UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY JEDNÉHO SÚBORU

DIPLOMOVÁ PRÁCA

2022 Bc. Daniel LISÝ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY JEDNÉHO SÚBORU

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2022 **Bc. Daniel LISÝ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Daniel Lisý
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Viacrozmerné neparametrické testy polohy jedného súboru
Multivariate nonparametric one-sample location tests

Anotácia: V prípade viacrozmerného súboru dát vidno už niekoľko desaťročí snahu budovať testy o jeho polohe, ktoré fungujú bez ohľadu na rozdelenie, z ktorého dáta pochádzajú (t. j. neparametrickosť). Tieto testy bývajú zvyčajne oveľa odolnejšie voči odľahlým dátam (t. j. robustnejšie) než klasické testy pre tento problém založené na predpoklade normality dát.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 17.12.2019

Dátum schválenia: 21.12.2019 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt

V tejto diplomovej práci sa venujeme viacrozmerným neparametrickým testom polohy jedného súboru. Navrhujeme vylepšenú verziu *Bickelovho testu pomocou dvojíc dát*. Odvádzame tiež novú verziu testu pomocou trojíc dát. Na základe počítačových simulácií porovnávame medzi sebou sily oboch vylepšených testov a *Hotellingovho testu*. Ďalej sa venujeme *Blumenovmu dvojrozmernému testu*, ktorý testuje polohu mediánu a navrhujeme jeho novú verziu pomocou rankov. Ranková verzia je použiteľná pre testovanie hypotézy o symetrii pre širšiu škálu rozdelení. Nakoniec simulačne porovnávame medzi sebou sily jednotlivých testov uvažovaných v práci.

Kľúčové slová: stred symetrie, medián, sila testu, asymptotická variančná matica

Abstract

In this thesis we consider various multivariate nonparametric one-sample location tests. We propose an improved version of *Bickel's test using pairs of data*. We develop a new version of this test using triplets of data. By means of computer simulations we compare the power of each of these two improved tests and *Hotelling's test*. Next we consider *Blumen's bivariate sign test*, which tests the location of the median, and we propose a new version of this test using ranks. The ranked version is applicable to a wider variety of distributions, for which we can test for symmetry. Finally we compare the power of each of the various tests considered in this thesis.

Keywords: center of symmetry, median, power of a test, asymptotic variance matrix

Poďakovanie

Rád by som sa poďakoval vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za jeho trpezlivosť, ochotu a cenné rady počas diplomového výskumu.

Obsah

Úvod **8**

1 Bickelov neparametrický test pomocou dvojíc dát a jeho rozšírenie **9**

1.1 Asymptotická variančná matica Bickelovho testového vektora h_2 pomocou dvojíc 10

1.2 Analógia Bickelovho testu založená na trojiciach 13

1.3 Odhady asymptotických variančných matíc 18

1.3.1 Odhad asymptotickej variančnej matice pre test s trojicami 19

1.3.2 Alternatívny odhad asymptotickej variančnej matice pre Bickelov test s dvojicami 21

1.4 Porovnanie sily Bickelovho testu s dvojicami vs. testu s trojicami 21

2 Blumenov test a jeho rozšírenie **26**

2.1 Blumenov test 26

2.2 Afínna invariantnosť Blumenovho testu 29

2.3 Ranková verzia Blumenovho testu 30

3 Porovnanie sily testov **34**

Záver **37**

Zoznam použitej literatúry **39**

Prílohy **40**

Úvod

V tejto práci sa budeme zaoberať neparametrickými štatistickými testami, ktoré použijeme na testovanie polohy dátového súboru. Medzi hlavné výhody neparametrického prístupu patrí [3] :

- Menšie nároky kladené na rozdelenie, z ktorého pochádzajú dáta.
- Robustnosť voči outlierom, teda v prípade nespoľahlivých alebo chybných dát nemusí zlyhať test.
- Často sú jednoduchšie aplikovateľné ako testy vyžadujúce normalitu.
- Vývoj výpočtového softvéru dnes umožňuje rýchly výpočet presných a približných p-hodnôt pre neparametrické testy.

V tejto práci sa venujeme Bickelovmu neparametrickému testu [1] o parametri polohy jedného súboru, ktorý pracuje s dvojicami dát. Rozoberáme spôsob, akým rozumne odhadovať istú asymptotickú variančnú maticu potrebnú pre tento test. V práci navrhujeme rozšírenú verziu testu, ktorá používa trojice dát.

V ďalšej časti sa venujeme Blumenovmu testu [2], ktorý sa používa pre dvojrozmerné dáta, a testuje polohu daného súboru. Dokazujeme afínnu invariantnosť tohto testu. V práci navrhujeme jeho vylepšenú verziu pomocou rankov.

Pomocou simulácií v softvéri R [4] testujeme správanie testovej štatistiky za platnosti nulovej hypotézy (presnejšie chí-kvadrátovosť) a porovnávame silu testov pri rôznych porušeniach nulovej hypotézy.

Cieľom práce je vylepšiť existujúce testy o parametri polohy jedného súboru viacrozmerných dát.

1 Bickelov neparametrický test pomocou dvojíc dát a jeho rozšírenie

V tejto kapitole sa venujeme Bickelovmu neparametrickému testu z článku [1].

Nech $X_a = (X_{1a}, \ldots, X_{pa})^T, 1 \leq a \leq n$, je náhodný výber z $F(x_1 - \theta_1, \ldots, x_p - \theta_p)$, kde F je p-rozmerná distribučná funkcia symetrická okolo počiatku súradnicovej sústavy. $F_1, F_2, \ldots F_p$ sú jej marginálne distribučné funkcie.
Testujeme hypotézu:

$$H_0 : \theta = 0 \qquad vs. \qquad H_1 : \theta \neq 0,$$

kde $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$. Podobne ako pri Hotellingovom teste, H_0 zamietame, ak platí: $n[h_2 - \mu]\hat{\pi}_2^{-1}[h_2 - \mu]^T > c$. Vektor h_2 má i-tu zložku definovanú nasledovne:

$$h_2^i = \frac{2}{n(n+1)} \sum_{1 \leq a \leq b \leq n} I^+(X_{ia} + X_{ib}) \, , \qquad i = 1, 2 \ldots, p, \tag{1}$$

kde $I^+(x)$ je indikátorová funkcia, čiže $I^+(x) = 1$, ak $x > 0$, $I^+(x) = 0$, ak $x \leq 0$. Vektor h_2 môžeme interpretovať ako priemer vektorizovanej indikátorovej funkcie každej dvojice dátových bodov, pričom dáta sa môžu opakovať. Počet takýchto dvojíc je $\binom{n+2-1}{2} = \frac{n(n+1)}{2}$. Ďalej μ je stredná hodnota h_2, a $\hat{\pi}_2$ je odhad variančnej matice výrazu $\sqrt{n}h_2$. V článku [1] sa na strane 162 bez odvodenia uvádza ij-ta zložka asymptotickej variančnej matice pre výraz $\sqrt{n}h_2$, ako:

$$\pi_2^{ij} := \lim_{n \to \infty} Var(\sqrt{n}h_2)_{ij} = E\left[\left(F_i(x_{i1}) - \frac{1}{2}\right)\left(F_j(x_{j1}) - \frac{1}{2}\right)\right],$$

kde E značí strednú hodnotu za platnosti H_0. V článku [1] na strane 166 autor uvádza odhad na výpočet tejto asymptotickej variančnej matice:

$$\hat{\pi}_2^{ij} := \frac{1}{n(n-1)(n-2)} \sum_{r \neq s \neq t} I^+(X_{ir} - X_{is}) I^+(X_{jr} - X_{jt}) \, , \quad \text{pre } i \neq j \tag{2}$$

$$\frac{1}{12} \qquad , \quad \text{pre } i = j.$$

Avšak keď sme simuláciami odhadovali variančnú maticu výrazu $\sqrt{n}h_2$, nevychádzali nám hodnoty podobné Bickelovým vzorcom pre π_2, či $\hat{\pi}_2$, preto sa v nasledujúcej časti pokúsime odvodiť vzorec pre asymptotickú variančnú maticu π_2 .

1.1 Asymptotická variančná matica Bickelovho testového vektora h_2 pomocou dvojíc

Variančnú maticu výrazu $\sqrt{n}h_2$ môžeme vyjadriť ako:

$$Var(\sqrt{n}h_2)_{ij} = nCov(h_2^i, h_2^j) = nE[\,(h_2^i - E[h_2^i]) \cdot (h_2^j - E[h_2^j])\,] = \tag{3}$$
$$= n[\,E(h_2^i h_2^j) - E(\,h_2^i E[h_2^j]\,) - E(\,h_2^j E[h_2^i]\,) - E(\,E[h_2^i]E[h_2^j]\,)\,].$$

Zo symetrie distribučnej funkcie F okolo počiatku vieme, že $E(h_2^i) = \frac{1}{2}$. Teda:

$$(3) = n\left[\,E(h_2^i h_2^j) - \frac{1}{2}E(h_2^i) - \frac{1}{2}E(h_2^j) - E\left(\frac{1}{2}\cdot\frac{1}{2}\right)\,\right] = n\left[E(h_2^i h_2^j) - \frac{1}{4}\right] =$$
$$= \left[E\left(\left[\frac{2}{n(n+1)}\sum_{1\le a\le b\le n} I^+(X_{ia}+X_{ib})\right]\left[\frac{2}{n(n+1)}\sum_{1\le c\le d\le n} I^+(X_{jc}+X_{jd})\right]\right) - \frac{1}{4}\right] =$$
$$= \frac{4n}{n^2(n+1)^2}\left(\sum_{1\le a\le b\le n}\sum_{1\le c\le d\le n} E\left(\left[I^+(X_{ia}+X_{ib})\right]\left[I^+(X_{jc}+X_{jd})\right]\right)\right) - \frac{n}{4}. \tag{4}$$

Naše dáta sú v tvare $X_a = (X_{1a}, \ldots, X_{pa})^T, 1 \le a \le n$. Prvý index nám udáva, o ktorú zložku vektora sa jedná, druhý index je poradie dátového bodu. Sumu stredných hodnôt $E\left(\left[I^+(X_{ia}+X_{ib})\right]\left[I^+(X_{jc}+X_{jd})\right]\right)$ rozdelíme na základe toho, či sa rovnajú jednotlivé indexy a, b, c, d, čím dostaneme konštelácie "rovností a nerovností", t. j. napríklad konštelácia $[a \neq b,\ a \neq c,\ a = d,\ b = c,\ b \neq d,\ c \neq d]$. V prílohe A sme odvodili vzorec pre počet takýchto konštelácií pre všeobecný počet premenných, ktorý by nám v tomto prípade pre 4 indexy dal počet konštelácií 15. Rôzne konštelácie môžu mať rôznu početnosť. Pre jednoduchosť, keďže sledujeme čisto indexy poradia, tak v ďalšom značení indexy zložky vektora i a j vynechávame.

1) Majme napríklad konšteláciu, kde sa ani jeden z indexov X-ov nerovná, teda napríklad $E[I^+(X_1 + X_2)I^+(X_3 + X_4)]$. V prvej zložke $I^+(X_\cdot + X_\cdot)$ môžeme vybrať ľubovoľnú dvojicu rôznych indexov, v druhej môžeme vybrať ľubovoľné rôzne indexy okrem vybraných v prvej zložke, preto početnosť tejto konštelácie je $\binom{n}{2}\binom{n-2}{2}$.

2) Pri konštelácii typu $E[I^+(X_1 + X_1)I^+(X_2 + X_3)]$ môžeme vybrať ľubovoľný index v prvej zložke, kde možností je n, v druhej zložke teda $\binom{n-1}{2}$, dokopy $n\binom{n-1}{2}$.

3) Pri konštelácii typu $E[I^+(X_1+X_2)I^+(X_1+X_3)]$, v ktorej sa rovnajú menšie indexy X-ov v prvej a druhej zložke, môžeme vybrať ľubovoľnú trojicu dát. Takýchto trojíc je $\binom{n}{3}$. Ešte máme voľnosť v tom, že v ktorej indikátorovej funkcii sa nachádza väčší index, preto dostávame dvakrát toľko kombinácií.

Pre ostatné typy konštelácií výpočty početností kvôli stručnosti vynechávame. Sú veľmi podobné.

