ON THE NONPARAMETRIC TEST FOR TWO SAMPLE PROBLEM BASED ON SPACINGS

I. BAIRAMOV and N. ÖZKAYA

Ankara University, Faculty of Science, Department of Statistics
06100 Tandoğan, Ankara
Turkey

Abstract

A new nonparametric test for two sample problem based on order statistics is suggested. This test can be successful in the case of small samples. The probabilities of the statistic are given. Comparisons with Kolmogorov-Smirnov, Mann Whitney-Wilcoxon, and Wilks' empty block test for small sample sizes are made by the simulation.

Key words: Order Statistics, Kolmogorov -Smirnov test, Sample bloks, covarage.

1. Introduction

Let $\Im_c$ be the class of all continuous distribution functions (d.f.'s), X_1, $X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ be random samples obtained from populations with d.f. F and Q respectively, $F, Q \in \Im_c$. Denote the probability measures as P_1 and P_2 according to F and Q respectively. Let $\wp$ be the class of probability measures according to $\Im_c$.

One of the main problem of nonparametric statistics is to test the null hypothesis $H_0 : F(x) = Q(x)$ against a composite alternative $H_1 : F(x) \neq Q(x)$; $\quad F, Q \in \Im_c$.

This also can be formulated as the following:

$$H_0 : (F, Q) \in \Im_0, \ H_1 : (F, Q) \in \Im^* \setminus \Im_0,$$

where

$$\Im^* = \{(F,Q) : (F,Q) \in \Im_c\}, \Im_0 = \{(F,Q) : (F,Q) \in \Im^*; F = Q\}.$$

The test function for testing H_0 against H_1 is determined as follows,

$$\Psi(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) = \begin{cases} 1, & \text{if } (X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in \Omega \\ 0, & \text{otherwise} \end{cases}$$

where $\Omega \in R^{n+m}$ is the critical set.

The function

$$\beta_\Psi(P_1, P_2) = E_{P_1 \times P_2} \Psi(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m)$$
$$= P_1 \times P_2 ((X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in \Omega), P_1, P_2 \in \wp$$

is the power function of the test, where $P_1 \times P_2$ denotes the product of probability measures P_1 and P_2; so that $P_1 \times P_2 ((X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in \Omega)$ is the probability of the event $((X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in \Omega)$, while X has distribution P_1 and Y has distribution P_2.

According to Neyman-Pearson theory the desirable property of the test Ψ is

$$P_1 \times P_1 \{(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in \Omega\} \leq \alpha, \text{ for all } P_1 \in \wp,$$

for a given level of significance α.

If

$$\lim_{n,m \to \infty} \beta_\Psi(P_1, P_2) = 1, \text{ for all } P_1 \neq P_2,$$

then the test Ψ is called consistent for checking H_0 against H_1.

If

$$E_{P_1 \times P_1} \Psi(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) = \alpha, \text{ for all } P_1 \in \wp, \qquad (1)$$

then the test Ψ is called nonparametric. If (1) holds asymptotically for large values of m and n then Ψ is called asymptotically nonparametric test.

Tests of this type have been devised by Dixon (1940), Mathisen (1943), Smirnov (1939), Wald and Wolfowitz (1940), Wilks (1961), Borovkov (1975). Most of other tests are somewhat the modification of Wilks' empty block test

or Kolmogorov-Smirnov's test. For more details one can see e.g. Borovkov (1984).

