-1})\right\}\right]
$$

$$
\exp\left\{-\frac{1}{2}\sum_{i=1}^{r}(u_i^2 + w_i^2) - \frac{1}{2p_{r+1}}(S_r + T_r)^2\right\} \prod_{i=1}^{r} du_i dw_i,
$$

where

$$
\begin{aligned}
S_0 &= 0, \qquad & S_i &= \sum_{j=1}^{i} u_j {p_j}^{1/2}, \qquad & i &= 1, \cdots, r + 1, \\
T_0 &= 0, \qquad & T_i &= \sum_{j=1}^{i} v_j {q_j}^{1/2}, \qquad & i &= 1, \cdots, r
\end{aligned} \tag{2.5}
$$

and for the domain of integration we have

$$
G^{+} = \{S_{i-1} + T_{i-1} < y, S_i + T_{i-1} < y, i = 1, \cdots, r\}. \tag{2.6}
$$

In the two sided case the following form holds, for $y > 0$,

$$
\lim_{n \to \infty} P\left[\left(\frac{n}{2}\right)^{1/2} D_{n,n} < y\right] \tag{2.7}
$$

$$
= \frac{1}{(2\pi)^r p_{r+1}} \int \underset{G}{\cdots} \int \prod_{i=1}^{r+1} \left(\sum_{\gamma=-\infty}^{\infty} \left[\exp\left\{-\frac{2}{p_i}(2\gamma y - u_i {p_i}^{1/2})\, 2\gamma y\right\}\right.\right.
$$

$$
\left.\left. - \exp\left\{-\frac{2}{p_i}[(2\gamma + 1)y + S_i + T_{i-1}][(2\gamma + 1)y + S_{i-1} + T_{i-1}]\right\}\right]\right)
$$

$$
\exp\left\{-\frac{1}{2}\sum_{i=1}^{r}(u_i^2 + w_i^2) - \frac{1}{2p_{r+1}}(S_r + T_r)^2\right\} \prod_{i=1}^{r} du_i dw_i
$$

with the same S_i and T_i as above, while for the domain of integration we have

$$G = \{-y < S_{i-1} + T_{i-1} < y_1 - y < S_i + T_{i-1} < y, i = 1, \cdots, r\}. \tag{2.8}$$

As in the one sample case the distributions are no longer independent of $F(x)$; they depend on the values $F(x_i)$ and $F(x_i + 0)$ at the points of discontinuity, but only on them.

The case of $P_i = 0$ for each i, that is, the case of a discrete distribution was considered by Š. Šujan [16]. For finite n the corresponding distributions can be obtained from the formulas given in the paper cited, but the above relations do not work for $p_i = 0$. According to the calculation of Šujan for purely discrete random variables, with our above notations the following hold

$$\begin{aligned} \lim_{n\to\infty} P\left[\left(\frac{n}{2}\right)^{1/2} D_{n,n}^{+} < y\right] \\ = \frac{1}{(2\pi)^{r-1} q_r^{1/2}} \int \underset{G_{r-1}}{\cdots} \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{r} w_i^2 - \frac{1}{2g_r} T_{r-1}^2\right\} \prod_{i=1}^{r-1} dw_i, \end{aligned} \tag{2.9}$$

where $G_{r-1}^{+} = \{T_i < y, i = 1, \cdots, r-1\}$.

The corresponding limit relation for $D_{n,n}$ has the same form with the single exception that the domain of integration has the two sided form

$$G_{r-1} = \{-y < T_i < y, i = 1, \cdots, r-1\}. \tag{2.10}$$

It was pointed out by Kolmogorov and proved by Noether [9] that for a critical value c the relation

$$P(D_n > c \mid F \text{ is discrete}) \leqq P(D_n > c \mid F \text{ is continuous}) \tag{2.11}$$

holds. This means that for given size α of a test based on the Kolmogorov statistic

$$D_n = \sup_{(x)} |F_n(x) - F(x)| \tag{2.12}$$

the critical region must be at least as large as in the continuous case. Table I gives numerical calculations carried out by Šujan showing how pessimistic the Kolmogorov-Smirnov two sample test may be in the discrete case.

