ON THE COMPARISON OF TWO SAMPLES WITH SLIGHTLY DIFFERENT SIZES

by

J. REIMANN and I. VINCZE

Introduction

In our following considerations we suggest the possibility of a two sample test of SMIRNOV-type for comparison of samples with slightly different sizes.

If we denote by $F_n(x)$ and $G_m(x)$ the empirical distribution functions of the two samples taken independently from populations with continuous distribution functions $F(x)$ and $G(x)$ resp., then the test of SMIRNOV is based on the statistics

$$\max_{(x)} \left(F_n(x) - G_m(x)\right)$$

or

$$\max_{(x)} |F_n(x) - G_m(x)|.$$

The distributions of these statistics under the hypotheses $F(x) \equiv G(x)$ and in case $m = n$ are due to GNEDENKO and KOROLYUK and have simple forms. In other cases the formulae are more complicated or only asymptotical representations are at disposal. (See: J. BLACKMAN [1], V. OZOLS [6], J. L. HODGES [3], V. S. KOROLYUK [5].

In practice the case $m = n$ is of great importance. At the design of experiments the equal size of samples often can be ensured and the corresponding statistics can be evaluated without much calculations by means of the following very simple method of GNEDENKO and KOROLYUK [2]:

Let us denote by

$$\zeta_1^* < \zeta_2^* < \ldots < \zeta_{n+m}^*$$

the union of the mentioned two samples $\xi_1, \xi_2, \ldots, \xi_n$ and $\eta_1, \eta_2, \ldots, \eta_m$ resp., arranged in order of magnitude. Let now be

$$\vartheta_i = \begin{cases} +1, \text{ if } \zeta_i^* = \xi_j \\ -1, \text{ if } \zeta_i^* = \eta_l. \end{cases}$$

As it is easy to see in case $m = n$ the known relations

$$\max_{(x)} \left(F_n(x) - G_n(x)\right) = \frac{\max_{(i)} (\vartheta_1 + \ldots + \vartheta_i)}{n} = \frac{\max_{(i)} S_i}{n}$$

and

$$\max_{(x)} |F_n(x) - G_n(x)| = \frac{\max_{(x)} |\vartheta_1 + \ldots + \vartheta_i|}{n} = \frac{\max_{(i)} |S_i|}{n}$$

hold.

The idea lies at hand to make use of the statistics $\max_{(i)} S_i$ and $\max_{(i)} |S_i|$ in case of different sample sizes too the distribution of which can be obtained easily. The necessity of making use of these statistics for nearly equal sample sizes arises for instance if some of the experiments cannot be used and to obtain equal sample sizes elements of one of the samples have to be omitted. This sometimes would mean the loss of valuable information.[1]

In the following we shall determine the distributions and limiting distributions of these statistics, or more precisely the distributions of the following statistics: supposed that $m > n$

$$B^+_{n,m} = \max_{(x)} \left(n\, F_n(x) - m\, G_m(x)\right),$$

and

$$B_{n,m} = \max_{(x)} \left| n\, F_n(x) - m\, G_m(x) + \frac{m-n}{2} \right| - \frac{m-n}{2}.$$

We shall give the limiting distributions in the case when the sizes of the two samples only "slightly differ", i. e. if $n \to \infty$ and $\frac{(m-n)^2}{m+n} \to 4c^2$, where $c > 0$ is a constant. We shall prove furthermore that in this case *the test based on the statistics $B_{n,m}$ is asymptotically consistent against all continuous alternatives, and the statistics $B^+_{n,m}$ is asymptotically consistent against all continuous alternatives $F(x) > G(x)$.*

Thus this test can be suggested in cases mentioned above. (In the finite case in lack of nearer investigations $c < 1$ may be used.)

Let $R^+_{n,m}$ and $T^+_{n,m}$ resp. denote the first and last of the indices i for which the sum S_i is maximal. Let us further denote by $R_{n,m}$ the first index for which $\left| S_i + \frac{m-n}{2} \right| - \frac{m-n}{2}$ is maximal. We shall determine the joint distributions and limiting distributions of the pairs of statistics $(B^+_{n,m}, R^+_{n,m})$, $(B^+_{n,m}, T^+_{n,m})$ and $(B_{n,m}, R_{n,m})$. These pairs of statistics evidently enable more efficient tests, but their tabulations afford considerable efforts.

In the case $m = n$, i, e, $c = 0$. we obtain the distributions of GNEDENKO and KOROLJUK and the distributions contained in article [8] resp.

We wish to mention that our method is connected with that of J. L. HODGES [3] used for the determination of the significance probabilities of the SMIRNOV two sample test. As standard methods are used and our reflecting procedure is simpler in the following above article will not be mentioned.

Our § 1 and § 2 cover the mentioned distribution and limiting distribution theorems, § 3 is devoted to the asymptotic consistency, while in § 4 a remark is made concerning the limiting stochastic process.

[1] See HODGES [3] § 4. p. 477.