Let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(m)}$ be the order statistics constructed from $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$, respectively. Kolmogorov-Smirnov test statistic is constructed by the distance $\sup_x |F_n(x) - G_m(x)|$, where $F_n(x)$ and $G_m(x)$ are empirical distribution functions of the samples $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$, respectively. Consider the sample blocks $\Delta_1 = \left(-\infty, X_{(1)}\right]$, $\Delta_2 = \left(X_{(1)}, X_{(2)}\right], ...,$ $\Delta_n = \left(X_{(n-1)}, X_{(n)}\right]$, $\Delta_{n+1} = \left(X_{(n)}, \infty\right)$. Let $s_r = s_r(n, m)$ be the number of blocks containing exactly r element of another sample $Y_1, Y_2, ..., Y_m$. A class of tests can be constructed using test statistic based on linear combination $\sum_{r=0}^{k} c_r s_r$, where $c_0, c_1, ..., c_k$ are given weight constants. When $k = 0$ then we have the Wilks' empty block test. Wilks' test is based on the random variables $r_1, r_2, ..., r_{n+1}$, where r_i shows the number of the observations of the sample $Y_1, Y_2, ..., Y_m$ falling into the random interval $\left(X_{(i-1)}, X_{(i)}\right]$, ($X_{(0)} = -\infty$, $X_{(n+1)} = \infty$). Let $u_1 = F(X_{(1)})$, $u_2 = F(X_{(2)}) - F(X_{(1)}), ...,$ $u_n = F(X_{(n)}) - F(X_{(n-1)})$, $u_{n+1} = 1 - F(X_{(n)})$ be the coverages of blocks. The major fact used in all classical tests based on blocks is that the conditional probability for random vector $(r_1, r_2, ..., r_{n+1})$ to have a specific value $(r'_1, r'_2, ..., r'_{n+1})$, given $(u_1, u_2, ..., u_n)$ is

$$\frac{m!}{r'_1! r'_2! ..., r'_{n+1}!} u_1^{r'_1} u_2^{r'_2} ... u_n^{r'_n} (1 - u_1 - u_2 - ... - u_n)^{r'_{n+1}}.$$

In this work we construct a nonparametric test for testing H_0 against H_1, based on the properties of order statistics to be an invariant confidence intervals containing the main distributed mass, i.e. if $X_1, X_2, ..., X_n, X_{n+1}$ is a random sample with continuous distribution function then

$$P\left\{X_{n+1} \in (X_{(i)}, X_{(j)})\right\} = \frac{j - i}{n + 1} \tag{2}$$

(see Madreimov, Petunin, 1982). The proposed test can be used successfully when sample sizes are small.

2. The test based on the order statistics

Let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ be the order statistics constructed from $X_1, X_2, ..., X_n$. Denote

$$\Delta_1 = \left(-\infty, X_{(1)}\right], \Delta_2 = \left(X_{(1)}, X_{(2)}\right], ..., \Delta_n = \left(X_{(n-1)}, X_{(n)}\right], \Delta_{n+1} = \left(X_{(n)}, \infty\right).$$

Define the following random variables (r.v.'s).

$$\xi_i^k = \begin{cases} 1, & Y_k \in \Delta_i \\ 0, & Y_k \notin \Delta_i \end{cases} \quad k = 1, 2, ...m, \ i = 1, 2, ..., n+1$$

Let

$$\zeta_i = \sum_{k=1}^{m} \xi_i^k.$$

It is clear that ζ_i is the number of observations $(Y_1, Y_2, ..., Y_m)$ falling into the interval $\Delta_{i.}$ When the hypothesis H_0 is true then

$$P\left\{Y_i \in \left(X_{(k)}, X_{(l)}\right)\right\} = \frac{l-k}{n+1}, \quad 1 \leq k < l \ \leq n, \quad i = 1, 2, ..., m$$

and for $l = k+1$ it becomes

$$p_k \equiv P\left\{Y_i \in \Delta_k\right\} = \frac{1}{n+1}, \ \ i = 1, 2, ..., m$$

Bairamov and Petunin (1991) introduced the notion of invariant confidence interval containing the main distributed mass of general population. Let $X_{1,}X_2, ..., X_n$ be independent r.v.'s with common d.f. F which belongs to a class of d.f.'s $\Im$. Suppose $f_1(u_{1,}u_2, ..., u_n)$ and $f_2(u_{1,}u_2, ..., u_n)$ are two Borel functions with the following property:

$$f_1(u_{1,}u_2, ..., u_n) \leq f_2(u_{1,}u_2, ..., u_n) \qquad (u_{1,}u_2, ..., u_n) \ \in R^n.$$

Let X_{n+1} be a new sample point which is independent of $X_{1,}X_2, ..., X_n$ and has the same d.f. F. If

$$P\left\{X_{n+1} \in (f_1(X_{1,}X_2, ..., X_n), f_2(X_{1,}X_2, ..., X_n))\right\} = \alpha \ \text{ for all } \quad F \in \Im,$$

then $(f_1(X_1,X_2,...,X_n), f_2(X_1,X_2,...,X_n))$ is called an invariant confidence interval containing the main distributed mass for the class of distributions $\Im$ with confidence level α .

