THE ROBUSTNESS AGAINST DEPENDENCE OF NONPARAMETRIC TESTS FOR THE TWO-SAMPLE LOCATION PROBLEM

Abstract. Nonparametric tests for the two-sample location problem are investigated. It is shown that the supremum of the size of any test can be arbitrarily close to 1. None of these tests is most robust against dependence.

1. Introduction. Situations with some kind of dependencies for the Mann–Whitney–Wilcoxon test were investigated by Hollander, Pledger and Lin [3], Pettit and Siskind [4], Serfling [6], Zieliński [8], [9]. In this paper we consider the robustness against dependence of a large family of nonparametric tests for the two-sample location problem, including the test mentioned above. We take advantage of a new description of dependence, called Rüschendorf’s $\varepsilon$-neighbourhoods, proposed in [2].

2. Problem and notation. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ denote two independent random samples from populations with continuous distribution functions $F_X(x) = F(x - \Delta)$ and $F_Y(y) = F(y)$ respectively. We verify the hypothesis $H : \Delta = 0$ against $K : \Delta > 0$ by means of a test $\phi$ of size $\alpha$. We assume that $\phi$ belongs to some family $\Phi$ of tests (see Sec. 3 for the definition of $\Phi$).

Let $P(F) = \{P : P(Z_i \leq z) = F(z), \ i = 1, \ldots, m + n\}$ describe all possible violations of independence. We denote $P(F)$ briefly by $P$. Let $P_C \subset P$ be the subfamily of all continuous distribution functions. Moreover, let $C_\varepsilon \subset P_C$ be a family of all c.d.f. which correspond to small dependencies (for more details see Sec. 4).

1991 Mathematics Subject Classification: Primary 62G35; Secondary 62G10.

Key words and phrases: robustness of tests, robustness against dependence, nonparametric tests for the two-sample location problem, size of test.
The following problems are considered in this paper:

(A) Given any $\phi \in \Phi$, compute the supremum of the size of $\phi$ under all kinds of dependencies, i.e. $\sup_{P \in P} \int_{\mathcal{X}} \phi \, dP$, where $\mathcal{X}$ denotes a sample space.

(B) Given any $\phi \in \Phi$, evaluate the robustness of the size of $\phi$ against small dependencies. We use the oscillation of the size over $C_\varepsilon$ as a measure of robustness (see [7]):

$$r_\varepsilon(\phi) = \sup_{P \in C_\varepsilon} \int_{\mathcal{X}} \phi \, dP - \inf_{P \in C_\varepsilon} \int_{\mathcal{X}} \phi \, dP.$$ 

(C) Find the most robust test in $\Phi$, i.e. a test $\phi_0$ such that $r_\varepsilon(\phi_0) \leq r_\varepsilon(\phi)$ ($\forall \phi \in \Phi$) for all $\varepsilon$.

3. The family $\Phi$. We restrict our consideration to a family $\Phi$ of one-sided nonparametric tests for the two-sample location problem given as follows:

**Definition.** $\phi \in \Phi$ if and only if

(i) $\phi(x_1 + \tau, \ldots, x_m + \tau, y_1 + \tau, \ldots, y_n + \tau) = \phi(x_1, \ldots, x_m, y_1, \ldots, y_n) \forall \tau,$

(ii) $\phi(x_1, \ldots, x_{i-1}, x_i + \delta, x_{i+1}, \ldots, x_m, y_1, \ldots, y_n) \geq \phi(x_1, \ldots, x_m, y_1, \ldots, y_n) \ (\forall \delta \geq 0), \ i = 1, \ldots, m,$

(iii) if $X_{1:m} > Y_{1:n}$ then $\phi(x_1, \ldots, x_m, y_1, \ldots, y_n) = 1,$

if $X_{m:m} < Y_{1:n}$ then $\phi(x_1, \ldots, x_m, y_1, \ldots, y_n) = 0,$

where $X_{i:m}$ and $Y_{i:n}$ denote the $i$th order statistics from the first and the second sample respectively.

The conditions (i)–(iii) seem to be quite natural. The Mann–Whitney–Wilcoxon test, the Fisher–Yates test, the Rosenbaum test and many other tests for the two-sample location problem belong to the family $\Phi$ (see [1]).

4. A description of dependence. By the Rüschendorf theorem (see [5]) we know that $h$ is the density of a probability measure on $[0, 1]^r$ with uniform marginals and continuous w.r.t. the Lebesgue measure $d\mu$ on $[0, 1]^r$ if and only if $h = 1 + S f$ where $f \in L^1([0, 1]^r)$ and $S : L^1 \rightarrow L^1$ is the linear operator given by

$$S f = f - \sum_{i=1}^r \int [0,1]^r \int f \, dz_1 \ldots \widehat{dz_i} \ldots dz_r + (r - 1) \int [0,1]^r \int f \, dz_1 \ldots dz_r$$

and $S f \geq -1$.