§ 1. Distribution theorems

With the notations of the introduction the following hold:

Theorem 1. *In the case* $F(x) \equiv G(x)$ *and for* $m > n$

$$\mathbf{P}(B^+_{n,m} = k) = \frac{2k+1+m-n}{m+k+1} \frac{\binom{m+n}{n-k}}{\binom{m+n}{n}}, \quad k = 0, 1, 2, \ldots n, \tag{1.1}$$

$$\mathbf{P}(B^+_{n,m} = k,\ R^+_{n,m} = r) = \begin{cases} 0, \text{ if } k < 0 \text{ or } r+k \text{ odd}, \\ \dfrac{m-n+1}{m+1}, \text{ if } k = 0,\ r = 0, \\ \dfrac{2k(m-n+k+1)}{r(2m-r+k+2)} \dfrac{\binom{r}{\frac{r+k}{2}}\binom{m+n-r}{n-\frac{r+k}{2}}}{\binom{m+n}{n}}, \\ \text{if } k = 1, 2, \ldots n;\ r = k, k+2, \ldots 2n-k, \end{cases} \tag{1.2}$$

$$\mathbf{P}(B^+_{n,m} = k,\ T^+_{n,m} = t) = \tag{1.3}$$

$$= \begin{cases} 0, \text{ if } k < 0 \text{ or } t+k \text{ odd}, \\ \dfrac{1}{t+1} \dfrac{m-n}{m+n-t} \dfrac{\binom{t}{\frac{t}{2}}\binom{m+n-t}{n-\frac{t}{2}}}{\binom{m+n}{n}}, \text{ if } k = 0, \qquad t = 0, 2, \ldots 2n, \\ \dfrac{2(k+1)(k+m-n)}{(t+k+2)(m+n-t)} \dfrac{\binom{t}{\frac{t+k}{2}}\binom{m+n-t}{n-\frac{t+k}{2}}}{\binom{m+n}{n}}, \text{ if } k = 1, 2, \ldots n, \\ t = k, k+2, \ldots, 2n-k. \end{cases}$$

Remarks. The proofs of formulae (1.1) and (1.2) are derived independently and thus by replacing $r = k + 2s$ the following combinatorical relation is obtained:

$$\sum_{s=0}^{n-k} \frac{k}{k+2s} \frac{m-n+k+1}{m-s+1} \binom{k+2s}{s} \binom{m+n-k-2s}{m-s} =$$

$$= \frac{m-n+2k+1}{m+k+1} \binom{m+n}{m+k},$$

which is valid for $m > n$ and $k = 1, 2, \ldots, n$. Analogous relation follow from (1.3).

Before turning to the two sided case, we wish to make some previous remarks. As the random function $nF_n(x) - mG_m(x)$ equals to 0 for $x = -\infty$ and equals to $n - m < 0$ for $x = +\infty$, the maximum of $|nF_n(x) - mG_m(x)|$ cannot be smaller than $|n - m|$. Consequently concerning the absolute deviation the following event may be of interest

$$-\frac{m-n}{2} - k < \min_{(x)}\left(nF_n(x) - mG_m(x)\right) + \frac{m-n}{2} \leqq$$

$$\leqq \max_{(x)}\left(nF_n(x) - mG_m(x)\right) + \frac{m-n}{2} \leqq \frac{m-n}{2} + k$$

or in other words the value k is the deviation from 0 in the positive direction and from $n - m$ in the negative direction. If now the absolute maximum of this deviation is denoted by $B_{n,m}$, i. e.

$$B_{n,m} = \max_{(x)}\left|nF_n(x) - mG_m(x) + \frac{m-n}{2}\right| - \frac{m-n}{2},$$

then the following theorem holds:

Theorem 2. If $F(x) \equiv G(x)$, $m > n$ *and with the notations* $s = 2k + m - n$, $p = m + n - r$

$$\mathbf{P}(B_{n,m} = k) = \frac{1}{\binom{m+n}{n}} \sum_{\gamma=-\infty}^{\infty} \left[\binom{m+n}{m+\gamma s} - \binom{m+n}{m+k+\gamma s}\right] =$$

$$= \frac{2^{m+n+1}}{s\binom{m+n}{n}} \sum_{\lambda=1}^{\infty} \cos^{m+n}\frac{\lambda\pi}{s} \sin\frac{k\lambda\pi}{s} \sin\frac{(s-k)\lambda\pi}{s}. \tag{1.4}$$

$$\mathbf{P}(B_{n,m} = k, R_{n,m} = r) = \frac{2}{\binom{m+n}{n}} \times$$

$$\times \left[\sum_{\gamma=-\infty}^{\infty} \frac{k+2\gamma s}{r}\binom{r}{\frac{1}{2}(r+s)+\gamma s} \sum_{\lambda=-\infty}^{\infty} \frac{s-k+1+2\lambda(s+2)}{p+s-k+2+\lambda(s+2)} \times\right.$$

$$\times \binom{p}{\frac{1}{2}(p+s-k)+\lambda(s+2)} + \sum_{\gamma=-\infty}^{\infty} \frac{s-k+2\gamma s}{r} \times \tag{1.5}$$

$$\left.\times \binom{r}{\frac{1}{2}(r+s-k)+\gamma s} \sum_{\lambda=-\infty}^{\infty} \frac{k+1+2\lambda(s+2)}{p+k+2+\lambda(s+2)} \binom{p}{\frac{1}{2}(p+k)+\lambda(s+2)}\right].$$

2. Let us turn now to the proof of our assertions. According to our introduction let $\zeta_1^* < \zeta_2^* < \ldots < \zeta_{n+m}^*$ be the union of the entirely independent elements of samples $\xi_1, \xi_2, \ldots, \xi_n$ and $\eta_1, \eta_2, \ldots, \eta_m$ arranged in order of magnitude. Of the definition it follows that the system $(\vartheta_1, \vartheta_2, \ldots, \vartheta_{n+m})$ consists of n $+1$-s and m -1-s. In consequence of the independency and the common distribution of the sample elements all of the $\binom{m+n}{n}$ possible arrangements of the $+1$-s and -1-s are of the same probability $\frac{1}{\binom{m+n}{n}}$.

