uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Vladislav Brion

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Mathematics)

GRADE / DEGREE

Department of Mathematics and Statistics

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Nonparametric Tests of Hypotheses for Umbrella Alternatives

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. M. Alvo

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. M. Mojirsheihani

Dr. M. Zarepour

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Nonparametric Tests of Hypotheses for Umbrella Alternatives

Vladislav Brion
Department of Mathematics
University of Ottawa, Canada

Supervised by

Mayer Alvo
Department of Mathematics
University of Ottawa, Canada

March 9, 2007

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for the
M.Sc. degree in Mathematics

Ottawa-Carleton Institute for Graduate Studies and Research in Mathematics and Statistics

Your file Votre référence
ISBN: 978-0-494-41618-1
Our file Notre référence
ISBN: 978-0-494-41618-1

Acknowledgement

I wish to express my hearty gratitude to Professor Mayer Alvo for his assistance in all stages of this research.

Abstract

A general method for nonparametric hypothesis testing for umbrella alternatives is proposed. Such alternatives arise in situations where the treatment effect changes in direction after reaching a peak. The approach consists of defining two sets of rankings, one corresponding to the observations and the other to the alternative. The test statistic measures the distance between the two sets. The limiting distributions are obtained under the null hypothesis when the location of the peak is known. The simulation study shows good power of the test. A method is proposed for estimating the location of the peak, if it is unknown.

Contents

Notation

$a_n(i)$ - scores;

c_1 – cardinality of non-permuted set E for the approach 1;

c_2 – cardinality of non-permuted set E for the approach 2;

c - common symbol of cardinality of non-permuted set E;

θ - population medians;

E - rank permutation set "most in agreement" with the alternative;

F_l - distribution for the population l;

g - subset in the non-permuted set E, $1 \le j \le n$;

h_i - coefficients in the linear rank statistic;

i - column in the set E, $1 \le i \le n$;

j - population rank in the set E, $1 \le j \le k$ for the first approach or observation rank, $1 \le j \le n$ for the second approach;

k - number of populations;

l - population number in the sets π and E; $1 \le l \le k$;

m_l - number of observations in the population l;

m - number of observations, if equal for all populations;

$\overline{m}$ - average number of observations; $\overline{m} = \frac{n}{k}$;

n - total number of observations, $n = \sum_{l=1}^{k} m_l = mk$;

n_r - number of observations in the first r populations, $n_r = \sum_{l=1}^{r} m_l, 1 \le r \le k$

p - peak population index, $1 \le p \le k$;

T_n - statistic;

$u(j,l)$ - frequency of occurrence of the population rank j in any column in the group of columns, corresponding to the population F_l in the set E for the approach 1;

$w(j,i)$ - frequency of occurrence of the observation rank j in the column i in the set E for the approach 2;

μ, ν - permutations of integers;

π - vector of observed ranks;

$\overline{\pi_l}$ - average of the ranks in population F_l, $\overline{\pi_l} = \frac{1}{m_l} \sum_{i=n_{l-1}+1}^{n_l} \pi(i)$;

List of tables

1. Introduction

1.1 The hypothesis testing problem

Nonparametric tests of hypotheses for umbrella alternatives belong to the group of tests, where the key interest is in the relative ordering of the distributions. The observed data consists of k independent random samples, one from each population. Under the common null hypothesis of these tests, there is no difference among the populations, and, therefore, the data can be considered as a single random sample from one population. The alternative hypothesis may contain various nonnull relationships among the distributions. An example of such alternatives is the case of non-equal medians for two or more treatments.

In this thesis we consider an umbrella alternative, under which, for example, the population medians exhibit a monotone pattern subject to a change in direction after reaching a peak.

Let the observed data consist of $n = \sum_{l=1}^{k} m_l$ observations, with m_l observations from population F_l, $l = 1,...,k$. The data satisfy the following assumptions:

- The n random variables $\{X_{11}, X_{12},..., X_{1m_1}, X_{21}, X_{22},..., X_{2m_2},..., X_{k1}, X_{k2},..., X_{km_k}\}$ are mutually independent;
- For any $l \in \{1,...,k\}$, the m_l random variables $\{X_{l1}, X_{l2},..., X_{lm_l}\}$ constitute a random sample from an absolutely continuous distribution function F_l;

The hypotheses considered are:

$$\begin{aligned} H_0 &: F_1(x) = ... = F_k(x) \\ H_1 &: F_1(x) \geq ... F_{p-1}(x) \geq F_p(x) \leq F_{p+1}(x) ... \leq F_k(x) \end{aligned} \tag{1.1}$$

with at least one strict inequality for some x.

A parametric formulation may be stated as follows:

$$F_l \equiv F(x - \theta_l) \tag{1.2}$$

for $l = 1,...,k$, where F is an absolutely continuous distribution function with median zero. In that case the hypotheses become:

$$\begin{aligned} H_0 &: \theta_1 = ... = \theta_k \\ H_1 &: \theta_1 \le ... \le \theta_{p-1} \le \theta_p \ge \theta_{p+1} \ge ... \ge \theta_k \end{aligned} \tag{1.3}$$

The location of the peak p may or may not be known. If the peak is unknown, the hypothesis H_1 is defined for some $p \in \{1,...,k\}$. The extremal cases $p = k$ or $p = 1$ are equivalent to the usual ordered alternatives. Therefore, the ordered alternatives are special cases of the umbrella alternatives.

1.2. Literature review

The term "umbrella alternative" was labeled by Mack and Wolfe (1981) who proposed testing such alternatives for known peak. The test statistic is the sum of Mann-Whitney counts for both sides of the umbrella (to the right of the peak these counts are reversed). The mean and variance under H_0 may be calculated directly, or by using a large-sample approximation. Large values of the statistic correspond to the case, where the sequence of population medians is close to the umbrella configuration. If the statistic exceeds some critical value, then H_0 should be rejected in favor of the umbrella alternative.

For the more common case of unknown umbrella peak, the various modifications were proposed by Mack and Wolfe (1981), Hettmansperger and Norton (1987), Millen and Wolfe (2005). All modifications are based on the statistic for the known peak with estimation of the most likely position of umbrella.

The test, which detects umbrella patterns, may be valid to detect monotone patterns. Moreover, in some cases the umbrella test has higher power. For example, the power of the Jonckheere-Terpstra (1954) test of ordered alternatives is significantly affected by violation from the monotone pattern at either the beginning or the end of the median sequence. For this case, Hollander and Wolfe (1999) proposed to apply the tests for umbrella alternatives instead of the Jonckheere-Terpstra test. Alvo and Cabilio (1991) considered testing for ordered alternatives when there are missing data.

1.3. Applications of the umbrella alternatives

Examples, where umbrella alternatives may be appropriate, are as follows:

- the effectiveness of a drug may increase with increasing level of a drug dosage and decrease (and even become toxic) after the optimal dosage is exceeded;
- the crop yield may peak with increasing quantities of fertilizer applied and then exhibit a decrease thereafter;
- the ability to learn may increase with age up to a certain time point and drop thereafter;
- a chemical reaction may be accelerated by increasing a catalyst concentration up to an optimal value; thereafter no increase in concentration will influence the reaction.

In chapter 2 we propose a procedure for determining a test statistic when the peak is known. In chapter 3 we obtain the test statistic, corresponding to the Spearman-Rho distance. In chapter 4 we find the limiting distributions of these statistics under the null hypothesis. The limiting distributions are based on assumption that the minimal sample size gets large. Finally, some simulation results are provided. We also consider the problem of finding the umbrella peak and obtain the power of the test, based on the Spearman-Rho statistic.

2. A new approach for finding test statistics when the peak is known

2.1 The test procedure

Motivated by the work of Critchlow (1992), Alvo and Pan (1996) proposed a general approach for testing hypotheses based on the ranks of the observations. The advantage of this approach is that it may be applied to various problems of hypothesis testing, including the special case of ordered alternatives. In this thesis, this approach will be used to test hypotheses, as described in section 1.1.

Let $P = \{\mu\}$ be the set of all permutations of integers $1,...,n$. For any two permutations μ and ν, define a distance $d(\mu,\nu)$ to measure the separation between μ and ν. Let:

$$n_r = \sum_{l=1}^{r} m_l,\ 1 \le r \le k \tag{2.1}$$

The proposed test procedure is based on the following steps:

1. Rank all n observations together so that the smallest gets rank 1 and the biggest gets rank n. Let $\pi = [\pi(1),...,\pi(m_1) \mid \pi(m_1+1),...,\pi(m_1+m_2) \mid ... \mid \pi(n_{k-1}+1),...,\pi(n)]$ represent the n - dimensional vector of ranks of the $\{X_{lj}\}; l = 1,...,k; j = 1,...,m_l$. The ranks are grouped by population. In the view of the continuity assumption, no ties occur among the observations with probability one.
2. Define $\{\pi\}$ to be the subclass of permutations "equivalent" to the observable permutation π in the sense that ranks occupied by independent identically distributed random variables are exchangeable. This subclass consists of all the permutations π where the rankings within each population are permuted among themselves only. The m_l ranks from population F_l will produce $m_l!$ permutations. Therefore, the cardinality of $\{\pi\}$ is $\prod_{l=1}^{k} m_l!$ or $(m!)^k$ for the case of equal number of observations, where $m_l = m$ for $l = 1,...,k$.

3. Define E to be the subclass of "extremal" permutations consisting of all permutations that are "most in agreement with the alternative H_1". This set consists of those permutations, which provide the strongest evidence in favor of the alternative. The next part contains a more detailed description of the construction of this set.
4. Compute the distance between sets $\{\pi\}$ and E to be the sum of all pairwise distances between them:

$$d(\{\pi\}, E) = \sum_{\mu \in \{\pi\}} \sum_{\nu \in E} d(\mu, \nu) \tag{2.2}$$

Small values of the distance (2.2) are inconsistent with the null hypothesis and consequently lead to rejection of H_0 in favor of H_1.

For both sets, $\{\pi\}$ and E, the next symbols will be used:

- i - will refer to the position of a single rank, $1 \le i \le n$.
- l - will refer to the population, $1 \le l \le k$.

2.2 The construction of the extremal set

The sets E and $\{\pi\}$ contain permutations of ranks, which may be arranged in columns. They differ in the following way. For the set $\{\pi\}$, permutations are applied to the initial ranking vector obtained by ranking all the observations in such a way that the ranks assigned to a population are preserved. For the set E, permutations are applied to all possible subsets (the number of these subsets depends on k and p), everyone of which ideally follows the umbrella configuration. For example, if $1 \le r < s \le p$, any rank in the column corresponding to the population F_r should be less than any rank in the column corresponding to the population F_s while the reverse is true if $p \le r < s \le k$. The umbrella configuration does not further restrict the position of the ranks. Either of the next two approaches may be used while building the set E, based on the analysis of relations between the ranks in the columns corresponding to the populations F_r and F_s; $1 \le r < p < s \le k$:

1. The ranks within a population are consecutive integers. Each population respects the order prescribed by the alternative;
2. The ranks within a population are arbitrary integers, which respect the order prescribed by the alternative on the populations.