Ak vypíšeme všetky typy takýchto konštelácií, dostaneme:

$$
\begin{aligned}
(4) = \frac{4n}{n^2(n+1)^2}(\quad & E[I^+(X_1+X_2)I^+(X_3+X_4)] \quad \binom{n}{2}\binom{n-2}{2} + && (5)\\
+ \quad & E[I^+(X_1+X_1)I^+(X_2+X_3)] \quad n\binom{n-1}{2} + && (6)\\
+ \quad & E[I^+(X_2+X_3)I^+(X_1+X_1)] \quad n\binom{n-1}{2} + && (7)\\
+ \quad & E[I^+(X_1+X_2)I^+(X_1+X_3)] \quad 2\binom{n}{3} + && (8)\\
+ \quad & E[I^+(X_1+X_3)I^+(X_2+X_3)] \quad 2\binom{n}{3} + && (9)\\
+ \quad & E[I^+(X_1+X_2)I^+(X_2+X_3)] \quad \binom{n}{3} + && (10)\\
+ \quad & E[I^+(X_2+X_3)I^+(X_1+X_2)] \quad \binom{n}{3} + && (11)\\
+ \quad & E[I^+(X_1+X_1)I^+(X_1+X_2)] \quad \binom{n}{2} + &&\\
+ \quad & E[I^+(X_1+X_2)I^+(X_1+X_1)] \quad \binom{n}{2} + &&\\
+ \quad & E[I^+(X_1+X_2)I^+(X_2+X_2)] \quad \binom{n}{2} + &&\\
+ \quad & E[I^+(X_2+X_2)I^+(X_1+X_2)] \quad \binom{n}{2} + &&\\
+ \quad & E[I^+(X_1+X_1)I^+(X_2+X_2)] \quad 2\binom{n}{2} + &&\\
+ \quad & E[I^+(X_1+X_2)I^+(X_1+X_2)] \quad \binom{n}{2} + &&\\
+ \quad & E[I^+(X_1+X_1)I^+(X_1+X_1)] \quad n \qquad) - \frac{n}{4}\,. &&
\end{aligned}
$$

Istou skúškou správnosti je, že súčet početností, ktorými sa násobia stredné hodnoty, je $\left(\frac{n(n-1)}{2}\right)^2$.

Keďže hľadáme asymptotickú variančnú maticu π_2^{ij}, tak chceme vypočítať limitu:

$$\lim_{n\to\infty} Var(\sqrt{n}h_2)_{ij}\,.$$

Keďže výraz $\frac{4n}{n^2(n+1)^2}$ je rádu $\frac{1}{n^3}$ a výrazy typu $E[I^+(\dots)I^+(\dots)]$ sú rádu 1, tak v limite členy, ktoré majú početnosť nižšieho rádu než n^3, vypadnú. Ostanú nám len konštelácie v riadkoch (5) až (11).

Ak sú vo výrazoch typu $E[I^+(X_1+X_2)I^+(X_3+X_4)]$ jednotlivé súčinitele nezávislé, tak zo symetrie distribučnej funkcie F okolo počiatku vyplýva, že sa stredné hodnoty rovnajú $\frac{1}{4}$ - to platí pre prvé tri prípady (5), (6), (7). Pri prípadoch (8) až (11), ktoré majú členy $\binom{n}{3}$, počítame rovnakú strednú hodnotu, ale prvé $I^+(\dots)$ a druhé $I^+(\dots)$ môžu byť závislé, keďže i-ta a j-ta zložka jedného dátového bodu môžu byť závislé, preto musíme túto strednú hodnotu vypočítať. V ďalšom $f_i(x)$ označujú marginálne hustoty rozdelenia, z ktorého dáta pochádzajú, a $f_{ij}(y_1, y_2)$ značí dvojrozmernú hustotu rozdelenia i-tej a j-tej zložky dátového bodu.

$$
\begin{aligned}
&E[I^+(X_{i1}+X_{i2})I^+(X_{j2}+X_{j3})] = Pr(X_{i1}+X_{i2}>0 \wedge X_{j2}+X_{j3}>0) = \\
&= \iiiint\limits_{\substack{x_{i1}+x_{i2}>0\\ x_{j2}+x_{j3}>0}} f_i(x_{i1}) f_{ij}(x_{i2},x_{j2}) f_j(x_{j3}) dx_{i1}dx_{i2}dx_{j2}dx_{j3} = \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i2},x_{j2}) \left[\int\limits_{-x_{i2}}^{\infty} f_i(x_{i1}) \left[\int\limits_{-x_{j2}}^{\infty} f_j(x_{j3})dx_{j3}\right] dx_{i1}\right] dx_{i2}dx_{j2} = \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i2},x_{j2}) \left[\int\limits_{-x_{i2}}^{\infty} f_i(x_{i1})\cdot[1-F_j(-x_{j2})]dx_{i1}\right] dx_{i2}dx_{j2} = \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i2},x_{j2})\cdot[1-F_i(-x_{i2})]\cdot[1-F_j(-x_{j2})]\, dx_{i1}dx_{i2}dx_{j2}, \qquad (12)
\end{aligned}
$$

čo vďaka symetrii F_i okolo 0 vieme upraviť na tvar:

$$
(12) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i2},x_{j2})F_i(x_{i2})F_j(x_{j2})dx_{i2}dx_{j2} = E[F_i(X_{i1})F_j(X_{j1})]\ .
$$

Teraz vypočítajme tri limity. Najprv pre výraz $\frac{4n}{n^2(n+1)^2}(E[I^+(X_1+X_2)I^+(X_3+X_4)]\binom{n}{2}\binom{n-2}{2} - \frac{n}{4}$, teda korešpondujúci s časťou rovnice označenej (5) vynásobenej zlomkom na začiatku a s odrátaním $\frac{n}{4}$ na konci. Vieme, že stredná hodnota v takomto prípade je vďaka symetrii F okolo počiatku a nezávislosti rovná $\frac{1}{4}$:

$$
\begin{aligned}
&\lim_{n\to\infty}\frac{4n}{n^2(n+1)^2}\left[\frac{1}{4}\right]\binom{n}{2}\binom{n-2}{2} - \frac{n}{4} = \lim_{n\to\infty}\frac{4n^2(n-1)(n-2)(n-3)}{16n^2(n+1)^2} - \frac{n}{4} = \qquad (13)\\
&= \lim_{n\to\infty}\frac{4n^2(n-1)(n-2)(n-3) - 4n^3(n+1)^2}{16n^2(n+1)^2} = \lim_{n\to\infty}\frac{(n-1)(n-2)(n-3) - n(n+1)^2}{4(n+1)^2} = \\
&= \lim_{n\to\infty}\frac{n^3 - n^3 - 6n^2 - 2n^2 + O(n^1)}{4n^2 + O(n^1)} = -2\ .
\end{aligned}
$$

Teraz počítame limitu pre výrazy typu $E[I^+(X_1+X_1)I^+(X_2+X_3)]$ s početnosťou $n\binom{n-1}{2}$. Také prípady máme dva, korešpondujú s časťami rovnice označenými (6) a (7). V tomto prípade je stredná hodnota, rovnako ako v predošlej limite, rovná $\frac{1}{4}$.

$$\lim_{n\to\infty} 2 \times \frac{4n}{n^2(n+1)^2}\left[\frac{1}{4}\right] n\binom{n-1}{2} = \lim_{n\to\infty} \frac{8n^2(n-1)(n-2)}{8n^2(n+1)^2} = 1\ . \tag{14}$$

Teraz počítame limitu pre výrazy s početnosťou $\binom{n}{3}$. Takýchto prípadov máme 6, korešpondujú s časťami rovnice označenými (8) až (11):

$$\lim_{n\to\infty} 6 \times \frac{4n}{n^2(n+1)^2} E[F_i(X_{i1})F_j(X_{j1})]\binom{n}{3} = \lim_{n\to\infty} \frac{24n^2(n-1)(n-2)}{6n^2(n+1)^2} E(F_i(X_{i1})F_j(X_{j1})) =$$
$$= 4E(F_i(X_{i1})F_j(X_{j1}))\ . \tag{15}$$

Ak teraz dáme dokopy výsledky limít (13), (14), (15), tak dostaneme:

$$\pi_2^{ij} = \lim_{n\to\infty} Var(\sqrt{n}h_2)_{ij} = -2 + 1 + 4E(F_i(X_{i1})F_j(X_{j1})) = 4E(F_i(X_{i1})F_j(X_{j1})) - 1,$$

čo vieme prepísať do tvaru:

$$\pi_2^{ij} = 4E\left[\left(F_i(X_{i1}) - \frac{1}{2}\right)\left(F_j(X_{j1}) - \frac{1}{2}\right)\right].$$

V pôvodnom článku [1] je vzorec uvedený na strane 162 bez odvodenia a navyše nesprávne, presnejšie, chýba vynásobenie číslom 4.

1.2 Analógia Bickelovho testu založená na trojiciach

V tejto časti navrhujeme novú verziu Bickelovho testu, kde tentoraz pracujeme nie s dvojicami ale s trojicami. Predpoklady na rozdelenie dát a hypotéza, ktorú testujeme, sú rovnaké ako na začiatku kapitoly 1 pre test s dvojicami. Základom novej testovej štatistiky bude vektor h_3, ktorého i-ta zložka je definovaná ako:

$$h_3^i = \frac{6}{n(n+1)(n+2)} \sum_{1\le a\le b\le c\le n} I^+(X_{ia} + X_{ib} + X_{ic})\ , \qquad i = 1, 2 \ldots, p\ . \tag{16}$$

Ide o priemer vektorizovanej indikátorovej funkcie všetkých kombinácií trojíc dát s opakovaním a zlomok pred sumou je prevrátená hodnota počtu týchto trojíc, ktorých je $\binom{n+3-1}{3}$.

Rovnako ako pre test s dvojicami chceme vypočítať variančnú maticu výrazu $\sqrt{n}h_3$ v limite pre $n \to \infty$. Vyjadríme jednotlivé zložky variančnej matice:

$$Var(\sqrt{n}h_3)_{ij} = nCov(h_3^i, h_3^j) = nE[\ (h_3^i - E[h_3^i]) \cdot (h_3^j - E[h_3^j])\ ] =$$
$$= n[\ E(h_3^i h_3^j) - E(\ h_3^i E[h_3^j]\ ) - E(\ h_3^j E[h_3^i]\ ) - E(\ E[h_3^i]E[h_3^j]\ )\ ] \tag{17}$$

Zo symetrie distribučnej funkcie F okolo počiatku vieme, že $E(h_3^i) = \frac{1}{2}$. Teda:

$$(17) = n\left[\ E(h_3^i h_3^j) - \frac{1}{2}E(h_3^i) - \frac{1}{2}E(h_3^j) - E\left(\frac{1}{2}\cdot\frac{1}{2}\right)\ \right] = n\left[E(h_3^i h_3^j) - \frac{1}{4}\right]. \tag{18}$$

Vypíšeme tento výraz podľa definície h_3:

$$(18) = n\left[E\left(\left[\frac{6}{n(n+1)(n+2)}\sum_{a\le b\le c} I^+(X_{ia}+X_{ib}+X_{ic})\right]\cdot\left[\frac{6}{n(n+1)(n+2)}\sum_{u\le v\le w} I^+(X_{ju}+X_{jv}+X_{jw})\right]\right) - \frac{1}{4}\right]. \tag{19}$$

Dostaneme dvojitú sumu:

$$(19) = \frac{36n}{n^2(n+1)^2(n+2)^2}\sum_{a\le b\le c}\sum_{u\le v\le w} E\left[I^+(X_{ia}+X_{ib}+X_{ic})I^+(X_{ju}+X_{jv}+X_{jw})\right] - \frac{n}{4}\,. \tag{20}$$

Rovnako ako v prípade testu založenom na dvojiciach chceme dvojitú sumu rozdeliť na jednotlivé prípady konštelácií "rovností a nerovností" indexov a, b, c, u, v, w . Podľa vzorca v prílohe A môžeme dostať potenciálne až 203 konštelácií, mnohé z nich však budú v limite zanedbateľné kvôli početnosti nižšieho rádu ako $O\left(\frac{n^2(n+1)^2(n+2)^2}{36n}\right) = O(n^5)$. Nenulové budú len tri typy konštelácií. Pre jednoduchosť, keďže sledujeme čisto indexy dát, tak v ďalšom značení indexy zložky dát vynechávame:

$$I) \quad E\left[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_6)\right].$$

V tomto type konštelácie sa ani jeden z indexov nerovná inému. Počet kombinácií, ako môžeme vybrať takúto šesticu indexov, je $\binom{n}{3}\binom{n-3}{3}$.

$$II) \quad E\left[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_5)\right].$$

V takomto type konštelácie máme jednu rovnosť medzi indexami v jednej indikátorovej funkcii. Početnosť týchto typov konštelácií vynásobíme číslom 4, lebo v oboch indikátorových funkciách môže byť rovnosť medzi prvým a druhým alebo druhým a tretím indexom, z čoho získame 4 kombinácie. Počet kombinácií, ako môžeme vybrať takúto päticu indexov, je $\binom{n}{3}\binom{n-3}{2}$.

$$III) \quad E\left[I^+(X_1+X_2+X_3)I^+(X_3+X_4+X_5)\right].$$

V tomto type konštelácie máme tiež jednu rovnosť medzi indexami, ale jeden index sa rovná druhému v inej indikátorovej funkcii. Koľko je takýchto typov konštelácií? Podľa pozície rovnajúcich sa indexov v indikátorovej funkcii nájdeme 9 prípadov. Pri každom z týchto prípadov môžeme mať dodatočnú voľnosť pri tom, ako zvolíme zvyšné indexy. Napríklad pri prvom prípade $E\left[I^+(\underline{\underline{X_1}}+X_2+X_3)I^+(\underline{\underline{X_1}}+X_4+X_5)\right]$ sme zvyšné indexy zvolili v štýle $(2,3|4,5)$, ale rovnako môžeme zvoliť $(2,4|3,5)$, $(2,5|3,4)$, $(3,4|2,5)$, $(3,5|2,4)$, $(4,5|2,3)$. Teraz vypíšeme pre všetkých 9 prípadov, aké kombinácie zvyšných premenných môžeme zvoliť:

