UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



Viacrozmerné neparametrické testy nezávislosti

DIPLOMOVÁ PRÁCA

2019 Ľudovít Horváth

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Viacrozmerné neparametrické testy nezávislosti

DIPLOMOVÁ PRÁCA

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

Bratislava 2019 **Ľudovít Horváth**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Viacrozmerné neparametrické testy nezávislosti
Multivariate non-parametric tests of independence

Anotácia: Dobre známe neparametrické testy nezávislosti v jednorozmernom svete sú napr. tie založené na Spearmanovom či Kendallovom koeficiente korelácie. Existuje však množstvo ich zovšeobecnení do sveta viacrozmerného a tiež viacrozmerné prístupy stojace na úplne iných myšlienkach. Téma vznikla už v 30. rokoch minulého storočia, avšak nové prístupy k nej sa objavujú doteraz. Zmyslom diplomovej práce by bolo zoznámiť sa s niektorými testovacími postupmi a pomocou počítačových simulácií na umelých dátach získať odporúčania pre bežného používateľa, od ktorej z metód možno čakať aké správanie v závislosti od typu vstupných dát.

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

Dátum zadania: 26.01.2018

Dátum schválenia: 26.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou ďakujem svojmu vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za odborné rady, pomoc, vytrvalosť a podnetné pripomienky, ktoré mi pomohli pri písaní diplomovej práce. Ďakujem aj svojej rodine a priateľom za ich trpezlivosť.

Abstrakt v štátnom jazyku

HORVÁTH, Ľudovít: Viacrozmerné neparametrické testy nezávislosti [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2019, 47 s.

V tejto práci predstavujeme štyri nové neparametrické testy nezávislosti. Vychádzame z testov z rokov 2003 a 2005 založených na bežnom priestorovom znamienku, ktoré zovšeobecnili dnes už klasické nápady Blomquista, Spearmana a Kendalla do viacrozmerných priestorov. Pôvodné testy používajú na dosiahnutie affinnej invariantnosti vnútornú štandardizáciu, keďže vonkajšia štandardizácia by zabezpečila invariantnosť len rotačnú. My však naše testy budujeme pomocou Ojovho priestorového znamienka. To je síce náročnejšie na výpočet než bežné priestorové znamienko, no je affinne ekvivariantné, a preto na zabezpečenie affinnej invariantnosti našich testov stačí len jednoduchá vonkajšia štandardizácia namiesto numericky náročnej vnútornej. Tým pádom majú naše nové testy oproti predchodcom aj omnoho jednoduchší tvar. Rovnako ako pôvodné testy majú i naše za platnosti nezávislosti rozdelenie chí-kvadrát, čo sme aj simulačne overili. Simulácie tiež ukázali, že v mnohých prípadoch sa naše testy chovajú lepšie ako pôvodné testy.

Kľúčové slová: Spearmanovo rho, Kendallovo tau, Ojovo priestorové znamienko, viacrozmenré neparametrické testy nezávislosti

Abstract

HORVÁTH, Ľudovít: Multiavriate non-parametric tests of independence [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2019, 47p.

In this work, we present four new non-parametric tests of independence. Our baselines are the test based on a common spatial sign published in 2003 and 2005, which has been generalized using nowadays "classic" ideas by Blomquist, Spearman, and Kendall for multidimensional spaces. The original tests use inner standardization to achieve affine invariance, as outer standardization would ensure just rotational invariance. On the other hand, we build our tests with Oja spatial sign, which is more complicated than normal spatial sign, but it is affinely equivariant and therefore to ensure affine invariance of our tests, it is enough to use simple outer standardization instead of numerically demanding inner one. Hence, our new tests can be expressed in a much simpler form than their predecessors. Like the original ones, our tests also follow chi-square distribution for the independent observation, which we confirmed by simulations, too. Moreover, the simulations showed that in many cases our tests behave better than the original ones.