Without loss of generality we assume that $F$ is the uniform distribution on $[0, 1]$. 


Moreover, define $\mathcal{R} = \{ f \in L^1([0, 1]^r) : Sf \geq -1 \}$, and let $[0, u]^r = \{ x \in [0, 1]^r : 0 \leq x_i \leq u_i, i = 1, \ldots, r \}$.

Basing on the above theorem and assumptions we may write that

$$\mathcal{P}_C = \left\{ P : P(u) = \int_{[0,u]^r} (1 + Sf) \, d\mu, f \in \mathcal{R} \right\}.$$ 

In [2] a new description of dependence, called Rüschendorf’s $\varepsilon$-neighbourhoods, was proposed and motivated. Following that paper let $\mathcal{R}_{\varepsilon} = \{ f \in L^1 : \|Sf\| \leq \varepsilon, Sf \geq -1 \}$, where $\| \cdot \|$ is the $L^1$ norm. Then

$$\mathcal{C}_{\varepsilon} = \left\{ P : P(u) = \int_{[0,u]^r} (1 + Sf) \, d\mu, f \in \mathcal{R}_{\varepsilon} \right\}$$

describes the family of distributions which correspond to small departures from independence, so called $\varepsilon$-dependence, i.e. if $\varepsilon$ is sufficiently small then the dependence measured by $\rho$-Spearman’s and many other measures is small as well, and conversely.

In order to solve our problems (A) and (B) it will be necessary for any $\phi \in \Phi$ to compute:

(A) $\sup_{f \in \mathcal{R}} \int_{[0,1]^r} (1 + Sf) \phi \, d\mu$,

(B) $r_{\varepsilon}(\phi) = \sup_{f \in \mathcal{R}_{\varepsilon}} \int_{[0,1]^r} (1 + Sf) \phi \, d\mu - \inf_{f \in \mathcal{R}_{\varepsilon}} \int_{[0,1]^r} (1 + Sf) \phi \, d\mu$.

It is easy to show (see Sec. 7) that the above expressions are equivalent to the following, more convenient in further investigations:

(A) $\sup_{g \in \mathcal{G}} \int_{[0,1]^r} (1 + g) \phi \, d\mu$,

(B) $r_{\varepsilon}(\phi) = \sup_{g \in \mathcal{G}_{\varepsilon}} \int_{[0,1]^r} (1 + g) \phi \, d\mu - \inf_{g \in \mathcal{G}_{\varepsilon}} \int_{[0,1]^r} (1 + g) \phi \, d\mu$,

where $\mathcal{G} = \{ g \in L^1([0, 1]^r) : g \geq -1, \int_{[0,1]^r} g \, d\mu = 0, \int_{[0,1]^{r-1}} g \, dz_1 \ldots \hat{z}_i \ldots dz_r = 0, \forall i = 1, \ldots, r \}$, and $\mathcal{G}_{\varepsilon} = \{ g \in \mathcal{G} : \| g \| \leq \varepsilon \}$.

5. Results. Now we can state the solutions of our problems (A)–(C).

THEOREM 1. Let $\phi \in \Phi$. Suppose that all kind of dependencies between samples and among observations in samples are allowed. Then the size of the test $\phi$ can be arbitrarily close to 1, i.e. $\sup_{P \in \mathcal{P}} \int_X \phi \, dP = 1$. 
**Theorem 2.** The robustness of any test $\phi \in \Phi$ against $\varepsilon$-dependence equals $\varepsilon/2$, i.e. $r_\varepsilon(\phi) = \varepsilon/2$.

From this theorem we get immediately:

**Corollary.** In the family $\Phi$ of one-sided nonparametric tests for the two-sample location problem, no test is most robust against dependence.

6. Proofs

**Proof of Theorem 1.** Let $\{G_N\}_{N=2}^\infty$ be the sequence of $r$-dimensional subsets of $[0, 1]^r$, $r = m + n$, given by

$$G_N = \bigcup \left[ \left( \frac{i}{N} - \frac{1}{N}, \frac{i+1}{N} - \frac{1}{N} \right)^m \times \left( \frac{j-1}{N}, \frac{j}{N} \right)^n \right],$$

where the union is extended over all $(i, j)$ of the form $(k \mod N, k)$, for $k = 1, \ldots, N$.