1. $E\left[I^+(\underline{\underline{X_1}}+X_2+X_3)I^+(\underline{\underline{X_1}}+X_4+X_5)\right]\ldots$ $(2,3|4,5),(2,4|3,5),(2,5|3,4),(3,4|2,5),(3,5|2,4),(4,5|2,3)$
2. $E\left[I^+(\underline{\underline{X_2}}+X_3+X_4)I^+(X_1+\underline{\underline{X_2}}+X_5)\right]\ldots(3,4|1,5),(3,5|1,4),(4,5|1,3)$
3. $E\left[I^+(\underline{\underline{X_3}}+X_4+X_5)I^+(X_1+X_2+\underline{\underline{X_3}})\right]\ldots(4,5|1,2)$
4. $E\left[I^+(X_1+\underline{\underline{X_2}}+X_3)I^+(\underline{\underline{X_2}}+X_4+X_5)\right]\ldots(1,3|4,5),(1,4|3,5),(1,5|3,4)$
5. $E\left[I^+(X_1+\underline{\underline{X_3}}+X_4)I^+(X_2+\underline{\underline{X_3}}+X_5)\right]\ldots(1,4|2,5),(1,5|2,4),(2,4|1,5),(2,5|1,4)$
6. $E\left[I^+(X_1+\underline{\underline{X_4}}+X_5)I^+(X_2+X_3+\underline{\underline{X_4}})\right]\ldots(1,5|2,3),(2,5|1,3),(3,5|1,2)$
7. $E\left[I^+(X_1+X_2+\underline{\underline{X_3}})I^+(\underline{\underline{X_3}}+X_4+X_5)\right]\ldots(1,2|4,5)$
8. $E\left[I^+(X_1+X_2+\underline{\underline{X_4}})I^+(X_3+\underline{\underline{X_4}}+X_5)\right]\ldots(1,2|3,5),(1,3|2,5),(2,3|1,5)$
9. $E\left[I^+(X_1+X_2+\underline{\underline{X_5}})I^+(X_3+X_4+\underline{\underline{X_5}})\right]\ldots$ $(1,2|3,4),(1,3|2,4),(1,4|2,3),(2,3|1,4),(2,4|1,3),(3,4|1,2)$

Čiže pre prípad III) nakoniec dostávame dokopy 30 kombinácií:

$$
\begin{aligned}
6\times E\left[I^+(\underline{\underline{X_1}}+X_2+X_3)I^+(\underline{\underline{X_1}}+X_4+X_5)\right]\\
3\times E\left[I^+(\underline{\underline{X_2}}+X_3+X_4)I^+(X_1+\underline{\underline{X_2}}+X_5)\right]\\
1\times E\left[I^+(\underline{\underline{X_3}}+X_4+X_5)I^+(X_1+X_2+\underline{\underline{X_3}})\right]\\
3\times E\left[I^+(X_1+\underline{\underline{X_2}}+X_3)I^+(\underline{\underline{X_2}}+X_4+X_5)\right]\\
4\times E\left[I^+(X_1+\underline{\underline{X_3}}+X_4)I^+(X_2+\underline{\underline{X_3}}+X_5)\right]\\
3\times E\left[I^+(X_1+\underline{\underline{X_4}}+X_5)I^+(X_2+X_3+\underline{\underline{X_4}})\right]\\
1\times E\left[I^+(X_1+X_2+\underline{\underline{X_3}})I^+(\underline{\underline{X_3}}+X_4+X_5)\right]\\
3\times E\left[I^+(X_1+X_2+\underline{\underline{X_4}})I^+(X_3+\underline{\underline{X_4}}+X_5)\right]\\
6\times E\left[I^+(X_1+X_2+\underline{\underline{X_5}})I^+(X_3+X_4+\underline{\underline{X_5}})\right]
\end{aligned}
$$

V každej z týchto kombinácií je možnosť vybrať päticu indexov, a to $\binom{n}{5}$ spôsobmi.

Typmi konštelácií $I)$, $II)$, $III)$ sme vyčerpali všetky konštelácie, v ktorých v menej ako troch indexoch z $\{a, b, c, u, v, w\}$ nastáva rovnosť. Ostatné konštelácie budú mať početnosti rádu n^4 alebo nižšieho.

Ak rozdelíme výraz vo vzorci (20) na typy konštelácií $I)$, $II)$, $III)$, dostaneme:

$$\begin{aligned}(20) = \frac{36n}{n^2(n+1)^2(n+2)^2}\Big(& E[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_6)] \quad \binom{n}{3}\binom{n-3}{3} + \\ + \quad & 4 \times E[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_5)] \quad \binom{n}{3}\binom{n-3}{2} + \\ + \quad & 30 \times E[I^+(X_1+X_2+X_3)I^+(X_3+X_4+X_5)] \quad \binom{n}{5} \quad \Big) \quad - \frac{n}{4} \quad + \\ + \quad & \text{výrazy v limite rovné 0.}\end{aligned}$$

Výraz $\frac{36n}{n^2(n+1)^2(n+2)^2}$ je rádu $\frac{1}{n^5}$ a výrazy typu $E[I^+(X_\cdot + X_\cdot + X_\cdot)I^+(X_\cdot + X_\cdot + X_\cdot)]$ sú rádu 1, preto v limite členy, ktoré majú početnosť nižšieho rádu než n^5, vypadnú, čo platí pre všetky ostatné konštelácie, ktoré sme preto nemuseli v limite uvažovať. Ak sú vo výrazoch typu $E[I^+(X_\cdot + X_\cdot + X_\cdot)I^+(X_\cdot + X_\cdot + X_\cdot)]$ jednotlivé súčinitele $I^+(\dots)$ nezávislé, tak zo symetrie F vyplýva, že tieto stredné hodnoty sa rovnajú $\frac{1}{4}$. To platí pre prvé dva prípady, pri prípade $E[I^+(X_{i1}+X_{i2}+X_{i3})I^+(X_{j3}+X_{j4}+X_{j5})]$ počítame strednú hodnotu súčinu so zohľadnením závislosti, keďže i-ta a j-ta zložka vektora jedného dátového bodu môžu byť závislé.

$$E[I^+(X_{i1}+X_{i2}+X_{i3})I^+(X_{j3}+X_{j4}+X_{j5})] = Pr(X_{i1}+X_{i2}+X_{i3} > 0 \ \wedge \ X_{j3}+X_{j4}+X_{j5} > 0) =$$

$$= \int\limits_{\substack{X_{i1}+X_{i2}+X_{i3}>0 \\ X_{j3}+X_{j4}+X_{j5}>0}} \cdots \int f_i(x_1)f_i(x_2)f_{ij}(x_{i3},x_{j3})f_j(x_4)f_j(x_5)dx_{i1}dx_{i2}dx_{i3}dx_{j3}dx_{j4}dx_{j5} =$$

$$= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i3},x_{j3}) \iint\limits_{x_{i1}+x_{i2}>-x_{i3}} f_i(x_{i1})f_i(x_{i2}) \iint\limits_{x_{j4}+x_{j5}>-x_{j3}} f_j(x_{j4})f_j(x_{j5})dx_{j4}dx_{j5}dx_{i1}dx_{i2}dx_{i3}dx_{j3} =$$

$$= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i3},x_{j3})f_i(x_{i2})f_j(x_{j4}) \int\limits_{-x_{i2}-x_{i3}}^{\infty} f_i(x_{i1}) \int\limits_{-x_{j3}-x_{j4}}^{\infty} f_j(x_{j5})dx_{j5}\ dx_{i1}dx_{j4}dx_{i2}dx_{i3}dx_{j3} =$$

$$= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i3},x_{j3})f_i(x_{i2})f_j(x_{j4}) \int\limits_{-x_{i2}-x_{i3}}^{\infty} f_i(x_{i1})\ [1-F_j(-x_{j3}-x_{j4})]\ dx_{i1}dx_{j4}dx_{i2}dx_{i3}dx_{j3} =$$

$$= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i3},x_{j3})f_i(x_{i2})f_j(x_{j4})\ F_j(x_{j3}+x_{j4}) \int\limits_{-x_{i2}-x_{i3}}^{\infty} f_i(x_{i1})dx_{i1}\ dx_{j4}dx_{i2}dx_{i3}dx_{j3} =$$

$$= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{ij}(x_{i3},x_{j3})f_i(x_{i2})f_j(x_{j4})\ F_j(x_{j3}+x_{j4})F_i(x_{i2}+x_{i3})\ dx_{j4}dx_{i2}dx_{i3}dx_{j3} =$$

$$= E[F(X_{i2}+X_{i3})F(X_{j3}+X_{j4})].$$

Teraz vypočítajme tri limity:

I) Najprv pre výrazy $\frac{36n}{n^2(n+1)^2(n+2)^2}E[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_6)]\binom{n}{3}\binom{n-3}{3} - \frac{n}{4}$ vieme, že stredná hodnota v takomto prípade je za platnosti H_0, vďaka nezávislosti, rovná $\frac{1}{4}$:

$$\lim_{n\to\infty} \frac{36n}{n^2(n+1)^2(n+2)^2}\left[\frac{1}{4}\right]\binom{n}{3}\binom{n-3}{3} - \frac{n}{4} =$$
$$\lim_{n\to\infty} \frac{36n}{n^2(n+1)^2(n+2)^2}\left[\frac{1}{4}\right]\frac{n(n-1)(n-2)}{6}\frac{(n-3)(n-4)(n-5)}{6} - \frac{n}{4} =$$
$$\lim_{n\to\infty} \frac{n}{4}\left(\frac{n(n-1)\dots(n-5)}{n^2(n+1)^2(n+2)^2} - 1\right) = \lim_{n\to\infty} \frac{n}{4}\left(\frac{n(n-1)\dots(n-5) - n^2(n+1)^2(n+2)^2}{n^2(n+1)^2(n+2)^2}\right) =$$
$$= \lim_{n\to\infty} \frac{1}{4}\left(\frac{(n^7-n^7) - n^6 - 2n^6 - 3n^6 - 4n^6 - 5n^6 - n^6 - n^6 - 2n^6 - 2n^6 + O(n^5)}{n^6 + O(n^5)}\right) =$$
$$\lim_{n\to\infty} \frac{1}{4}\left(\frac{-21n^6 + O(n^5)}{n^6 + O(n^5)}\right) = -\frac{21}{4}.$$

II) Teraz vypočítajme limitu pre výraz $4\times\frac{36n}{n^2(n+1)^2(n+2)^2}E[I^+(X_1+X_2+X_3)I^+(X_4+X_5+X_5)]\binom{n}{3}\binom{n-3}{2}$, rovnako ako v predošlom prípade stredná hodnota sa vďaka nezávislosti rovná $\frac{1}{4}$:

$$\lim_{n\to\infty} 4\times\frac{36n}{n^2(n+1)^2(n+2)^2}\left[\frac{1}{4}\right]\binom{n}{3}\binom{n-3}{2} =$$
$$\lim_{n\to\infty} 4\times\frac{36n}{n^2(n+1)^2(n+2)^2}\left[\frac{1}{4}\right]\frac{n(n-1)(n-2)(n-3)(n-4)}{12} = 3.$$

III) Počítajme limitu pre výraz $30\times\frac{36n}{n^2(n+1)^2(n+2)^2}E[I^+(X_1+X_2+X_3)I^+(X_3+X_4+X_5)]\binom{n}{5}$:

$$\lim_{n\to\infty} 30\times\frac{36n}{n^2(n+1)^2(n+2)^2}E\left[\dots\right]\frac{n(n-1)\dots(n-4)}{120} = 9E[\dots].$$

Ak teraz dáme dokopy limity **I**), **II**), **III**), dostávame, že:

$$\lim_{n\to\infty} Var(\sqrt{n}h_3)_{ij} = -\frac{21}{4} + 3 + 9E[F_i(X_{i1}+X_{i2})F_j(X_{j2}+X_{j3})] =$$
$$= 9E[F_i(X_{i1}+X_{i2})F_j(X_{j2}+X_{j3})] - \frac{9}{4}.$$

To vieme vďaka symetrii F_i, F_j okolo nuly upraviť na tvar:

$$\pi_3^{ij} := \lim_{n\to\infty} Var(\sqrt{n}h_3)_{ij} = 9E\left[\left(F_i(X_{i2}+X_{i3}) - \frac{1}{2}\right)\left(F_j(X_{j3}+X_{j4}) - \frac{1}{2}\right)\right].$$

Vidíme, že tvar asymptotickej variančnej matice pre test s trojicami je v mnohom podobný matici pre Bickelov test s dvojicami, čo naznačuje, že by sa potenciálne dal zovšeobecniť pre k-tice, kde k je väčšie ako 3.

1.3 Odhady asymptotických variančných matíc

V tejto kapitole sa venujeme odhadu asymptotických variančných matíc pre Bickelovo h_2, vychádzajúce z dvojíc dát, a novonavrhnuté h_3, vychádzajúce z trojíc dát.