Key words: Spearman's rho, Kendall's tau, Oja spatial sign, tests of independence

Obsah

Úvod **8**

1 Základné pojmy **10**

1.1 Priestorové znamienko 10

1.2 Priestorové poradie 10

1.3 Afínna invariantnosť 11

1.4 Ojov medián 11

2 Pôvodné testy nezávislosti **12**

2.1 Afínne invariantná transformácia a vnútorná štandardizácia 12

2.2 Spearmanovo rho 14

2.3 Kendallovo tau 14

3 Naše testy nezávislosti **16**

3.1 Nové Spearmanovo rho 21

3.2 Nové Kendallovo tau 24

3.3 Nový Blomquistov test 27

3.4 Nový test pomocou Ojovho znamienkového poradia 29

4 Porovnanie jednotlivých metód **32**

4.1 Vlastnosti testov pre malé sady dát 32

4.1.1 Spearmanovo rho 32

4.1.2 Kendalovo tau 34

4.1.3 Blomquistov test 36

4.1.4 Test pomocou Ojovho znamienkového poradia 37

4.2 Testy na závislých dátach pre $p = q = 2$ a $n = 30$ 39

4.2.1 Parameter $a = 0.2$ 39

4.2.2 Parameter $a = 0.3$ 41

4.2.3 Parameter $a = 0.4$, $a = 0.5$ 42

Záver **45**

Zoznam použitej literatúry **47**

Úvod

V dnešnej dobe, ktorá nám prináša veľké množstvo dát, nás často zaujíma otázka, aké sú závislosti medzi týmito dátami. To je dôvod, prečo sa aj my budeme venovať tejto téme. Keďže sa nejedná len o jednorozmerné údaje, ale o sériu viacerých údajov, je vhodné tieto dáta študovať ako celok, nie po jednotlivých zložkách. Vezmime si napríklad zdravotné údaje, kde každého pacienta nereprezentuje len jedna hodnota, ale celá množina meraní, predstavujúca jeho zdravotný stav. Ak teda budeme skúmať, či má naňho liečba nejaký vplyv, nebudeme skúmať len zmeny jednotlivých zložiek, ale zmenu ako celok. Keby sme sa na zmenu pozerali jednotlivo, nemuseli by sme zachytiť užitočné vzťahy medzi jednotlivými premennými a tiež by sme len ťažko vedeli identifikovať extrémne hodnoty. Táto situácia môže nastať napríklad v takom prípade, keby máme dáta pacienta, ktorý má všetky hodnoty nízke, avšak jednu relatívne vysokú, vzhľadom na ostatné. Ak by sme to však posudzovali jednotlivo, tento pacient by nám ako extrémna hodnota nevyšiel, keďže by jeho výsledky jednotlivo neboli extrémne.

My sa teda v diplomovej práci budeme venovať viacrozmerným neparametrickým testom nezávislosti, kde pri testovaní budeme využívať viaceré testové štatistiky. Bude to hlavne viacrozmerná verzia testu nezávislosti založeného na Spearmanovom rho a viacrozmerná verzia testu nezávislosti založeného na Kendallovom tau [1,str.603]. Vlastnosti týchto testov sú už veľmi dobré popísané a dokázané, preto budeme následne z týchto testov vychádzať pri nami odvodených nových testových štatistikách. Pôvodne je v týchto testoch použité priestorové znamienko [1], ktoré my nahradíme Ojovým znamienkom [5], odvodíme vlastnosti nových testových štatistík, ktoré následne porovnáme s pôvodnými testovými štatistikami. Pôvodné testy majú vlastnosť afínnej invariantnosti, čo budeme vyžadovať aj od nových testov, keďže výsledok testu nesmie závisieť od toho, aký súradnicový systém použijeme. Ak si to napríklad vezmeme z fyzikálneho hľadiska, náš test musí dať rovnaké výsledky bez ohľadu na to, či sme teplotu merali v stupňoch Celzia alebo vo Fahrenheitoch.