The symbols of the set E for these approaches will be E_1 and E_2 respectively.

The first approach is the partial case of the second approach that contains more rank permutations, which are "most in agreement with H_1". Consequently, the test based on the second approach should be more precise. On the other hand, the second approach is more complicated, and the statistics, based on this approach, may not have a closed form. In this thesis, both of the approaches will be considered in some detail.

2.3 Examples of the sets $\{\pi\}$, E_1 and E_2

In this section we illustrate the construction of various permutation sets.

Example 1. To illustrate the construction of set $\{\pi\}$, suppose that the ranks of the observed data in the notation above are: $\{1, 4| 2, 6, 8 | 5, 7\}$. Then the set $\{\pi\}$ consists of the following permutations:

[1, 4| 2, 6, 8 | 5, 7],
[4, 1| 2, 6, 8 | 5, 7],
[1, 4| 2, 8, 6 | 5, 7],
[4, 1| 2, 8, 6 | 5, 7],
[1, 4| 6, 2, 8 | 5, 7],
[4, 1| 6, 2, 8 | 5, 7],
[1, 4| 6, 8, 2 | 5, 7],
[4, 1| 6, 8, 2 | 5, 7],
[1, 4| 8, 6, 2 | 5, 7],
[4, 1| 8, 6, 2 | 5, 7],
[1, 4| 8, 2, 6 | 5, 7],
[4, 1| 8, 2, 6 | 5, 7],
[1, 4| 2, 6, 8 | 7, 5],
[4, 1| 2, 6, 8 | 7, 5],
[1, 4| 2, 8, 6 | 7, 5],
[4, 1| 2, 8, 6 | 7, 5],
[1, 4| 6, 2, 8 | 7, 5],
[4, 1| 6, 2, 8 | 7, 5],
[1, 4| 6, 8, 2 | 7, 5],
[4, 1| 6, 8, 2 | 7, 5],
[1, 4| 8, 6, 2 | 7, 5],
[4, 1| 8, 6, 2 | 7, 5],
[1, 4| 8, 2, 6 | 7, 5],
[4, 1| 8, 2, 6 | 7, 5]

Example 2. To illustrate the construction of sets E_1 and E_2, let $m = 2, k = 3, n = 6$. We note that in E_1 we have consecutive integers for each population. All possible non-permuted rank sequences, which are included in sets E_1 and E_2 are as follows:

E_1:	E_2:
[1, 2| 5, 6| 3, 4],	[1, 2| 5, 6| 3, 4],
[3, 4| 5, 6| 1, 2]	[1, 3| 5, 6| 2, 4],
	[1, 4| 5, 6| 2, 3],
	[2, 3| 5, 6| 1, 4],
	[2, 4| 5, 6| 1, 3],
	[3, 4| 5, 6| 1, 2]

We then must include all permutations within populations for each of the vectors above. We note that in E_1 we have consecutive integers for each population.

Example 3. In the case of unequal number of observations the extremal set will be more complicated. To construct the set E_1, first consider the order of the populations. Let $k = 5$, $m_1 = 5$, $m_2 = 3$, $m_3 = 7$, $m_4 = 11$, $m_5 = 9$, $n = 35$:

- Let j be a population rank in some subset, where $1 \le j \le k$. For any subset, the population rank $j = k$ may occupy only the $l = p$ group of the columns.
- Build the set E as the union of subsets; each of which has its own sequence of population ranks. Each sequence should follow the umbrella configuration. It may be made by selection of the $k-1$ population ranks from 1 to $k-1$ to the left $p-1$ groups of columns (the remaining $k-p$ population ranks will occupy the right $k-p$ groups of columns):

1,	2,	5,	4,	3
1,	3,	5,	4,	2
1,	4,	5,	3,	2
2,	3,	5,	4,	1
2,	4,	5,	3,	1
3,	4,	5,	2,	1

The cardinality of this set is

$$c_1 = \binom{k-1}{p-1} \tag{2.3}$$

- The length of the group of columns l is equal to m_l. For each subset, start from the population $j = 1$ and write in the group of columns l, occupied by this population, ranks from 1 to m_l. The population $j = 2$ will have ranks starting from $m_l + 1$. For our example, the ranks are as follows (the population ranks are in the brackets):

Population:	1	2	3	4	5
	1 – 5 (1)	6 – 8 (2)	29 – 35 (5)	18 – 28 (4)	9 – 17 (3)
	1 – 5 (1)	15 – 17 (3)	29 – 35 (5)	18 – 28 (4)	6 – 14 (2)
	1 – 5 (1)	26 – 28 (4)	29 – 35 (5)	15 – 25 (3)	6 – 14 (2)
	10 – 14 (2)	15 – 17 (3)	29 – 35 (5)	18 – 28 (4)	1 – 9 (1)
	10 – 14 (2)	26 – 28 (4)	29 – 35 (5)	15 – 25 (3)	1 – 9 (1)
	21 – 25 (3)	26 – 28 (4)	29 – 35 (5)	10 – 20 (2)	1 – 9 (1)

- Permute ranks in every population.

For the same reasons, as in the case of the set $\{\pi\}$, the cardinality of the set E is:

$$c_1 \prod_{l=1}^{k} m_l ! \tag{2.4}$$

Example 4. For set E_2, the subsets are constructed directly from the observation ranks. For any subset, the columns corresponding to the population p, will contain the ranks from $n - m_p + 1$ to n. The remaining $n - m_p$ ranks should be selected among n_{p-1} groups of columns, left from the peak of umbrella. The rank order in the right part of umbrella will be automatically determined. Therefore, the cardinality of set E_2:

$$c_2 = \binom{n - m_p}{n_{p-1}} \tag{2.5}$$

For the case of equal numbers of observations the cardinality is:

$$c_2 = \binom{(k-1)m}{(p-1)m} \tag{2.6}$$

The cardinality of the set E that includes permutations within each population is:

$$c_2 \prod_{l=1}^{k} m_l ! \tag{2.7}$$

2.4 Examples of distance functions

Some examples of distance functions are as follows:

- Spearman's Rho $d_s(\mu,\nu) = \frac{1}{2}\sum_{i=1}^{n}[\mu(i)-\nu(i)]^2$
- Kendall's Tau $d_k(\mu,\nu) = \sum_{1=i_1<i_2=n} \{1-\text{sgn}[\mu(i_1)-\mu(i_2)]\text{sgn}[\nu(i_1)-\nu(i_2)]\}$
- Spearman's Footrule $d_F(\mu,\nu) = \sum_{i=1}^{n}|\mu(i)-\nu(i)|$
- Hamming $d_H(\mu,\nu) = \sum_{i=1}^{n} I\left[\mu(i) \neq \nu(i)\right]$, where I is the indicator function.

All these distances satisfy the condition for right invariance, i.e.

$d(\mu,\nu) = d(\mu\tau,\nu\tau) \quad \forall \mu,\nu,\tau \in P$ and, consequently, do not depend on labeling of the objects ranked. In this thesis we will consider only the Spearman's Rho distance.

3. The test statistic

3.1 The common form of the distance

We start from the calculation of the common form of the Spearman-Rho distance, which is valid for both definitions of the set E.

$$d_s(\{\pi\},E)=\tfrac{1}{2}\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}d_s(\mu,\nu)=\tfrac{1}{2}\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\sum_{i=1}^{n}[\mu(i)-\nu(i)]^2$$

$$=\tfrac{1}{2}\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\sum_{i=1}^{n}[\mu^2(i)+\nu^2(i)]-\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\sum_{i=1}^{n}\mu(i)\nu(i) \tag{3.1}$$

$$=\tfrac{1}{2}\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\sum_{l=1}^{k}\sum_{i=n_{l-1}+1}^{n_l}[\mu^2(i)+\nu^2(i)]-\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\sum_{l=1}^{k}\sum_{i=n_{l-1}+1}^{n_l}\mu(i)\nu(i)=d_1-d_2$$

where

$$d_1=c\frac{n(n+1)(2n+1)}{6}\prod_{l=1}^{k}(m_l!)^2 \tag{3.2}$$

$$d_2=\sum_{l=1}^{k}\sum_{i=n_{l-1}+1}^{n_l}\left[\sum_{\mu\in\{\pi\}}\mu(i)\right]\left[\sum_{\nu\in E}\nu(i)\right] \tag{3.3}$$

and c is equal to either c_1 or c_2, depending on the choice of the set E.

Now, we consider the next sum for all columns corresponding to the population F_l:

$$\sum_{i=m_{l-1}+1}^{m_l}\sum_{\mu\in\{\pi\}}\sum_{\nu\in E}\mu(i)\nu(i) \tag{3.4}$$

Here $\mu(i) = \pi(i)$ and for any i from $n_{l-1}+1 \le i \le n_l$, the set $\{\pi\}$ contains $\prod_{l=1}^{k} m_l!$ ranks from $\pi(n_{l-1}+1)$ to $\pi(n_l)$ with equal occurrence of these ranks. Therefore, the number of each of these ranks in column i is $\frac{1}{m_l}\prod_{l=1}^{k} m_l!$. Similarly, for every subset g (which contains its own rank sequence following the umbrella configuration) with its rank exchanges, the frequency of occurrence of each rank from $v_g(n_{l-1}+1)$ to $v_g(n_l)$ in any column i is $\frac{1}{m_l}\prod_{l=1}^{k} m_l!$. Due to the rank exchange, every column $n_{l-1}+1 \le i \le n_l$ (there are m_l such columns) in the set $\{\pi\}$, as well as in the set E, contains the same ranks. In order to calculate (3.4), we note:

$$\sum_{i=m_{l-1}+1}^{m_l}\sum_{\mu\in\{\pi\}}\sum_{v\in E}\mu(i)v(i) = m_l\frac{1}{m_l}\prod_{l=1}^{k} m_l!\sum_{i=n_{l-1}+1}^{n_l}\pi(i)\frac{1}{m_l}\prod_{l=1}^{k} m_l!\sum_{g=1}^{c}\sum_{i=n_{l-1}+1}^{n_l} v_g(i)$$

$$= \frac{1}{m_l}\prod_{l=1}^{k}(m_l!)^2\sum_{i=n_{l-1}+1}^{n_l}\pi(i)\sum_{g=1}^{c}\sum_{i=n_{l-1}+1}^{n_l} v_g(i) \tag{3.5}$$

Consequently:

$$d_2 = \prod_{l=1}^{k}(m_l!)^2\sum_{l=1}^{k}\left[\frac{1}{m_l}\sum_{i=n_{l-1}+1}^{n_l}\pi(i)\sum_{g=1}^{c}\sum_{i=n_{l-1}+1}^{n_l} v_g(i)\right]$$

$$= \prod_{l=1}^{k}(m_l!)^2\sum_{l=1}^{k}\left[\frac{1}{m_l}\sum_{i=n_{l-1}+1}^{n_l}\pi(i)\sum_{g=1}^{c} v_{gl}\right] = \prod_{l=1}^{k}(m_l!)^2\sum_{l=1}^{k}\left[\overline{\pi}_l\sum_{g=1}^{c} v_{gl}\right] \tag{3.6}$$

where

$$v_{gl} = \sum_{i=n_{l-1}+1}^{n_l} v_g(i) \tag{3.7}$$

and

$$\overline{\pi}_l = \frac{1}{m_l}\sum_{i=n_{l-1}+1}^{n_l}\pi(i) \tag{3.8}$$

is the average of the ranks within the population F_l.