Let us now consider the partial sum

$$S_i = \vartheta_1 + \vartheta_2 + \ldots + \vartheta_i, \qquad (S_0 = 0)$$

which gives the difference between the number of ξ_j-s and η_l-s smaller than ζ_i^* i. e. S_i is equal to $nF_n(\zeta_i^* + 0) - m\,G_m(\zeta_i^* + 0)$.

Therefore in proving relations (1.1), (1.2) and (1.3) of theorem 1, we have to determine the probabilities of the events

(2.1) $$\max_{1 \leqq i \leqq 2n} S_i = k, \qquad k = 0, 1, 2, \ldots n$$

(2.2′) $$S_i \leqq 0 \quad \text{for} \quad 1 \leqq i \leqq 2n \quad \text{in case} \quad k = 0, r = 0$$

(2.2″)
$$\begin{cases} S_i < k \quad \text{for} \quad 1 \leqq i \leqq r-1 \\ S_r = k \\ S_r \leqq k \quad \text{for} \quad r+1 \leqq i \leqq 2n-k \quad \text{in case} \quad k = 1, 2, \ldots n \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad r = k,\ k+2, \ldots 2n-k \end{cases}$$

(2.3′)
$$\begin{cases} S_i \leqq 0 \quad \text{for} \quad 1 \leqq i \leqq t-1 \\ S_t = 0 \\ S_i < 0 \quad \text{for} \quad t+1 \leqq i \leqq 2n \text{ in case } k = 0, t = 0, 2, 4 \ldots 2n \end{cases}$$

(2.3″)
$$\begin{cases} S_i \leqq k \quad \text{for} \quad 1 \leqq i \leqq t-1 \\ S_t = k \\ S_i < k \quad \text{for} \quad t+1 \leqq i \leqq 2n-k \quad \text{in case} \quad k = 1, 2, \ldots n, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = k, k+2, \ldots 2n-k. \end{cases}$$

In the same way the proof of relation (1.4) of theorem 2 requires the probability of the event

(2.4) $$\max_{0 \leqq i \leqq m+n} \left| S_i + \frac{m-n}{2} \right| - \frac{m-n}{2} = k,$$

while for relation (1.5) of theorem 2, the probability of the event

$$(2.5)\qquad \begin{cases} \left| S_i + \dfrac{m-n}{2} \right| - \dfrac{m-n}{2} < k \text{ for } 1 \leqq i \leqq r-1, \\ \left| S_r + \dfrac{m-n}{2} \right| - \dfrac{m-n}{2} = k, \\ \left| S_i + \dfrac{m-n}{2} \right| - \dfrac{m-n}{2} \leqq k \text{ for } r+1 \leqq i \leqq m+n-k \end{cases}$$

is needed.

In order to determine the probabilities of the events (2.1) — (2.5) we consider the following random walk on the points of the straight line: Let us start in the origin and arrive after $n + m$ steps to the point $-(m-n)$. According to our assumptions each of the possible $\binom{n+m}{n}$ paths have the same probability, so we have to determine the number of paths satisfying the restrictions given by relations (2.1) — (2.5). In determining the probabilities belonging to the events (2.1) — (2.3) we shall make use of the method applied in [8], in case (2.4) and (2.5) we shall refer to a lemma due to ELLIS.

3. Relation (2.1). The number of paths reaching the point $+k$ is counted. If we consider such a path and reflect it from the point reaching the height $+k$ for the first time about the point $+k$, then we obtain a path which startes from the origin and reaches after $n + m$ steps the height $2k + (m - n)$. The number of steps made in the positive direction is $m + k$, in the negative direction $n - k$, thus the number of all such pathes is equal to $\binom{n+m}{n-k}$. Therefore the number of paths not reaching the height $+k$ is $\binom{n+m}{n} - \binom{n+m}{n-k}$ which equals $\binom{n+m}{n-k} \mathbf{P}(\max_{(i)} S_i < k)$ and a substraction leads to relation (1.1).

4. The case (2.2′), i. e. $k = 0$; $r = 0$. In this case our assertion follows directly from the following known lemma (see e. g. [7] exercise 37 p. 74, solution p. 604): The probability of the event, that in a random sequence consisting of α -1-s and β $+1$-s, the number of $+1$-s never exceeds that of the -1-s (i. e. no partial sum exceeds 0) is equal to $\dfrac{\alpha - \beta + 1}{\alpha + 1}$.

5. In determining the probability of the event (2.2″) we shall proceed as in paper [8] for $m = n$ (loc. cit. § 3, p. 190—191). According to this the number of paths reaching the height k for the first time at the r-th step is

$$\frac{k}{r}\binom{r}{\frac{r+k}{2}} \qquad \text{for } k > 0.$$

For the further part of the path, i. e. for the succeeding $m + n - r$ steps, it is required that the height $+k$ must not be exceeded. According to the lemma in 4. with $\alpha = m - \frac{r-k}{2}$ and $\beta = n - \frac{r+k}{2}$ only the $\frac{2(m-n+k+1)}{2m-r+k+2}$-th portion of the possible $\binom{m+n-r}{n-\frac{r+k}{2}}$ paths satisfy this condition ($k = 1, 2, \ldots$). As each of the considered first r steps and following $m + n - r$ steps may be combined we obtain the last formula of (1.2).

In the case of $T^{+}_{n,m}$ of theorem 1 the same procedure may be carried through, but starting at the endpoint $(n - m)$ and arriving to the origin and considering in this case the *first* maximum place.