V prvej kapitole sa budeme venovať vysvetleniu základných pojmov, ktoré budeme potrebovať v ďalších kapitolách. Je to konkrétne priestorové znamienko, centrované poradie, affínna invariantnosť a Ojov medián. V druhej kapitole si vysvetlíme testy nezávislosti, ktoré sú odvodené od priestorového znamienka, opíšeme ako sa

konštruujú a ukážeme ako vyzerá ich testová štatistika. Vysvetlíme ako vyzerá vnútorná štandardizácia dát, ktorá nám zabezpečí, že tieto testy budú affínne invariantné. Sú to konkrétne už hore spomínané zovšeobecnené Spearmanovo rho a zovšeobecnené Kendallovo tau. V tretej kapitole si predstavíme ako vyzerá Ojovo priestorové znamienko, z ktorého si následne odvodíme nové centrované poradie. Pomocou vonkajšej štandardizácie zabezpečíme affínnu invariantnosť, keďže Ojovo priestorové znamienko je už affínne ekvivariantné. Pomocou Ojovho priestorového znamienka si odvodíme naše nové Spermanovo rho, Kendallovo tau a ďalšie dva nové testy. Následne si overíme predpoklad, že tieto testy majú za platnosti nezávislosti rovnaké asymptotické rozdelenie ako pôvodné testy. Overíme si to simuláciami, kde budeme generovať náhodné dáta, z ktorých vyrátame testovú štatistiku a porovnáme ju s príslušným chí-kvadrátom. V poslednej kapitole budeme jednotlivé testy porovnávať, aby sme zistili ich vlastnosti. Určíme, ktoré je kedy vhodnejšie použiť. Budeme ich skúmať vzhľadom na rôzne rozdelenia dát, rôzne počty dát a rôzne závislosti medzi nimi.

1 Základné pojmy

V celej našej diplomovej práci budeme skúmať závislosť medzi x a y v rámci náhodného vektora $\begin{pmatrix} x \\ y \end{pmatrix}$, kde x je $p-$rozmerný vektor a y je $q-$rozmerný vektor. Naša sada dát, čiže n iid vektorov, teda bude vyzerať nasledovne:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix},$$

kde $x_i = (x_i^{(1)}, ..., x_i^{(p)})^T$ a $y_i = (y_i^{(1)}, ..., y_i^{(q)})^T$.

1.1 Priestorové znamienko

V jednorozmere je priestorové znamienko +1 alebo -1, vzhľadom na to, ktorým smerom od počiatku sa nachádza naša hodnota. Pre viacrozmerné dáta je táto funkcia definovaná nasledovne:

$$S(x) = \begin{cases} \|x\|^{-1}x & (\text{ak } x \neq 0), \\ 0 & (\text{ak } x = 0\ ), \end{cases}$$

kde $\|x\|$ predstavuje L_2-normu (Euklidovskú vzdialenosť bodu x od nuly). Funkčná hodnota je iba smer, čo je bod na povrchu $p-$rozmernej gule s polomerom 1.

1.2 Priestorové poradie

V jednorozmernom svete nám poradie $R(x_i)$ predstavuje poradie člena X_i v dátovej sade x_1,x_2, ... ,$x_n \in \mathbb{R}$, ktoré sú zoradené od najmenšieho po najväčšie. Je zrejmé, že stredná hodnota $R(x_i)$ je $(n+1)/2$, a teda ak definujeme $R_c(X_i) = \frac{2}{n}(R(X_i) - (n+1)/2)$ dostávame centrované priestorové poradie. To, že priestorové poradie je centrované znamená, že platí $\sum_{i=1}^{n} R_c(x_i) = 0$. V zvyšnej časti diplomovej práce budeme používať pre označenie centrovaného priestorového poradia $R_c(x_i)$ len skrátené značenie $R(x_i) = R_i$, vynecháme spodný index c.

Prechod do viacrozmeru vychádza z prepisu jednorozmerného poradia pomocou jednorozmerného znamienka. Je jednoducho vidieť, že $R_i = \frac{1}{n}\sum_j sign(x_i - x_j)$. Ako sa

uvádza v [1], viacrozmerné centrované poradie je teda definované ako

$$R_i = ave_j\{S_{ij}\}$$

kde $ave_i\{t_i\} = \frac{1}{n}\sum_{i=1}^{n} t_i$ a $S_{ij} = S(x_i - x_j)$ pre $i, j = 1, ...n$. S_{ij} je teda priestorové znamienko z rozdielu i-tého a j-tého merania.