Expression (3.6) may be written as:

$$\sum_{i=1}^{n} h_i \pi(i) \tag{3.9}$$

where

$$h_i = \frac{1}{m_l} \sum_{g=1}^{c} v_{gl} \tag{3.10}$$

is a coefficient, depending on i. For any population F_l, all columns $n_{l-1}+1 \le i \le n_l$ in the sum (3.6) have the same coefficient, denoted as h_l.

We will use the form (3.9) in section 4, when asymptotic distributions will be discussed.

Note that:

$$\sum_{l=1}^{k} \sum_{g=1}^{c} v_{gl} = \sum_{g=1}^{c} \sum_{l=1}^{k} v_{gl} = c\frac{n(n+1)}{2} \tag{3.11}$$

Let $\bar{\pi}$ be the total rank average. Since

$$\bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} \pi(i) = \bar{v} = \frac{1}{n}\sum_{i=1}^{n} v(i) = \frac{n+1}{2} \tag{3.12}$$

it follows:

$$\begin{aligned} d_2 &= \prod_{l=1}^{k} (m_l!)^2 \left\{ \sum_{l=1}^{k} \left[\left(\bar{\pi}_l - \bar{\pi}\right) \sum_{g=1}^{c} v_{gl} \right] + \sum_{l=1}^{k} \bar{\pi} \sum_{g=1}^{c} v_{gl} \right\} \\ &= \prod_{l=1}^{k} (m_l!)^2 \left\{ \sum_{l=1}^{k} \left[\left(\bar{\pi}_l - \bar{\pi}\right) \sum_{g=1}^{c} v_{gl} \right] + c\frac{n(n+1)^2}{4} \right\} \end{aligned} \tag{3.13}$$

and the distance has the form:

$$
\begin{aligned}
d_s(\{\pi\},E) &= c\frac{n(n+1)(2n+1)}{6}\prod_{l=1}^{k}(m_l!)^2 \\
&-\prod_{l=1}^{k}(m_l!)^2\left\{\sum_{l=1}^{k}\left[\left(\overline{\pi}_l-\frac{n+1}{2}\right)\sum_{g=1}^{c}v_{gl}\right]+c\frac{n(n+1)^2}{4}\right\} \\
&= c\frac{n(n^2-1)}{12}\prod_{l=1}^{k}(m_l!)^2-\prod_{l=1}^{k}(m_l!)^2\sum_{l=1}^{k}\left[\left(\overline{\pi}_l-\frac{n+1}{2}\right)\sum_{g=1}^{c}v_{gl}\right] \\
&= c\frac{n(n^2-1)}{12}\prod_{l=1}^{k}(m_l!)^2-\prod_{l=1}^{k}(m_l!)^2 S
\end{aligned}
\tag{3.14}
$$

where

$$
S=\sum_{l=1}^{k}\left[\left(\overline{\pi}_l-\frac{n+1}{2}\right)\sum_{g=1}^{c}v_{gl}\right] \tag{3.15}
$$

Let $w(v(i),i)$ be the frequency of occurrence of rank $v(i)$ in the column i of the set E. Then:

$$
0\le w(v(i),i)\le c;\ \ \sum_{i=1}^{n}w(v(i),i)=c,\forall v(i);\ \ \sum_{v(i)}w(v(i),i)=c,\forall i \tag{3.16}
$$

Now, the expression $\sum_{g=1}^{c}v_{gl}$ in (3.15) may be written in a different form as:

$$
\sum_{g=1}^{c}v_{gl}=\sum_{g=1}^{c}\sum_{i=n_{l-1}+1}^{n_l}v_g(i)=\sum_{i=n_{l-1}+1}^{n_l}\sum_{v(i)}v(i)w(v(i),l) \tag{3.17}
$$

It should be underlined that unlike the ordered alternative, expression (3.15) cannot directly measure how "big" the distance between two sets is, if the location of the peak of umbrella p is unknown. For fixed m_l and k, the cardinality of the set E, and, therefore, the sum (3.17), depends on p. The maximal cardinality occurs, if the umbrella is symmetrical, the minimal cardinality corresponds to the ordered alternative.

3.2 The distance for the first approach

This approach was defined in section 2.2. We recall that two groups of columns in different parts of any subset of the set E are strongly separated by values of their ranks. Based on this approach, the value of $\sum_{i=n_{l-1}+1}^{n_l} \sum_{v(i)} v(i)w(v(i),l)$ will be found.

Theorem. For given j and l, the observation rank sequence $n_{l-1}+1,...,n_l$ in any subset of the set E is the same, i.e. it does not depend on the position of other population ranks.

Proof: Let the population rank j be on the l position and $l < p$. The populations whose ranks are smaller than j will occupy the first $l-1$ columns, as well as the last $j-1-(l-1)=j-l$ columns, starting from the $k-(j-l)+1$. The number of observation ranks in these populations is $\sum_{r=1}^{l-1} m_r + \sum_{r=k-j+l+1}^{k} m_r = n_{l-1}+n-n_{k-j+l}$, starting from rank 1. Therefore, the ranks in population j are from $n_{l-1}+n-n_{k-j+l}+1$ to $n_{l-1}+n-n_{k-j+l}+m_l = n_l+n-n_{k-j+l}$. This result does not depend on the sequence of the small population ranks.

Let $l > p$. The populations whose ranks exceed j will occupy the last $k-l$ columns, starting from $l+1$, and the first $j-1-(k-l)=j+l-k-1$ columns. The number of observation ranks in these populations is: $\sum_{r=l+1}^{k} m_r + \sum_{r=1}^{j+l-k-1} m_r = n_{j+l-k-1}+n-n_l$, starting from rank 1. Therefore, the ranks in population j are from $n_{j+l-k-1}+n-n_l+1$ to $n_{j+l-k-1}+n-n_l+m_l = n_{j+l-k-1}+n-n_{l-1}$. Again, this result does not depend on the sequence of the small populations.

Consequently, the sum of the observation ranks of population j in the columns corresponding to the population F_l, for any subset is:

for $l < p$

$$\frac{(n_{l-1}+n-n_{k-j+l}+1)+(n_{l-1}+n-n_{k-j+l}+m_l)}{2}m_l$$

$$=\left(n_l+n-n_{k-j+l}+\frac{1-m_l}{2}\right)m_l. \tag{3.18}$$

For $l > p$

$$\frac{(n_{j+l-k-1}+n-n_l+1)+(n_{j+l-k-1}+n-n_l+m_l)}{2}m_l$$

$$=\left(n_{j+l-k-1}+n-n_{l-1}+\frac{1-m_l}{2}\right)m_l, \tag{3.19}$$

whereas for $l = p$

$$\frac{(n-m_p+1)+n}{2}m_p=\left(n+\frac{1-m_p}{2}\right)m_p. \tag{3.20}$$

Note that (3.20) is a partial case of (3.18) and (3.19) for $j = k$ and $l = p$.

Let $u(j,l)$ be the number of occurrences of population rank j in the columns, corresponding to the population F_l in all subsets of the set E. Using (3.18), (3.19) and (3.20), we have:

$$\sum_{i=n_{l-1}+1}^{n_l}\sum_{v(i)}v(i)w(v(i),l)=c_1\frac{1-m_l}{2}m_l+\begin{Bmatrix}\sum_j u(j,l)(n_l+n-n_{k-j+l})m_l, & 1\le l<p\\ \sum_j u(j,l)(n_{j+l-k-1}+n-n_{l-1})m_l, & p<l\le k\\ c_1 n m_p, & l=p\end{Bmatrix} \tag{3.21}$$

Consequently, the expression (3.15) will be:

$$S(\{\pi\},E_1)=\sum_{l=1}^{p} m_l\left(\overline{\pi}_l-\frac{n+1}{2}\right)\left\{\sum_j\left(u(j,l)(n_l+n-n_{k-j+l})\right)+c_1\frac{1-m_l}{2}\right\}$$

$$+\sum_{l=p+1}^{k} m_l\left(\overline{\pi}_l-\frac{n+1}{2}\right)\left\{\sum_j\left(u(j,l)(n_{j+l-k-1}+n-n_{l-1})\right)+c_1\frac{1-m_l}{2}\right\} \qquad (3.22)$$

Some special cases are:

1. Ordered alternative. Here $c_1=1$, $u(j,l)=1$, $p=k$ and $j=l$. Therefore:

$$S(\{\pi\},E_1)=\sum_{l=1}^{p}\sum_{i=n_{l-1}+1}^{n_l}\left(\pi(i)-\frac{n+1}{2}\right)\left\{n_l+\frac{1-m_l}{2}\right\}=\tfrac{1}{2}\sum_{l=1}^{p}\sum_{i=n_{l-1}+1}^{n_l}\pi(i)\{n_l+n_{l-1}\} \qquad (3.23)$$

It is equivalent to the result obtained by Alvo and Pan (1996).

2. Equal numbers of observations. For this case expression (3.22) may be significantly simplified. We consider it in the next part.

3.3 The distance for the first approach: equal numbers of observations

If the numbers of observations are equal for all populations, $m_l = m, \forall l$ then $n_r = rm$.

Therefore, from (3.22):

$$S(\{\pi\}, E_1) = \sum_{l=1}^{k} m\left(\overline{\pi_l} - \frac{n+1}{2}\right)\left(\sum_j jmu(j,l) + c_1\frac{1-m}{2}\right)$$

$$= \sum_{l=1}^{k}\left(m\left(\overline{\pi_l} - \frac{n+1}{2}\right)\sum_j jmu(j,l)\right) \tag{3.24}$$

In order to evaluate this sum, we need to calculate the sum of population ranks in the columns corresponding to the population F_l for all subsets of the set E_1, which is equal to $\sum_j ju(j,l)$.

Our analysis starts from evaluation of $u(j,l)$.

