ON TWO-SAMPLE TESTS BASED ON ORDER STATISTICS

I. VINCZE
MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES

1. Introduction

1.1. The present paper contains some remarks on the comparison of two samples in one and in two dimensions based on order statistics. The aim is to point out some possibilities for refinement of two-sample tests and to obtain statistics whose exact distribution can be calculated easily.

1.2. Tests based on order statistics have advantages and disadvantages over tests which utilize the specific form of the distributions. The advantages include (a) they are quick to use, in general, and (b) they do not presuppose the examination as to whether the samples agree with the assumed distributions. On the other hand these tests, in general, are (a) not so efficient and (b) biased when taking into account a wide range of alternatives. However in many cases of practical applications we have some restricted sets of types of alternatives on the one hand and we may refine the test on the other hand. These two circumstances provide a possibility of diminishing the distance between parametrical and nonparametrical procedures. This process has already been treated in the literature; the following considerations are some approaches from one side of the question.

We speak first of the possibility of refinement by the *use of a pair of statistics instead of one statistic.* The practical application of a pair of statistics seems to be easy enough, although to go over to the use of three statistics seems to be far too complicated.

1.3. In section 2 we consider the refinement property of the test based on a pair of statistics, which will be obvious qualitatively. For a more quantitative treatment, that is, for the problem of efficiency, we shall return later in some cases.

In section 3 we present some joint distributions which are related to the two-sample test of Smirnov and to the Galton test. We mention also a modification of the Smirnov statistic for the case of nearly equal sample sizes for which the distribution can be calculated easily. We then make some further remarks concerning the joint distribution and the limiting processes of the problems treated.

Finally, in section 4 we make two remarks on the two-dimensional case. In this section there is a limiting distribution theorem for the maximum of the sums of

independent random variables, for the case where the number of terms is also a random variable.

2. On best critical regions

2.1. We use the following notation. Let ξ and η be random variables with continuous distribution functions $F(x)$ and $G(x)$, respectively. Then let $\xi_1, \xi_2, \cdots, \xi_n$ and $\eta_1, \eta_2, \cdots, \eta_m$ be independent observations on ξ and η, respectively, that is, samples taken from populations with distribution functions $F(x)$ and $G(x)$. In the following we are interested only in the relative permutations of the ordered sample elements. Therefore we introduce the union of the two samples in order of magnitude,

$$\tau_1^* < \tau_2^* < \cdots < \tau_{n+m}^* \tag{1}$$

and define the random variables

$$\vartheta_i = \begin{cases} +1, & \tau_i^* = \xi_j, \\ -1, & \tau_i^* = \eta_e. \end{cases} \tag{2}$$

In this case the sample space reduces to the $\binom{n+m}{n}$ possible arrangements of the set $(\vartheta_1, \vartheta_2, \cdots, \vartheta_{n+m})$. Under the assumption $F(x) \equiv G(x)$, each order has the common probability $\binom{n+m}{n}^{-1}$.

We now consider the hypothesis $H_0: G(x) \equiv F(x)$, and an alternative H_1: $G(x) \equiv F_1(x) = \psi[F(x)]$, where $\psi(0) = 0$, $\psi(1) = 1$, and $\psi(y)$ is monotonically increasing in $0 \leqq y \leqq 1$, with $\psi(y_0) \neq y_0$ for some y_0. Then we may construct, with the aid of the method of Neyman and Pearson, a best test for deciding between H_0 and H_1. For this purpose we have to calculate the probability ratios

$$\frac{P(\vartheta_1 = \epsilon_1, \vartheta_2 = \epsilon_2, \cdots, \vartheta_{n+m} = \epsilon_{n+m} | H_1)}{P(\vartheta_1 = \epsilon_1, \vartheta_2 = \epsilon_2, \cdots, \vartheta_{n+m} = \epsilon_{n+m} | H_0)}, \tag{3}$$