6. In derivation of the further probabilities we shall make use of the following general lemma due to ELLIS (see e. g. JORDAN [4] p. 404—408):

Let us consider the random walk on the integer points a of the interval $(0, s)$ of the straight line. Let us start at the point $a = i$ $(0 < i < s)$ and arrive after N steps to the point $a = j$ $(0 < j < s)$ passing neither the origin nor the point s. The number of such paths is given by

$$f(s, N, i, j) = \sum_{j=-\infty}^{\infty} \left[\binom{N}{\frac{1}{2}(N-i+j)+\gamma s} - \binom{N}{\frac{1}{2}(N+i+j)+\gamma s} \right] =$$

$$= \frac{2^{N+1}}{s} \sum_{\lambda=1}^{s-1} \cos^{N-1} \frac{\lambda\pi}{s} \sin \frac{i\lambda\pi}{s} \sin \frac{j\lambda\pi}{s}.$$

This expression is obtained by putting in the cited formulae $p = q = \frac{1}{2}$ and multiplying them by 2^N, i. e. by the number of all possible paths in the case investigated there and finally changing to our notations.

In our case the random walk takes place in the interval $(-(m-n)-k, k)$ starting at point 0 and arriving after $m + n$ steps to the point $-(m - n)$. For applying our above formulae the interval has to be translated by $(m - n) + k$. Thus the probability of the event (2.3) is obtained by replacing in above formulae $s = 2k + m - n$, $N = m + n$, $i = m + n - k$, $j = k$

$$\frac{f(2k+m-n, m+n, k+m-n, k)}{\binom{m+n}{n}},$$

which gives our formulae in (1.4).

Formula (1.5) is obtained in the following way: The number of paths starting from the origin and reaching the point $k - 1$ in $r - 1$ steps without passing the points $-(m - n) - k$ and $+k$ is

$$f(2k+m-n, r-1, k+m-n, 2k+m-n-1) = f(s, r-1, s-k, s-1)$$

using the notation $s = 2k + m - n$.

The next step of each path must lead from $k-1$ to k. Now the number of paths starting from k and reaching after $m+n-r$ steps the point $-(m-n)$ without having passed $-(m-n)-k-1$ or $k+1$ is

$$f(s+2, m+n-r, s+1, k+1).$$

Therefore the number of paths reaching the point k at the r-th step for the first time without previously having passed $-(m-n)-k$ and in the following reaching neither point $k+1$ nor $-k-(m-n)+1$ is the product of above two expressions.

We may determine in the same way the number of paths starting from the origin, arriving at the $(m+n)$-th step to the point $-(m-n)$ and reaching the lowest position $-(m-n)-k$ for the first time at the r-th step, without having passed the height k before, further without reaching the heights $k+1$ and $-k-(m-n)-1$. Then we obtain

$$f(s, r-1, s-k, 1) \cdot f(s+2, m+n-r, 1, k+1).$$

Expression (1.5) is obtained from above quantities after the following modifications:

$$f(s, r-1, s-k, s-1) =$$

$$= \sum_{\gamma=-\infty}^{\infty} \left[\binom{r-1}{\frac{1}{2}(r+k)-1+\gamma s} - \binom{r-1}{\frac{1}{2}(r-k)-1+(\gamma+1)s} \right] =$$

$$= \sum_{\gamma=-\infty}^{\infty} \left[\binom{r-1}{\frac{1}{2}(r+k)-1+\gamma s} - \binom{r-1}{\frac{1}{2}(r+k)-(\gamma+1)s} \right].$$

As we have only a finite nonvanishing number of terms we may replace in the second term $\gamma+1$ by $-\gamma$ thus the result is only the reverse order translated by 1 in the second terms. Finally each difference in the sum equals the corresponding term in (1.2). For $f(s, r-1, s-k, 1)$, $f(s+2, m+n-r, s+1, k+1)$ and $f(s+2, m+n-r, 1, k+1)$ analogous modifications lead to our results in (1.5).

§ 2. Limiting distribution theorems

7. Under the conditions $F(x) \equiv G(x)$ and if $\frac{m-n}{\sqrt{m+n}} \to 2c$ $(m \geqq n, c \geqq 0)$ the following limiting relations are valid (in each case let be $y \geqq 0$, $0 \leqq z \leqq 1$):

Theorem 3.

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}^{+}}{\sqrt{n+m}} < y\right) = 1 - e^{-2y^2-4cy}.$$

Theorem 4.

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B^+_{n,m}}{\sqrt{n+m}} < y, \frac{R^+_{n,m}}{n+m} < z\right) = \lim_{n\to\infty} \mathbf{P}\left(\frac{B^+_{n,m}}{\sqrt{n+m}} < y, \frac{T^+_{n,m}}{n+m} < z\right) =$$

$$= \sqrt{\frac{2}{\pi}} \int_0^y \int_0^z \frac{u(u+2c)}{[v(1-v)]^{3/2}} e^{-\frac{(u+2cv)^2}{2v(1-v)}} du\, dv .$$

Theorem 5.