Let $l < p$. If population rank j occupies the columns corresponding to the population F_l, for all possible subsets of the set E_1, the $k-j-1$ ranks from $j+1$ to $k-1$ should be allocated among $p-l-1$ groups of the columns corresponding to the populations between F_{l+1} and F_{p-1}. Independently, the $j-1$ population ranks from 1 to $j-1$ should be allocated among the $l-1$ groups of the columns corresponding to the populations between F_1 and F_{l-1}. The remaining ranks will be allocated automatically. Consequently, the frequency of occurrence of the rank population j in the columns corresponding to the population F_l in all subsets of the set E_1 is:

$$u(j,l) = \begin{cases} \binom{k-j-1}{p-l-1}\binom{j-1}{l-1}, & l \le j \le k-p+l \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le l \le p-1 \tag{3.25}$$

Now, we consider the case $l > p$. If population rank j occupies the columns corresponding to the population F_l, for all possible subsets of the set E_1, the $k-j-1$ ranks from $j+1$ to $k-1$ should be allocated among $l-p-1$ groups of the columns corresponding to the populations

between F_{p+1} and F_{l-1}. Independently, the $j-1$ population ranks from 1 to $j-1$ should be allocated among the $l-1$ groups of the columns corresponding to the populations between F_{l+1} and F_k. The remaining ranks will be allocated automatically. Consequently, the frequency of occurrence of the rank population j in the columns corresponding to the population F_l in all subsets of the set E_1 is:

$$u(j,l)=\left\{\begin{array}{ll}\binom{k-j-1}{l-p-1}\binom{j-1}{k-l}, & k-l+1\le j\le k-l+p\\ 0, & \text{otherwise}\end{array}\right\},\quad p+1\le l\le k \tag{3.26}$$

The expression $\sum_j ju(j,l)$ can be calculated using the identity in Feller (1972, p.65 exercise 14):

$$\sum_{j=0}^{k}\binom{a+k-j-1}{a-1}\binom{b+j-1}{b-1}=\binom{a+b+k-1}{a+b-1} \tag{3.27}$$

For $1\le l<p$, using (3.27):

$$\begin{aligned}&\sum_{j=l}^{k-p+l} j\binom{k-j-1}{p-l-1}\binom{j-1}{l-1}=\sum_{j=l}^{k-p+l} l\binom{k-j-1}{p-l-1}\binom{j}{l}=l\sum_{j=0}^{k-p}\binom{k-j-l-1}{p-l-1}\binom{j+l}{l}\\ &=l\binom{k}{p}=c_1 k\frac{l}{p}\end{aligned} \tag{3.28}$$

By the same reasoning for $p<l\le k$:

$$\begin{aligned}&\sum_{j=k-l+1}^{k-l+p} j\binom{k-j-1}{l-p-1}\binom{j-1}{k-l}=\sum_{j=k-l+1}^{k-l+p}(k-l+1)\binom{k-j-1}{l-p-1}\binom{j}{k-l+1}\\ &=(k-l+1)\sum_{j=0}^{p-1}\binom{l-j-2}{l-p-1}\binom{j+k-l+1}{k-l+1}=(k-l+1)\binom{k}{k-p+1}=c_1 k\frac{k-l+1}{k-p+1}\end{aligned} \tag{3.29}$$

For the peak $l = p$:

$$\sum_{j=k}^{k} jw(j,l) = c_1 k \tag{3.30}$$

The expression (3.30) will be included in (3.28) as a special case with $l = p$. The sum $\sum_j jw(j,l)$ is linear with respect to l.

Based on (3.28) and (3.29), the expression (3.22) is:

$$S(\{\pi\}, E_1) = c_1 n \left\{ \left(\sum_{l=1}^{p} \left(\overline{\pi}_l - \frac{n+1}{2} \right) \frac{l}{p} + \sum_{l=p+1}^{k} \left(\overline{\pi}_l - \frac{n+1}{2} \right) \frac{k-l+1}{k-p+1} \right) \right\} \tag{3.31}$$

The result is a linear rank statistic.

3.4 The distance for the second approach

For this approach, two groups of columns corresponding to two different two populations are mixed by values of their ranks. In this section the term "rank" is equivalent to the term "observation rank" and its symbol is j, $1 \le j \le n$. As in the previous section, in order to find

$$\sum_{i=n_{l-1}+1}^{n_l} \sum_{\nu(i)} \nu(i) w(\nu(i), i) = \sum_{i=n_{l-1}+1}^{n_l} \sum_{j} j w(j, i),$$

we need to evaluate $w(j,i)$.

Let rank j occupy column i. The ranks from $n-m_p+1$ to n are out of consideration since they occupy the same columns from $n_{p-1}+1$ to n_p corresponding to the population F_p, for all subsets of the set E.

Let $i \le n_{p-1}$. The first $j-1$ ranks should be allocated among $i-1$ columns, to the left of column i. Independently, the $n-m_p-j$ ranks from $j+1$ to $n-n_p$ should be allocated among $n_{p-1}-i$ columns between the columns $i+1$ and n_{p-1}. The remaining ranks will be allocated automatically. Consequently, the frequency of occurrence of rank j in column i in all subsets of the set E is:

$$w(j,i) = \begin{cases} \binom{n-m_p-j}{n_{p-1}-i}\binom{j-1}{i-1}, & i \le j \le n-n_p+i \\ 0, & \text{otherwise} \end{cases}, \quad i \le n_{p-1} \tag{3.32}$$

Now, we consider the case $i > n_p$. The $j-1$ ranks from 1 to $j-1$ should be allocated among $k-i$ columns, from the right of column i. Independently, the $n-m_p-j$ ranks from $j+1$ to $n-m_p$ should be allocated among $i-n_p-1$ columns between the columns n_p+1 and $i-1$. The remaining ranks will be allocated automatically. Consequently, the frequency of occurrence of rank j in column i in all subsets of the set E is:

$$w(j,i)=\left\{\begin{array}{ll}\binom{n-m_p-j}{i-n_p-1}\binom{j-1}{n-i}, & n-i+1\le j\le n+n_{p-1}-i+1\\ 0, & \text{otherwise}\end{array}\right\},\quad n_p<i \tag{3.33}$$

The value of $\sum_j jw(j,i)$ will be found by using (3.27) and (2.5)

For $i\le n_{p-1}$:

$$\sum_{j=i}^{n-n_p+i} j\binom{n-m_p-j}{n_{p-1}-i}\binom{j-1}{i-1}=\sum_{j=i}^{n-n_p+i} i\binom{n-m_p-j}{n_{p-1}-i}\binom{j}{i}=\sum_{j=0}^{n-n_p} i\binom{n-m_p-j-i}{n_{p-1}-i}\binom{j+i}{i} \tag{3.34}$$

$$=i\binom{n-m_p+1}{n_{p-1}+1}=(n-m_p+1)c_2\frac{i}{n_{p-1}+1}$$

For $i>n_p$:

$$\sum_{j=n-i+1}^{n+n_{p-1}-i+1} j\binom{n-m_p-j}{i-n_p-1}\binom{j-1}{n-i}=\sum_{j=n-i+1}^{n+n_{p-1}-i+1}(n-i+1)\binom{n-m_p-j}{i-n_p-1}\binom{j}{n-i+1}$$

$$=\sum_{j=0}^{n_{p-1}}(n-i+1)\binom{i-1-m_p-j}{i-n_p-1}\binom{j+n-i+1}{n-i+1}=(n-i+1)\binom{n-m_p+1}{n-n_p+1} \tag{3.35}$$

$$=(n-i+1)\frac{n-m_p+1}{n-n_p+1}\binom{n-m_p}{n-n_p}=(n-m_p+1)c_2\frac{n-i+1}{n-n_p+1}$$

For $n_{p-1}<i\le n_p$ (the columns belong to the peak population F_p)

$$c_2(i+n-n_p) \tag{3.36}$$

Note that (3.36) is a separate case and cannot be considered as a partial case of (3.34) or (3.35).

Consequently, the expression of $\sum_{i=n_{l-1}+1}^{n_l}\sum_j jw(j,i)$ is:

for $l < p$

$$\sum_{i=n_{l-1}+1}^{n_l}(n-m_p+1)c_2\frac{i}{n_{p-1}+1}=c_2\frac{n-m_p+1}{n_{p-1}+1}\frac{n_{l-1}+1+n_l}{2}m_l$$
$$=c_2\frac{n-m_p+1}{n_{p-1}+1}m_l\left(n_l+\frac{1-m_l}{2}\right). \tag{3.37}$$

For $l > p$

$$\sum_{i=n_{l-1}+1}^{n_l}(n-m_p+1)c_2\frac{n-i+1}{n-n_p+1}=c_2\frac{n-m_p+1}{n-n_p+1}\frac{n-n_{l-1}+n-n_l+1}{2}m_l$$
$$=c_2\frac{n-m_p+1}{n-n_p+1}m_l\left(n-n_l+m_l+\frac{1-m_l}{2}\right), \tag{3.38}$$

whereas for $l = p$

$$c_2\left(n+\frac{1-m_p}{2}\right)m_p. \tag{3.39}$$

Consequently, the expression (3.15) has a form:

$$S(\{\pi\},E_2)=c_2\left\{m_p\left(\bar{\pi}_p-\frac{n+1}{2}\right)\left[n+\frac{1-m_p}{2}\right]\right.$$
$$+\frac{n-m_p+1}{n_{p-1}+1}\sum_{l=1}^{p-1}\left[m_l\left(\bar{\pi}_l-\frac{n+1}{2}\right)\left(n_l+\frac{1-m_l}{2}\right)\right] \tag{3.40}$$
$$\left.+\frac{n-m_p+1}{n-n_p+1}\sum_{l=p+1}^{k}\left[m_l\left(\bar{\pi}_l-\frac{n+1}{2}\right)\left(n-n_l+m_l+\frac{1-m_l}{2}\right)\right]\right\}$$

Special cases of (3.40) are:

1. Ordered alternative. Here $c_2 = 1$ and $p = k$, $n - m_k = n_{k-1}$ and $\frac{n - m_k + 1}{n_{k-1} + 1} = 1$. Therefore:

$$S(\{\pi\}, E_2) = m_p\left(\overline{\pi}_p - \frac{n+1}{2}\right)\left[n + \frac{1-m_k}{2}\right]$$

$$+\frac{n - m_k + 1}{n_{k-1} + 1}\sum_{\substack{l=1 \\ l \neq p}}^{k-1}\left[m_l\left(\overline{\pi}_l - \frac{n+1}{2}\right)(n_l + \frac{1-m_l}{2})\right] = \tfrac{1}{2}\sum_{l=1}^{k}\left[m_l\overline{\pi}_l(n_l + n_{l-1})\right] \tag{3.41}$$

This result is equivalent to (3.23) since for the ordered alternative we do not have two parts of the set E and the two approaches of this set are equivalent.