Let $\{g_N\}_{N=2}^\infty$ be the sequence of real functions on $[0, 1]^r$ defined as follows:

$$g_N(z) = \begin{cases} 
nr - 1 - 1 & \text{for } z \in G_N, \\
-1 & \text{for } z \notin G_N.
\end{cases}$$

We show that $g_N \in \mathcal{G}$ ($\forall N \geq 2$):

(a) $g_N \geq -1$ by the definition,

(b) $\int_{[0,1]^r} g_N \, d\mu = N \frac{Nr - 1}{N^r} + (-1) \left( 1 - N \frac{1}{N^r} \right) = 0$,

(c) $\int_{[0,1]^{r-1}} g_N \, dz_1 \ldots dz_i \ldots dz_r = N^{r-1} \frac{N - 1}{N^{r-1}} + (-1) \left( 1 - \frac{1}{N^{r-1}} \right) = 0$

for $i = 1, \ldots, r$.

So $g_N \in \mathcal{G}$ for every $N \geq 2$.

Now take $\phi \in \Phi$ and denote by $\alpha$ its size. Suppose that $\phi$ is a non-randomized test with a critical region $K_\alpha$. It is easy to check that for each $N \geq 2$,

$$G_N \setminus \left[ \left( \frac{0}{N}, \frac{1}{N} \right)^m \times \left( \frac{N - 1}{N}, 1 \right)^n \right]$$

$\subset \{(z_1, \ldots, z_r) : 0 \leq z_j \leq z_i \leq 1, \; i = 1, \ldots, m; \; j = m+1, \ldots, m+n \} \subset K_\alpha$

for every $\phi \in \Phi$ (see conditions specified in Sec. 3). So we get
Two-sample location problem

\[ \sup_{\phi \in \Phi} \int \phi d\mu \geq \sup_{\phi \in \Phi} \int (1 + g) \phi d\mu \]

\[ = \sup_{g \in \mathcal{G}} \int (1 + g) d\mu \geq \int (1 + g_N) d\mu \]

\[ = \alpha + \left\{ \frac{(N - 1) N r^{-1} - 1}{N r} + (-1) \left( \alpha - (N - 1) \frac{1}{N r} \right) \right\} \]

\[ = \alpha + \frac{N - 1}{N} - \alpha = \frac{N - 1}{N} . \]

Choosing \( N \) large enough one can come arbitrarily close to 1.

**Remark.** For simplicity we have assumed in the proof that \( \phi \) is a non-randomized test. The theorem is true for randomized tests as well.

**Proof of Theorem 2.** We take a non-randomized test \( \phi \in \Phi \) and denote by \( \alpha \) its size and by \( K_\alpha \) its critical region. Let \( \{G_N\} \) be as in the proof of Theorem 1. Consider the sequence \( \{g'_N\} \) of real functions on \([0, 1]^r\) defined by

\[ g'_N(z) = \begin{cases} 
\frac{\epsilon}{2} N^{r-1} & \text{for } z \in G_N, \\
-\frac{\epsilon}{2} N^{r-1} - 1 & \text{for } z \notin G_N,
\end{cases} \]

for \( N \geq N_0 = (2/(2 - \epsilon))^{1/(r-1)} \). It is easily seen that \( g'_N \in \mathcal{G}_\epsilon \) (\( \forall N \geq N_0 \)).

So we get

\[ \sup_{g \in \mathcal{G}_\epsilon} \int (1 + g) \phi d\mu \]

\[ = \sup_{g \in \mathcal{G}_\epsilon} \int (1 + g_N) d\mu \]

\[ = \alpha + \left\{ \frac{(N - 1) \epsilon N r^{-1} - 1}{N r} + (-1) \left( \alpha - (N - 1) \frac{1}{N r} \right) \right\} \]

\[ = \alpha + \frac{\epsilon}{2} (1 - \alpha) \quad \text{as } N \to \infty . \]

In order to show that also \( \sup_{g \in \mathcal{G}_\epsilon} \int (1 + g) \phi d\mu \leq \alpha + \frac{\epsilon}{2} (1 - \alpha) \), we consider the operator \( Tg = \int_{K_\alpha} g d\mu \). It is a bounded linear operator, so we get \( T \|g\| \leq \|T\| \|g\| \quad (\forall g \in \mathcal{G}) \).

By the proof of Theorem 1,

\[ \|T\| = \sup_{g \in \mathcal{G}} \frac{\|Tg\|}{\|g\|} = \frac{1 - \alpha}{2} . \]
(this is evident, because $\varepsilon$ cannot be greater than 2). So we get

\[
(\forall g \in \mathcal{G}_\varepsilon) \quad T g \leq \frac{1 - \alpha}{2} \|g\| \leq \frac{1 - \alpha}{2} \varepsilon
\]

and therefore

\[
\sup_{g \in \mathcal{G}_\varepsilon} \int_{K_\alpha} (1 + g) d\mu = \sup_{g \in \mathcal{G}_\varepsilon} (\alpha + T g) \leq \alpha + \frac{1 - \alpha}{2} \varepsilon.
\]