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}} < y\right) = e^{2c^2} \sum_{i=-\infty}^{\infty} [e^{-2[2iy+(2i+1)c]^2} - e^{-2[(2i+1)(y+c)]^2}] =$$

$$= \sqrt{\frac{\pi}{8}} \frac{e^{2c^2}}{y+c} \left[\sum_{\lambda=1}^{\infty} e^{-\frac{\lambda^2\pi^2}{8(y+c)^2}} \left(\cos\frac{c\,\lambda\pi}{y+c} - (-1)^\lambda\right)\right].$$

Remark. In case $c = 0$ we obtain from theorem 5 the following forms of the Kolmogorov distribution:

$$\sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2y^2}$$

and

$$\frac{\sqrt{2\pi}}{2y} \sum_{\lambda=0}^{\infty} e^{-\frac{(2\lambda+1)^2\pi^2}{8y^2}}$$

resp.

Using the notations

$$f_c(y,z) = \frac{1}{z^{3/2}} \sum_{i=-\infty}^{\infty} [(1-4i)y - 4ic] e^{-\frac{[(4i-1)y+4ic]^2}{2z}},$$

$$\varphi_c(y,z) = \frac{1}{z^{3/2}} \sum_{j=-\infty}^{\infty} [(4j+1)y + (4j+2)c] e^{-\frac{[(4j+1)y+(4j+2)c]^2}{2z}}$$

the following theorem is valid:

Theorem 6.

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}} < y, \frac{R_{n,m}}{n+m} < z\right) =$$

$$= \sqrt{\frac{2}{\pi}} \int_0^y \int_0^z [f_c(u,v)\varphi_c(u,1-v) + f_c(u,1-v)\varphi_c(u,v)]\, du\, dv .$$

Remark. In the case $c = 0$ we obtain the following joint distribution theorem:

Corollary. *If* $m-n=o(\sqrt{m+n})$, *then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}}<y, \frac{R_{n,m}}{m+n}<z\right)=\sqrt{\frac{8}{\pi}}\int_0^y\int_0^z f(u,v)f(u,1-v)\,du\,dv,$$

where

$$f(y,z)=\frac{1}{z^{3/2}}\sum_0^\infty (-1)^i(2i+1)e^{-\frac{(2i+1)y^2}{2z}},$$

which contains as a special case for $m-n=0$ the result of [8] (see p. 188. theorem 4).

8. **Proofs.** As to the measure theoretical background of our limiting distribution theorems we refer to the proof in [8] (see loc. cit. § 4, p. 197.).

For derivation of the formulae of theorems 3—6, the following notations are introduced:

$$m+n=2N,$$

$$m-n=2l\sim 2c\sqrt{2N},$$

$$r\sim 2Nz,$$

$$k\sim y\sqrt{2N}$$

and from these follow

$$m+n-r\sim 2N(1-z),$$

$$m=N+O(\sqrt{N}),$$

$$n=N+O(\sqrt{N}),$$

$$s=2k+(m-n)\sim 2(y+c)\sqrt{2N},$$

$$dy\sim\frac{t}{\sqrt{2N}},\quad dz\sim\frac{1}{N}\,(dr=2(!)).$$

We shall make use of the following well-know relations:

$$\lim_{N\to\infty}\frac{\binom{2N}{N-l}}{\binom{2N}{N}}=e^{-2c^2}, \tag{8.1}$$

$$\binom{2N}{N}\sim\frac{2^{2N+\frac{1}{2}}}{\sqrt{2\pi N}}, \tag{8.2}$$

$$\lim_{N\to\infty}\left(\cos\frac{a}{2\sqrt{N}}\right)^N=e^{-\frac{a^2}{8}}. \tag{8.3}$$

a) **Proof of theorem 3.** The distribution function of the corresponding final case is (according) to formula (1.1) of theorem 1)

$$\mathbf{P}(B^+_{n,m} < k) = 1 - \frac{\binom{m+n}{m-k}}{\binom{m+n}{n}} = 1 - \frac{\binom{2N}{N-l-k}}{\binom{2N}{N}} \frac{\binom{2N}{N}}{\binom{2N}{N-l}}.$$

Applying (8.1) we obtain:

$$\mathbf{P}\left(\frac{B^+_{n,m}}{\sqrt{m+n}} < y\right) \to 1 - e^{-2y^2-4yc}.$$

b) In deriving theorem 4 we start from (1.2) of theorem 1 in case $k > 1$:

$$\mathbf{P}(B^+_{n,m} = k, R^+_{n,m} = r) = \frac{2k(m-n+k+1)}{r(2m-r+k+2)} \frac{\binom{r}{\frac{r+k}{2}}\binom{m+n-r}{n-\frac{r+k}{2}}}{\binom{m+n}{n}},$$

where $r = k,\ k+2,\ \ldots\ 2n-k$.

Using above relations we obtain

$$\frac{2k}{r}\,\frac{m-n+k+1}{2m-r+k+2} \sim \frac{y(y+2c)}{z(1-z)} \cdot \frac{1}{N}.$$

Of (8.1) and (8.2) it follows that

$$\frac{\binom{r}{\frac{r}{2}+\frac{k}{2}}}{\binom{r}{\frac{r}{2}}} \to e^{-\frac{y^2}{2z}} \qquad \text{if } r = 2Nz \to \infty.$$

$$\frac{\binom{m+n-r}{n-\frac{r+k}{2}}}{\binom{m+n-r}{\frac{1}{2}(m+n-r)}} \to e^{-\frac{(y+2c)^2}{2(1-z)}},$$

$$\frac{\binom{m+n}{\frac{1}{2}(m+n)}}{\binom{m+n}{n}} \to e^{2c^2},$$

$$\frac{\binom{r}{\frac{r}{2}}\binom{m+n-r}{\frac{1}{2}(m+n-r)}}{\binom{m+n}{\frac{1}{2}(m+n)}} \sim \sqrt{\frac{2}{\pi}}\,\frac{1}{[z(1-z)]^{1/2}}\,\frac{1}{\sqrt{2N}}.$$

By multiplication of above relations we obtain

$$\mathbf{P}(B^{+}_{n,m}=k, R^{+}_{n,m}=r) \sim \sqrt{\frac{2}{\pi}}\,\frac{y(y+2c)}{[z(1-z)]^{3/2}}\,e^{-\frac{(y+2c)^2}{z(1-z)}}\,\frac{1}{N}\cdot\frac{1}{\sqrt{2N}},$$

which gives the density function of the joint distribution function in theorem 4. The same procedure leads to result for $T^{+}_{n,m}$ as well.

c) We may obtain in the same way as in a) theorem 5 from the first formula of (1.4) in theorem 2.