1.3 Afínna invariantnosť

Keďže chceme aby naše testy dávali rovnaké výsledky bez ohľadu na to aký súradnicový systém použijeme, musia byť affínne invariantné. Vysvetlenie tohto pojmu predstavuje nasledovná definícia.

Definícia 1.1. *Testová štatistika $T(x_1, x_2, ..., x_n)$ je nazývaná affínne invariantná ak*

$$T(Dx_1 + b, Dx_2 + b, ..., Dx_n + b) = T(x_1, x_2, ..., x_n),$$

pre každú $p \times p$ nesingulárnu maticu D, pre každý $p-$rozmerný vektor b a pre každé $p-$rozmerné dáta $x_1, x_2, ..., x_n \in \mathbb{R}^p$.

1.4 Ojov medián

Definícia 1.2. *Nech $x_1, x_2, ..., x_n \in \mathbb{R}^k$, potom objem $k-$rozmerného simplexu určeného $k+1$ bodmi $x_{i_1}, x_{i_2}, ..., x_{i_k}, x$ je definovaný ako*

$$V(x_{i_1}, x_{i_2}, ..., x_{i_k}, x) = \frac{1}{k!} abs\{det(\begin{smallmatrix} 1 & \cdots & 1 & 1 \\ x_{i_1} & \cdots & x_{i_k} & x \end{smallmatrix})\}$$

Definícia 1.3. *Nech $x_1, x_2, ..., x_n \in \mathbb{R}^k$, potom affínne invariantný mnohorozmerný Ojov medián $\hat{\mu}$ je definovaný ako*

$$\min_{\mu \in \mathbb{R}^k} \sum_{i_1 < ... < i_k} V(x_{i_1}, x_{i_2}, ..., x_{i_k}, \mu)$$

Je to teda taký vektor, ktorý minimalizuje súčet objemov všetkých simplexov s vrcholmi $x_{i_1}, x_{i_2}, ..., x_{i_k}, \mu$.

2 Pôvodné testy nezávislosti

V tejto kapitole sa budeme venovať pôvodným testom nezávislosti, ktoré sú predstavené v [1]. Testy sú založené na znamienkovej funkcii a z nej odvodenej funkcii centrovanej hodnosti. Sú odvodené od ich jednorozmerných predchodcov, kde $p = q = 1$.
Jednorozmerné Kendallovo tau(1938) je

$$ave_{ij}\{sign(x_i - x_j)sign(y_i - y_j))\},$$

vo viacrozmernom prípade ho definujeme ako

$$ave_{ij}\{S_{ij}^{x} S_{ij}^{y\top}\}.$$

Jednorozmerné Spearmanovo rho(1904) je

$$ave_i\{(rank(x_i) - \frac{n+1}{2})(rank(y_i) - \frac{n+1}{2})\},$$

vo viacrozmernom prípade ho definujeme ako

$$ave_i\{R_i^x R_i^{y\top}\}.$$

2.1 Afínne invariantná transformácia a vnútorná štandardizácia

Od testov očakávame, že budú afínne invariantné, preto potrebujeme naše dáta vnútorne štandardizovať. Všeobecná myšlienka ako skonštruovať testovú štatistiku, je použiť $p-$rozmernú hodnotu funkcie $T(y)$, ktorá sa označuje ako funkcia skóre, na získanie z individuálnych skóre $T(y_i)$ $i = 1, 2, ..., n$. V našom prípade je funkcia skóre buď priestorové znamienko $S(x)$ alebo centrované poradie $R(x)$. To, že naše testy sú affínne invariantné zabezpečíme pomocou vnútornej štandardizácie, ktorej postup je vysvetlený v [4]:

- *Vnútorná štandardizácia skóre*: Treba nájsť transformačnú maticu $K^{-1/2}$ takú, že ak sa $\hat{T}_i = T(K^{-1/2}y_i)$, potom platí

$$p \cdot ave_i\{\hat{T}_i\hat{T}_i^{\top}\} = ave_i\{\hat{T}_i^{\top}\hat{T}_i\}I_p,$$

výberová kovariančná matica z $\hat{T}_1, \hat{T}_2, ..., \hat{T}_n$ bude diagonálna.