It is clear from (2) that the random interval $(X_{(k)}, X_{(l)}), (1 \le k < l \le n)$ is an invariant confidence interval for class of all continuous distribution functions. If f_1 and f_2 are continuous, symmetric and different on every set with a nonzero lebesgue measure functions of n arguments, only $f_1(X_1,X_2,...,X_n) = X_{(i)}$, $f_2(X_1,X_2,...,X_n) = X_{(j)}$, $1 \le i < j \le n$ form an invariant confidence interval for $\Im_c$ (see Bairamov, Petunin ,1991).

Consider the following statistic

$$\eta = \sum_{k=1}^{n+1} \left(\frac{\zeta_k}{m} - p_k \right)^2$$

which can be interpreted as the sum of squared errors between frequencies and the probabilities of the events $Y_k \in (X_{(i-1)}, X_{(i)})$ $i = 1, 2, ..., n+1$.

The following lemma will be useful for further discussions.

Lemma 1. For $0 \le i_k \le m$, $(k = 1, 2, ..., n+1)$, $i_1 + i_2 + ... + i_{n+1} = m$ it is true that

$$P\{\zeta_1 = i_1, \zeta_2 = i_2, ..., \zeta_{n+1} = i_{n+1}\} =$$

$$= \frac{n!m!}{i_1!i_2!...i_{n+1}!} \int\limits_{x_1<x_2<...<x_n} \cdots \int Q^{i_1}(x_1) \left[Q(x_2) - Q(x_1)\right]^{i_2} ... \left[Q(x_n) - Q(x_{n-1})\right]^{i_n} \times$$

$$\times \left[1 - Q(x_n)\right]^{i_{n+1}} dF(x_1)dF(x_2)...dF(x_n).$$

Corollary 1. Assume that H_0 is true. Then for $0 \le i_k \le m$, $(k = 1, 2, ..., n+1)$, $i_1 + i_2 + ... + i_{n+1} = m$ it is true that

$$P\{\zeta_1 = i_1, \zeta_2 = i_2, ..., \zeta_{n+1} = i_{n+1}\} = \frac{1}{\binom{n+m}{n}}.$$

Note that Corollary 1 coincides with assertion 14.3.1 of Wilks (1962, p. 442).

Theorem 1. Assume that H_0 is true. Let $n+1 \geq m, l_1 \equiv \min\limits_{(i_1,i_2,...i_n)\in\wp_{1,2,...,n}}(i_1^2+ i_2^2+...+i_{n+1}^2) = m$, $l_2 \equiv \max\limits_{(i_1,i_2,...i_n)\in\wp_{1,2,...,n}}(i_1^2+i_2^2+...+i_{n+1}^2) = m^2$, $0 \leq i_k \leq m$, $k = 1,2,...,n+1$, where $\wp_{1,2,...,n}$ denotes the class of all permutations of $1,2,...,n$. Then the distribution of statistic η has the following form

$$P\{\eta \leq x\} = P\left\{\sum_{k=1}^{n+1}\left(\frac{\zeta_k}{m} - p_k\right)^2 \leq x\right\} =$$

$$= \begin{cases} 0, & x \leq \frac{l_1}{m^2} - a \\ \sum\limits_{j=l_1}^{[m^2(x+a)]} \frac{1}{C_{n+m}^n} P_{m,n,j}\ , & x \in \left[\frac{l_1}{m^2} - a, \frac{l_2}{m^2} - a\right] \\ 1, & x \geq \frac{l_2}{m^2} - a \end{cases}$$

where $P_{m,n,j}$ is the number of sequences $\{i_1, i_2, ..., i_{n+1}\}$ satisfying $i_1 + i_2 + ... + i_{n+1} = m$ and $i_1^2 + i_2^2 + ... + i_{n+1}^2 = j$, where $[x]$ is a minimal integer less than x, $p_k \equiv a = \frac{1}{n+1}$.