In deriving the second form of the distribution function in theorem 5, we may make use of (8.3)

$$\left(\cos\frac{\lambda\pi}{s}\right)^{m+n} \to e^{-\frac{\lambda^2\pi^2}{s(y+c)^2}}, \qquad \text{where} \quad s \sim 2(y+c)\sqrt{m+n},$$

further

$$\sin\frac{k\,\lambda\pi}{s}\sin\frac{(k+m-n)\,\lambda\pi}{s} = \frac{1}{2}\left(\cos\frac{m-n}{s}\lambda\pi - (-1)^{\lambda}\right) \sim$$

$$\sim \frac{1}{2}\left[\cos\frac{c\,\lambda\pi}{y+c} - (-1)^{\lambda}\right]$$

and

$$\frac{2^{n+m+1}}{s\binom{m+n}{n}} \sim \sqrt{\frac{\pi}{2}}\,e^{2c^2}.$$

d) From formula (1.5) of theorem 2 we may obtain the density function of theorem 6 in the same way as for the one sided case in b).

§ 3. Proof of the asymptotic consistency

Let us suppose that instead of the null hypothesis H_0: $F(x) \equiv G(x)$ the alternative hypothesis $H_1 : G(x) = F_1(x) \not\equiv F(x)$ holds, where the distribution functions are continuous. We shall prove that in the case of a test based on the statistics $\dfrac{B_{n,m}}{\sqrt{m+n}}$ and on the level (the error of first kind) α, the probability of rejecting H_0 if it is not true tends to 1 in case $n, m \to \infty$, $(m-n)^2 \sim 4c^2(m+n)$.

Let us denote by $\Delta = \max_{(x)} | F(x) - F_1(x) |$ and let be x_0 a point for which $| F(x_0) - F_1(x_0) | = \Delta$; as mentioned before

$$\mathbf{P}\left(\frac{B_{n,m}}{\sqrt{m+n}} > d_\alpha \,|\, H_0\right) = \alpha$$

holds, where in case of $B_{n,m} > d_\alpha \sqrt{m+n}$ H_0 is rejected.

The probability of this event is evidently not less than that of the event that for the point x_0 the following relation holds

$$\left| n F_n(x_0) - m G_m(x_0) + \frac{m-n}{2} \right| - \frac{m-n}{2} > d_\alpha \sqrt{m+n}\,.$$

Now it will be shown that under the validity of H_1 the probability of the latter event tends to 1. Making use of the fact that $F_1(x_0) = F(x_0) \pm \Delta$ this event may be written in the following form:

$$\left| \frac{n}{m+n} \sqrt{m+n}\,(F_n(x_0) - F(x_0)) - \frac{m}{m+n} \sqrt{m+n}\,(G_m(x_0) - F_1(x_0)) + \right.$$

$$\left. + \frac{n-m}{\sqrt{m+n}} F(x_0) \pm \sqrt{m+n}\,\frac{n}{m+n}\,\Delta + \frac{m-n}{\sqrt{m+n}} \right| - \frac{m-n}{2\sqrt{m+n}} > d_\alpha\,.$$

The terms on the left are — except that of $\frac{n}{\sqrt{m+n}}\,\Delta$ which tends to infinity — bounded with probability near to 1. Hence the probability that this event will occur if H_1 is valid tends to 1.

In the same way the consistency of the $B^+_{n,m}$ statistics can be proved, under the alternative hypothesis.

§ 4. Remark on the limiting process

Let us suppose now that $F(x) \equiv G(x) \equiv x$ in $0 \leqq x \leqq 1$, i. e. let us consider the case of the uniform distribution in the interval (0, 1). For the stochastic process

$$\varphi_{n,m}(x) = \frac{n(F_n(x) - x) - m(G_m(x) - x)}{\sqrt{n+m}}$$

defined in the interval (0, 1)

$$\mathbf{M}(\varphi_{n,m}(x)) = 0\,,$$

$$\mathbf{M}[\varphi_{n,m}(x)\,\varphi_{n,m}(x')] = x(1-x') \qquad 0 \leqq x \leqq x' \leqq 1$$

hold for any x, x'. Hence the limiting process is a Gaussian one in (0, 1) with the same expected value 0 and covariance function as above. Our statistic $nF_n(x) - mG_m(x)/\sqrt{m+n}$ has the limiting expected value $-2cx$ and covariance function $x - xx'(1 + 4c^2)$.

(Received March 30, 1960.)