TABLE I

NUMERICAL EXAMPLE OF PESSIMISM OF KOLMOGOROV-SMIRNOV TWO SAMPLE TEST FOR THE CASE $n = 3$, $r = 2$, $c = 0$, $q_2 = 1 - q_1$

q_1	$P(D_{3,3} > 0)$
1/4	0.308
1/3	0.332
1/2	0.363
F continuous	0.75

The last row corresponds to the continuous case. The example chosen is a very extreme case. It can be shown that the less pessimistic case is $p_1 = p_2 = \frac{1}{2}$.

3. On distributions of statistics connected with the two sample problem

In the last two decades a lot of effort has been expended to determine the exact probabilities of the Kolmogorov-Smirnov two sample statistic, which was known for special sample sizes only ($m = kn$, k positive integer). In this respect, the paper of Steck [15] can be considered as containing far reaching results. Using ideas of Maag and Stephens, [7] and also of Lehmann, the Smirnov statistics were expressed in terms of the ranks of one sample. Steck then expressed explicitly the distribution in the form of a determinant, when one underlying distribution is the power of the other, $G(x) = [F(x)]^k$. Further, by giving determinant formulas for the frequency content under the null hypothesis of any parallelepiped in the sample space of the ranks of one sample, he obtained the null joint distribution of the one sided statistic and thus the null distribution of the two sided statistic, for arbitrary sample sizes.

Since the application of a pair of statistics as test statistic was introduced by Vincze [19], some authors have determined joint distribution laws considering the question from this point of view. V. Sujan [17] obtained the joint distribution of the maximum deviation and the number of runs; she, among others, proved that these two statistics are asymptotically independent. Mohanty and Pestros [8] considered joint distributions belonging to different rank statistics. In my paper the test was examined when the alternative is specified, in which case the likelihood ratios are completely ordered. The test for one sided alternatives was constructed by the authors mentioned using a partial ordering in the range space of the joint statistic based on certain relations of likelihood ratios.

The considerations mentioned led to a development in the theory of simple random walks, and an interesting new method was constructed by Dwass [4] in the technique of generating functions.

For the former point we refer to the paper of Sen [13] in which, among other results, a systematic treatment is given of certain useful path transformations.

In [4] the random walk $\{\vartheta_1, \vartheta_2, \cdots\}$ is considered with $P(\vartheta_i = +1) = p$, $P(\vartheta_i = -1) = 1 - p = q$, the ϑ_i being independent. Assuming that $p > q$ the random walk returns to the origin at most finitely often with probability one. Denote by T the last index for which the partial sum $s_i = \vartheta_1 + \vartheta_2 + \cdots + \vartheta_i$ vanishes (that is, $s_T = 0$, T even). Let U be a function defined on the random walk which is completely determined by the first T of the ϑ_j. Let us define

$$U_n(\vartheta_1, \vartheta_2, \cdots, \vartheta_{2n}) = U(\vartheta_1, \vartheta_2, \cdots, \vartheta_T), \qquad \text{when} \quad T = 2n. \tag{3.1}$$

But the U_n can be considered as those rank statistics which play a role in the two sample problems. The relation

$$E(U) = \sum_{n=0}^{\infty} E(U_n)P(T = 2n) = (1 - 2p) \sum_{n=0}^{\infty} \binom{2n}{n} (pq)^n E(U_n) \tag{3.2}$$

given by Dwass [4] makes it possible for him to derive previously known and also new distributions in a very simple way.