Proof. From (2) one can write

$$P\left\{\xi_i^k = 1\right\} = P\left\{X \in \left(X_{(i-1)}, X_{(i)}\right)\right\} = \frac{1}{n+1} \equiv a,\ k = 1,2,...,m\ ,\ i = 1,2,...,n+1$$

$X_{(0)} = -\infty$, $X_{(n+1)} = \infty$. One has

$$P\left(\sum_{k=1}^{n+1}\zeta_k^2 = j\right) = \sum_{\substack{i_1+i_2+...+i_{n+1}=m \\ i_1^2+i_2^2+...+i_{n+1}^2=j}} 1 = \frac{1}{C_{n+m}^n} P_{m,n,j}\ ,\ j = m, m+1, ..., m^2. \tag{3}$$

It is clear that $P_{m,n,j} = 0$ for a values of $j \in [m, m+1, m+2, ..., m^2]$ for which does not exists a sequence $\{i_1, i_2, ..., i_{n+1}\}$ satisfying $i_1+i_2+...+i_{n+1} = m$ and $i_1^2 + i_2^2 + ... + i_{n+1}^2 = j$.

One deserves that η may be written as follows

$$\eta = \sum_{k=1}^{n+1}\left(\frac{\zeta_k}{m} - p_k\right)^2 = \sum_{k=1}^{n+1}\frac{\zeta_k^2}{m^2} - 2a\sum_{k=1}^{n+1}\frac{\zeta_k}{m} + (n+1)a^2 = \sum_{k=1}^{n+1}\frac{\zeta_k^2}{m^2} - a. \tag{4}$$

Therefore using (3) and (4) one has

$$F_\eta(x) = P(\eta \leq x) = P\left(\frac{1}{m^2}\sum_{k=1}^{n+1}\zeta_k^2 - a \leq x\right) = P\left(\sum_{k=1}^{n+1}\zeta_k^2 \leq m^2(x+a)\right),$$

$$= \begin{cases} 0, & x \le \frac{l_1}{m^2} - a \\ \sum\limits_{j=l_1}^{[m^2(x+a)]} \frac{1}{C_{n+m}^n} P_{m,n,j} \ , & x \in \left[\frac{l_1}{m^2} - a, \frac{l_2}{m^2} - a\right] \\ 1, & x \ge \frac{l_2}{m^2} - a \end{cases}$$

$$= \begin{cases} 0, & x \le \frac{1}{m} - \frac{1}{n+1} \\ \sum\limits_{j=m}^{[m^2(x+a)]} \frac{1}{C_{n+m}^n} P_{m,n,j} \ , & x \in \left[\frac{1}{m} - \frac{1}{n+1}, 1 - \frac{1}{n+1}\right] \\ 1, & x \ge 1 - \frac{1}{n+1} \end{cases}$$

Thus the theorem is proved.

Intuitively the large values of η conforms with H_1 . Hence we can reject H_0 for large values of η and the declared α. So we propose the following test function as

$$\Psi^* \left(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m\right) = \begin{cases} 1, & \eta > x_\alpha(m) \\ 0, & \eta \le x_\alpha(m) \end{cases} .$$

Denote critical region as $W_\alpha = \{(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) : \eta > x_\alpha(m)\}$.

Considering $X_1, X_2, ..., X_n$ as a training sample, $Y_1, Y_2, ..., Y_m$ as a control sample and selecting $x_\alpha(m)$ such that $x_\alpha(m) \to 0$ as $m \to \infty$ one can prove the following

Theorem 3. Let $F \ne Q$ and for any $x_1 < x_2 < ... < x_n$ there exist $x_k, \{k = 2, 3...n\}$ such that $Q(x_k) - Q(x_{k-1}) \ne F(x_k) - F(x_{k-1})$. Then the test W_α for testing $H_0 : F = Q$, against $H_1 : F \ne Q$ is consistent , i.e.

$$\lim_{m \to \infty} P_F \times P_Q \left\{ \sum_{i=1}^{n+1} \left(\frac{\zeta_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} = 1, \ \text{for} \ F \ne Q.$$

Proof. Let $-\infty < x_1 < x_2 < ... < x_n < \infty$ and

$$\Delta_1^* = (-\infty, x_1] \, , \Delta_2^* = (x_1, x_2] \, , ..., \Delta_{n+1}^* = (x_n, \infty)$$

be the nonrandom intervals. Define the r.v.'s

$$\xi_i^{*k} = \begin{cases} 1, \ Y_k \in \Delta_i^* \\ 0, \ Y_k \notin \Delta_i^* \end{cases} , \ k = 1, 2, ..., m \quad i = 1, 2, ..., n+1$$

and ${\zeta^*}_i = \sum_{k=1}^{m} {\xi_i^*}^k$. It is clear that

$$P\left\{{\xi_i^*}^{\,k} = 1\right\} = P\left\{Y_k \in (x_{i-1}, x_i)\right\} \equiv p_i^* = \int_{x_{i-1}}^{x_i} dQ(x) = Q(x_i) - Q(x_{i-1})$$