2. Equal numbers of observations. Here c_2 is defined in (2.6), $m_l = m$ and the expression (3.40) is:

$$S(\{\pi\}, E_2) = c_2 m\left\{\left(\overline{\pi}_p - \frac{n+1}{2}\right)\left[km + \frac{1-m}{2}\right]\right.$$

$$+\frac{(k-1)m+1}{(p-1)m+1}\sum_{l=1}^{p-1}\left(\overline{\pi}_l - \frac{n+1}{2}\right)(lm + \frac{1-m}{2}) \tag{3.42}$$

$$\left.+\frac{(k-1)m+1}{(k-p)m+1}\sum_{l=p+1}^{k}\left(\overline{\pi}_l - \frac{n+1}{2}\right)(km - lm + m + \frac{1-m}{2})\right\}$$

4. Asymptotic distributions under the null hypothesis

4.1 Hajek-Sidak theorem for limiting null distributions

In the previous chapters we defined a test statistic for umbrella alternatives. In this chapter we will find its limiting distribution under the null hypothesis. The limiting distribution of a statistic T is defined as asymptotically normal with mean μ and variance σ^2,

i.e. $T \sim N(\mu,\sigma^2)$ as $n \to \infty$, if:

$$\frac{T-\mu}{\sigma} \xrightarrow{d} N(0,1);\ n \to \infty$$

where n is sample size.

The evaluation of the limiting distribution of the Spearman statistics is based on the theorem proposed by Hajek and Sidak. The theorem considers a subfamily of rank statistics, named simple linear rank statistic, which may be expressed in the form:

$$T = \sum_{i=1}^{n} h_i a_n(\pi(i)) \tag{4.1}$$

where $a_n(i)$ is an arbitrary score function and h_i are coefficients.

From the common expression of the distance (3.15), we can conclude that the Spearman-Rho distances for both of approaches are simple linear rank statistics. Put

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_i, \qquad \overline{a_n} = \frac{1}{n}\sum_{i=1}^{n} a_n(i) \tag{4.2}$$

It follows from Hajek and Sidak:

$$\mathrm{E}T = n\overline{a_n}\,\overline{h} \tag{4.3}$$

The next theorem, due to Hajek-Sidak (1999), (Theorem 1 in 6.1.6) evaluates the asymptotic distribution of the statistic (4.1), if the following condition is satisfied:

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{n}[h_i-\bar{h}]^2}{\max_{1<i<n}[h_i-\bar{h}]^2}=\infty \tag{4.4}$$

Theorem 4.1 *Let* $\varphi(u)$ *be a square integrable function, i.e.:*

$$\int_0^1[\varphi(u)-\bar{\varphi}]^2du<\infty, \qquad \bar{\varphi}=\int_0^1\varphi(u)du \tag{4.5}$$

and for simple linear rank statistic (4.1) and some square integrable function $\varphi(u)$,

$$\lim_{n\to\infty}\int_0^1\{\pi(1+[un])-\varphi(u)\}^2du=0 \tag{4.6}$$

holds.

Then, under H_0 *and (4.6), the statistic (4.1) for (4.4) is asymptotically normal* (μ,σ^2) *with*

$$\mu=n\bar{h}\bar{a} \tag{4.7}$$

$$\sigma^2=\sum_{i=1}^{n}[h_i-\bar{h}]^2\int_0^1[\varphi(u)-\bar{\varphi}]^2du \tag{4.8}$$

The expression $[un]$ in (4.6) means the greatest integer less than or equal to un. The function $\varphi(u)$ may be taken as:

$$\varphi(u)=\begin{cases}u, & \text{if } 0\le u\le 1\\ 0, & \text{otherwise.}\end{cases} \tag{4.9}$$

In this case $\bar{\varphi}=\int_0^1\varphi(u)du=\frac{1}{2}$ and $\int_0^1[\varphi(u)-\bar{\varphi}]^2du=\frac{1}{12}$. By lemma 1 (Hajek, Sidak and Sen (1999), page 195), the condition (4.6) is satisfied, if

$$a_n(i)=\varphi(\frac{i}{n+1}) \tag{4.10}$$

We use this theorem for evaluation of the asymptotic null distribution of the Spearman-Rho statistic for both approaches.

4.2 Asymptotic distribution of the Spearman-Rho statistic: first approach

Let

$$T_1 = \frac{S(\{\pi\}, E_1)}{n(n+1)c_1} = \frac{1}{n(n+1)c_1}\sum_{l=1}^{k} m_l\left(\overline{\pi}_l - \frac{n+1}{2}\right)h_i$$
$$= \frac{1}{n(n+1)c_1}\sum_{l=1}^{k}\sum_{i=n_{l-1}+1}^{n_l}\left(\pi(i) - \frac{n+1}{2}\right)h_i \tag{4.11}$$

Then, $T_1 \equiv S(\{\pi\}, E_1)$, where the symbol "$\equiv$" means that both terms differ only by some constants, depending on sample size n. From (4.9) and (4.10), the score function is:

$$a_n(i) = \frac{i - \overline{\pi}}{n+1} = \frac{i}{n+1} - \frac{1}{2} \tag{4.12}$$

The coefficients h_i are obtained from (3.22):

$$h_i = \left\{ \begin{array}{ll} \frac{1}{nc_1}\left\{\sum_j \left(u(j,l)(n_l + n - n_{k-j+l})\right) + c_1\frac{1-m_l}{2}\right\}, & m_{l-1}+1 \le i \le m_l;\ 1 \le l \le p \\ \frac{1}{nc_1}\left\{\sum_j \left(u(j,l)(n_{j+l-k-1} + n - n_{l-1})\right) + c_1\frac{1-m_l}{2}\right\}, & m_{l-1}+1 \le i \le m_l;\ p<l \le k \end{array} \right\} \tag{4.13}$$

Then

$$T_1 = \sum_{i=1}^{n} h_i a_n(\pi(i)) \tag{4.14}$$

is a simple linear rank statistic and condition (4.6) of theorem 4.1 is satisfied.

From (4.13) we have:

$$h_i < \frac{1}{nc_1}\sum_j u(j,l)(n+n) = O(1) \tag{4.15}$$

The expression (4.4) may be written as:

$$\frac{\sum_{i=1}^{n}[h_i-\bar{h}]^2}{\max_{1<i<n}[h_i-\bar{h}]^2}=\frac{\sum_{l=1}^{k}m_l[h_l-\bar{h}]^2}{O(1)}\geq\frac{\min(m_l)\sum_{l=1}^{k}[h_l-\bar{h}]^2}{O(1)} \tag{4.16}$$

where

$$h_l=h_i,\ n_{l-1}<i\leq n_l \tag{4.17}$$

Consequently:

$$\lim_{\min(m_l)\to\infty}\frac{\sum_{i=1}^{n}[h_i-\bar{h}]^2}{\max_{1<i<n}[h_i-\bar{h}]^2}=\infty \tag{4.18}$$

Thus, theorem 4.1 is applicable to the statistic (4.11), if the minimal sample size is big enough. Due to the fact that

$$\overline{a_n}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n+1}-\frac{1}{2}\right)=0 \tag{4.19}$$

we conclude that $\mu=0$ (4.20)

The variance of the statistic (4.11) may be found from (4.8) and (4.9):

$$\sigma^2=\frac{1}{12}\sum_{i=1}^{n}[h_i-\bar{h}]^2 \tag{4.21}$$

In general case, the expression (4.21) does not have a closed form.

Now, consider the partial case of the statistic T_1 for equal numbers of observations. From (3.31) we have:

$$T_1 = \frac{m}{n+1}\left(\sum_{l=1}^{p}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\frac{l}{p} + \sum_{l=p+1}^{k}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\frac{k-l+1}{k-p+1}\right)$$

$$= \frac{1}{n+1}\left(\sum_{l=1}^{p}\sum_{i=m(l-1)+1}^{ml}\left(\pi(i) - \frac{n+1}{2}\right)\frac{l}{p} + \sum_{l=p+1}^{k}\sum_{i=m(l-1)+1}^{ml}\left(\pi(i) - \frac{n+1}{2}\right)\frac{k-l+1}{k-p+1}\right) \quad (4.22)$$

$$= \sum_{i=1}^{n} h_i a_n(\pi(i))$$

where $a_n(i)$ is defined in (4.12) and:

$$h_i = \begin{cases} \dfrac{l}{p}, & 1 \le l \le p,\ m(l-1) < i \le ml \\ \dfrac{k-l+1}{k-p+1}, & p < l \le k,\ m(l-1) < i \le ml \end{cases} \quad (4.23)$$

All properties proved for the statistic (4.11) are valid for the statistic (4.22). From (4.2):

$$\overline{h} = \frac{1}{n}\left(m\sum_{l=1}^{p}\frac{l}{p} + m\sum_{l=p+1}^{k}\frac{k-l+1}{k-p+1}\right) = \frac{1}{k}\left(\sum_{l=1}^{p}\frac{l}{p} + \sum_{l=p}^{k-p}\frac{l}{k-p+1}\right) = \frac{k+1}{2k} \quad (4.24)$$

$$\sum_{i=1}^{n} h_i^2 = m\left(\frac{\sum_{l=1}^{p} l^2}{p^2} + \frac{\sum_{l=1}^{k-p} l^2}{(k-p+1)^2}\right) \quad (4.25)$$

$$= m\left(\frac{(p+1)(2p+1)}{6p} + \frac{(k-p)(2k-2p+1)}{6(k-p+1)}\right)$$

And the variance is:

$$\sigma^2 = \frac{1}{12}\sum_{i=1}^{n}[h_i - \bar{h}]^2 = \frac{1}{12}\left[\sum_{i=1}^{n} h_i^2 - km(\bar{h})^2\right]$$

$$= \frac{m}{12}\left(\frac{(p+1)(2p+1)}{6p} + \frac{(k-p)(2k-2p+1)}{6(k-p+1)} - \frac{(k+1)^2}{4k}\right) \tag{4.26}$$