where $\epsilon_i = +1$ occurs n times and $\epsilon_i = -1$ occurs m times. For the numerator we have the formula

$$\begin{aligned} n! \binom{m}{\alpha_1} \binom{m-\alpha_1}{\alpha_2} \cdots \binom{m-\alpha_1-\alpha_2-\cdots-\alpha_{r-1}}{\alpha_r} \cdots \binom{\alpha_{n+1}}{\alpha_{n+1}} \\ \int_0^1 \int_0^{y_n} \cdots \int_0^{y_3} \int_0^{y_2} [\psi(y_1)]^{\alpha_1} [\psi(y_2) - \psi(y_1)]^{\alpha_2} \cdots [\psi(y_r) - \psi(y_{r-1})]^{\alpha_r} \\ \cdots [1 - \psi(y_n)]^{\alpha_{n+1}} dy_1 dy_2 \cdots dy_n, \end{aligned} \tag{4}$$

if $\epsilon_{\alpha_1+1} = \cdots = \epsilon_{\alpha_1+\alpha_2+\cdots+\alpha_r+r} = \cdots = +1$ and thus the other ϵ_i are equal to -1. The denominator has the value $\binom{n+m}{n}^{-1}$.

Calculating these probability ratios, we may construct the best critical region, and using a suitable randomization we can reach a given exact level $1 - \beta$ where β is small.

Nevertheless statisticians do not choose this way for comparison of two samples. The first reason is that, even if such a definite alternative exists, the critical region is not, in general, simple enough for practical use. Also they do not choose this method because of numerical difficulties in the calculation of probabilities (3). In connection with the latter, we should like to remark: it is not often mentioned in the literature that a Monte Carlo method can be used, not only for the solution of numerical problems of analysis such as differential equations, but also to aid statisticians. In those cases we have a probabilistic model with certain probability formulas to be calculated. This numerical calculation can sometimes be carried out by using a Monte Carlo method. For example, if we have, in practice, a given kind of alternative which often occurs, then the computation of this a large number of times leads to the determination of the probability ratios (3) so that we can construct the best critical region.

2.2. The method used by statisticians is to choose functions of the sample elements with good properties, for example, with suitable statistical characteristics, with simple distributions, and so forth. Classical and often used statistics are those of Smirnov, the one- and two-sided maximum deviations,

$$
\begin{aligned}
D_{n,m}^{+} &= \max_{(x)}[F_n(x) - G_m(x)], \\
D_{n,m} &= \max_{(x)}|F_n(x) - G_m(x)|,
\end{aligned} \tag{5}
$$

where $F_n(x)$ and $G_m(x)$ denote the empirical distribution functions corresponding to the two samples. The test based on $D_{n,m}$ is asymptotically consistent against all continuous alternatives. Given an alternative, then, for the best critical region on the "one-dimensional" $D_{n,m}$ space, we obtain a linear point set which we may assume to be $D_1 < D_{n,m} < D_2$, say. If we now choose a second statistic $E_{n,m} = E_{n,m}(\vartheta_1, \vartheta_2, \cdots, \vartheta_{n+m})$, we have in the (D, E) plane a strip ($D_1 < D_{n,m} < D_2$; $-\infty < E_{n,m} < +\infty$). However, with the aid of the joint distribution function

$$
H(x,y) = P\{D_{n,m} < x, E_{n,m} < y\}, \tag{6}
$$

we obtain a best critical region in the two-dimensional (D, E) space which will differ, in general, from the strip. This means that we have a better test and an improvement on the test of Smirnov. There arises the question of whether this refinement has a significant effect or is only qualitative. The answer depends on the alternative and on the chosen statistic, but from heuristic reasoning we can conclude that in certain cases the use of a pair of statistics is more efficient than using only one statistic.