Bickelov test s dvojicami pracuje s vektorom h_2, ktorý je definovaný v (1). V kapitole 1 sme odvodili vzorec pre asymptotickú variančnú maticu π_2 výrazu $\sqrt{n}h_2$. Prvok v i-tom riadku a j-tom stĺpci je:

$$\pi_2^{ij} = 4E\left[\left(F_i(X_{i1}) - \frac{1}{2}\right)\left(F_j(X_{j1}) - \frac{1}{2}\right)\right].$$

V článku [1] sa uvádza odhad $\hat{\pi}_2$ tejto asymptotickej variančnej matice:

$$\hat{\pi}_2^{ij} = \frac{1}{n(n-1)(n-2)} \sum_{r\neq s\neq t} I^+(X_{ir} - X_{is})I^+(X_{jr} - X_{jt}) \;, \quad \text{pre } i \neq j$$

$$\frac{1}{12} \qquad\qquad\qquad\qquad\qquad\qquad , \quad \text{pre } i = j.$$

Simuláciami tejto variančnej matice sme zistili, že tento odhad ale nedáva správne hodnoty blízke k π_2, ktorú sme odvodili v časti 1.1. Človeku ľahko napadne, že v Bickelovom článku [1] chýba vynásobenie celej matice číslom 4 a následne taktiež vo vzorci (2) chýba odčítanie -1 na konci.

Ak teda upravíme Bickelov vzorec odhadu $\hat{\pi}_2$ na základe správneho vzorca, tak dostaneme:

$$\hat{\pi}_2^{ij} = 4\left[\frac{1}{n(n-1)(n-2)} \sum_{\substack{r,s,t \\ r\neq s, s\neq t}}^{n} I^+(X_{ir} - X_{is})I^+(X_{js} - X_{jt})\right] - 1 \;, \qquad pre \qquad i \neq j$$

$$\hat{\pi}_2^{ij} = \frac{1}{3} \qquad\qquad\qquad\qquad\qquad\qquad , \qquad pre \qquad i = j.$$

Konvergenciu tohto odhadu k π_2 sme overili simuláciami.

Následne testovou štatistikou je výraz:

$$n\left(h_2 - \frac{1}{2}\right)^T \hat{\pi}_2^{-1}\left(h_2 - \frac{1}{2}\right),$$

kde $\frac{1}{2}$ je stredná hodnota h_2 za platnosti H_0. Testovú štatistiku porovnávame s chí-kvadrátom s p stupňami voľnosti (pozri článok [1]).

Simuláciami sme sa pokúsili odhadnúť pravdepodobnosť chyby 1. druhu tohto testu, dáta sme generovali z dvojrozmerného normálneho rozdelenia $N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4.80 & 5.12\\5.12 & 6.20\end{bmatrix}\right)$ a

testovali sme, či sa stred symetrie rovná $(0,0)^T$. Počet dát n sme menili a pre každé n sme generovali 10 000 simulácií. Takto špecifické rozdelenie používame kvôli konzistentnosti pri ďalších testoch, kde sa vlastnosti tohto rozdelenia, ako napríklad korelácia, využijú. V Tabuľke 1 si môžeme všimnúť, že odhadnutá pravdepodobnosť chyby 1. druhu vychádza pri malom n výrazne nad nominálnou hodnotou 5%.

n	30	40	60	100
Odhadnutá P(chyba 1. druhu)	0.080	0.069	0.058	0.057

Tabuľka 1: Simulovaná pravdepodobnosť chyby 1. druhu Bickelovho testu s dvojicami v závislosti od počtu dát n

1.3.1 Odhad asymptotickej variančnej matice pre test s trojicami

Test s trojicami pracuje s vektorom h_3, ktorý je definovaný v (16).

Vzorec pre asymptotickú variančnú maticu výrazu $\sqrt{n}h_3$, ktorý sme odvodili v kapitole 1 je:

$$\pi_3^{ij} = 9E\left[\left(F_i(x_{i2}+x_{i3})-\frac{1}{2}\right)\left(F_j(x_{j3}+x_{j4})-\frac{1}{2}\right)\right].$$

Chceme nájsť odhad $\hat{\pi}_3$ pre túto maticu. Očividne by malo byť π_3^{ij} blízke hodnote

$$9E\left[\left(\hat{F}_i(x_{i2}+x_{i3})-\frac{1}{2}\right)\left(\hat{F}_j(x_{j3}+x_{j4})-\frac{1}{2}\right)\right], \tag{21}$$

kde $\hat{F}_i(y), \hat{F}_j(y)$ predstavujú empirické distribučné funkcie, ktoré sa dajú prepísať pomocou indikátorovej funkcie I^+, ktorú sme definovali na začiatku kapitoly 1, a to nasledovne:

$$\hat{F}_i(y) = \frac{1}{n}\sum_{d=1}^{n} I^+(y - X_{id}), \tag{22}$$

analogicky pre $\hat{F}_j(y)$. Strednú hodnotu vo výraze (21) odhadneme empiricky priemerom cez všetky kombinácie indexov a, b, c výrazov $\left(\hat{F}_i(x_{ia}+x_{ib})-\frac{1}{2}\right)\left(\hat{F}_j(x_{jb}+x_{jc})-\frac{1}{2}\right)$, teda dostaneme odhad:

$$\hat{\pi}_3^{ij} = \frac{9}{n(n-1)(n-2)} \sum_{\substack{a,b,c=1 \\ a\neq b, a\neq c, b\neq c}}^{n} \left(\left[\frac{1}{n}\sum_{d=1}^{n} I^+(X_{ia}+X_{ib}-X_{id})\right]-\frac{1}{2}\right)\left(\left[\frac{1}{n}\sum_{e=1}^{n} I^+(X_{jb}+X_{jc}-X_{je})\right]-\frac{1}{2}\right). \tag{23}$$

Ako výsledok dostávame testovú štatistiku:

$$n(h_3 - 1/2)^T \hat{\pi}_3^{-1}(h_3 - 1/2),$$

kde $\frac{1}{2}$ je stredná hodnota h_3 za platnosti H_0 . Testovú štatistiku porovnávame s chí-kvadrátom s p stupňami voľnosti. Simuláciami sme sa pokúsili odhadnúť pravdepodobnosť chyby 1. druhu tohto testu. Dáta sme generovali z dvojrozmerného normálneho rozdelenia $N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4.80 & 5.12\\5.12 & 6.20\end{bmatrix}\right)$ a testovali sme, či sa stred symetrie rovná $(0,0)^T$. Počet dát n sme menili a pre každé n sme generovali 10 000 simulácií. Výsledky sú v Tabuľke 2:

n	30	40	60	100	150
Odhadnutá P(chyba 1. druhu)	0.024	0.030	0.035	0.044	0.046

Tabuľka 2: Simulovaná pravdepodobnosť chyby 1. druhu Bickelovského testu s trojicami v závislosti od počtu dát n

Silu pre dáta odchýlené od testovaného stredu symetrie $(0,0)^T$ sme odhadovali generovaním dát z normálneho rozdelenia $N_2\left(\begin{bmatrix}0\\posun\end{bmatrix}, \begin{bmatrix}4.80 & 5.12\\5.12 & 6.20\end{bmatrix}\right)$. Počet dát je $n = 100$. Generovali sme 10 000 simulácií a sledovali sme, v akej miere bola zamietnutá hypotéza $H_0 : \mu = (0,0)^T$ pri danom posune skutočného stredu symetrie v y-ovej osi od bodu $(0,0)^T$. Z Tabuľky 3 sa zdá, že test funguje dobre. Avšak ak porovnáme odhadnutú pravdepodob-

posun	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
miera zamietnuntia H_0	0.041	0.116	0.369	0.729	0.936	0.993	1	1

Tabuľka 3: Simulovaná sila nášho testu pomocou trojíc v závislosti od veľkosti "posunu" od platnosti H_0

nosť chyby 1. druhu testu s pomocou dvojíc s pravdepodobnosťou chyby 1. druhu testu pomocou trojíc, zistíme, že Bickelova verzia testu s dvojicami zamieta H_0 v prípade, že platí, s oveľa väčšou pravdepodobnosťou ako je hladina významnosti testu 5%, a to najmä pri menšom počte dát n. Pri našom teste pomocou trojíc bola pravdepodobnosť chyby 1. druhu vždy pod 5%, a to možno vďaka tomu, že sme odhad $\hat{\pi}_3$ pre π_3 skonštruovali principiálne inak ako skonštruoval Bickel odhad $\hat{\pi}_2$ pre π_2. Preto skúsime vytvoriť odhad variančnej matice π_2 pre test s dvojicami rovnakým spôsobom, ako sme vytvorili $\hat{\pi}_3$ pri teste s trojicami.

1.3.2 Alternatívny odhad asymptotickej variančnej matice pre Bickelov test s dvojicami

Rozumným odhadom variančnej matice π_2 by bolo zjavne

$$4E\left[\left(\hat{F}_i(X_{i1}) - \frac{1}{2}\right)\left(\hat{F}_j(X_{j1}) - \frac{1}{2}\right)\right],$$

kde $\hat{F}_i(y), \hat{F}_j(y)$ predstavujú empirické distribučné funkcie, ktoré sa dajú prepísať v tvare indikátorovej funkcie ako v rovnici (22). Strednú hodnotu odhadneme empiricky priemerom cez všetky dátové body X_{ik}, ktorých máme dokopy n. Pri výpočte hodnoty empirickej distribučnej funkcie $\hat{F}(X_{ik})$ vynechávame k-te meranie X_{ik}, v ktorom hodnotu $\hat{F}(X_{ik})$ počítame aby sme dostali nevychýlený odhad pre π_2. Čiže máme odhad $\tilde{\pi}_2$, ktorého i-ty riadok a j-ty stĺpec je definovaný ako:

$$\tilde{\pi}_2^{ij} := \frac{4}{n}\sum_{k=1}^{n}\left(\frac{1}{n-1}\sum_{\substack{l=1\\l\neq k}} I^+(X_{ik} - X_{il}) - \frac{1}{2}\right)\left(\frac{1}{n-1}\sum_{\substack{l=1\\l\neq k}} I^+(X_{jk} - X_{jl}) - \frac{1}{2}\right).$$

V prípade, že počítame odhad variančnej matice v teste s dvojicami takýmto spôsobom, tak dostaneme odhadnutú pravdepodobnosť chyby 1. druhu (ak ju simulujeme rovnakým spôsobom ako v Tabuľke 1) nasledovne:

n	30	40	60	100
Odhadnutá P(chyba 1. druhu)	0.038	0.043	0.044	0.046

Tabuľka 4: Simulovaná pravdepodobnosť chyby 1. druhu Bickelovho testu s dvojicami v závislosti od počtu dát n, ak variančnú maticu odhadujeme pomocou $\tilde{\pi}_2$

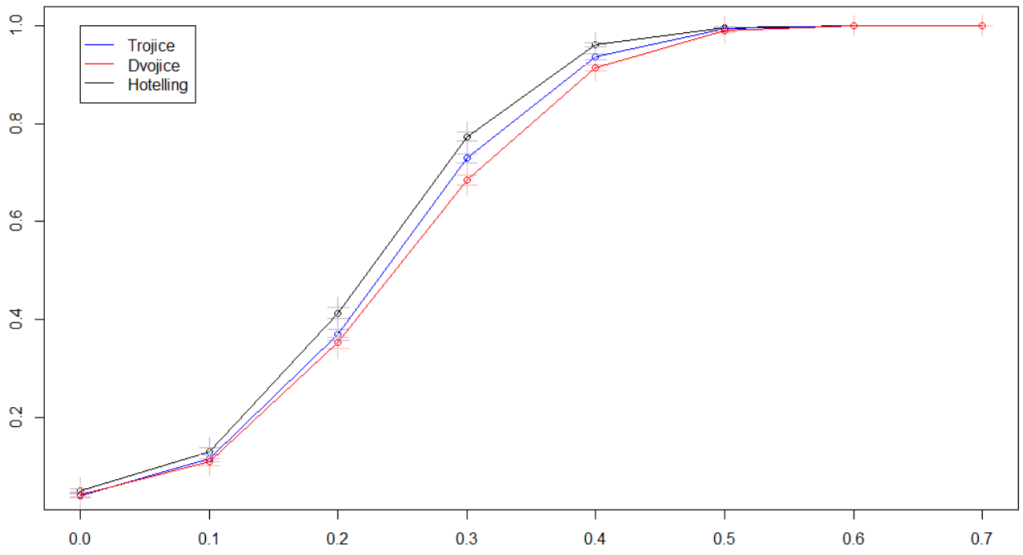
Náš odhad $\tilde{\pi}_2$ variančnej matice π_2 v teste s dvojicami dáva odhad pravdepodobnosti chyby 1. druhu omnoho priaznivejší, nakoľko neprekračuje nominálnu hladinu významnosti 5%.

1.4 Porovnanie sily Bickelovho testu s dvojicami vs. testu s trojicami

V tejto časti porovnávame silu nami upraveného Bickelovho testu s dvojicami, pre ktorý sme vytvorili nový odhad $\tilde{\pi}_2$ variančnej matice π_2, s nami navrhnutým testom s trojicami.

V nasledujúcom grafe je vykreslený odhad síl testov s dvojicami a trojicami pri rôznom posune skutočného stredu symetrie od hypotetizovaného stredu symetrie $(0,0)^T$. Do grafu sme pridali aj tradičný Hotellingov test, ktorý predpokladá normalitu dát.

Dáta sme najprv generovali z dvojrozmerného normálneho rozdelenia $N_2\left(\begin{bmatrix} 0 \\ posun \end{bmatrix}, \begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}\right)$, kde počet dát v simulácii je $n = 100$, a počet simulácií je 10 000. Na Obrázku 1 rovnako ako na ďalších obrázkoch je krížikmi nad a pod vykreslenou hodnotou odhadnutej sily vyznačený 97.5%-ný interval spoľahlivosti pre tento odhad. Ak sa tieto intervaly v niektorej hodnote *posunu* pre rôzne dva testy neprekrývajú, tak dostaneme štatisticky významnú dominanciu jedného testu nad druhým na hladine významnosti 95%.