4.3 Asymptotic distribution of the Spearman-Rho statistic: second approach

Let

$$T_2 = \frac{S(\{\pi\}, E_2)}{n(n+1)c_2} = \frac{1}{n(n+1)} \sum_{l=1}^{k} m_l \left(\overline{\pi}_l - \frac{n+1}{2} \right) h_i = \frac{1}{n(n+1)} \sum_{l=1}^{k} \sum_{i=n_{l-1}+1}^{n_l} \left(\pi(i) - \frac{n+1}{2} \right) h_i \quad (4.27)$$

where:

$$h_i = \begin{Bmatrix} \frac{1}{n} \frac{n-m_p+1}{n_{p-1}+1} \left(n_l + \frac{1-m_l}{2} \right), & 1 \le l < p, \; n_{l-1} < i \le n_l \\ \frac{1}{n} \frac{n-m_p+1}{n-n_p+1} \left(n - n_l + m_l + \frac{1-m_l}{2} \right), & p < l \le k, \; n_{l-1} < i \le n_l \\ \frac{1}{n} \left(n + \frac{1-m_p}{2} \right), & l = p, \quad n_{l-1} < i \le n_l \end{Bmatrix} \quad (4.28)$$

and the score function is the same, as in (4.12). Then, by (3.40), $T_2 \equiv S(\{\pi\}, E_2)$. The statistic T_2 differs from the statistic T_1 only by coefficients h_i. The values of h_i satisfy the following restrictions:

$$h_i = \frac{1}{n} \frac{n-m_p+1}{n_{p-1}+1} \left(n_l + \frac{1-m_l}{2} \right) < \frac{n-m_p+1}{n} \frac{n_l}{n_{p-1}+1} < 1, \qquad 1 \le l < p$$

$$h_i = \frac{1}{n} \frac{n-m_p+1}{n-n_p+1} \left(n - n_l + m_l + \frac{1-m_l}{2} \right) < \frac{n-m_p+1}{n} \frac{n-n_l+m_l}{n-n_p+1} < 1, \quad p < l \le k \quad (4.29)$$

$$h_i = \frac{1}{n} \left(n + \frac{1-m_p}{2} \right) < 1, \qquad l = p$$

Thus, the expressions (4.16) and (4.18) are valid for the statistic T_2. Therefore, the Hajek – Sidak theorem may be applied to this statistic, if the minimal sample size is big enough. The variance

does not have a closed form and may be found from (4.21). The result may be simplified, if we consider the partial case of the statistic T_2 for equal numbers of observations. From (3.42), the coefficients h_i are:

$$h_i = \begin{Bmatrix} \dfrac{(k-1)m+1}{(p-1)m+1}\left(\dfrac{l}{k}+\dfrac{1-m}{2n}\right), & 1 \le l < p,\ m(l-1) < i \le ml \\ \dfrac{(k-1)m+1}{(k-p)m+1}\left(\dfrac{k-l+1}{k}+\dfrac{1-m}{2n}\right), & p < l \le k,\ m(l-1) < i \le ml \\ 1+\dfrac{1-m}{2n}, & l = p,\quad m(l-1) < i \le ml \end{Bmatrix} \tag{4.30}$$

In order to find the variance of T_2, we need to perform the same calculations.

$$\begin{aligned} \sum_{l=1}^{k} h_l &= \frac{1}{n}\sum_{l=1}^{p-1}\left\{lm+\frac{1-m}{2}\right\}\frac{(k-1)m+1}{(p-1)m+1}+\frac{1}{n}\sum_{l=1}^{k-p}\left\{lm+\frac{1-m}{2}\right\}\frac{(k-1)m+1}{(k-p)m+1} \\ &+\frac{1}{n}\left(km+\frac{1-m}{2}\right)=\frac{1}{n}\frac{(k-1)m+1}{(p-1)m+1}\frac{\dfrac{m+1}{2}+\dfrac{(2p-3)m+1}{2}}{2}(p-1) \\ &+\frac{1}{n}\frac{(k-1)m+1}{(k-p)m+1}\frac{\dfrac{m+1}{2}+\dfrac{(2k-2p-1)m+1}{2}}{2}(k-p)+\frac{1}{n}\left(km+\frac{1-m}{2}\right) \\ &=\frac{1}{n}\left((k-1)m+1\right)\left[\frac{p-1}{2}+\frac{k-p}{2}\right]+\frac{1}{2}\left((k-1)m+1+km\right)=\frac{k(n+1)}{2n} \end{aligned} \tag{4.31}$$

Therefore,

$$\bar{h} = \frac{n+1}{2n} \tag{4.32}$$

The variance of T_2 is:

$$\sigma^2 = \frac{1}{12}\sum_{i=1}^{n}[h_i - \bar{h}]^2 = \frac{1}{12}\sum_{i=1}^{n} h_i^2 - km(\bar{h})^2 = \frac{m}{12}\left[s_1^2 + s_2^2 + s_3^2 - k(\bar{h})^2\right] \tag{4.33}$$

where:

$$s_1^2 = \left(\frac{(k-1)m+1}{(p-1)m+1}\right)^2\left[\frac{p(p-1)(2p-1)}{6k^2} + \frac{1-m}{2n}\frac{p(p-1)}{k} + \left(\frac{1-m}{2n}\right)^2(p-1)\right] \tag{4.34}$$

$$s_2^2 = \left(\frac{(k-1)m+1}{(k-p)m+1}\right)^2\{\frac{(k-p)(k-p+1)(2k-2p+1)}{6k^2}$$

$$+\frac{1-m}{2n}\frac{(k-p)(k-p+1)}{k} + \left(\frac{1-m}{2n}\right)^2(k-p)\} \tag{4.35}$$

$$s_3^2 = \left[1 + \frac{1-m}{2n}\right]^2 \tag{4.36}$$

5. Simulation

5.1 Objective

In the previous chapter we determined the asymptotic distribution of our test statistic under the null distribution based on the Spearman distance. In order to determine the small sample size performance, we will perform a simulation study. The main goals of the simulation are:

- Estimation of peak location, if unknown;
- Evaluation of the power of the test.
- Evaluation of the Type 1 error.

In order to perform our analyses, we generated random samples from the same distributions, whose population medians have umbrella configuration with known peak. We consider four different distributions:

- Normal. The normal random samples were generated directly;
- Exponential. The exponential random samples were generated directly;
- Double exponential. Uniform random samples between 0 and 1 were generated. They were converted to be double exponentially distributed as follows:

$$y = \begin{cases} \ln(2u), & u < 0.5 \\ -\ln(2(1-u)), & u \geq 0.5 \end{cases} \tag{5.1}$$

where u is uniformly distributed.

In order to obtain $x \sim \text{dexp}(mean_i, 1)$, the next transformation was used:

$$x_i = mean_i + \frac{y}{\sqrt{2}} \tag{5.2}$$

- Logistic. Uniform random samples between 0 and 1 were generated. They were converted to be logistic distributed as follows:

$$y = \ln \frac{u}{1-u}, \tag{5.3}$$

where u is uniformly distributed.

In order to obtain $x \sim \text{logistic}(mean_l, 1)$, the next transformation was used:

$$x_l = mean_l + y\frac{\sqrt{3}}{\pi} \tag{5.4}$$

We start our simulation for the partial case of equal numbers of observations. For the first approach, the simulation was performed for the statistic:

$$T_1 = \frac{m}{n+1}\left\{\sum_{l=1}^{p}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\frac{l}{p} + \sum_{l=p+1}^{k}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\frac{k-l+1}{k-p+1}\right\} \tag{5.5}$$

For the second approach, the simulation was performed for the statistic:

$$T_2 = \frac{m}{n+1}\left\{\left(\overline{\pi}_p - \frac{n+1}{2}\right)\left[km + \frac{1-m}{2}\right] + \frac{(k-1)m+1}{(p-1)m+1}\sum_{l=1}^{p-1}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\left(lm + \frac{1-m}{2}\right) + \right.$$

$$\left. + \frac{(k-1)m+1}{(k-p)m+1}\sum_{l=p+1}^{k}\left(\overline{\pi}_l - \frac{n+1}{2}\right)\left(km - lm + m + \frac{1-m}{2}\right)\right\} \tag{5.6}$$

The procedure of simulation is as follows:

For the four separate distributions of random data, we consider values of $m = 3,...,15$ and 10000 simulations:

- Generate random data with means $0, 0.5, 1, 0.5, 0$ when $k = 5$, $p = 3$, variance $= 1$ and equal number of observations m;
- Find the peak location. The statistics (5.5) and (5.6) are applied for all possible locations of the peak (from 1 to k), i.e. as for the case of unknown peak. Among k distances, corresponding to k possible peak locations, the minimal distance, i.e. maximal value of statistics (5.6) and (5.6) points to the peak. The corresponding counter is incremented by 1.

In the cases of ties, i.e. non-unique maximal distance (it may happen especially for small values of m, because the statistic is based on ranks), the corresponding counters are incremented by $1/\text{number of ties}$. Dividing the result by 10000, for any m, we obtain k probabilities of the peak location at the corresponding population. The probability for $l = 3$ (true peak location), which is compared with the probabilities for other l (false peak locations), describes the validity of the test.

- Obtain the power of the test. Define the power function as the probability of rejection H_0 in favor to H_1 for the corresponding l. Assuming that the limiting distribution of the rank statistics is normal, for given $\alpha = 0.05$, the power function, which depends on m and l, is equal to the relative number of cases when H_1 was accepted, i.e.

  $\{\text{number of } (\frac{T_n - \mu_n}{\sigma_n} > 1.645)\}/10000$, where T_n is defined in (5.5) and (5.6). The power of the test is equal to the power function for $l = 3$ (true peak location). We want to have high power for large values of m. In addition, this power should be significantly higher, than the values of the power function for other l. These values also grow with increasing values of m.

Repeat the next steps 10000 times for evaluation of the Type 1 error:

- Generate random data with the same means 0, when $k = 5$, variance $= 1$ and equal number of observations m;
- Obtain the Type 1 error, i.e. the value of the power function.

5.2 Results for choosing of the peak location

The probabilities for estimation of the peak locations for various m are placed in tables 5.1 – 5.4. Each table corresponds to a different distribution. The true peak is located at F_3.

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.1160	0.1819	0.4132	0.1804	0.1083	0.1151	0.1740	0.4329	0.1716	0.1063
4	0.0843	0.1789	0.4638	0.1832	0.0896	0.0822	0.1668	0.4922	0.1710	0.0877
5	0.0711	0.1703	0.5183	0.1695	0.0706	0.0686	0.1551	0.5537	0.1538	0.0687
6	0.0600	0.1609	0.5658	0.1567	0.0563	0.0579	0.1440	0.6022	0.1416	0.0543
7	0.0437	0.1524	0.6051	0.1544	0.0443	0.0418	0.1332	0.6484	0.1340	0.0425
8	0.0354	0.1478	0.6334	0.1452	0.0381	0.0341	0.1283	0.6763	0.1262	0.0351
9	0.0312	0.1379	0.6699	0.1305	0.0304	0.0291	0.1185	0.7113	0.1124	0.0286
10	0.0221	0.1280	0.6982	0.1248	0.0268	0.0202	0.1078	0.7431	0.1050	0.0239
11	0.0205	0.1200	0.7265	0.1147	0.0181	0.0194	0.1003	0.7696	0.0930	0.0175
12	0.0147	0.1077	0.7480	0.1122	0.0173	0.0141	0.0871	0.7899	0.0924	0.0165
13	0.0144	0.1047	0.7611	0.1065	0.0132	0.0130	0.0859	0.8031	0.0863	0.0117
14	0.0127	0.0987	0.7786	0.0973	0.0126	0.0118	0.0764	0.8254	0.0757	0.0107
15	0.0095	0.0905	0.7989	0.0919	0.0090	0.0083	0.0687	0.8410	0.0739	0.0080

Table 5.1. Probability of choosing the peak for each population for normal random data.