3. Joint distributions

3.1. *The case $m = n$; the first maximum point.*

3.1.1. Gnedenko and Korolyuk [6] determined the exact distribution of the Smirnov statistic for the case $m = n$. In this case it is known that

$$D_{n,n}^{+} = \frac{1}{n} \max_{(i)} S_i,$$

(7)

$$D_{n,n} = \frac{1}{n} \max_{(x)} |S_i|,$$

where $S_i = \vartheta_1 + \vartheta_2 + \cdots + \vartheta_i$, with $1 \leqq i \leqq 2n$.

Let us now denote by $R_{n,n}^{+}$ and $R_{n,n}$ the smallest value of i for which the maximum occurs, that is, for which $S_i = nD_{n,n}^{+}$ and $|S_i| = nD_{n,n}$ but $S_{i-j} < S_i$ and $|S_{i-j}| < |S_i|$, for $j = 1, 2, \cdots, i - 1$. In a previous paper [13] the author determined the joint distributions of the pairs of statistics $(D_{n,n}^{+}, R_{n,n}^{+})$ and $(D_{n,n}, R_{n,n})$.

For the construction of a general critical region, the knowledge of the conditional expected value of $R_{n,n}^{+}$ and $R_{n,n}$ for a given deviation may be useful. In the one-sided case the form is simple (see [14]).

$$E\left(\frac{1}{2n} R_{n,n}^{+} | D_{n,n}^{+} = \frac{k}{n}\right) = \begin{cases} \frac{1}{2}\left(1 + \frac{1}{n}\right), & k = 0, \\ \frac{1}{2k+1}\left(1 + \frac{1}{2n}\right), & k = 1, 2, \cdots, n. \end{cases} \tag{8}$$

3.1.2. We want to mention an interesting property of the distribution of $R_{n,n}^{+}$. This is that

$$P\{R_{n,n}^{+} = 2r - 1\} = P\{R_{n,n}^{+} = 2r\}, \qquad r = 1, 2, \cdots, n, \tag{9}$$

which is a slightly modified version of a theorem of E. S. Andersen (see for examples [4], p. 86).

K. Sarkadi [12] has given a very simple proof of this theorem with the aid of the random walk model and has shown that these probabilities are decreasing in r. This latter result is analogous to that of Birnbaum and Pyke [1] for the one-sample case.

3.1.3. Let us assume now that our variables are distributed uniformly in the interval (0, 1). In this case the limiting process $\xi(t)$ of the random process

$$\xi_n(t) = \sqrt{\frac{n}{2}} [F_n(t) - G_n(t)], \qquad 0 \leqq t \leqq 1, \tag{10}$$

under the null hypothesis $F(x) \equiv G(x)$ is Gaussian with expected value

$$E[\xi(t)] = 0, \qquad 0 \leqq t \leqq 1, \tag{11}$$

and covariance function

$$E[\xi(t)\xi(t')] = t(1 - t'), \qquad 0 \leqq t \leqq t' \leqq 1. \tag{12}$$

The marginal distributions of the limiting distributions of $(D_{n,n}^{+}, R_{n,n}^{+})$ and $(D_{n,n}, R_{n,n})$ as known are the Kolmogorov-Smirnov distributions for the first variables, and the uniform distribution for $(R_{n,n}^{+}/2n)$. For the two-sided case, so far as I know, the distribution function of the "absolute" maximum point is not

in the literature. If we denote by ρ the value of t where the Gaussian process attains the maximum of its absolute value, then

$$(13)\quad P\{\rho < z\} = \left(\frac{\pi}{2}\right)^{1/2} \int_0^z \frac{dv}{[v(1-v)]^{3/2}} \int_0^\infty u^2 f\left(\frac{u}{v^{1/2}}\right) f\left[\frac{u}{(1-v)^{1/2}}\right] du,$$

where

$$(14)\quad f(u) = \left(\frac{8}{\pi}\right)^{1/2} \sum_{i=0}^{\infty} (-1)^i (2i+1) \exp\left[-\frac{u^2}{2}(2i+1)^2\right].$$

The z-axis is a tangent of the distribution curve of infinite order and the curve is central symmetric relative to the point $(1/2, 1/2)$. We will come back to a further investigation of a series expansion of this function.