REFERENCES

[1] Blackman, J.: „Correction to »An extension of the Kolmogorov distribution«." *Annals of Mathematical Statistics* **29** (1958) 318—322.
[2] Гнеденко, Б. В. — Королюк В. С.: „О максимальном расхождении двух эмпирических распределений". *Доклады Академии Наук СССР* **80** (1951) 525—528.
[3] Hodges, J. L.: „The significance probability of the Smirnov two-sample test. *Arkiv för Matematik* **3** (1958) 469—486.
[4] Jordan, K.: *Fejezetek a klasszikus valószínűségszámításból.* Akadémiai Kiadó, Budapest, 1956.
[5] Королюк, В. С.: „Асимптотический анализ распределений максимальных уклонений в схеме Бернулли." *Теория вероятностей и ее применения* **4** (1959) 369—397.
[6] Озолс, В.: „О векторандах и непараметрическом критерии согласия для двух конечных выборок". *Известия АН СССР* **8** (1956) 150—158.
[7] Rényi A.: *Valószínűségszámítás.* Tankönyvkiadó, Budapest, 1955.
[8] Vincze, I.: „Einige zweidimensionale Verteilungs- und Grenzverteilungssätze in der Theorie der geordneten Stichproben." *MTA Matematikai Kutató Intézetének Közleményei* **2** (1957) 183—209.

ТЕОРЕМЫ О РАСПРЕДЕЛЕНИИ И ПРЕДЕЛЬНОМ РАСПРЕДЕЛЕНИИ, СВЯЗАННЫЕ С ДВУМЯ ВЫБОРКАМИ С НЕЗНАЧИТЕЛЬНО РАЗЛИЧНЫМ ЧИСЛОМ ЭЛЕМЕНТОВ

J. REIMANN и I. VINCZE

Резюме

Пусть $\xi_1, \xi_2, \ldots, \xi_n$ и $\eta_1, \eta_2, \ldots, \eta_m$ — выборки относительно случайных величин ξ и η с непрерывными функциями распределения $F(x)$ и $G(x)$, а $F_n(x)$ и $G_m(x)$ соответствующие эмпирические функции распределения.

Авторы определяют теоремы распределения и предельного распределения относительно следующих статистик:

$$B_{n,m}^{+} = \max_{(x)} [n F_n(x) - m G_m(x)]$$

$$B_{n,m} = \max_{(x)} \left| n F_n(x) - m G_m(x) + \frac{m-n}{2} \right| - \frac{m-n}{2} .$$

Относительно числа элементов они предполагают, что $m > n$ и в случае $n \to \infty$, $\frac{(m-n)^2}{m+n} \to 4c^2$, где $c > 0$ постоянная, т. е. числа элементов «незначительно» различны.

Пусть $R_{n,m}^{+}$ и $T_{n,m}^{+}$ означают нижнюю и верхнюю грань мест максимумов относительно статистики $B_{n,m}^{+}$, т. е. порядок первого и последнего элемента в соединенной последовательности элементов для которых имеет место максимум. Аналогичным образом пусть $R_{n,m}$ обозначает нижнюю грань мест максимумов относительно статистики $B_{n,m}$. Тогда при предположении $F(x) \equiv G(x)$ имеют место следующие теоремы о распределении и предельном распределении:

а) Теоремы о распределении:

Теорема 1.

$$(1.1)\qquad \mathbf{P}(B^{+}_{n,m}=k)=\frac{2k+1+m-n}{m+k+1}\cdot\frac{\binom{m+n}{n-k}}{\binom{m+n}{n}}\qquad (k=0,1,2,\ldots,m).$$

$$(1.2)\qquad \mathbf{P}(B^{+}_{n,m}=k,R^{+}_{n,m}=r)=\begin{cases}0 \text{ если } k<0 \text{ или } k+r \text{ нечетно},\\ \dfrac{m-n+1}{m+1}, \text{ если } k=0,\ r=0,\\ \dfrac{2k(m-n+k+1)}{r(2m-r+k+2)}\cdot\dfrac{\binom{r}{\frac{r+k}{2}}\binom{m+n-r}{n-\frac{r+k}{2}}}{\binom{m+n}{n}},\\ \text{если } k=1,2,\ldots,n;\ r=k,k+2,\ldots,2n-k.\end{cases}$$

$$(1.3)\qquad \mathbf{P}(B^{+}_{n,m}=k,T^{+}_{n,m}=t)=\begin{cases}\dfrac{1}{t+1}\ \dfrac{m-n}{m+n-t}\ \dfrac{\binom{t}{\frac{t}{2}}\binom{m+n-t}{n-\frac{t}{2}}}{\binom{m+n}{n}},\\ \text{если } k=0;\ t=0,2,\ldots,2n,\\ \dfrac{2(k+1)(k+m-n)}{(t+k+2)(m+n-t)}\cdot\dfrac{\binom{t}{\frac{t+k}{2}}\binom{m+n-t}{n-\frac{t+k}{2}}}{\binom{m+n}{n}},\\ \text{если } k=1,2,\ldots,n;\ r=k,k+2,\ldots,2n-k.\end{cases}$$

Теорема 2. *При обозначении* $s=2k+m-n$

$$(1.4)\qquad \mathbf{P}(B_{n,m}<k)=\frac{1}{\binom{m+n}{n}}\sum_{\gamma=-\infty}^{\infty}\left[\binom{m+n}{m+\gamma s}-\binom{m+n}{m+k+\gamma s}\right]=$$

$$=\frac{2^{m+n+1}}{s}\sum_{\lambda=1}^{\infty}\cos^{n+m}\frac{\lambda\pi}{s}\sin\frac{k\lambda\pi}{s}\sin\frac{(s-k)\lambda\pi}{s},$$

$$\mathbf{P}(B_{n,m}=k\,,\ R_{n,m}=r)=\frac{2}{\binom{m+n}{n}}\left[\sum_{\gamma=-\infty}^{\infty}\frac{k+2\gamma s}{r}\binom{r}{\frac{1}{2}(r+s)+\gamma s}\times\right.$$