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0838	0.1756	0.4786	0.1729	0.0889	0.0822	0.1658	0.5029	0.1620	0.0870
4	0.0667	0.1651	0.5484	0.1596	0.0599	0.0642	0.1500	0.5816	0.1465	0.0575
5	0.0473	0.1566	0.5946	0.1524	0.0490	0.0454	0.1433	0.6304	0.1342	0.0466
6	0.0377	0.1376	0.6530	0.1372	0.0343	0.0349	0.1193	0.6943	0.1181	0.0333
7	0.0300	0.1310	0.6891	0.1207	0.0291	0.0283	0.1108	0.7301	0.1040	0.0267
8	0.0253	0.1150	0.7178	0.1207	0.0211	0.0225	0.0956	0.7625	0.0998	0.0196
9	0.0168	0.1107	0.7445	0.1131	0.0149	0.0153	0.0872	0.7943	0.0898	0.0134
10	0.0124	0.0997	0.7768	0.0987	0.0122	0.0110	0.0756	0.8260	0.0758	0.0116
11	0.0082	0.0906	0.7980	0.0917	0.0114	0.0077	0.0697	0.8414	0.0715	0.0097
12	0.0082	0.0837	0.8203	0.0812	0.0066	0.0064	0.0634	0.8643	0.0605	0.0054
13	0.0064	0.0786	0.8357	0.0728	0.0065	0.0057	0.0573	0.8768	0.0543	0.0059
14	0.0036	0.0715	0.8511	0.0692	0.0046	0.0032	0.0505	0.8938	0.0488	0.0037
15	0.0044	0.0613	0.8639	0.0674	0.0029	0.0037	0.0445	0.9026	0.0469	0.0023

Table 5.2. Probability of choosing the peak for each population for double exponential data.

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.1015	0.1731	0.4424	0.1799	0.1028	0.1006	0.1636	0.4647	0.1692	0.1017
4	0.0802	0.1739	0.4960	0.1766	0.0731	0.0785	0.1623	0.5229	0.1655	0.0707
5	0.0601	0.1644	0.5546	0.1620	0.0587	0.0575	0.1502	0.5897	0.1459	0.0566
6	0.0494	0.1555	0.5936	0.1569	0.0443	0.0473	0.1373	0.6326	0.1410	0.0416
7	0.0353	0.1471	0.6318	0.1458	0.0399	0.0332	0.1254	0.6786	0.1251	0.0377
8	0.0310	0.1368	0.6631	0.1356	0.0333	0.0280	0.1202	0.7044	0.1164	0.0309
9	0.0250	0.1224	0.6994	0.1262	0.0268	0.0240	0.1053	0.7404	0.1060	0.0243
10	0.0199	0.1154	0.7279	0.1167	0.0200	0.0180	0.0963	0.7705	0.0975	0.0177
11	0.0152	0.1108	0.7475	0.1118	0.0146	0.0141	0.0886	0.7943	0.0903	0.0127
12	0.0130	0.1010	0.7714	0.1006	0.0138	0.0111	0.0793	0.8149	0.0823	0.0124
13	0.0103	0.0927	0.7921	0.0930	0.0118	0.0094	0.0741	0.8319	0.0747	0.0099
14	0.0078	0.0901	0.8132	0.0810	0.0078	0.0068	0.0683	0.8540	0.0641	0.0068
15	0.0068	0.0817	0.8258	0.0802	0.0055	0.0059	0.0612	0.8696	0.0582	0.0051

Table 5.3. Probability of choosing the peak for each population for logistic data.

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0836	0.1527	0.5292	0.1474	0.0869	0.0813	0.1404	0.5582	0.1351	0.0842
4	0.0630	0.1321	0.6130	0.1343	0.0574	0.0607	0.1188	0.6470	0.1184	0.0549
5	0.0440	0.1155	0.6775	0.1172	0.0456	0.0415	0.1006	0.7131	0.1013	0.0434
6	0.0341	0.1022	0.7252	0.1031	0.0353	0.0316	0.0846	0.7621	0.0892	0.0325
7	0.0235	0.0857	0.7747	0.0928	0.0231	0.0215	0.0693	0.8116	0.0760	0.0215
8	0.0174	0.0786	0.8054	0.0789	0.0196	0.0160	0.0623	0.8422	0.0623	0.0172
9	0.0126	0.0671	0.8354	0.0714	0.0133	0.0108	0.0516	0.8707	0.0555	0.0114
10	0.0129	0.0617	0.8544	0.0601	0.0108	0.0114	0.0483	0.8878	0.0434	0.0091
11	0.0078	0.0567	0.8764	0.0526	0.0064	0.0073	0.0422	0.9059	0.0386	0.0060
12	0.0070	0.0445	0.8959	0.0481	0.0045	0.0059	0.0325	0.9233	0.0351	0.0032
13	0.0049	0.0425	0.9071	0.0405	0.0048	0.0043	0.0278	0.9360	0.0274	0.0045
14	0.0025	0.0382	0.9225	0.0340	0.0028	0.0020	0.0267	0.9454	0.0234	0.0025
15	0.0024	0.0329	0.9279	0.0338	0.0029	0.0019	0.0207	0.9531	0.0218	0.0025

Table 5.4. Probability of choosing the peak for each population for exponential data.

The conclusions from choosing of the peak location are:

- In all tables, while m is increasing, the probability of choosing the true peak location becomes significantly bigger, than other probabilities of choosing the false peak locations. As it was expected, the second approach, which included more possible rank sequences following the umbrella alternative, gives better results than the first one and it may be recommended as the test for choosing the peak location.
- The high probability of choosing of the true peak location for exponential data may be caused by poor overlapping among ranks from different populations. The possible reason is the fact that the exponential distribution is non-symmetrical;
- Among symmetrical distributions, the test is more precise, if the data has a double-exponential distribution. The worst case is for the normal distribution.

5.3 The values of the power of the test

Tables 5.5 – 5.8 contain the power function (probability of rejecting H_0 when the peak is assumed to be at population F_i). Each table corresponds to a different distribution. The values at column 3 indicate the power of the test.

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0374	0.1756	0.3356	0.1691	0.0333	0.0374	0.1875	0.3530	0.1830	0.0333
4	0.0357	0.2238	0.4452	0.2280	0.0337	0.0357	0.2307	0.4469	0.2363	0.0337
5	0.0394	0.2677	0.5197	0.2627	0.0355	0.0394	0.2729	0.5203	0.2668	0.0355
6	0.0374	0.3008	0.6001	0.3027	0.0371	0.0374	0.3129	0.5999	0.3155	0.0371
7	0.0392	0.3459	0.6667	0.3394	0.0330	0.0392	0.3620	0.6625	0.3516	0.0330
8	0.0363	0.3854	0.7161	0.3807	0.0426	0.0363	0.3983	0.7157	0.3943	0.0426
9	0.0364	0.4234	0.7709	0.4156	0.0400	0.0364	0.4385	0.7673	0.4306	0.0400
10	0.0376	0.4587	0.8111	0.4570	0.0364	0.0376	0.4765	0.8088	0.4714	0.0364
11	0.0400	0.4898	0.8481	0.4880	0.0380	0.0400	0.5047	0.8455	0.5079	0.0380
12	0.0344	0.5184	0.8695	0.5300	0.0417	0.0344	0.5373	0.8658	0.5446	0.0417
13	0.0377	0.5542	0.8992	0.5469	0.0398	0.0377	0.5705	0.8974	0.5676	0.0398
14	0.0387	0.5800	0.9083	0.5688	0.0372	0.0387	0.5986	0.9072	0.5885	0.0372
15	0.0404	0.6006	0.9346	0.6064	0.0383	0.0404	0.6225	0.9318	0.6255	0.0383

Table 5.5. Power function for normally distributed data

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0310	0.2127	0.4349	0.2095	0.0321	0.0310	0.2279	0.4493	0.2259	0.0321
4	0.0286	0.2856	0.5612	0.2740	0.0296	0.0286	0.2923	0.5614	0.2818	0.0296
5	0.0335	0.3347	0.6358	0.3338	0.0331	0.0335	0.3402	0.6387	0.3408	0.0331
6	0.0337	0.3856	0.7252	0.3834	0.0333	0.0337	0.4004	0.7220	0.3983	0.0333
7	0.0373	0.4365	0.7879	0.4229	0.0298	0.0373	0.4542	0.7844	0.4411	0.0298
8	0.0343	0.4803	0.8294	0.4802	0.0359	0.0343	0.4966	0.8286	0.4949	0.0359
9	0.0329	0.5204	0.8743	0.5252	0.0354	0.0329	0.5398	0.8718	0.5458	0.0354
10	0.0370	0.5688	0.9051	0.5620	0.0323	0.0370	0.5880	0.9045	0.5801	0.0323
11	0.0350	0.6016	0.9254	0.6034	0.0380	0.0350	0.6208	0.9262	0.6255	0.0380
12	0.0356	0.6356	0.9458	0.6317	0.0343	0.0356	0.6563	0.9445	0.6516	0.0343
13	0.0378	0.6731	0.9588	0.6675	0.0349	0.0378	0.6925	0.9577	0.6866	0.0349
14	0.0355	0.7002	0.9693	0.7033	0.0360	0.0355	0.7199	0.9688	0.7219	0.0360
15	0.0337	0.7217	0.9779	0.7249	0.0353	0.0337	0.7433	0.9767	0.7452	0.0353

Table 5.6. Power function for double-exponentially distributed data

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0320	0.1812	0.3729	0.1886	0.0359	0.0320	0.1959	0.3877	0.2018	0.0359
4	0.0341	0.2394	0.4764	0.2432	0.0355	0.0341	0.2465	0.4765	0.2514	0.0355
5	0.0351	0.2947	0.5666	0.2916	0.0322	0.0351	0.2985	0.5679	0.2956	0.0322
6	0.0358	0.3308	0.6447	0.3353	0.0359	0.0358	0.3427	0.6433	0.3484	0.0359
7	0.0339	0.3758	0.7171	0.3772	0.0318	0.0339	0.3913	0.7133	0.3918	0.0318
8	0.0377	0.4174	0.7592	0.4226	0.0371	0.0377	0.4318	0.7587	0.4357	0.0371
9	0.0400	0.4506	0.8106	0.4533	0.0361	0.0400	0.4670	0.8087	0.4703	0.0361
10	0.0345	0.4876	0.8474	0.4903	0.0361	0.0345	0.5012	0.8450	0.5078	0.0361
11	0.0374	0.5285	0.8734	0.5167	0.0362	0.0374	0.5483	0.8719	0.5357	0.0362
12	0.0367	0.5553	0.9018	0.5618	0.0384	0.0367	0.5751	0.9017	0.5780	0.0384
13	0.0368	0.5927	0.9216	0.5915	0.0367	0.0368	0.6116	0.9202	0.6092	0.0367
14	0.0387	0.6192	0.9355	0.6190	0.0372	0.0387	0.6394	0.9343	0.6382	0.0372
15	0.0347	0.6464	0.9528	0.6483	0.0361	0.0347	0.6666	0.9510	0.6677	0.0361