An example in [19] shows that when we use a pair of statistics instead of one statistic in the case of a given simple alternative, the probability of the error of second kind is reduced to one half or one quarter of its previous value. In any case, depending on the alternative, the second kind error can be diminished from extremely high values to acceptable ones from the practical point of view. The examples mentioned concern the pair $(D^+_{n,n}, R^+_{n,n})$, where $D^+_{n,n}$ is the one sided maximal deviation between the two empirical distribution functions, while $R^+_{n,n}$ is the index of the sample element in the ordered union of two samples for which the maximum $D^+_{n,n}$ first occurs. The test based on $(D^+_{n,n}, R^+_{n,n})$ is compared with the test based on $D^+_{n,n}$ only for the equal sample sizes, $n = 10, 30, 50$. We thought that the power of the Kolmogorov-Smirnov test would be improved in this way for arbitrary sample sizes, that is, for the tests based on $(D^+_{n,m}, R^+_{n,m})$ and $D^+_{n,m}$, respectively. Surprisingly, Steck in his paper [15] showed that the situation is not so simple; it depends on the relationship between m and n. If they are relatively prime then there is precisely one value of $R^+_{n,m}$ which is associated with a given value of $D^+_{n,m}$, hence any test based on $(D^+_{n,m}, R^+_{n,m})$ is equivalent to the one based on $D^+_{n,m}$. The interesting and surprising consequence is the following: in a given case with $\sup_{(x)} [F(x) - G(x)] = 0.2$ using $(D^+_{50,50}, R^+_{50,50})$ the second kind error turns out to be 0.047, while by using only $D^+_{50,50}$ the corresponding value is 0.173, nearly four times as much. When we turn to samples $n = 50$ and $m = 51$ my method does not lead to any improvement of the two sample Smirnov test. What is the probability of the second kind error, how does it relate to the above two values, what is the asymptotic relation of the corresponding power functions? These are questions of interest.

I should like to mention that in our paper with Reimann [10], statistics of the following type were considered

$$nF_n(x) - mG_m(x), \tag{3.3}$$

the distribution of which can be determined easily, and does not show the same irregularity for m and n relatively prime as shown by Steck.

4. Some questions connected with the one sample problem

While an exact formula for the distribution of the Smirnov statistic

$$D^+_n = \sup_{(x)} [F_n(x) - F(x)] \tag{4.1}$$

was known very early (Smirnov 1944, independently Birnbaum and Tingey 1951), the distribution of the two sided statistic

$$D_n = \sup_{(x)} |F_n(x) - F(x)| \tag{4.2}$$

was obtained only within the last five years. Durbin [3] derived the generating function for the probabilities that the empirical distribution function lies between two parallel straight lines. He obtained recursion formulas and for a particular case, that is, for a certain integral value of a parameter, he has exact formulas. Epanechnikov [5] by means of a recurrence relation determined the exact probabilities. Independently Steck in his paper [14] gives, as he says, "a neat determinant for the probability that the order statistics for a sample of uniform random variables all lie in a multidimensional rectangle. An immediate application of this result gives the probability that the empirical distribution function lies between two other distribution functions." These authors have obtained a result for which a great deal of effort was expended in the last two decades.

Turning to the one sided case, the use of the ballot lemma (see Takács [18]) enables us to get very simple derivations of the distribution of D_n^+ and of distributions of a number of related statistics.

The following modified and extended form of the ballot lemma suggested by me leads almost immediately to the Smirnov-Birnbaum-Tingey theorem.

Let $A_0, A_1, A_2, \cdots, A_n$ be a complete system of events such that $P(A_0) = p$, $P(A_i) = q, i = 1, \cdots, n, p + nq = 1$. Denoting by ν_i the frequency of the event A_i in n trials, the following relation holds

$$P\left(\sum_{j=1}^{i} \nu_j < i, \qquad i = 1, \cdots, n\right) = p. \tag{4.3}$$

It is easy to see that an elementary proof can be obtained for this theorem from the following nice lemma due to and proved by Tusnády [2].

Let the points $P_1, P_2, \cdots, P_n$ be given on a directed circle of unit circumference and let us choose the positive number q such that $0 < nq < 1$. To an arbitrary point Q of the circle, construct the points $Q_1, Q_2, \cdots, Q_n$ consecutively in the positive direction with $\widehat{QQ_i} = iq$. Let the point Q be called a point of first category if the arc QQ_k contains less than k of points $P_1, P_2, \cdots, P_n$ for $k = 1, \cdots, n$. Then the measure of the set of points of first category is $1 - nq$.