Obr. 1: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného normálneho rozdelenia

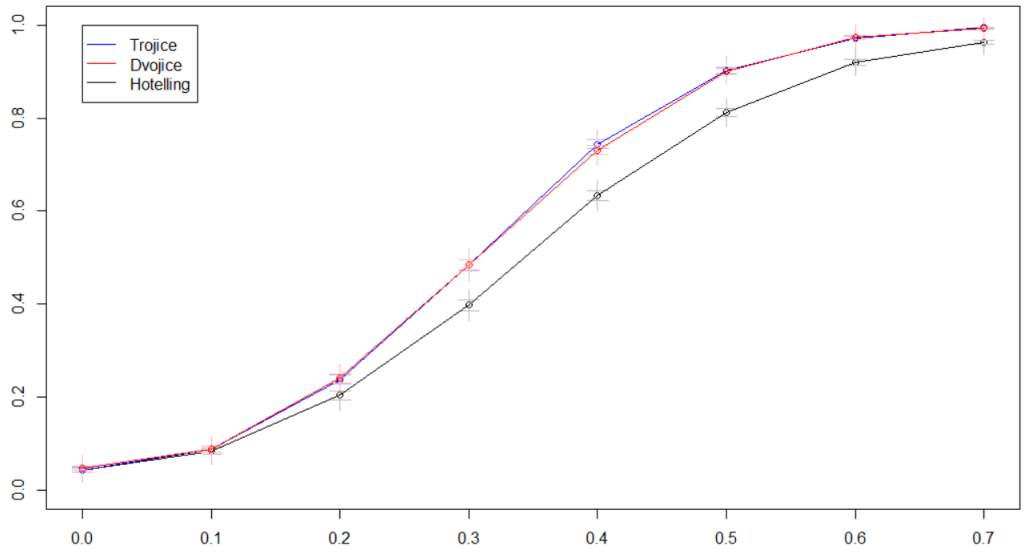
Na Obrázku 1 vidíme, že tak ako sa dalo očakávať, pre normálne rozdelenie má najvyššiu silu Hotellingov test. Nami navrhnutý test pomocou trojíc má mierne navrch nad Bickelovým testom pomocou dvojíc.

Pre odhad sily testu v nasledujúcom grafe sme dáta generovali z dvojrozmerného Studentovho rozdelenia s 3-mi stupňami voľnosti s vektorom vychýlenia $\alpha = \begin{bmatrix} 0 \\ posun \end{bmatrix}$ a so škálovou maticou $\Sigma = \begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}$, teda z rozdelenia, ktoré je definované ako

posun	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Hotelling	0.051	0.131	0.413	0.773	0.960	0.996	1	1
Dvojce	0.044	0.110	0.353	0.685	0.914	0.988	1	1
Trojice	0.0413	0.116	0.369	0.729	0.936	0.993	1	1

Tabuľka 5: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného normálneho rozdelenia

$\frac{Y}{\sqrt{\frac{U}{\nu}}}$, kde $Y \sim N_2(\alpha, \Sigma)$ a $U \sim \chi_\nu$. V tomto prípade sa U riadi chí-kvadrát rozdelením s 3-mi stupňami voľnosti. Toto rozdelenie má ťažšie chvosty oproti normálnemu v predchádzajúcej simulácii.



Obr. 2: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s 3-mi stupňami voľnosti

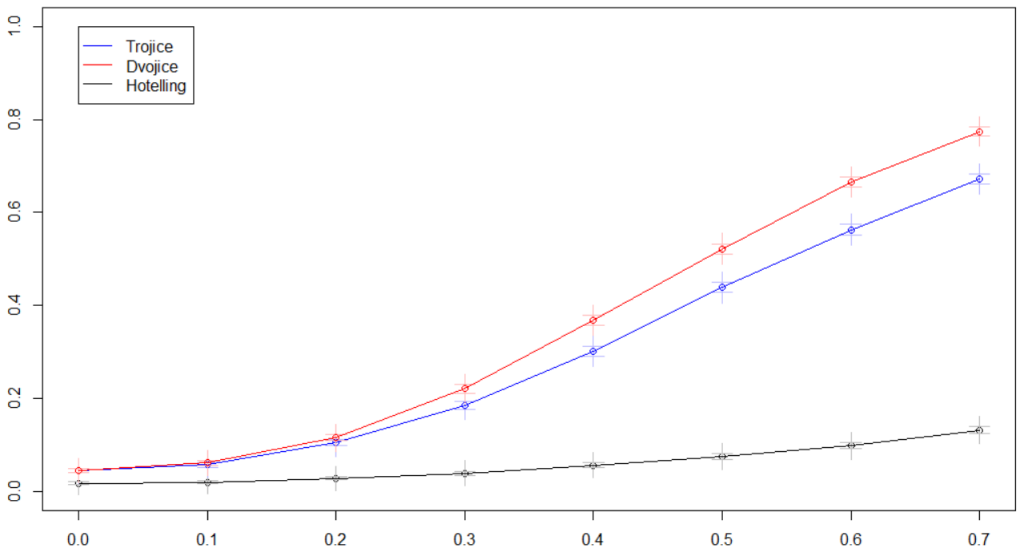
Na Obrázku 2 vidíme, že Hotellingov test si už pri takomto rozdelení dát vedie horšie ako testy pomocou dvojíc a trojíc, ktoré sú na tom veľmi podobne.

V nasledujúcom grafe sme dáta generovali z dvojrozmerného Studentovho rozdelenia s jedným stupňom voľnosti s vektorom vychýlenia $\begin{bmatrix} 0 \\ posun \end{bmatrix}$, a so škálovou maticou

posun	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Hotelling	0.044	0.085	0.204	0.398	0.633	0.813	0.920	0.963
Dvojce	0.048	0.088	0.241	0.485	0.731	0.900	0.974	0.993
Trojice	0.044	0.087	0.238	0.484	0.744	0.903	0.973	0.994

Tabuľka 6: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s 3-mi stupňami voľnosti

$\begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}$. Toto rozdelenie má ešte ťažšie chvosty než predchádzajúce rozdelenie.



Obr. 3: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s jedným stupňom voľnosti

Na Obrázku 3 vidíme, že ak dáta generujeme z rozdelenia s veľmi ťažkými chvostami, tak nami upravený Bickelov test s dvojicami má stále navrch nad nami navrhnutým testom s trojicami.

posun	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Hotelling	0.018	0.020	0.029	0.040	0.057	0.076	0.098	0.132
Dvojce	0.045	0.062	0.116	0.221	0.368	0.522	0.666	0.774
Trojice	0.045	0.057	0.105	0.185	0.301	0.439	0.563	0.672

Tabuľka 7: Porovnanie odhadnutej sily nami upraveného Bickelovho testu s dvojicami, nášho testu s trojicami a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s jedným stupňom voľnosti

2 Blumenov test a jeho rozšírenie

2.1 Blumenov test

Táto časť vychádza z článku [2].

Nech $X_a = (X_{1a}, X_{2a})^T,\ a = 1, 2, \ldots, n$, pochádzajú z dvojrozmerného rozdelenia, pre ktoré existuje bod $\mu = (\mu_x, \mu_y)^T$, pre ktorý platí:

$$Pr[a(X_{11} - \mu_x) + b(X_{21} - \mu_y) > 0] = \frac{1}{2}\,,$$

pre všetky reálne, nie naraz nulové a, b. Tento bod má vlastnosť, že keď rozdelíme priestor na dve polovice priamkou prechádzajúcou cez tento bod, tak pravdepodobnosť, že dátový bod sa vygeneruje na jednej alebo druhej polovici, je rovnaká. Tento bod μ nazveme mediánom tohto rozdelenia. Testujeme hypotézu:

$$H_0 : \mu = (0, 0)^T \quad vs. \quad H_1 : \mu \neq (0, 0)^T.$$

Blumenova procedúra testu spočíva v nasledovných krokoch:

1. Predelením ich vzdialenosťou od bodu $(0, 0)^T$ projektujeme dáta $X_1, X_2, \ldots, X_n$ na jednotkovú kružnicu okolo bodu $(0, 0)^T$.

2. Body na kružnici so zápornou y-ovou súradnicou preklopíme stredovo súmerne okolo bodu $(0, 0)^T$ nad x-ovú os.

3. Takto máme zoradené vektory podľa ich pozície na kružnici, počínajúc na tretej hodine a idúc proti smeru hodinových ručičiek. Následne vektory rozložíme rovnomerne na hornej polkružnici.

4. Body, ktoré sa pôvodne nachádzali na spodnej polovici, opäť preklopíme okolo stredu $(0, 0)^T$ naspäť.

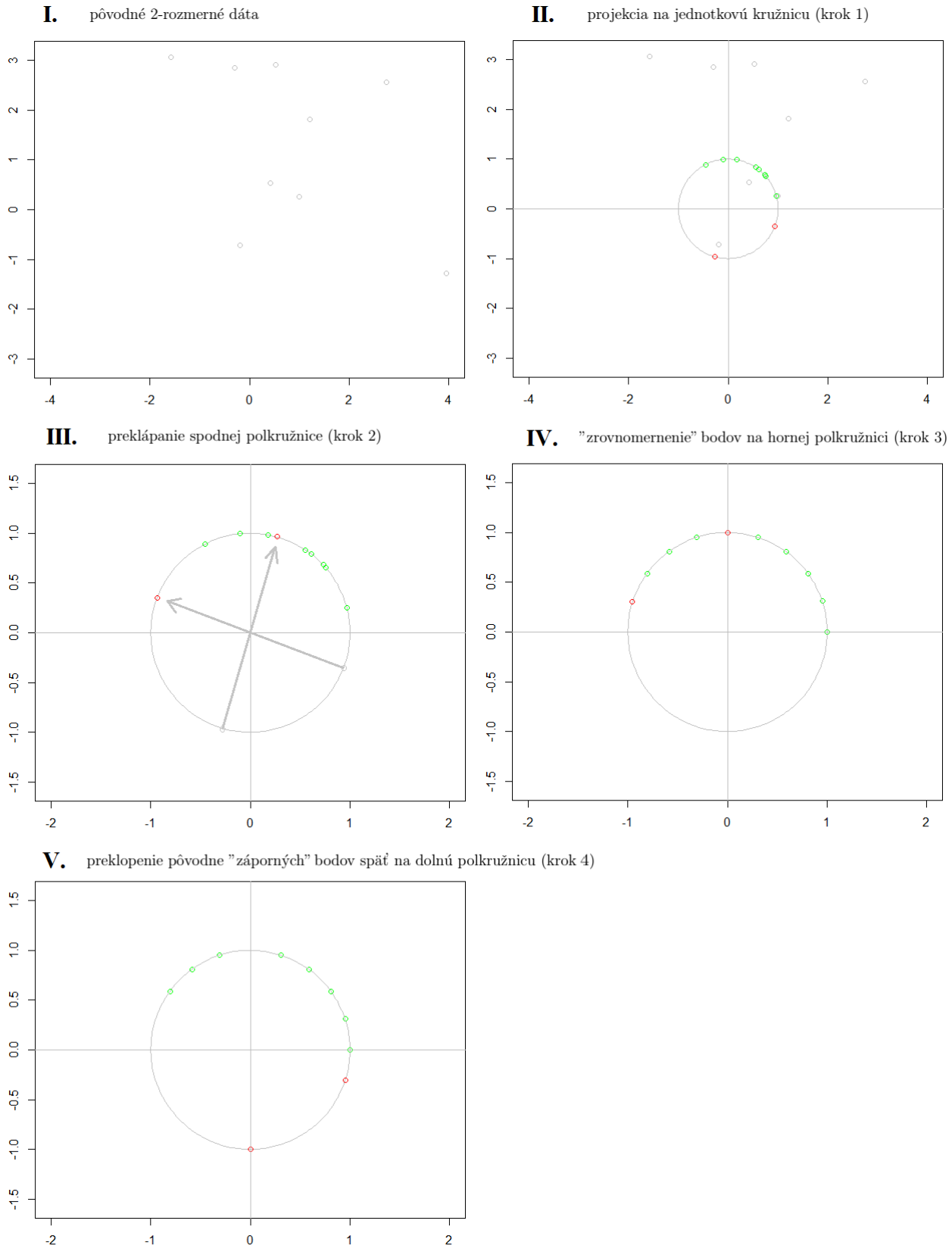
5. Sčítame všetky vektory, testová štatistika je norma na druhú výsledného vektora $(V_1, V_2)^T$ predelená $\frac{n}{2}$, teda:

$$T = \frac{2}{n}(V_1^2 + V_2^2)\,, \tag{24}$$

kde

$$V_1 = \sum_{j=0}^{n-1} a_j cos\left(\frac{\pi j}{n}\right) \quad , \quad V_2 = \sum_{j=0}^{n-1} a_j sin\left(\frac{\pi j}{n}\right) \ ,$$

kde a_j je buď -1 alebo 1 podľa toho, či máme alebo nemáme záporné znamienko pri pôvodnej y-ovej súradnici vektora, ktorý bol v usporiadaní v 3. kroku na j-tej pozícií.