3.2. *Samples with slightly different sizes.* The exact calculation of the Smirnov probabilities is rather complicated. Exact formulas are given Blackman [2] for the case $m = kn$ only. A method for calculating these probabilities is due to Ozols [9] and the determination of the significance probabilities is given by Hodges [7]. He uses a random walk which is essentially the examination of the partial sums $S_i = \vartheta_1 + \vartheta_2 + \cdots + \vartheta_i$, for $i = 1, 2, \cdots, n + m$. In a paper with J. Reimann [10] we suggested the use of the statistics

$$(15)\quad \begin{aligned} B^+_{n,m} &= \max_{(x)} [nF_n(x) - mG_m(x)], \\ B_{n,m} &= \max_{(x)} \left| nF_n(x) - mG_m(x) + \frac{m-n}{2} \right| - \frac{m-n}{2}, \end{aligned}$$

where $m > n$. These statistics have the form

$$(16)\quad \begin{aligned} B^+_{n,m} &= \max_{1 \leqq i \leqq n+m} S_i \\ B_{n,m} &= \max_{1 \leqq i \leqq n+m} \left| S_i + \frac{m-n}{2} \right| - \frac{m-n}{2}. \end{aligned}$$

Exact formulas can be obtained easily for the probabilities analogous to that of the Gnedenko-Korolyuk probabilities. We proved that in the case when the sample sizes differ only slightly, that is, if $(m-n)/(m+n)^{1/2} \to 2c$ with $c \geqq 0$ as $n \to \infty$, then the test based on $B_{n,m}$ is asymptotically consistent against all continuous alternatives and $B^+_{n,m}$ is asymptotically consistent against all continuous alternatives $G(x) \equiv F_1(x) \geqq F(x)$ and $F_1(x_0) > F(x_0)$ for some x_0.

In the one-sided case it seems advisable to consider the last maximum point, that is, the largest value of i for which S_i takes its maximum. Denoting this value of i by $T^+_{n,m}$, we determined the joint distribution and limiting distribution of the pair of statistics $[B^+_{n,m}/(n+m)^{1/2},\ T^+_{n,m}/(n+m)]$. In the two-sided case we considered the first maximum only.

The limiting process of the random function

$$(17)\quad \xi_{n,m}(t) = \frac{nF_n(t) - mG_m(t)}{(n+m)^{1/2}}$$

is Gaussian with expectation $-2ct$ and covariance function $t - tt'(1 + 4c^2)$ with $0 \leqq t \leqq t' \leqq 1$.

3.3. *Refinement of the Galton Statistic.* In the case $m = n$, the Galton statistic, as is well known, is the number of $\xi_i^* - S$ smaller than the corresponding $\eta_i^* - S$ in the ordered samples

$$\begin{aligned} &\xi_1^* < \xi_2^* < \cdots < \xi_n^*, \\ &\eta_1^* < \eta_2^* < \cdots < \eta_n^*, \end{aligned} \tag{18}$$

Let us denote this statistic by γ, then γ can be interpreted by considering a particle walking randomly on the straight line, starting at the origin and returning after $2n$ steps to it. Then 2γ is the "time," that is, the number of steps spent by the particle above 0.

This statistic is not very efficient and in order to refine it we proceed in two possible directions.