Hence

\[
\sup_{g \in \mathcal{G}_\varepsilon} \int_{[0, 1]^r} (1 + g) d\mu = \alpha + \frac{\varepsilon}{2} (1 - \alpha).
\]

Now we consider $\inf_{g \in \mathcal{G}_\varepsilon} \int_{[0, 1]^r} (1 + g) d\mu$. Let us define a sequence $\{G'_N\}$ by

\[
G'_N = \bigcup \left[ \left( \frac{i - 1}{N}, \frac{i}{N} \right)^m \times \left( \frac{j}{N}, \frac{j + 1}{N} \right)^n \right] \subset [0, 1]^r,
\]

where the union is extended over all $(i, j)$ of the form $(k, k \text{mod} N)$ for $k = 1, \ldots, N$.

Let $\{g'_N\}$ be the following sequence of real functions on $[0, 1]^r$:

\[
g'_N(z) = \begin{cases} 
\frac{\varepsilon}{2} N^{r-1} & \text{for } z \in G'_N, \\
-\frac{\varepsilon}{2} N^{r-1} - 1 & \text{for } z \notin G'_N,
\end{cases}
\]

where $N \geq N_0$. It is easily seen that $g'_N \in \mathcal{G}_\varepsilon (\forall N \geq N_0)$. So

\[
\inf_{g \in \mathcal{G}_\varepsilon} \int_{[0, 1]^r} (1 + g) d\mu = \inf_{g \in \mathcal{G}_\varepsilon} \int_{K_\alpha} (1 + g) d\mu \geq \int_{K_\alpha} (1 + g'_N) d\mu
\]

\[
= \alpha + \frac{\varepsilon}{2} N^{r-1} - \frac{\varepsilon}{2} N^{r-1} - 1 \left( \alpha - \frac{1}{N^r} \right)
\]

\[
\to \alpha (1 - \varepsilon/2) \quad \text{as } N \to \infty.
\]

Similarly, we can prove the opposite inequality:

\[ \inf_{g \in \mathcal{G}_\varepsilon} \int_{[0, 1]^r} (1 + g) d\mu \geq \alpha (1 - \varepsilon/2) \]

and therefore

\[
\inf_{g \in \mathcal{G}_\varepsilon} \int_{[0, 1]^r} (1 + g) d\mu = \alpha (1 - \varepsilon/2).
\]
Thus finally
\[
    r_{\varepsilon}(\phi) = \sup_{P \in \mathcal{C}} \int_X \phi \, dP - \inf_{P \in \mathcal{C}} \int_X \phi \, dP
\]
\[
    = \sup_{g \in \mathcal{G}^\varepsilon_{[0,1]^r}} \int_{[0,1]^r} (1 + g) \phi \, d\mu - \inf_{g \in \mathcal{G}^\varepsilon_{[0,1]^r}} \int_{[0,1]^r} (1 + g) \phi \, d\mu
\]
\[
    = \alpha + \varepsilon \left( 1 - \varepsilon \right) - \alpha \left( 1 - \frac{\varepsilon}{2} \right) = \varepsilon \frac{\varepsilon}{2},
\]
which completes the proof. \(\blacksquare\)

As before, the theorem is also true for randomized tests.

7. Complements. In Section 4 we have stated that in our problem we could consider the families \(\mathcal{G}\) and \(\mathcal{G}_\varepsilon\) instead of \(\mathcal{R}\) and \(\mathcal{R}_\varepsilon\). This follows from

LEMMA. Let \(S\) be the operator defined in Section 4. Then \(S(\mathcal{R}) = \mathcal{G}\) and \(S(\mathcal{R}_\varepsilon) = \mathcal{G}_\varepsilon\).

Proof. It suffices to show that \(S(\mathcal{R}) = \mathcal{G}\). The same proof remains valid for the second assertion.

Suppose \(f \in \mathcal{R}\). Then \(Sf \in L^1([0,1]^r), Sf \geq -1\) and \(\int_{[0,1]^r} Sf \, d\mu = \int_{[0,1]^r} Sf \, d\mu = 0\). So \(S(\mathcal{R}) \subseteq \mathcal{G}\).

Now take any \(g \in \mathcal{G}\). Then
\[
    Sg = g - \sum_{i=1}^r \int_{[0,1]^r} g \, dz^1 \ldots \hat{d}z^i \ldots dz^r + (r - 1) \int_{[0,1]^r} g \, dz^1 \ldots dz^r = g
\]
and hence \(S(\mathcal{R}) \supseteq \mathcal{G}\). \(\blacksquare\)

References


PRZEMYSŁAW GRZEGORZEWSKI
INSTITUTE OF MATHEMATICS
WARSAW UNIVERSITY OF TECHNOLOGY
PL. POLITECHNIKI 1
00-661 WARSZAWA, POLAND

Received on 25.10.1993