(1.5)

$$\times\sum_{\lambda=-\infty}^{\infty}\frac{s-k+1+2\lambda(s+2)}{m+n-r+s-k+2+\lambda(s+2)}\binom{m+n-r}{\frac{1}{2}(m+n-r)+\frac{1}{2}(s-k)+\lambda(s+2)}+$$

$$+\sum_{\gamma=-\infty}^{\infty}\frac{s-k+2\gamma s}{r}\binom{r}{\frac{1}{2}(r+s-k)+\gamma s}\times$$

$$\left.\times\sum_{\lambda=-\infty}^{\infty}\frac{k+1+2\lambda(s+2)}{m+n-r+k+2+\lambda(s+2)}\binom{m+n-r}{\frac{1}{2}(m+n-r)+\frac{1}{2}k+\lambda(s+2)}\right].$$

Замечание: в случае $m=n$ из (1.1) и (1.4) получаются распределения Гнеденко—Королюк-а.

b) Теоремы о предельном распределении:

Если $F(x)\equiv G(x)$ и $\dfrac{m-n}{\sqrt{n+m}}\to 2c\quad(m\geqq n, c\geqq 0)$,

то в случае $y\geqq 0$, $1\geqq z\geqq 0$

Теорема 3. $\lim\limits_{n\to\infty}\mathbf{P}\left(\dfrac{B^{+}_{n,m}}{\sqrt{n+m}}<y\right)=1-e^{-2y^2-4cy}$.

Теорема 4.

$$\lim_{n\to\infty}\mathbf{P}\left(\frac{B^{+}_{n,m}}{\sqrt{n+m}}<y,\frac{R^{+}_{n,m}}{n+m}<z\right)=\lim_{n\to\infty}P\left(\frac{B^{+}_{n,m}}{\sqrt{n+m}}<y,\frac{T^{+}_{n,m}}{n+m}<z\right)=$$

$$=\sqrt{\frac{2}{\pi}}\int_0^y\int_0^z\frac{u(u+2c)}{[v(1-v)]^{3/2}}e^{-\frac{(u+2cv)^2}{2v(1-v)}}\,du\,dv.$$

Теорема 5.

$$\lim_{n\to\infty}\mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}}<y\right)=e^{2c^2}\sum_{i=-\infty}^{\infty}\left[e^{-2[2iy+(2i+1)c]^2}-e^{-2[(2i+1)(y+c)]^2}\right]=$$

$$=\sqrt{\frac{\pi}{8}}\,\frac{e^{2c^2}}{y+c}\left[\sum_{\lambda=1}^{\infty}e^{-\frac{\lambda^2\pi^2}{8(y+c)^2}}\left(\cos\frac{c\,\lambda\pi}{y+c}-(-1)^{\lambda}\right)\right].$$

Замечание: в случае $c = 0$ из теоремы 5 получается следующая форма распределения Колмогоров-а:

$$\sum_{i=0}^{\infty} (-1)^i e^{-2i^2y^2} = \frac{\sqrt{2\pi}}{2y} \sum_{\lambda=0}^{\infty} e^{-\frac{(2\lambda+1)^2\pi^2}{8y^2}}.$$

Теорема 6. *При обозначениях*

$$f_c(y,z) = \frac{1}{z^{3/2}} \sum_{i=-\infty}^{\infty} [(1-4i)y - 4ic]\, e^{-\frac{[(4i-1)y+4ic]^2}{2z}}$$

$$\varphi_c(y,z) = \frac{1}{z^{3/2}} \sum_{j=-\infty}^{\infty} [(4j+1)y + (4j+2)c]\, e^{-\frac{[(4j+1)y+(4j+2)c]^2}{2z}}$$

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}} < y, \frac{R_{n,m}}{n+m} < z\right) =$$

$$= \sqrt{\frac{2}{\pi}} \int_0^y \int_0^z [f_c(u,v)\varphi_c(u,1-v) + f_c(u,1-v)\varphi_c(u,v)]\, du\, dv.$$

Замечание: в случае $c = 0$, когда $m - n = O(\sqrt{m+n})$,

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n,m}}{\sqrt{n+m}} < y, \frac{R_{n,m}}{m+n} < z\right) = \sqrt{\frac{8}{\pi}} \int_0^y \int_0^z f(u,v) f(u,1-v)\, du\, dv,$$

где

$$f(y,z) = \frac{1}{z^{3/2}} \sum_{i=0}^{\infty} (-1)^i (2i+1) e^{-\frac{(2i+1)^2y^2}{2z}}.$$

Эта формула в качестве специального случая содержит теорему 4 на стр. 188 работы [8].

Авторы доказывают, что критерий, основывающий на статистике $B_{n,m}$, асимптотично состоятелен относительно всякой непрерывной альтернативной гипотезы, а критерий основываюшийся на статистике $B^+_{n,m}$, асимптотично состоятелен относительно непрерывной альтернативной гипотезы.

Авторы замечают, что предельный процесс стохастического процесса

$$\psi_{n,m}(x) = \frac{n F_n(x) - m G_m(x)}{\sqrt{n+m}}$$

является Гауссовым с математическим ожиданием $-2cx$ и корреляционной функцией $x - x x'$ $(1 + 4c^2)$ $(0 \leqq x \leqq x' \leqq 1)$.