Obr. 4: Ilustrácia procedúry Blumenovho testu

Na Obrázku 4 ilustrujeme prvé 4 kroky Blumenovho testu. Testovú štatistiku z kroku 5 porovnávame s chí-kvadrát rozdelením s dvoma stupňami voľnosti.

Chí kvadrátovosť za platnosti H_0 sme overili simuláciami. Dáta sme generovali z rozdelenia $N_2[(0,0)^T, I_2]$, kde n je počet dát v súbore, a m je počet vykonaných simulácií, potom odhadnutá P(chyba 1. druhu) bude počet zamietnutí H_0 deleno celkový počet testov m, pričom H_0 zamietame, ak testová štatistika vyjde väčšia ako 95%-ný kvantil chí-kvadrát rozdelenia s dvomi stupňami voľnosti.

	n=15	n=30	n=50	n=100
Odhadnutá P(chyba 1. druhu)	0.0509	0.0576	0.0506	0.0524

Tabuľka 8: Simulovaná pravdepodobnosť chyby 1. druhu Blumenovho testu

Z Tabuľky 8 je vidno, že odhadnutá pravdepodobnosť chyby 1. druhu sa nevychyľuje príliš od nominálnej hodnoty 5%.

2.2 Afínna invariantnosť Blumenovho testu

V pôvodnom článku [2] autor tvrdí, že test je invariantný voči zmenám rozsahu a rotácii osí. V tejto časti sa pokúsime ukázať, že testová štatistika je v skutočnosti invariantná voči akejkoľvek afínnej transformácií.

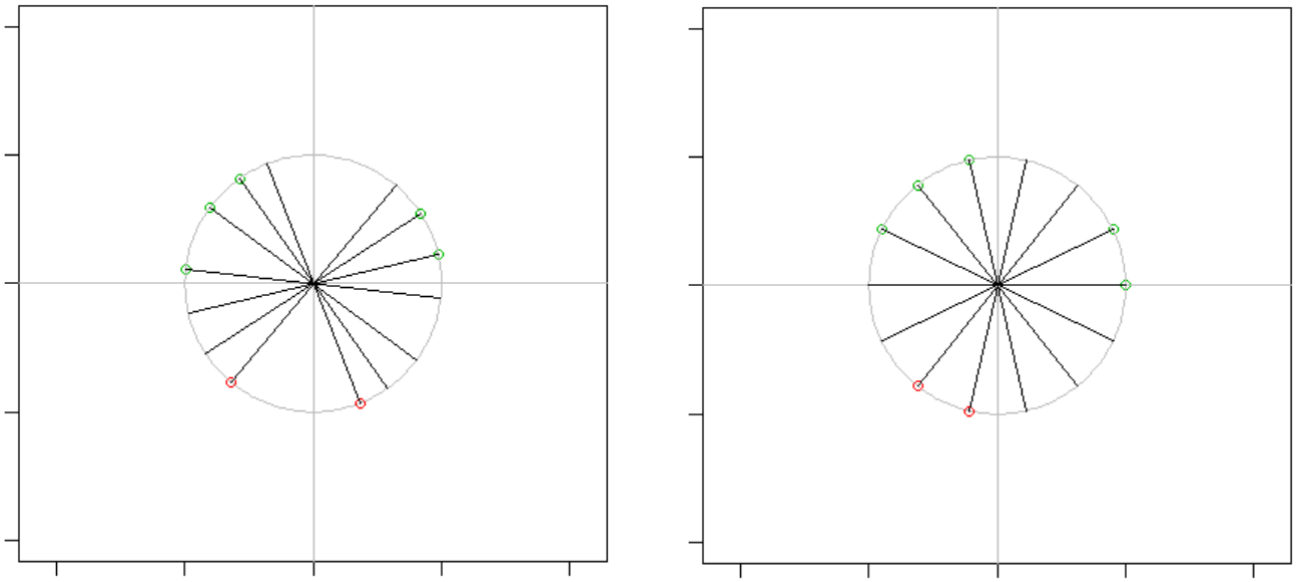
Hovoríme, že test je afínne invariantný, ak testová štatistika pri vstupných dátach $X_1, X_2, \ldots, X_n$ sa rovná testovej štatistike so vstupnými dátami $AX_1, AX_2, \ldots, AX_n$, kde sme pôvodné dáta prenásobili 2x2 regulárnou, pevne danou maticou A.

Uvedomme si, že kroky 2 až 4 v testovej procedúre sa dajú nahradiť nasledovne:

1. Pre každý bod skonštruujeme priamku prechádzajúcu cez tento bod a stred $(0,0)^T$. Táto priamka nám definuje protiľahlý bod na kružnici.

2. Takto získané priamky rozložíme rovnomerne na kružnici, ekvivalentne si môžeme predstaviť zrovnomernenie $2n$-uholníka tvoreného pôvodnými a ich protiľahlými bodmi.

Následne testovú štatistiku počítame z výsledného $2n$-uholníka, pričom uvažujeme len pôvodné body na kružnici. Uvedomme si tiež, že je jedno podľa akého bodu zrov-



Obr. 5: Ilustrácia alternatívnej procedúry pre kroky 2. až 4. Blumenovho testu

nomerňujeme ostatné body, keďže testová štatistika je plne determinovaná "vzorom" pôvodných n bodov na výslednom $2n$-uholníku.

Akákoľvek afínna transformácia nemôže vymeniť poradie týchto priamok, z čoho vyplýva, že testová štatistika ostáva nezmenená.

2.3 Ranková verzia Blumenovho testu

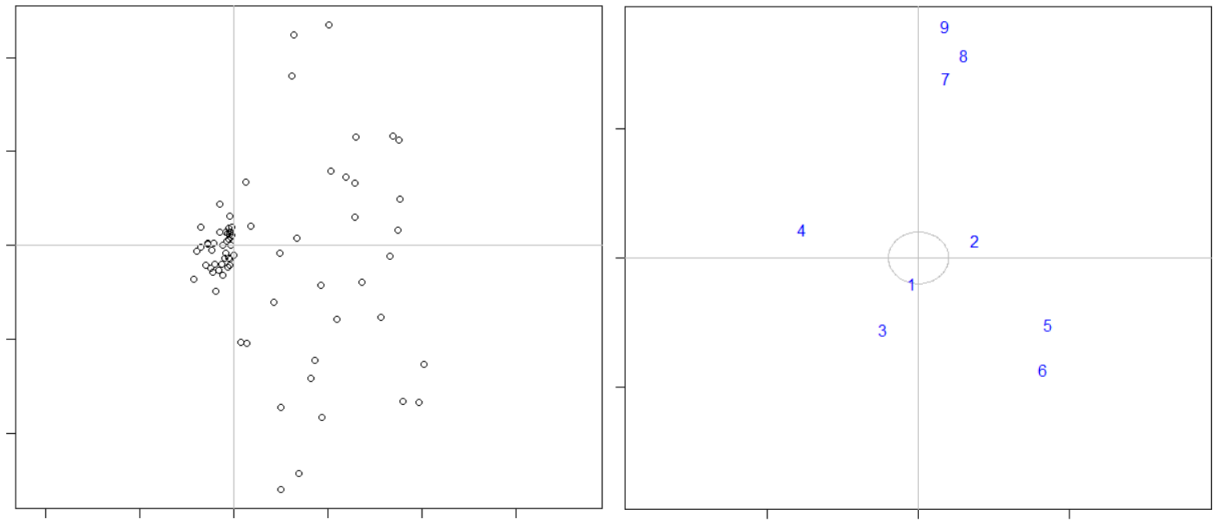
V tejto podkapitole navrhujeme novú verziu Blumenovho testu pridaním rankov k pozorovaniam.

Čo by sa stalo, kebyže chceme Blumenovým testom testovať hypotézu nie o mediáne dát ale o strede symetrie? Teda testujeme:

H_0 : rozdelenie je symetrické okolo $(0,0)^T$ $\quad vs. \quad$ H_1 : rozdelenie nie je symetrické okolo $(0,0)^T$.

Predstavme si, že máme dáta z rozdelenia, ktoré je smerovo rovnomerné okolo $(0,0)^T$, teda každý smer od $(0,0)^T$ je rovnako pravdepodobný, no pre rôzne smery je priemerná vzdialenosť od $(0,0)^T$ rozdielna. Pri takýchto rozdeleniach môže dôjsť k tomu, že stred symetrie μ rozdelenia je výrazne rozdielne od $(0,0)^T$, pričom bod $(0,0)^T$ je stále mediánom rozdelenia. V takýchto prípadoch teda Blumenov test nebude schopný rozlíšiť neplatnosť nulovej hypotézy o symetrii, lebo dáta projektujeme na kružnicu. Preto navrhneme modifikovanú

verziu Blumenovho testu s pridanými rankami. Ranky sú dané poradím vzdialenosti dát od bodu $(0,0)^T$, vďaka čomu budeme schopný použiť test na testovanie stredu symetrie na širšej škále rozdelení.



Obr. 6: Naľavo: príklad v rôznych smeroch heteroskedastického rozdelenia, napravo: ilustrácia rankovania dátových bodov

Testová štatistika je nasledovná:

$$T = \frac{1}{Var(V_1)}(V_1^2 + V_2^2),$$

kde do členov V_1 a V_2 pridávame aj príslušné ranky jednotlivých bodov, teda:

$$V_1 = \sum_{j=0}^{n-1} cos\left(\frac{\pi j}{n}\right) a_j rank_j \quad , \quad V_2 = \sum_{j=0}^{n-1} sin\left(\frac{\pi j}{n}\right) a_j rank_j.$$

Čiže sme upravili pôvodnú testovú štatistiku (24), v ktorej $Var(V_1)$ sa rovnalo $n/2$, teraz vypočítame $Var(V_1)$ pre náš prípad:

$$E(V_1) = \sum_{j=0}^{n-1} \left[cos\left(\frac{\pi j}{n}\right)\right] E(a_j rank_j) = 0.$$

Keďže a_j a $rank_j$ sú za platnosti H_0 nezávislé, tak $E(a_j rank_j) = E(a_j)E(rank_j) = 0$, keďže za platnosti H_0 platí $E(a_j) = 0$. Keďže $E(a_i rank_i) = 0$ a a_j, a_i sú nezávislé od $rank_j, rank_i$, tak:

$$Cov(a_j rank_j, a_i rank_i) = E[(a_j rank_j - E[a_j rank_j])(a_i rank_i - E[a_i rank_i])] =$$
$$= E(a_j a_i rank_j rank_i) = E(a_j)E(a_i)E(rank_j rank_i) = 0 \cdot E(...) = 0.$$

Teraz počítajme varianciu:

$$Var(V_1) = Var\left(\sum_{j=0}^{n-1} cos\left(\frac{\pi j}{n}\right) a_j rank_j\right) = \sum_{j=0}^{n-1} cos\left(\frac{\pi j}{n}\right)^2 Var(a_j rank_j). \tag{25}$$

Vypočítajme $Var(a_j rank_j)$:

$$Var(a_j rank_j) = E(a_j^2)E(rank_j^2) - E(a_j)^2(E(rank_j))^2, \tag{26}$$

keďže $E(a_j) = 0$ a $E(a_j^2) = 1$, tak:

$$(26) = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{6}(n+1)(2n+1).$$

Keďže:

$$cos\left(\frac{\pi j}{n}\right)^2 = \frac{1}{2}\left(cos\left(\frac{2\pi j}{n}\right) + 1\right),$$

tak:

$$\sum_{j=0}^{n-1} cos\left(\frac{\pi j}{n}\right)^2 = \sum_{j=0}^{n-1} \frac{1}{2} cos\left(\frac{2\pi j}{n}\right) + \sum_{j=0}^{n-1} \frac{1}{2} = 0 + \frac{n}{2}.$$

Kosínusový člen sa v sume vynuluje, keďže predstavuje rovnomerne rozdelené body na celej kružnici. Zostáva nám v (25) $\frac{n}{2} Var(a_j rank_j)$, teda:

$$Var(V_1) = \frac{1}{12} n(n+1)(2n+1).$$

Výsledok platí analogicky pre V_2.

Skutočnosť, že testová štatistika sa za platnosti H_0 riadi chí-kvadrát rozdelením s dvomi stupňami voľnosti, sa dá overiť simuláciami, čo sme aj overili (Tabuľka 9, stĺpec 0.0).

Môžeme teraz porovnať silu oboch testov pri porušení H_0, a sledovať, ktorý test si poradí so situáciou lepšie.

Dáta sú najprv generované z normálneho rozdelenia s $N_2[(0, \mu_y)^T, I]$, teda porušenie H_0 je posun skutočného stredu generovaného rozdelenia od $(0, 0)^T$, počet vygenerovaných bodov je 40, počet simulácií je 10 000.

V Tabuľke 9 si môžeme všimnúť, že pri tomto rozdelení dát má ranková verzia Blumenovho testu vyššiu silu.