PROOF (Tusnády). To each point P_i a chain

$$C_i = \{R_{i,1}, R_{1,2}, \cdots, R_{i,\nu(i)}\} \tag{4.4}$$

will be ordered in the following way: these points are consecutive but in the negative direction on the circle; further the arc $P_i R_{i,j}$ contains at least j of the points $P_1, P_2, \cdots, P_n$, including the point P_i for $j = 1, \cdots, \nu(i)$, while less than j for $j = \nu(i) + 1$.

It can easily be seen that if a chain covers the point P_j then it covers C_j as well.

A chain is called maximal if no other chain covers it. Two maximal chains are disjoint, hence the total length of all maximal chains is nq.

A point Q on the circle is now of the first category if and only if no chain covers it. Consequently the measure of all points of first category is $1 - nq$, as stated above.

Unfortunately a "two sided ballot lemma," which would be a tool for the derivation of the distribution of D_n, for example, does not yet exist.

Added in proof. See S. G. Mohanty, "Combinatorial methods in probability and statistics," lecture presented at the 58th Session of the Indian Science Congress Association, 1971.

A number of interesting articles have appeared in recent years which concern the power or asymptotic power of Kolmogorov-Smirnov tests. It is beyond our scope to mention these, but as a nice summarization of certain results in this field, I would mention the book of Hájek and Šidák [6].

5. Distribution of a Rényi type statistic due to Révész

Let the null hypothesis be that the density function is fully specified by $f(x)$, a function satisfying certain conditions (for example, $f'(x)$ exists and $|f'(x)|$ is bounded).

Let

$$x_0 < x_1 < x_2 < \cdots < x_{a_n} \tag{5.1}$$

be a division of the real axis or of the interval where $f(x) > 0$. The a_n, $n = 1, 2, \cdots$, are restricted by the relations

$$n^{1/3} \log n < a_n < n^{1-\varepsilon}, \qquad \varepsilon > 0. \tag{5.2}$$

Denoting by k_{i+1} the number of elements out of the sample $(X_1, X_2, \cdots, X_n)$ in the interval (x_i, x_{i+1}), $i = 1, \cdots, a_n - 1$, the empirical density function is defined by

$$f_n(x) = \frac{k_{i+1}}{n(x_{i+1} - x_i)}, \quad x_i \leqq x < x_{i+1}, i = 0, 1, \cdots, a_n - 1. \tag{5.3}$$

Révész [11] proved the following limiting relations which are *distribution free*

$$\lim_{n\to\infty} P\left[\left(\frac{n}{a_n}\right)^{1/2} \sup_{x_\alpha < x < x_\beta} \frac{f_n(x) - f(x)}{f(x)} < (2 \log a_n - \log\log a_n + y)^{1/2}\right] = \exp\{-\exp\{-y/2\}/2\pi^{1/2}\}, \tag{5.4}$$

$$\lim_{n\to\infty} P\left[\left(\frac{n}{a_n}\right)^{1/2} \sup_{x_\alpha < x < x_\beta} \left|\frac{f_n(x) - f(x)}{f(x)}\right| < (2 \log a_n - \log\log a_n + y)^{1/2}\right] = \exp\{-\exp\{-y/2\}/\pi^{1/2}\}. \tag{5.5}$$

The values $x_\alpha = x(\alpha, n)$ and $x_\beta = x(\beta, n)$ form an interval for which $f(x) \geqq (\log n)^{-1/3}$.

Further

$$\int_{x_\alpha}^{x_\beta} f(x)\, dx \to 1, \qquad \text{as} \quad n \to \infty. \tag{5.6}$$

6. On a two dimensional analogue of the Gnedenko-Korolyuk distribution

The difficulty encountered in constructing distribution free methods for samples taken on two or more variate random variables is well known. Recently Bickel [1] has given a distribution free version of the Smirnov two sample statistic in the p variate case. We shall consider the original Smirnov statistics in two dimensions

$$\begin{aligned} D^+_{n,n} &= \sup_{(x,y)} [F_n(x, y) - G_n(x, y)], \\ D_{n,n} &= \sup_{(x,y)} |F_n(x, y) - G_n(x, y)| \end{aligned} \tag{6.1}$$

in case of equal sample sizes. Our consideration leads to an immediate extension of the random walk model. As is known and will be illustrated below, the distribution of $D^+_{n,n}$ or $D_{n,n}$ does depend on the common theoretical continuous distribution function $G(x, y) \equiv F(x, y)$ of the two samples. Our aim is to propose a problem which will concern the independent case $F(x, y) = H_1(x)H_2(y)$ and which shows the difficulties even in this simple—distribution free—case.