- *Vnútorné centrovanie a štandardizácia skóre*: Treba nájsť vektor posunutia $\hat{\mu}$ a transformačnú maticu $K^{-1/2}$ tak, aby pre $\hat{T}_i = T(K^{-1/2}(y_i - \hat{\mu}))$ platilo

$$ave_i\{\hat{T}_i\} = 0 \qquad\qquad p \cdot ave_i\{\hat{T}_i\hat{T}_i^\top\} = ave_i\{\hat{T}_i^\top\hat{T}_i\}I_p.$$

V závislosti od použitej funkcie $T(...)$ dostaneme rôzne transformačné matice K, ale nie je zaručené, že takáto matica existuje, hlavne, čo sa týka prípadu $T(x) = R(x)$. My budeme na transformáciu dát používať len metódu štandardizácie skóre, čiže bez centrovanie, keďže naše funkcie $S(x), R(x)$ sú symetricky rozdelené okolo nuly. Pri centrovanom poradí to vyplýva priamo z definície a pri priestorovom znamienku to vyplýva z dôvodu, že ho počítame na diferenciách jednotlivých dát, čo uvidíme v kapitole 2.3.

Algoritmus na dosiahnutie odhadu transformačnej matice K vzhľadom na počiatok používa nasledovné dva kroky:

0.

$$K = I_p$$

1.

$$\hat{T}_i \leftarrow T(K^{-1/2}y_i) \qquad i = 1, 2, ..., n \qquad \hat{T} = (\hat{T}_1, \hat{T}_2, ..., \hat{T}_n)^\top$$

2.

$$K \leftarrow \frac{p}{tr(\hat{T}^\top\hat{T})}K^{1/2}\hat{T}^\top\hat{T}K^{1/2}$$

Konvergencia tohto algoritmu nie je zaručená, ale ako sa uvádza v [1] z empirických skúseností pre skóre funkciu rovnú priestorovému znamienku a centrovanému poradiu konverguje vždy.

2.2 Spearmanovo rho

V tejto už známej metóde sa skúma nezávislosť medzi dvoma zložkami x a y náhodného vektora $\binom{x}{y}$ a to na základe dát $\binom{x_1}{y_1}$,$\binom{x_2}{y_2}$,...,$\binom{x_n}{y_n}$, kde x_i je $p-$rozmerný vektor a y_i je $q-$rozmerný vektor. Testová štatistika je založená na funkcií centrovaného poradia $R(x)$. Ako prvé si transformujeme dáta pomocou vnútornej štandardizácie z kapitoly 2.1, kde za skórovú funkciu dosadíme funkciu centrovaného poradia a výslednou transformačnou maticou $K^{-1/2}$ transformujeme pôvodné dáta. Dáta po transformácií majú tvar $K_x^{-1/2}x_1, K_x^{-1/2}x_2, ..., K_x^{-1/2}x_n$ a $K_y^{-1/2}y_1, K_y^{-1/2}y_2, ..., K_y^{-1/2}y_n$. Tento proces transformácie sa robí samostatne pre x_i a samostatne pre y_i. Z transformovaných dát následne vyrátame testovú štatistiku:

$$\rho^2 = \frac{npq}{(c_x c_y)^2}\|ave_i\{R_i^x R_i^{y^T}\}\|^2,$$

kde

$$c_x^2 = ave_i\{R_i^{x^T} R_i^x\},$$
$$c_y^2 = ave_i\{R_i^{y^T} R_i^y\},$$

$R_i^x = R(K_x^{-1/2}x_i)$ a označenie $\|X\|^2$ znamená $tr(X^TX)$, t.j. Frobeniovu normu, kde $tr()$ predstavuje stopu matice. Ak sú tieto dve sady dát nezávislé, a marginálne rozdelenia x_i a y_i sú elipticky symetrické, potom testová štatistika má chi-kvadrát rozdelenie s pq stupňami voľnosti. Toto tvrdenie aj s dôkazom môžeme nájsť v [6].