μ_y	0.0	0.1	0.2	0.3	0.5
Originálny Blumen	0.0474	0.0720	0.1420	0.2803	0.6643
Ranková verzia Blumena	0.0461	0.0748	0.1662	0.3415	0.7639

Tabuľka 9: Porovnanie odhadnutej sily testov pri meniacom sa parametri k

V Tabuľke 10 uvažujeme dáta $(x_i, y_i)^T$ generované z rozdelenia $N_2[(0,0)^T, I]$. Následne každý bod i, pre ktorý platí $x_i > 0$ prenásobíme konštantou k, čím dostaneme rozdelenie, v ktorom hustota naľavo od y-ovej osy je zhodná s $N_2[(0,0)^T, I]$, zatiaľčo napravo $N_2[(0,0)^T, k^2 I]$, pričom k meníme. Testujeme hypotézu nie o mediáne, pre ktoré bol pôvodný Blumenov test navrhnutý, ale hypotézu o strede symetrie: H_0 : rozdelenie je symetrické okolo $(0,0)^T$.

k	1	2	3	5	10
Originálny Blumen	0.0474	0.0480	0.0485	0.0484	0.0499
Rankovaná verzia Blumena	0.0461	0.2124	0.3716	0.4892	0.5409

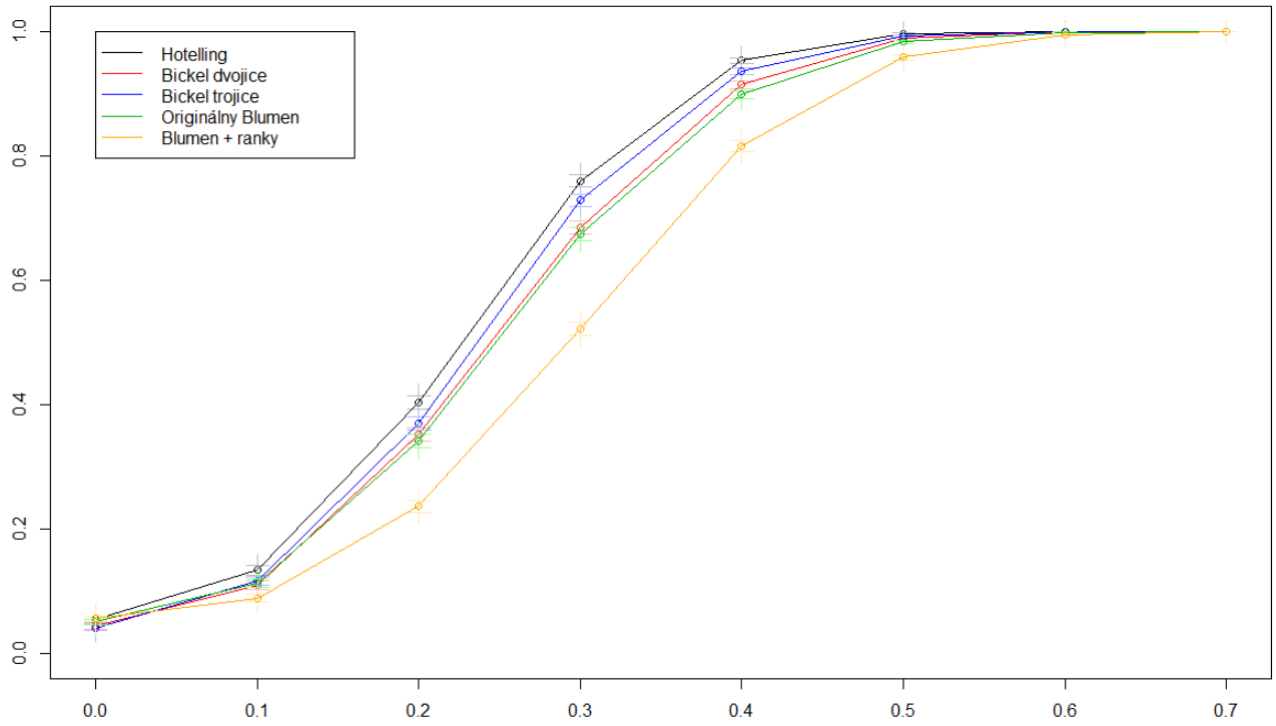
Tabuľka 10: Porovnanie odhadnutej sily testov pri meniacom sa parametri k

V Tabuľke 10 je jasne vidno, že originálny Blumenov test nie je schopný si všimnúť neplatnosť hypotézy o symetrii, narozdiel od rankovej verzie.

3 Porovnanie sily testov

V tejto časti porovnáme odhadnutú silu testov, ktorým sme sa venovali, a to Bickelov test pomocou dvojíc dát, náš test pomocou trojíc dát, Blumenov test, rankovú verziu Blumenovho testu a ešte klasický Hotellingov test. V časti 1.3 sme porovnávali Bickelov test pomocou dvojíc s našim testom pomocou trojíc a Hotellingov test. Silu testov sme odhadovali pomocou generovania dát z troch rôznych rozdelení: 1) z normálneho, 2) z dvojrozmerného Studentovho s 3-mi stupňami voľnosti, 3) z dvojrozmerného Studentovho s jedným stupňom voľnosti. Tieto tri rozdelenia majú progresívne ťažšie chvosty. Teraz do grafov s porovnaním sily v časti 1.3 vykreslíme aj odhadnutú silu pre Blumenov test a rankovú verziu Blumenovho testu.

1) Dáta sme generovali z dvojrozmerného normálneho rozdelenia $N_2\left(\begin{bmatrix} 0 \\ posun \end{bmatrix}, \begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}\right)$, kde počet dát v simulácii je $n = 100$, a počet simulácií je 10 000.

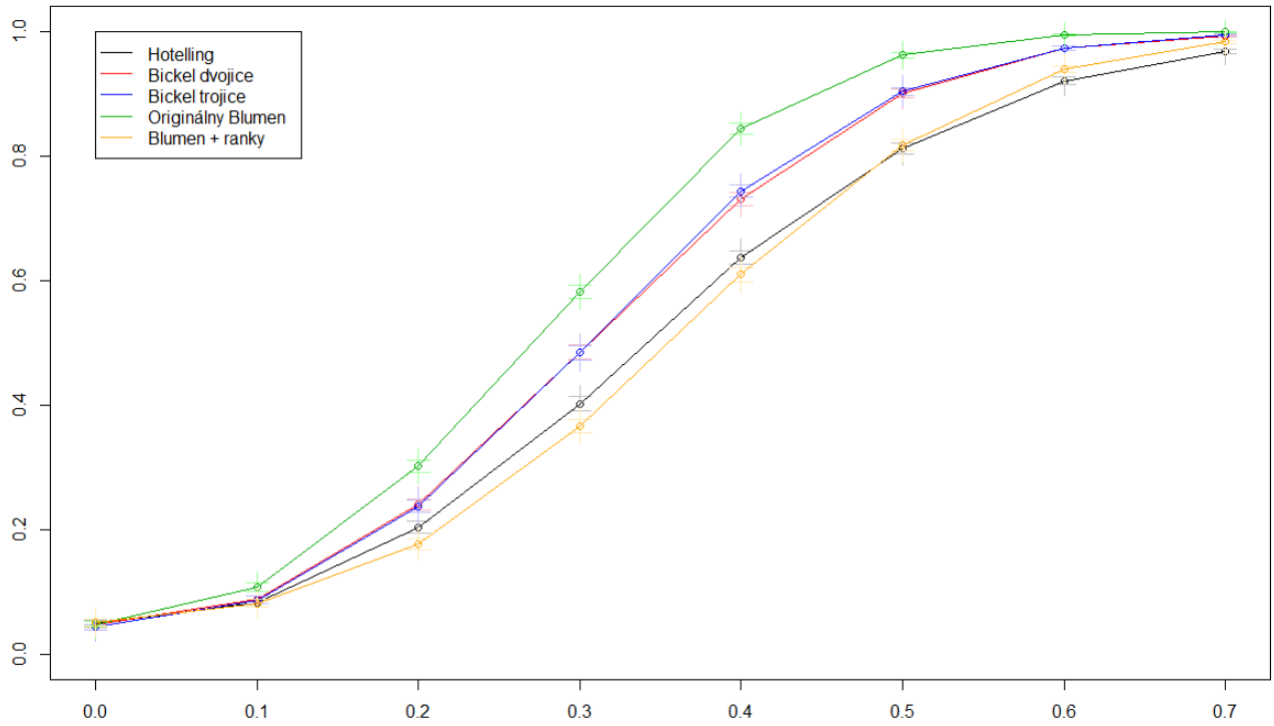


Obr. 7: Porovnanie odhadnutej sily Bickelovho testu s dvojicami, nášho testu s trojicami, Blumenovho testu, rankovej verzie Blumenovho testu a Hotellingovho testu pre dáta generované z dvojrozmerného normálneho rozdelenia

Pri tomto rozdelení na Obrázku 7 vidíme, že Blumenov test si vedie približne rov-

nako ako Bickelov test s dvojicami, pričom ranková verzia Blumenovho testu dopadla zo všetkých ďaleko najhoršie.

Pre odhad sily testu v nasledujúcom grafe sme dáta generovali z dvojrozmerného Studentovho rozdelenia s 3-mi stupňami voľnosti s vektorom vychýlenia $\begin{bmatrix} 0 \\ posun \end{bmatrix}$, a so škálovou maticou $\begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}$. Toto rozdelenie má ťažké chvosty. Výsledok odhadov sily je vidno na Obrázku 8.

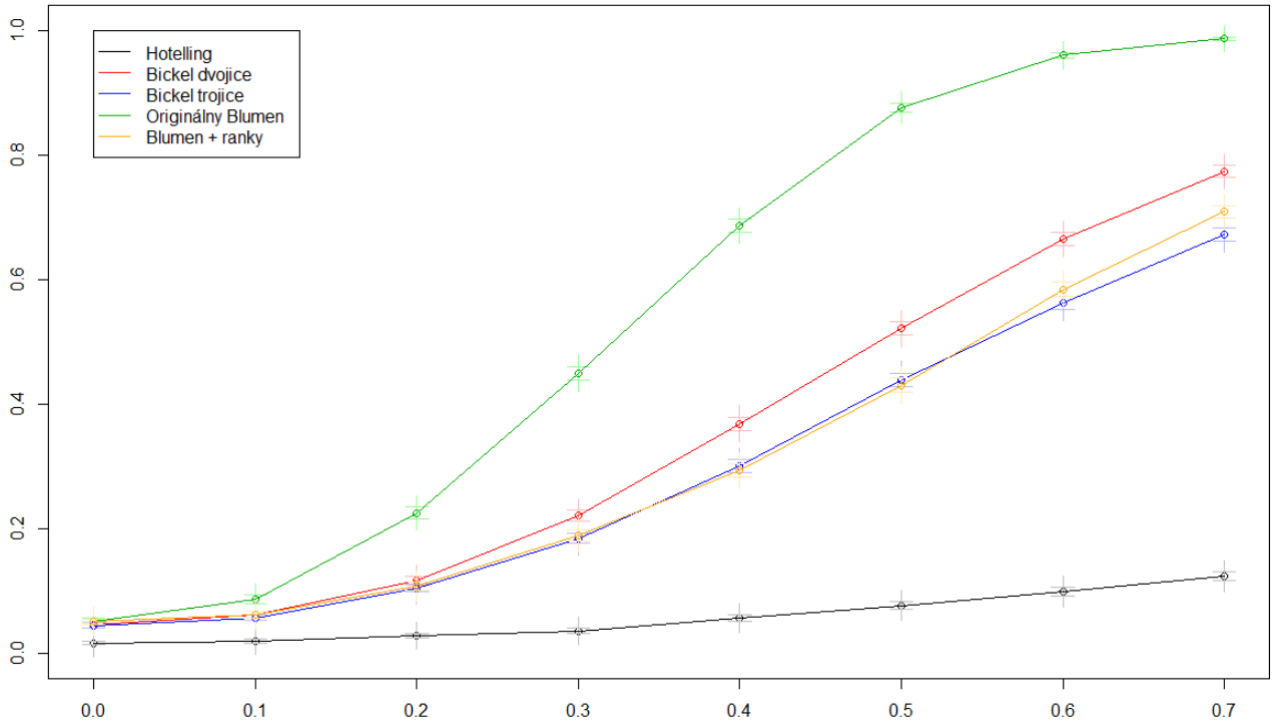


Obr. 8: Porovnanie odhadnutej sily Bickelovho testu s dvojicami, nášho testu s trojicami, Blumenovho testu, rankovej verzie Blumenovho testu a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s 3-mi stupňami voľnosti

Pri tomto rozdelení vidíme na Obrázku 8, že originálny Blumenov test má zjavne navrch nad všetkými ostatnými. Ranková verzia a Hotellingov test, ktorý predpoladá normalitu dát, dopadli najhoršie. Testy na Bickelov štýl sú niekde medzi.

V nasledujúcom grafe sme dáta generovali z dvojrozmerného Studentovho rozdelenia s jedným stupňom voľnosti, s vektorom vychýlenia $\begin{bmatrix} 0 \\ posun \end{bmatrix}$ a so škálovou maticou

$\begin{bmatrix} 4.80 & 5.12 \\ 5.12 & 6.20 \end{bmatrix}$. Toto rozdelenie má ešte ťažšie chvosty ako predošlé Studentovo rozdelenie.