Let (X_i, Y_i) and (X'_i, Y'_i), $i = 1, \cdots, n$, be two samples with

$$P(X_i < x, Y_i < y) = P(X'_i < x, Y'_i < y) = F(x, y). \tag{6.2}$$

Since $F(x, y)$ is continuous there is a "two dimensional ordering" of the two samples with probability one in the following way: let $\eta^*_1 < \eta^*_2 < \cdots < \eta^*_{2n}$ be the ordered union of the samples $(Y_1, Y_2, \cdots, Y_n)$ and $(Y'_1, Y'_2, \cdots, Y'_n)$ and let us denote by ξ_i the corresponding X or X' of η^*_i, that is, we have

$$(\xi_1, \eta^*_1), (\xi_2, \eta^*_2), \cdots, (\xi_{2n}, \eta^*_{2n}). \tag{6.3}$$

Taking now the ordered version of the ξ_i

$$\xi^*_1 < \xi^*_2 < \cdots < \xi^*_{2n}, \tag{6.4}$$

the following random variables will be introduced

$$\vartheta_{i,j} = \begin{cases} +1 & \text{if } \eta^*_i = Y_h \text{ and } \xi_i = \xi^*_j = X_h \\ -1 & \text{if } \eta^*_i = Y'_\ell \text{ and } \xi_i = \xi^*_j = X'_h \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

Now we have an arrangement of $+1$ and of -1, each n in number, in a $2n \times 2n$ table containing in each row and in each column exactly one element. This corresponds to and is analogous with the random walk used first by Gnedenko and Korolyuk and independently by Drion. Let us introduce the "partial" sums in the following way

$$s_{0,0} = 0, \quad s_{k,\ell} = \sum_{i \leqq k} \sum_{j \leqq \ell} \vartheta_{i,j}, 1 \leqq k \leqq 2n, 1 \leqq \ell \leqq 2n, \quad s_{2n,2n} = 0. \tag{6.6}$$

As can be seen very easily

$$D_{n,n}^{+} = \frac{1}{n} \max_{(k,\ell)} s_{k,\ell}, \qquad D_{n,n} = \frac{1}{n} \max_{(k,\ell)} |s_{k,\ell}|. \tag{6.7}$$

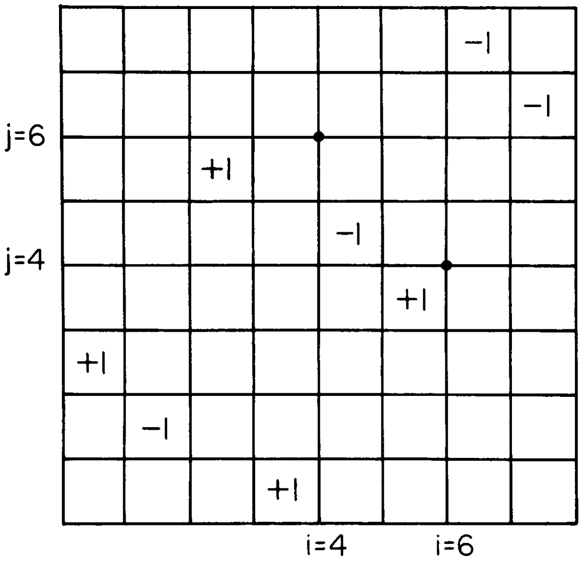

FIGURE 1
Array of $\vartheta_{i,j}$ for $n = 4$.

For example, in Figure 1 we have $n = 4$ and as can be justified very simply the following relations hold

$$D_{4,4}^{+} = \tfrac{1}{4} s_{4,6} = \tfrac{1}{4} s_{6,4} = \tfrac{1}{2}. \tag{6.8}$$

There are altogether $(2n)!$ possible arrangements within the square and each of them allows $\binom{2n}{n}$ possible allocations of the $+1$ and the -1, that is, the number of all possible configurations is $(2n)!\binom{2n}{n} = [(2n)!/n!]^2$.