Then one can write

$$P_F \times P_Q \left\{ \sum_{i=1}^{n+1} \left(\frac{\zeta_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} =$$

$$= n! \int \cdots \int P\left\{ \sum_{i=1}^{n+1} \left(\frac{\zeta_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \,\middle|\, X_{(1)} = x_1, ..., X_{(n)} = x_n \right\} \times$$

$$\times dF(x_1)...dF(x_{n)}$$

$$= n! \int \cdots \int_{x_1<x_2<...<x_n} P\left\{ \sum_{i=1}^{n+1} \left(\frac{{\zeta^*}_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} dF(x_1)...dF(x_{n)}$$

$$= n! \int \cdots \int_{x_1<x_2<...<x_n} P\left\{ \sum_{i=1}^{n+1} \left(\frac{{\zeta^*}_i}{m} - p_i^* + p_i^* - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} dF(x_1)...dF(x_{n)}$$

$$= n! \int \cdots \int_{x_1<x_2<...<x_n} P\left\{ \left(\sum_{i=1}^{n+1} \left(\frac{{\zeta^*}_i}{m} - p_i^* \right)^2 + 2 \sum_{i=1}^{n+1} \left(\frac{{\zeta^*}_i}{m} - p_i^* \right) \left(p_i^* - \frac{1}{n+1} \right) + \right.\right.$$

$$\left.\left. + \sum_{i=1}^{n+1} \left(p_i^* - \frac{1}{n+1} \right)^2 \right) > x_\alpha(m) \right\} dF(x_1)...dF(x_n)$$

By the law of large numbers $\frac{{\zeta^*}_i}{m} \to p_i^*$, a.s. as $m \to \infty$.

$$P\left\{ \sum_{i=1}^{n+1} \left(\frac{\zeta_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} \underset{m\to\infty}{\to} n! \int \cdots \int_{x_1<x_2<...<x_n} P\left\{ \sum_{i=2}^{n+1} \left(p_i^* - \frac{1}{n+1} \right)^2 > 0 \right\} \times$$

$$\times dF(x_1)...dF(x_{n)} \qquad (5)$$

Then one obtains from (5) for large n,

$$\lim_{m\to\infty} P\left\{ \sum_{i=1}^{n+1} \left(\frac{\zeta_i}{m} - \frac{1}{n+1} \right)^2 > x_\alpha(m) \right\} =$$

$$= n! \int \cdots \int_{x_1<x_2<\ldots<x_n} P\left\{ \sum_{i=2}^{n+1} \left(Q(x_i) - Q(x_{i-1}) - \frac{1}{n+1} \right)^2 > 0 \right\} dF(x_1)...dF(x_{n)}$$

$$= n! \int \cdots \int_{x_1<x_2<\ldots<x_n} P\left\{ \sum_{i=2}^{n+1} \left(Q(x_i) - Q(x_{i-1}) - F_n^*(x_i) + F_n^*(x_{i-1}) \right)^2 > 0 \right\} dF(x_1)...dF(x_{n)}$$

$$\backsim n! \int \cdots \int_{x_1<x_2<\ldots<x_n} P\left\{ \sum_{i=2}^{n+1} \left(Q(x_i) - Q(x_{i-1}) - F(x_i) + F(x_{i-1}) \right)^2 > 0 \right\} dF(x_1)...dF(x_{n)} = 1$$

Thus the theorem is proved.

Now consider the test $W_\alpha = \{(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) : \eta > x_\alpha(m)\}$, where x_α is determined from the equality

$$P\{(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) \in W_\alpha / H_0\} = 1 - F_\eta(x_\alpha).$$

The numerical values of $F_\eta(x)-$d.f. of η for $n = m = 15$ is provided in Table 3.