3.3.1. In the above scheme let us translate the ξ_i^* by κ steps to the left

$$\begin{aligned} &\xi_1^*, \cdots, \xi_\kappa^*, \xi_{\kappa+1}^*, \cdots, \xi_n^*, \\ &\qquad \eta_1^*, \cdots, \eta_{n-\kappa}^*, \cdots, \eta_n^*. \end{aligned} \tag{19}$$

Let us denote by κ_0^+ the smallest κ for which each $\xi_{\kappa+i}^*$ is smaller than the η_i^* below it. Then the relation (see [3])

$$\kappa_0^+ = nD_{n,n}^+ = n \max_{(x)} [F_n(x) - G_n(x)] \tag{20}$$

is valid, that is, κ_0^+ has a Gnedenko-Korolyuk distribution. Thus we have a simple method for determining the maximum deviation.

Analogously if κ_0^- is defined by translating the ξ_i^* to the right, then

$$\kappa_0^- = -nD_{n,n}^- = -n \min_{(x)} [F_n(x) - G_n(x)] \tag{20'}$$

holds.

Let us now denote by κ_ν^+ the number of steps to the left required to have the number of the η_i^* exceeding the $\xi_{\kappa_\nu+i}^*$ be exactly ν, if this is possible. The distribution

$$P\{\nu = a | \kappa_\nu^+ = k\} \tag{21}$$

is given by Mihalevič [8]. Here $\nu = a$ is the time spent by the particle above the height $\kappa_\nu^+ = k$. If ν is small, then the corresponding κ_ν^+ can be treated as the "real" maximum and the corresponding pair of statistics $(\kappa_\nu^+, \kappa_{\nu'}^-)$ is the range containing most of the random walk. Here also is a possibility of refining the Smirnov statistics. However the joint distribution of the statistics is, so far as I know, not yet known.

The limiting distribution of (21) is not given in [8]. It is

$$\lim_{n\to\infty} P\left\{ z \leqq \frac{\nu}{2n} < z + dz \middle| \frac{\kappa_\nu^+}{\sqrt{n}} = y \right\} \sim \int_z^1 \frac{y}{[v(1-v)]^{3/2}} \exp\left[-y^2 \frac{v}{1-v} \right] dv\, dz \tag{22}$$

which shows that the measure of the t-set, for which the limiting process $\xi(t)$, with $0 \leqq t \leqq 1$, exceeds y, has infinite density at the point $z = 0$.

3.3.2. There is another improvement of the Galton statistic, by E. Csáki and I. Vincze [3]. Let us denote by $\lambda - 1$ the number of the i for which either $S_{i-1} = -1$, $S_i = 0$, $S_{i+1} = +1$, or $S_{i-1} = +1$, $S_i = 0$, $S_{i+1} = -1$ occurs. Then, considering the random walk, λ means the number of "waves" of the random path, that is, if for, and only for, $2\alpha_1, 2(\alpha_1 + \alpha_2), \cdots, 2(\alpha_1 + \alpha_2 + \cdots + \alpha_{\lambda-1}) < 2n$ the mentioned event occurs, then the first $2\alpha_1$ steps lead above the point 0, the next $2\alpha_2$ steps below it, or conversely, and so forth.

We have, for the distribution of λ,

$$P\{\lambda = l\} = \frac{2l}{n} \frac{\binom{2n}{n-l}}{\binom{2n}{n}}, \qquad l = 1, 2, \cdots, n. \tag{23}$$

This expression, multiplied by $\binom{2n}{n}/2$, agrees with the number of paths starting at the origin and arriving for the first time after $2n$ steps at the height $2l$. Indeed, one way of proving this formula is a one-to-one transformation of both kinds of paths. There is another method which leads also to the determination of the joint distribution. It is easy to see that this probability has the form

$$P\{\lambda = l\} = \frac{2}{\binom{2n}{n}} \sum_{\substack{\alpha_1 + \cdots + \alpha_l = n \\ \alpha_i \geqq 1}} \frac{1}{\alpha_1 + 1}\binom{2\alpha_1}{\alpha_1} \frac{1}{\alpha_2 + 1}\binom{2\alpha_2}{\alpha_2} \cdots \frac{1}{\alpha_l + 1}\binom{2\alpha_l}{\alpha_l}. \tag{24}$$