Unfortunately the different arrays may have different probabilities depending on $F(x, y)$.

Consider the two extreme cases, $Y = X$ and $Y = -X$.

If $Y = X$, then $(\vartheta_{1,1}, \vartheta_{2,2}, \cdots, \vartheta_{2n,2n})$ are the nonzero terms in the $2n \times 2n$ square, where each of the $\binom{2n}{n}$ possible arrays of the $+1$ and the -1 is of equal probability. In this case the relations hold

$$P\left(D_{n,n}^{+} < \frac{k}{n}\right) = 1 - \frac{\binom{2n}{n-k}}{\binom{2n}{n}}, \qquad k = 0, 1, 2, \cdots, n, \tag{6.9}$$

and

$$(6.10) \qquad P\left(D_{n,n} < \frac{k}{n}\right) = \frac{1}{\binom{2n}{n}} \sum_{j=-\infty}^{\infty} - 1^j \binom{2n}{n-jk}, \qquad k = 1, \cdots, n,$$

that is, the Gnedenko-Korolyuk distributions are valid.

In the second case, when $Y = -X$, the nonzero terms are $(\vartheta_{1,2n}, \vartheta_{2,2n-1}, \cdots, \vartheta_{2n,1})$. We mention without proof that in this case the distribution of the Kuiper statistics determined by Maag and Stephens [7] is valid. In the two sided case we have

(6.11)

$$P\left(D_{n,n} < \frac{k}{n}\right) = 1 - \frac{2}{\binom{2n}{n}}\left[k \sum_{j=1}^{\infty} \binom{2n}{n-jk} - (k+1) \sum_{j=1}^{\infty} \binom{2n}{n-j(k+1)}\right],$$

$$k = 2, 3, \cdots, n.$$

This was derived for finding the distribution of the maximum deviation when two samples of the same size n are distributed uniformly on the circumference of a circle.

These two cases already show the dependence on the theoretical distribution function. Let us turn now to the independent case. The following problem is raised.

The distribution of the maximum deviation and of the absolute maximum deviation is to be determined when the two random variables are independent

$$(6.12) \qquad \begin{aligned} &P\left[D_{n,n}^{+} < \frac{k}{n} \middle| F(x, y) = H_1(x)H_2(y)\right] = ?\\ &P\left[D_{n,n} < \frac{k}{n} \middle| F(x, y) = H_1(x)H_2(y)\right] = ?\end{aligned}$$

In this case each array has the same probability, as given above, and the determination of the probabilities can be reduced to the enumeration of those paths, that is, arrays, for which the maximum is k/n or $\max_{(i,j)} s_{i,j} = k$. As the above example shows, the Markovian property does not hold. There does not exist a "first" maximum, instead simultaneously two places with $s_{i,j} = k$. In this way, however, a reflection can be made and a path with $\{s_{0,0} = 0, s_{2n,2n} = 0\}$ can be transformed into a path with $\{s_{0,0} = 0. \quad s_{2n,2n} = 2k\}$; however this is not a one to one mapping of the paths.

Concerning the reflection, let us consider the pair of indices (i^*, j^*) for which $s_{i^*,j^*} = k = \max s_{i,j}$ and change $\vartheta_{i,j}$ into $-\vartheta_{i,j}$, for which either $i > i^*$

or $j > j^*$ or both. Denote by α and β the number of $+1$ and -1, respectively, in the set $\{\vartheta_{i,j}, i \leqq i^*, j \leqq j^*\}$. Then $\alpha - \beta = k$, and the $+1$, of which there are $n - \alpha$, will be replaced by -1, and the -1, of which there are $n - \beta$, will be replaced by $+1$. In this way the number of $+1$ will amount to $\alpha + n - \beta$, while the number of -1 will be $\beta + n - \alpha$. Consequently

$$s_{2n,2n} = \alpha + n - \beta - (\beta + n - \alpha) = 2(\alpha - \beta) = 2k. \tag{6.13}$$

Now the number of paths with $s_{0,0} = 0$, $s_{2n,2n} = 2k$ is $(2n)!\binom{2n}{n-k}$ which must be smaller than the number of arrays with $D^+_{n,n} \geqq k/n$. Consequently the relation holds

$$P\left(D^+_{n,n} < \frac{k}{n}\right) \leqq 1 - \frac{\binom{2n}{n-k}}{\binom{2n}{n}}. \tag{6.14}$$

This relation follows also from a consideration of Nedoma (personal communication).