Table 5.7. Power function for logistic data

	The first approach					The second approach				
m	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
3	0.0438	0.2616	0.5216	0.2627	0.0463	0.0438	0.2818	0.5391	0.2821	0.0463
4	0.0436	0.3398	0.6605	0.3492	0.0428	0.0436	0.3493	0.6622	0.3601	0.0428
5	0.0443	0.4115	0.7608	0.4080	0.0475	0.0443	0.4198	0.7599	0.4178	0.0475
6	0.0463	0.4670	0.8322	0.4591	0.0478	0.0463	0.4832	0.8273	0.4763	0.0478
7	0.0469	0.5158	0.8822	0.5245	0.0449	0.0469	0.5364	0.8766	0.5436	0.0449
8	0.0505	0.5758	0.9169	0.5708	0.0494	0.0505	0.5920	0.9123	0.5898	0.0494
9	0.0468	0.6201	0.9501	0.6243	0.0465	0.0468	0.6379	0.9461	0.6448	0.0465
10	0.0505	0.6576	0.9621	0.6638	0.0483	0.0505	0.6785	0.9583	0.6812	0.0483
11	0.0505	0.6977	0.9734	0.6941	0.0513	0.0505	0.7184	0.9701	0.7151	0.0513
12	0.0497	0.7304	0.9846	0.7368	0.0511	0.0497	0.7499	0.9820	0.7554	0.0511
13	0.0477	0.7752	0.9896	0.7706	0.0504	0.0477	0.7923	0.9880	0.7890	0.0504
14	0.0494	0.7955	0.9939	0.7962	0.0473	0.0494	0.8141	0.9925	0.8114	0.0473
15	0.0494	0.8177	0.9947	0.8277	0.0498	0.0494	0.8344	0.9938	0.8430	0.0498

Table 5.8. Power function for exponentially distributed data.

The conclusions from evaluation of power of the test are:

- The possible problem for evaluation of the power of the test, which occurs especially for large values of m, is as follows: the hypothesis H_1 may be accepted for some (not unique) peak locations simultaneously. For large m, the value of the power function in the third column is bigger, than other values in the same row. As expected, the power increases significantly, as m increases. Both approaches give approximately the same power for the test. It may be concluded, that both statistics are equally powerful. However, the statistic based on the first approach is preferable, because it is easier to implement.
- As for the test of choosing of peak location, the extreme high power of the test for the exponentially distributed data is caused from non-symmetrical configuration of the exponential distribution;
- The symmetrical distributions, ordered by simulation results from the best to the worst are: double exponential, logistic, and normal.

5.4 The values of the Type 1 error

Table 5.9 contains the values of the Type 1 error, which is equal to the value of the power function under the null hypothesis.

m	normal	logistic	dexp	exp
3	0.0417	0.0418	0.0448	0.0403
4	0.0477	0.0459	0.0453	0.0415
5	0.0464	0.0469	0.0475	0.0481
6	0.0473	0.0453	0.0506	0.0476
7	0.0470	0.0466	0.0482	0.0442
8	0.0495	0.0464	0.0476	0.0441
9	0.0501	0.0488	0.0508	0.0500
10	0.0488	0.0480	0.0480	0.0470
11	0.0448	0.0491	0.0498	0.0505
12	0.0481	0.0490	0.0502	0.0494
13	0.0493	0.0492	0.0498	0.0475
14	0.0480	0.0496	0.0486	0.0463
15	0.0468	0.0528	0.0500	0.0478

Table 5.9. The Type 1 error for differently distributed data

We can conclude that the real value of the Type 1 error does not exceed the significance of the test.

5.5 The case of unequal numbers of observations

Now, we consider the general case of unequal numbers of observations. We generate normal random samples with $k = 5$ and $p = 3$. We provide simulations only for the first approach, based on the statistic T_1. The distance is calculated from (4.11). As in the case of equal numbers of observations, we will check, how well this statistic estimates the peak location and find the power of the test. Its variance may be calculated in one of two ways:

- Directly, using the values of coefficients h_i from (4.13);
- By approximation. Define $\overline{m} = \frac{n}{k}$ be the average number of observations. We consider the statistic T_1 as having approximately an equal numbers of observations. Its variance is calculated from (4.26). We will find powers, corresponding to both variance calculations. Their comparison will show, how well such an approximation is.

Consider the five cases of various combinations of m_1, m_2, m_3, m_4, m_5. For each case $\overline{m} = 10$. The corresponding number of observations is:

Case	m_1	m_2	m_3	m_4	m_5
1	10	8	14	7	11
2	6	10	10	15	9
3	9	14	12	7	8
4	12	6	7	10	15
5	14	8	6	9	13

The power of the tests based on the different variance evaluation is placed in the table 5.10.

	The direct evaluation					The approximate evaluation				
Case	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
1	0.0422	0.5338	0.8841	0.5373	0.0337	0.0428	0.5415	0.8716	0.5629	0.0348
2	0.0944	0.4357	0.7619	0.3036	0.0115	0.0956	0.4159	0.7682	0.2878	0.0115
3	0.0178	0.3594	0.8263	0.4932	0.0709	0.0180	0.3432	0.8173	0.4975	0.0731
4	0.0705	0.4857	0.7495	0.4396	0.0211	0.0725	0.5302	0.7745	0.4686	0.0219
5	0.0364	0.4322	0.7072	0.4541	0.0447	0.0370	0.4759	0.7485	0.4857	0.0456

Table 5.10. The power of the test for unequal numbers of observations

Table 5.11 contains results for choosing the peak location

Case	F_1	F_2	F_3	F_4	F_5
1	0.0167	0.1008	0.77810	0.08780	0.0166
2	0.0384	0.1521	0.67200	0.11710	0.0204
3	0.0145	0.1045	0.72870	0.12350	0.0288
4	0.0406	0.1467	0.63370	0.15000	0.0290
5	0.0364	0.1622	0.58305	0.17475	0.0436

Table 5.11. The probability of choosing the peak

We repeat the simulation for cases with $\overline{m} = 15$. The corresponding number of observations is:

Case	m_1	m_2	m_3	m_4	m_5
1	15	13	19	12	16
2	11	15	15	20	14
3	14	19	17	12	13
4	17	11	12	15	20
5	19	13	11	14	18

The power of the tests for these cases is placed in table 5.12.

	The direct evaluation					The approximate evaluation				
Case	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5
1	0.0425	0.6637	0.9588	0.6614	0.0364	0.0427	0.6678	0.9553	0.6749	0.0366
2	0.0833	0.5786	0.9077	0.4749	0.0183	0.0837	0.5656	0.9091	0.4614	0.0184
3	0.0212	0.5337	0.9351	0.6383	0.0631	0.0214	0.5212	0.9329	0.6415	0.0637
4	0.0683	0.6570	0.9120	0.5951	0.0217	0.0697	0.6825	0.9203	0.6151	0.0220
5	0.0372	0.6133	0.8956	0.6168	0.0439	0.0374	0.6391	0.9066	0.6338	0.0441

Table 5.12. The power of the test for unequal numbers of observations

Table 5.13 contains results of choosing the peak location.

Case	F_1	F_2	F_3	F_4	F_5
1	0.0066	0.0745	0.8471	0.0654	0.0064
2	0.0170	0.1089	0.7831	0.0818	0.0092
3	0.0068	0.0747	0.8228	0.0875	0.0082
4	0.0175	0.1089	0.7706	0.0949	0.0081
5	0.0133	0.1152	0.7313	0.1252	0.0150

Table 5.13. The probability of choosing the peak

The conclusions for the case of unequal sample sizes are:

- The power is relatively low, if the peak population has a relatively small number of observations, as in cases 4 and 5 in Table 5.10. If the peak population has a relatively large number of observations, the power is higher, as in cases 4 and 5 in Table 5.12.
- The probability of choosing the peak location grows with increasing sample size for the peak population.

6. Future work

In chapter 2 we provided examples for various distance functions. This thesis considers the Spearman-Rho distance only. The analysis of this distance is the easiest, because it has the form of a simple linear rank statistic. This fact enables us to apply the Hajek-Sidak limit theorem. The second approach of the set E provided better results for obtaining the true peak location. This procedure is based on the statistic only and does not use the asymptotic theory. Hence, if the peak is known, the second approach is not necessary, because the powers of both approaches are equal.

The next work may be developed in the following directions:

- Using different distance function. The statistic based on the Kendall-Tau distance is non-linear. Thus, in order to apply the limit theorems, we need to find the projection of this statistic onto the family of linear rank statistic. This procedure is very complicated for the case of unequal number of observations. In addition, the distance for the second approach of the set E has a huge expression that makes it useless. The statistic based on the first approach may be used for the case of known peak location;
- The approach put forward in this thesis may be used in other more complicated situations.

References

[1] M.Alvo and J.Pan (1996). A General Theory of Hypothesis Testing Based on Rankings. *Journal of Statistical Planning and Inference*, 61, 219-248.

[2] D.E.Critchlow (1992). On Rank Statistics: an Approach via Metrics on the Permutation Group. *Journal of Statistical Planning and Inference,* 32, 325-345

[3] W.Feller (1972). *Probability Theory*. Wiley and Sons, New York.

[4] J.Hajek, Z.Sidak and P.K.Sen (1999). Theory of Rank Tests. Academic Press, New York.

[5] T.P.Hettmansperger and R.M.Norton (1987). Tests for Patterned Alternatives in k-Sample Problems. *Journal of the American Statistical Association*, 82, 292-299

[6] M.Hollander and D.A.Wolfe (1999). *Nonparametric Statistical Methods*. Wiley and Sons, New York.

[7] A.R.Jonckheere (1954). A Test of Significance for the Relation Between m Rankings and k Ranked Categories. *British Journal of Statistical Phychology*, 7, 93-100.

[8] G.A.Mack and D.A.Wolfe (1981). K-sample Rank Tests for Umbrella alternatives. *Journal of the American Statistical Association*, 76, 175-181.

[9] B.A.Millen and D.A.Wolfe (2005). A Class of Nonparametric Tests for Umbrella Alternatives. *Journal of Statistical Research*, 39, 7-24.