Here, $F_\eta(x_\alpha)$ should be close to one, in order to make small probability of error of type one. As the next step the d.f.

$$F_\eta(x) = \sum_{j=l_1}^{[m^2(x+a)]} \frac{1}{C_{n+m}^n} P_{m,n,j} \ , \ x \in \left[\frac{l_1}{m^2} - a, \frac{l_2}{m^2} - a \right]. \qquad (6)$$

has been tabulated for different values of n and m $(n + 1 \geq m)$. The values $P_{m,n,j}$ in (6) calculated using the following iterative relation

$$\sum_{i_1=0}^{m} \sum_{i_2}^{m-i_1} \ldots \sum_{i_{n+1}=0}^{m-i_1-i_2-\ldots-i_n} I(i_1, i_2, ..., i_{n+1}, m, n),$$

where

$$I(i_1, i_2, ..., i_{n+1}, m, n) = \begin{cases} 1, & if \quad i_1^2 + i_2^2 + ... + i_{n+1}^2 = j, \ i_1 + i_2 + ... + i_{n+1} = m \\ 0, & otherwise \end{cases}.$$

Using Minitab package program the samples from different distributions are obtained. These samples are given in Table 1. Comparison of results for two sample problem made for test W_α,based on test statistics η with Mann

Whitney-Wilcoxon and Kolmogorov Smirnov tests. Using software SPSS it is observed that the test W_α for small sample size gives more efficient results.

Table 1. The values of order statistics drawn from different distribution $n = 15$.

For an evaluation of the results obtained by application of η statistic and for comparison with the other tests statistics we provide the following numerical examples:

Table 2. Comparison of test based on η statistics with Mann Whitney

Wilcoxon, Kolmogorov Smirnov and Wilks' empty block tests for $\alpha = 0.05$.

Comparison with the Wilks' empty block test is also made. Given in Table 2 critical region $s_0 \geq s_{table}$ for Wilks' empty block test is constructud as follows:

Suppose $X_1, X_2, ..., X_n$ is a sample obtained from distributions with d.f. $F(x)$ and $Y_1, Y_2, ..., Y_m$ is a sample obtained from distributions with d.f. $Q(x)$. Let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ be the order statistics constructed from $X_1, X_2, ..., X_n$.

The intervals $\left(-\infty, X_{(1)}\right], \left(X_{(1)}, X_{(2)}\right], ..., \left(X_{(n-1)}, X_{(n)}\right], \left(X_{(n)}, \infty\right)$ are called sample blocks $B_1^{(1)}, B_1^{(2)}, ..., B_1^{(n+1)}$ respectively. Let s_i be the number of the blocks $B_1^{(1)}, B_1^{(2)}, ..., B_1^{(n+1)}$ which contain i elements from $Y_1, Y_2, ..., Y_m$. That is, s_0 the number which contain 0 elements from $Y_1, Y_2, ..., Y_m$. $(s_0, s_1, ..., s_m)$ is a multidimensional random variable which must satisfy the conditions

$$s_0 + s_1 + ... + s_m = n + 1$$

$$s_1 + 2s_2 + ... + ms_m = m.$$

Assume that H_0 is true. Probability function of s_0 is

$$P(s_0) = \frac{\left(\begin{array}{c} n+1 \\ s_0 \end{array} \right) \left(\begin{array}{c} m-1 \\ n-s_0 \end{array} \right)}{\left(\begin{array}{c} n+m \\ n \end{array} \right)}$$

$s_0 = k, k+1, ..., n$, where $k = \max(0, n-m+1)$.

The test has its critical region $W_{a\alpha}$.

$$W_{a\alpha} = \{(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m) : s_0 > s_{0\alpha}(\alpha, n, m)\}.$$

$s_{0\alpha}(\alpha, n, m)$ is the smallest integer for

$$P\left(W_{a\alpha} \setminus (F_1, F_2) \in \Im_0\right) \leq \alpha. \tag{7}$$

For $n = 15$ and $m = 15$ $s_{0\alpha}(\alpha, n, m) = s_{table} = 10$ found from (7). Using Table 1 we obtained values of s_0. In the table 2 results of Wilks empty block test is presented.

Now using Table 2 we try to explain how to use the test based on the η statistic.

Case 1. (See Table 2) Let $X_1, X_2, ..., X_n$ $(n = 15)$ be a sample obtained from distribution with d.f. $F(x)$ and $Y_1, Y_2, ..., Y_m$ $(m = 15)$ be a sample obtained from distribution with d.f. $Q(x)$. Here $F(x)$ is a uniform over $(0, 1)$ distribution function and $Q(x)$ is the standard Normal distribution function .