Let us denote by $L(v)$ the generating function

$$L(v) = \sum_{n=l}^{\infty} P\{\lambda_n = l\} \binom{2n}{n} v^n, \tag{25}$$

and take further

$$l(v) = \frac{1 - v - (1 - 4v)^{1/2}}{v} = \sum_{\alpha=1}^{\infty} \frac{1}{\alpha + 1}\binom{2\alpha}{\alpha} v^\alpha, \tag{26}$$

as is known. Now it follows that

$$L(v) = 2[l(v)]^l \tag{27}$$

which results, after a slight calculation, in the required mentioned probability.

We now treat the joint distribution of the Galton statistic γ and the number of waves λ,

$$P\{\gamma = g, \lambda = l\}. \tag{28}$$

For this probability we have a similar formula to that above but containing two parts, the first of which corresponds to the paths starting in the positive direction, in which case we have for the summation $\alpha_1 + \alpha_3 + \cdots = g$ and $\alpha_2 + \alpha_4 + \cdots = n - g$, and the second similarly for the paths starting in negative direction

$\alpha_1 + \alpha_3 + \cdots = n - g$ and $\alpha_2 + \alpha_4 + \cdots = g$. The calculation results in the expression

$$P\{\gamma = g, \lambda = l\} = \frac{l^2}{2g(n-g)} \frac{\binom{2g}{g - \frac{l}{2}}\binom{2n-2g}{n-g-\frac{l}{2}}}{\binom{2n}{n}}, \tag{29}$$

if l is even. If l is odd, we have

$$P\{\gamma = g, \lambda = l\} = \frac{l^2 - 1}{4g(n-g)} \frac{1}{\binom{2n}{n}} \left[\binom{2g}{g - \frac{l+1}{2}}\binom{2n-2g}{n-g-\frac{l-1}{2}} + \binom{2g}{g - \frac{l-1}{2}}\binom{2n-2g}{n-g-\frac{l+1}{2}} \right]. \tag{30}$$

Here, if $g = 0$ or n, then $l = 1$, while if $1 \leqq g \leqq n - 1$, then $l = 2, 3, \cdots$, and min $(2g + 1, 2n - 2g + 1)$.

This pair of statistics has the advantage that the test based on it is (at the same time) "two-sided."

The computations for the case of slightly different sizes may be carried out in the same way as for the case of the $B^{+}_{n,m}$ and $B_{n,m}$ statistics (see [10], [3]).

For the limiting distribution we have

$$\lim_{n\to\infty} P\left\{\frac{\gamma}{n} < z, \frac{\lambda}{\sqrt{n}} < y\right\} = \frac{1}{2\sqrt{\pi}} \int_0^z \int_0^y \frac{u^2}{[v(1-v)]^{3/2}} \exp\left[\frac{-u^2}{4v(1-v)}\right] du\, dv, \tag{31}$$

$$0 \leqq y, 0 \leqq z \leqq 1.$$

4. Comparison of two-dimensional samples

4.1. *Application of the Smirnov statistics.* Let us denote by $F(x, y)$ and $G(x, y)$ the common distribution functions of the independent vector variables (ξ_i, η_i) for $i = 1, 2, \cdots, n$ and (ξ_i', η_i') for $i = 1, 2, \cdots, m$, respectively.

In order to decide whether the null hypothesis H_0: $G(x, y) \equiv F(x, y)$ holds or not, the following simple procedure can be used: let us project the sample points on a straight line of the (x, y) plane, with its angle α with the x-axis chosen randomly from $(0, 2\pi)$. Let $z = 0$ be the projection of the point $(0, 0)$ and let z be the signed distance from it. Let us denote the projections of the sample points of the two samples in order of magnitude by $\tau_1^*, \tau_2^*, \cdots, \tau_n^*$ and $\tau_1'^*, \tau_2'^*, \cdots, \tau_m'^*$. Then the τ_i^* and the $\tau_i'^*$ are the ordered elements of two independent random samples. The corresponding marginal distributions are denoted by $F^{(\alpha)}(z)$ and $G^{(\alpha)}(z)$, respectively. Now the comparison of the two samples may be carried out with the aid of the Smirnov statistic

$$D_{n,m}^{(\alpha)} = \max_{(z)} |F_n^{(\alpha)}(z) - G_m^{(\alpha)}(z)|, \tag{32}$$

whose distribution under the null hypothesis is independent of α.