Obr. 9: Porovnanie odhadnutej sily Bickelovho testu s dvojicami, nášho testu s trojicami, Blumenovho testu, rankovej verzie Blumenovho testu a Hotellingovho testu pre dáta generované z dvojrozmerného Studentovho rozdelenia s jedným stupňom voľnosti

Na obrázku 9 jasne vidíme, že zo všetkých testov je najlepší originálny Blumenov test. Pri rozdelení s takto ťažkými chvostami si Hotellingov test vedie veľmi slabo. Ranková verzia Blumenovho testu si vedie porovnateľne ako náš test pomocou trojíc. Z týchto výsledkov sa zdá, že na ťažké chvosty je dobré použiť originálny Blumenov test, ktorý má vysokú silu, aj nízku výpočtovú náročnosť, avšak medzi jeho nevýhody patrí použiteľnosť len na úzku škálu rozdelení, keďže sa dá použiť len na dvojrozmerné dáta s mediánom.

Záver

V práci sme sa venovali dvom neparametrickým testom: 1) Bickelovmu testu z článku [1], ktorý je použiteľný pre viacrozmerné dáta a 2) Blumenovmu testu z článku [2], ktorý je použiteľný pre dvojrozmerné dáta. Cieľom práce bolo nájsť vylepšenia týchto testov.

V kapitole 1 sme sa pokúsili na náhodne vygenerovaných dátach aplikovať Bickelov test pomocou dvojíc tak, ako je uvedený v článku [1]. Avšak Bickelov odhad $\hat{\pi}_2$ asymptotickej variančnej matice π_2 sa nezhodoval s asymptotickou variančnou maticou odhadnutou simuláciami, preto sme v časti 1.1 odvodili vzorec pre π_2 a na základe neho sme opravili chyby v Bickelových vzorcoch (na strane 162 článku [1] vzorec (2.3),(b) chýba vynásobenie číslom 4). V časti 1.3 sme na základe správneho vzorca pre π_2 opravili Bickelov odhad $\hat{\pi}_2$ (na strane 166 článku [1] pod vzorcom (5.1) sa uvádza odhad, v ktorom chýba vynásobenie číslom 4 a odčítanie -1 na konci).

V časti 1.2 navrhujeme test s použitím trojíc dát, ktorý je analogický Bickelovmu testu, ktorý používa dvojice dát. Odvodili sme tvar asymptotickej variančnej matice π_3, ktorý vyšiel v mnohom podobný ako pri teste pomocou dvojíc. Pre maticu π_3 sme vytvorili odhad $\hat{\pi}_3$, pomocou ktorého sme simulovali správanie testu za platnosti nulovej hypotézy.

Odhadnutá pravdepodobnosť chyby 1. druhu testu pomocou trojíc vyšla veľmi priaznivá, keďže neprekračovala nominálnu hodnotu 5%. To sa nedalo povedať pre Bickelov test pomocou dvojíc s použitím Bickelovho odhadu $\hat{\pi}_2$, pri ktorom odhadnutá pravdepodobnosť chyby prvého druhu vychádza výrazne nad 5%, hlavne pri malom počte dát n. Tento nesúlad sme podozrievali, že nastal kvôli principiálne odlišným odhadom $\hat{\pi}_2$ a $\hat{\pi}_3$, preto sme pre maticu π_2 vytvorili nový odhad $\tilde{\pi}_2$, pomocou ktorého mal vylepšený Bickelov test pomocou dvojíc odhadnutú pravdepodobnosť chyby 1. druhu omnoho prijateľnejšiu.

Následne sme v časti 1.4 porovnali sily upraveného Bickelovho testu pomocou dvojíc, nášho testu pomocou trojíc a Hotellingovho testu. Silu sme porovnávali na rozdeleniach s rôzne ťažkými chvostami. Aj keď síce pri niektorých rozdeleniach vyšla sila nášho testu pomocou trojíc lepšie ako pri ostatných, tak je škoda, že pri rozdelení s najťažšími chvostami mal Bickelov test pomocou dvojíc výrazne navrch.

V kapitole 2 sme popisovali procedúru Blumenovho dvojrozmerného testu a zistili sme, že jednotlivé kroky procedúry sa dajú zjednodušiť, na základe čoho sme v časti 2.2 ľahko ukázali afínnu invariantnosť tohto testu.

Keďže Blumenov test testuje polohu mediánu, tak sme sa zamysleli nad tým, ako by mohol Blumenov test zlyhať pri testovaní symetrie a našli sme príklad rozdelenia, ktoré nie je symetrické, no medián má nulový. Preto sme v časti 2.3 pre tento i podobné prípady navrhli vylepšenú verziu testu za pomoci rankov.

V kapitole 3 porovnávame sily všetkých testov uvažovaných v tejto práci opäť na rozdeleniach s rôzne ťažkými chvostami, presnejšie na dvojrozmernom Studentovom a dvojrozmernom normálnom rozdelení. V porovnaní pri rozdelení s ťažkými chvostami jasne zvíťazil originálny Blumenov test, medzi ktorého výhody ťiež patrí nízka výpočtová náročnosť a afínna invariantnosť. Avšak treba si uvedomiť aj nevýhody jednotlivých testov, akou je napríklad v prípade práve originálneho Blumenovho testu jeho použiťelnosť len na dvojrozmerné dáta. Taktiež v niektorých prípadoch je originálny Blumenov test porazený aj rankovou verziou testu, ktorá si zase niekedy vedie ako jedna z najhorších.

Náš test pomocou trojíc si vedie najlepšie spomedzi neparametrických testov pri normálnom rozdelení, no pri ťažších chvostoch ho porážajú originálny Blumenov test, aj jeho Bickelova dvojicová analógia, takže sa bohužiaľ nedá vyhlásiť za všestranné vylepšenie oproti pôvodnej dvojicovej verzii od Bickela.

Veľká časť nášho výskumu bola venovaná programovaniu algoritmov pre jednotlivé testy. Jednoznačne najnáročnejším na efektívne naprogramovanie bol náš test pomocou trojíc, ktorý je výpočtovo náročnejší ako Bickelov test pomocou dvojíc.

Práca by sa dala rozšíriť nájdením vzorca pre asymptotické variančné matice π_k pre Bickelov test pomocou k-tíc, kde k je väčšie ako 3. Ďalej by sa dala rozšíriť porovnaním výpočtovej náročnosti jednotlivých testov, taktiež porovnaním sily testov uvažovaných v práci na širšej škále rozdelení.

Zoznam použitej literatúry

[1] Bickel P.J. - On Some Asymptotically Nonparametric Competitors of Hotelling's T^2, The Annals of Mathematical Statistics, Vol. 36, No. 1 (Feb., 1965), 160–173, 1964

[2] Blumen I. - A New Bivariate Sign Test, Journal of the American Statistical Association, 53:282, 448–456, 1958

[3] Hollander M., Wolfe D. A., Chicken E. - Nonparametric Statistical Methods, Hoboken, 2014

[4] R Core Team: R: *A language and environment for statistical* computing, R Foundation for Statistical Computing, Vienna, Austria, dostupné na internete (15.1.2021): https://www.R-project.org/

Príloha A

V tejto prílohe sa zaoberáme otázkou, aký je počet konštelácií "rovností a nerovností" k premenných. Majme napríklad $k = 3$ premenné, t. j. a, b, c, potom $[a \neq b,\ a \neq c,\ b = c]$ je príklad takejto konštelácie.

Pre všeobecné k sa počet konštelácií riadi vzorcom:

$$PK_k = \sum_{i=0}^{k-1} \binom{k-1}{i} PK_{k-1-i} \tag{27}$$

Odvodenie vzorca

Ak $k = 2$, tak máme dve premenné a, b, ich konštelácie sú: 1) $[a \neq b]$, 2) $[a = b]$, teda počet konštelácií $PK_2 = 2$.

Počet všetkých konštelácií pre tri premenné PK_3 sa rovná 5. Vypíšeme ich:

1. $[a \neq b,\ a \neq c,\ b \neq c]$
2. $[a \neq b,\ a \neq c,\ b = c]$
3. $[a \neq b,\ a = c,\ b \neq c]$
4. $[a = b,\ a \neq c,\ b \neq c]$
5. $[a = b,\ a = c,\ b = c]$.

Ak k zvýšime na $k = 4$, tak pridávame štvrtú premennú d. Táto premenná sa môže rovnať alebo nerovnať pôvodným a, b, c. V prípade, že sa nerovná ani jednej z nich, tak kladieme minimálne podmienky na zvyšné premenné a, b, c, teda pre prípad $[d \neq a, d \neq b, d \neq c, \dots]$ dostaneme všetkých 5 možností pre a, b, c:

1. $[d \neq a, d \neq b, d \neq c, a \neq b,\ a \neq c,\ b \neq c]$
2. $[d \neq a, d \neq b, d \neq c, a \neq b,\ a \neq c,\ b = c]$
3. $[d \neq a, d \neq b, d \neq c, a \neq b,\ a = c,\ b \neq c]$
4. $[d \neq a, d \neq b, d \neq c, a = b,\ a \neq c,\ b \neq c]$
5. $[d \neq a, d \neq b, d \neq c, a = b,\ a = c,\ b = c]$.

Početnosť takýchto prípadov teda vo všeobecnosti bude PK_{k-1}.

V prípade, že sa d rovná exkluzívne jednej z a, b alebo c, tak efektívne znížime počet premenných o jedna, lebo napríklad v prvom prípade $a = d$, a je ekvivalentné d pričom d sa nerovná ani jednej z ostatných premenných. Týmto kladieme na premennú a podmienku, preto je teraz v analogickej pozícií ako v prípade, keď nekladieme žiadne podmienky na premenné b, c, o úroveň nižšie. Inak povedané, počet premenných je nižší o jedna, preto dostaneme PK_{k-2}:

6. $[d = a, d \neq b, d \neq c, a \neq b, a \neq c, c \neq b]$

7. $[d = a, d \neq b, d \neq c, a \neq b, a \neq c, c = b]$.

Počet týchto prípadov vynásobíme počtom kombinácií, ako sa môže premenná d rovnať exkluzívne jednej z pôvodných premenných, čo je všeobecne $\binom{k-1}{1}$, v našom prípade $\binom{3}{1} = 3$. Teda vypíšeme aj ostatné prípady. Pre $d = b$ máme:

8. $[d \neq a, d = b, d \neq c, a \neq b, a \neq c, c \neq b]$

9. $[d \neq a, d = b, d \neq c, a \neq b, a = c, c \neq b]$.

Pre $d = c$ máme:

10. $[d \neq a, d = b, d \neq c, a \neq b, a \neq c, c \neq b]$

11. $[d \neq a, d = b, d \neq c, a \neq b, a = c, c \neq b]$.

Počet prípadov, kedy sa nová premenná d rovná exkluzívne s dvomi premennými z pôvodných premenných, je všeobecne $\binom{k-1}{2}$, v našom prípade $\binom{3}{2} = 3$. V takomto prípade kladieme už silnejšiu podmienku, napríklad ak $[d = a, d = b, d \neq c, \dots]$, tak nám z podmienok vyplýva, že $[\dots, a = b, a \neq c, c \neq b]$. Tým, že a, b, d sa rovnajú a zároveň sa nerovnajú s ostatnými premennými c, dostávame jediný prípad.

Ostatné prípady pre $k = 4$, najprv pre $d = a, d = b$ máme:

12. $[d = a, d = b, d \neq c, a = b, a \neq c, c \neq b]$.

Pre $d = a, d = c$ máme:

13. $[d = a, d \neq b, d = c, a \neq b, a = c, c \neq b]$.

Pre $d = b, d = c$ máme:

14. $[d \neq a, d = b, d = c, a \neq b, a \neq c, c = b]$.

Počet prípadov, kedy sa nová premenná d rovná so všetkými z pôvodných premenných, je všeobecne $\binom{k-1}{k-1}$, čo nám dáva náš posledný prípad pre počet konštelácií "rovností nerovností" pre počet premenných $k = 4$:

15. $[d = a, d = b, d = c, a = b, a = c, c = b]$.

Teda počet konštelácií "rovností a nerovností", ak máme $k = 4$ premenných, je $PK_4 = 15$.

Za logikou vzorca sa skrýva skutočnosť, že ak máme k-ticu, tak z pohľadu ľubovoľnej jednej premennej z k-tice, môžeme konštelácie rozdeliť na kategórie podľa toho, s koľkými z ostatných premenných sa naša premenná exkluzívne rovná.

Ak sa rovná exkluzívne s i premennými z ostatných, tak takýchto kombinácií ostatných premenných, s ktorými sa rovná, je $\binom{k-1}{i}$, pričom ostatných $k - 1 - i$ premenných si žije vlastným životom a môžu teda vytvoriť každú konšteláciu pre $(k - 1 - i)$-ticu, ktorých počet je PK_{k-1-i}.

Tento vzorec je v našej práci užitočný z pohľadu poskytnutia odhadu, aké zložité by bolo odvodenie vzorca pre asymptotickú variančnú maticu, tak ako sme to robili pre test pomocou trojíc v časti 1.3. Na základe vzorca sme zistili, že prinajhoršom budeme uvažovať o rozdelení sumy (20) na najviac $PK_6 = 203$ konštelácií, pričom v skutočnosti sú mnohé konštelácie z hľadiska ďalších výpočtov identické, takže sme mnohé zhrnuli dokopy a vďaka ich nízkej početnosti sme ich v limite nemuseli zvážiť.

V prípade, že by sme odvádzali vzorec pre test pomocou štvoríc, tak v tejto časti odvodenia by sme pracovali s najviac $PK_8 = 4140$ konšteláciami. Dá sa očakávať, že bude stačiť uvažovať len časť z nich a opäť budú mnohé z nich z hľadiska výpočtu ekvivalentné.