We want to test hypothesis $H_0 : F(x) = Q(x)$ by the help of η statistic. Consider

$$\Delta_1 = \left(-\infty, X_{(1)}\right], \Delta_2 = \left(X_{(1)}, X_{(2)}\right], ..., \Delta_n = \left(X_{(n-1)}, X_{(n)}\right], \Delta_{n+1} = \left(X_{(n)}, \infty\right).$$

Denote by ζ_i the number of observations $Y_1, Y_2, ..., Y_{15}$ falling to Δ_i, $i = 1, 2, ..., 15$.

That is $\zeta_1 = 10$, $\zeta_2 = 1, \zeta_3 = 0, \zeta_4 = 1, \zeta_5 = 0, \zeta_6 = 0, \zeta_7 = 0, \zeta_8 = 1$, $\zeta_9 = 0, \zeta_{10} = 0, \zeta_{11} = 0, \zeta_{12} = 0, \zeta_{13} = 2, \zeta_{14} = 0, \zeta_{15} = 0, \zeta_{16} = 0$. By definition the numerical value of η statistic has been calculated as $\eta = \sum\limits_{k=1}^{n+1} \left(\frac{\zeta_k}{m} - a\right)^2$. Thus

$$\eta = \left(\frac{10}{15} - \frac{1}{16}\right)^2 + 3\left(\frac{1}{15} - \frac{1}{16}\right)^2 + \left(\frac{2}{15} - \frac{1}{16}\right)^2 + 11\left(0 - \frac{1}{16}\right)^2 = 0.4130$$

and η_{table} can be found from the Table 3 (see Appendix) for $n = 15$ and $m = 15$. $F_\eta\left(m^2(x_\alpha + \frac{1}{n+1})\right) = 0.95$ in order to have $m^2\left(x_\alpha + \frac{1}{n+1}\right) = 60 \Rightarrow 15^2\left(x_\alpha + \frac{1}{16}\right) = 60 \Rightarrow x_\alpha = \eta_{\text{table}} = 0.2041$ and since $\eta > \eta_{\text{table}}$, H_0 rejected.

Since, these samples are obtained from different distributions, it is expected that H_0 will be rejected. $X_1, X_2, ..., X_n \sim U(0,1)$ and $Y_1, Y_2, ..., Y_m \sim N(0,1)$. Other cases given in Table 2 is analyzed analogously.

References

1. Bairamov, I.G. and Petunin, Yu.I. (1991) Structure of Invariant Confidence Intervals Containing the Main Distributed Mass. *Theor. Prob. Appl.* Vol. **35**, No:1, pp. 15-26.

2. Borovkov, A.A. (1975) Asymptotically Optimal Tests For Compound Hypotheses. *Theor. Prob. Appl.* Vol.**20**, No:1, pp.447-469.

3. Borovkov, A.A. (1984) *Mathematical Statistics.* Moskow, Nauka (in-Russian).

4. Dixon , W.J. (1940) A Criterion for Testing the Hypothesis That Two Samples are From the same Population, *Ann. Math. Stat.* , Vol. **11**, pp. 199-204.

5. Madreimov, I. and Y.I. Petunin (1983) A Characterization of the Uniform Distribution with the Aid of Order Statistics. *Theor. Prob. and Math. Statist.* 27, 105-110.

6. Mathisen, H.C. (1943) A Method of Testing the Hypothesis That Two Samples are From the Same Population, *Ann. Math. Stat.* , Vol. **14**, pp. 188-194.

7. Smirnov, N. (1939) On The Estimation of the Discrepancy Between Emprical Curves of Distribution For Two Independent Samples , *Bull. Math. Univ.* Moscow, Vol. **2**, No: 2, pp. 3-16.

8. Wald, A. and Wolfowitz, J. (1940) On a Test Whether Two Samples are From the Same Population, *Ann. Math. Stat.* , Vol. **11**, pp. 147-162.

9. Wilks, S.S. (1962) A Combinatorial Test For the Problem of Two Samples From Continuous Distributions, Proc. Fourth Berkeley Symp. *on Math. Stat. and Prob.*, University of California Press.