Under certain conditions (refer to Rényi [11] and Gilbert [5]), which require that a two-dimensional distribution be uniquely determined by its infinitely many marginal distributions, this test is asymptotically consistent against all continuous alternatives with probability 1 relative to the measure defined on the α-set. This means that if $G(x, y) \not\equiv F(x, y)$ only with probability 0 can we choose α in such a way that for the corresponding marginal distributions $F^{(\alpha)}(z) \equiv G^{(\alpha)}(z)$ holds.

Connected with this question we raised a problem (see [15]) which seems to be unsolved at present. It is to determine the distribution (or the limiting distribution) of the maximum deviation when we let α vary in the interval $(0, 2\pi)$. The answer will depend on the actual distribution function,

$$P\left\{\max_{(\alpha)} D_{n,m}^{(\alpha)} < z\right\} = \psi[z; F(x, y)]. \tag{33}$$

A. N. Kolmogorov has suggested the extension of this question to the problem of looking for the extreme functions,

$$\sup_{(F)} \inf_{(F)} \psi[z; F(x, y)]. \tag{34}$$

However, this test does not seem to be very efficient. We may repeat the procedure by choosing another angle α', but then the determination of the exact level of the test seems to be rather difficult.

4.2. *A theorem on two samples of equal sizes.* Using the notation given in section 4.1, let $m = n$ and $F(\infty, y) = H(y)$, then (see [16])

THEOREM 1. *If an $\eta = y$ is chosen at random according to the measure defined by $H(y)$ and if we denote by $P_{n,n}^{(k)}$ and $Q_{n,n}^{(k)}$ the probabilities of the events*

$$\left\{\max_{(x)} [F_n(x, y) - G_n(x, y)] < \frac{k}{n}\right\} \tag{35}$$

and

$$\left\{\max_{(x)} |F_n(x, y) - G_n(x, y)| < \frac{k}{n}\right\} \tag{36}$$

then we have

$$P_{n,n}^{(k)} = \frac{1}{(2n+1)\binom{2n}{n}} \sum_{i=0}^{n} \sum_{j=\max(0,i-k)}^{n} \binom{2n-i-j}{n-i} \left[\binom{i+j}{i} - \binom{i+j}{i-k}\right] \tag{37}$$

and

$$Q_{n,n}^{(k)} = \frac{1}{(2n+1)\binom{2n}{n}} \sum_{i=0}^{n} \sum_{j=\max(0,i-k)}^{\min(n,i+k)} \binom{2n-i-j}{n-i} \sum_{h=-\infty}^{\infty} (-1)^h \binom{i+j}{i+hk}. \tag{38}$$

Our theorem has a combinatorial formulation

THEOREM 2. *Consider a random sequence* $(\vartheta_1, \vartheta_2, \cdots, \vartheta_{2n})$ *consisting of repeated* $n+1$ *and* $n-1$, *where each order has the common probability* $\binom{2n}{n}^{-1}$. *Let us choose the value* λ *from the set* $(0, 1, 2, \cdots, 2n)$, *each value with the probability* $(2n+1)^{-1}$ *and the indices* $\alpha_1, \alpha_2, \cdots, \alpha_\lambda$ *from* $(1, 2, \cdots, 2n)$ *at random, where each set has the same probability. Then the relations*

$$P_{n,n}^{(k)} = P\left\{\max_{0 \leqq l \leqq \lambda} (\vartheta_{\alpha_1} + \vartheta_{\alpha_2} + \cdots + \vartheta_{\alpha_l}) < k\right\} \tag{39}$$

and

$$Q_{n,n}^{(k)} = \left\{\max_{0 \leqq l \leqq \lambda} |\vartheta_{\alpha_1} + \vartheta_{\alpha_2} + \cdots + \vartheta_{\alpha\lambda}| < k\right\} \tag{40}$$

hold for $k = 0, 1, 2, \cdots, n$.