For $n = 2$ the number of possible arrays is 144, the distribution of $D^+_{4,4}$ is contained in Table II.

TABLE II

DISTRIBUTION OF $D^+_{4,4}$

k	$P\left(D_{n,n} < \frac{k}{n}\right)$	$1 - \frac{\binom{2n}{n-k}}{\binom{2n}{n}}$
1	$\frac{15}{144}$	$\frac{1}{3} = \frac{48}{144}$
2	$\frac{92}{144}$	$\frac{5}{6} = \frac{120}{144}$
3	1	1

In the case of $n = 3$ the number of possible arrangements is 14,400 for which a computer is needed.

REFERENCES

[1] P. J. BICKEL, "A distribution free version of the Smirnov two sample test in the p-variate case," *Ann. Math. Statist.*, Vol. 40 (1969), pp. 1–23.

[2] E. CSÁKI and G. TUSNÁDY, "On the number of intersections and the ballot theorem," *Studia Sci. Math. Hungar.*, to be published.

[3] J. DURBIN, "The probability that the sample distribution function lies between two parallel straight lines," *Ann. Math. Statist.*, Vol. 39 (1968), pp. 398–411.

[4] M. Dwass, "Simple random walk and rank order statistics," *Ann. Math. Statist.*, Vol. 38 (1967), pp. 1042–1053.

[5] V. A. Epanechnikov, "The significance level and power of the two-sided Kolmogorov test in the case of small samples," *Teor. Verojatnost. i Primenen.*, Vol. 13 (1968), pp. 725–730.

[6] J. Hájek and Z. Šidák, *Theory of Rank Tests*, New York, Academic Press, 1967.

[7] U. R. Maag and M. A. Stephens, "The V_{NM} two sample test," *Ann. Math. Statist.*, Vol. 39 (1968), pp. 923–935.

[8] S. G. Mohanty and C. I. Petros, "Joint distributions of several nonparametric statistics and tests based on them," unpublished.

[9] G. E. Noether, "Note on the Kolmogorov statistics in the discrete case," *Metrika*, Vol. 7 (1969), pp. 115–116.

[10] J. Reimann and I. Vincze, "On the comparison of two samples with slightly different sizes," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 5 (1960), pp. 293–309.

[11] P. Révész, "Testing of density functions," *Mathematica* (*Cluj*), 1971, in print.

[12] P. Schmid, "On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 1011–1027.

[13] K. Sen, "Paths of an odd number of steps with final position unspecified," *J. Indian Statist. Assoc.*, Vol. 7 (1969), pp. 107–135.

[14] G. P. Steck, "Rectangle probabilities for uniform order statistics and the probability that the empirical distribution lies between two distribution functions," *Ann. Math. Statist.*, Vol. 42 (1971), pp. 1-11.

[15] ———, "The Smirnov two sample tests as rank tests," *Ann. Math. Statist.*, Vol. 40 (1969), pp. 1449–1466.

[16] Š. Šujan, "On some problems concerning Kolmogorov-Smirnov test in the case of discrete theoretical distribution," thesis, Comenius University, Bratislava, 1970.

[17] V. Sujan, "The Kolmogorov-Smirnov statistics and the runs," unpublished.

[18] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*, New York, Wiley, 1967.

[19] I. Vincze, "Some questions connected with two sample tests of Smirnov type," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, Vol. 1 (1967), pp. 654–666.

[20] ———, "On Kolmogorov-Smirnov type distribution theorems," *Proceedings of the First International Symposium on Nonparametric Techniques*, New York, Cambridge University Press (1970), pp. 385–401.