For the limiting distributions if $n \to \infty$ *and* $k/\sqrt{2n} \to r$ *we have*

$$\lim_{n\to\infty} P_{n,n}^{(k)} = \int_0^1 \Phi\left(\frac{r}{[v(1-v)]^{1/2}}\right) dv - \exp[-2r^2] \int_0^1 \Phi\left(\frac{2v-1}{[v(1-v)]^{1/2}} r\right) dv, \tag{41}$$

$$\lim_{n\to\infty} Q_{n,n}^{(k)} = \sum_{j=-\infty}^{\infty} (-1)^j \exp[-2j^2r^2] \int_0^1 \left[\Phi\left(\frac{1-2j(1-v)}{[v(1-v)]^{1/2}} r\right) - \Phi\left(\frac{-1-2j(1-v)}{[v(1-v)]^{1/2}} r\right)\right] dv, \tag{42}$$

where $r > 0$ *and*

$$\Phi(r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} \exp\left[-\frac{u^2}{2}\right] du. \tag{43}$$

REFERENCES

[1] Z. W. BIRNBAUM and R. PYKE, "On some distributions related to the statistic D_n^+," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 179–187.

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

[3] E. CSÁKI and I. VINCZE, "On the Galton statistic," to appear in *Publ. Math. Inst. Hungar. Acad. Sci.*

[4] W. FELLER, *An Introduction to Probability Theory and its Applications*, New York, Wiley, 1952 (2nd ed.).

[5] W. M. GILBERT, "Projections of probability distributions," *Acta Math. Acad. Sci. Hungar.*, Vol. 6 (1955), pp. 195–198.

[6] B. V. GNEDENKO and V. S. KOROLYUK, "On the maximum discrepancy between two empirical distributions," *Dokl. Akad. Nauk SSSR*, Vol. 80 (1951), pp. 525–528.

[7] J. L. HODGES, JR., "The significance probability of the Smirnov two-sample test," *Ark. Mat.*, Vol. 3 (1958), pp. 469–486.

[8] V. S. MIHALEVIČ, "On the mutual disposition of two empirical distribution functions," *Dokl. Akad. Nauk SSSR*, Vol. 85 (1952), pp. 485–488.

[9] V. OZOLS, "On vectorands and the nonparametric test of goodness of fit for the two-sample case," *Izv. Akad. Nauk LSSR*, Vol. 8 (1956), pp. 153–158.

[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] A. RÉNYI, "On projections of probability distributions," *Acta. Math. Acad. Sci. Hungar.*, Vol. 3 (1952), pp. 131–142.

[12] K. SARKADI, "On Galton's rank order test," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 6 (1961), pp. 125–128.

[13] I. VINCZE, "Einige zweidimensionale Verteilungs und Grenzverteilungssatze in der Theorie der geordneten Stichproben," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 2 (1957), pp. 183–209.

[14] ———, "On some joint distributions and joint limiting distributions in the theory of order statistics, II," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 4 (1959), pp. 29–47.

[15] ———, "On some distributions and limiting distributions connected with two sample tests," *Futan Univ. J.* (Shanghai), Vol. 5 (1960), pp. 1–13. (In Chinese, with English summary.)

[16] ———, "On the deviation of two empirical distribution functions in two dimensions," *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl*, Vol. 10 (1960), pp. 361–372. (In Hungarian.)