2.3 Kendallovo tau

V tejto známej metóde sa podobne ako v Spermanovom rho používa centrované poradie. Okrem toho tu však používame aj priestorové znamienko, ktoré počítame z diferencií transformovaných dát $S_{ij}^x = K_x^{-1/2}S(x_i - x_j)$ pre všetky $i = 1, .., n$ a $j = 1, .., n$. c_x^2 a c_y^2 sú počítané ako v Spearmanovhom rho. Transformačné matice K_x však nie sú také isté pre c_x^2 a S_{ij}^x, keďže sú počítané z iných dát, pri inej skórovej funkcií. Pre c_x^2 sú počítané z dát $(x_i - x_j)$ pre všetky i, j pri skórovej funkcii $R(x)$ a pre S_{ij}^x sú počítané z dát $(x_i - x_j)$ pre všetky i, j pri skórovej funkcii $S(x)$. Rovnako to platí aj pre c_y^2 a S_{ij}^y. Testová štatistika potom vyzerá nasledovne:

$$\tau^2 = \frac{npq}{(2c_x c_y)^2} \| ave\{S_{ij}^x S_{ij}^{y^T}\} \|^2,$$

kde c_x^2 a c_y^2 sú počítané ako v Spermanovom rho. Ak sú tieto dve sady dát nezávislé a marginálne rozdelenie x_i a y_i sú elipticky symetrické, potom testová štatistika má chi-kvadrát rozdelenie s pq stupňami voľnosti. Podobne ako predošlé tvrdenie, aj toto môžeme nájsť v [6].

3 Naše testy nezávislosti

V tejto kapitole sa už budeme venovať nami vytvorenými novými testami nezávislosti, ktoré sú odvodené od ich predchodcov z kapitol 2.2, 2.3 nahradením bežného priestorového znamienka pomocou Ojovho priestorového znamienka a zmenou typu štandardizácie. Nato, aby sme si mohli definovať nové testy nezávislosti, potrebujeme si najskôr definovať Ojove priestorové znamienko, ktoré je odvodené z Ojovho mediánu.

Na definovanie Ojovho priestorového znamienka použijeme definíciu z [5]. Predpokladajme, že máme $p-$rozmerné vektory $x_1, x_2, ..., x_N$. Pre lepšiu prehľadnosť, označme

$$Q = \{q = (i_1, i_2, ..., i_{p-1}) : 1 \leq i_1 < ... < i_{p-1} \leq n\},$$

ako množinu všetkých rôznych podmnožín veľkosti $p-1$ množiny $\{1, 2, ..., n\}$. Jej počet prvkov je teda $N_q = \binom{n}{p-1}$.

Definícia 3.1. *Nech $x_1, x_2, ..., x_n \in \mathbb{R}^p$, sgn je klasické jednorozmerné priestorové znamienko a e_q, $q = (i_1, i_2, ..., i_{p-1})$, je $p-$rozmerný vektor kofaktorov x v $det(x_{i_1}...x_{i_{k-1}}x)$. Potom* **Ojove priestorové znamienko** *je definovaná ako*

$$OS(x) = \frac{1}{N_q} \sum_{q \in Q} sgn(e_q^T x) e_q.$$

Ako prvé nahradíme priestorové znamienko Ojovým priestorovým znamienkom. Táto zmena sa prejaví aj na centrovanom poradí, keďže to sa počíta pomocou priestorového znamienka. V ďalšej časti budeme používať označenie $OR(x)$ pre Ojove centrované poradie a $OS(x)$ pre Ojove priestorové znamienko.

Ojove centrované poradie bude mať teda nasledovný tvar:

$$OR(x_i) = ave_j\{OS(x_i - x_j)\},$$

Ako sa uvádza v [5], Ojove priestorové znamienko aj Ojovo centorvané poradie sú affínne equivariantné. To znamená, že ak

$$x_i^* = Ax_i + b$$

a A je plnej hodnosti, potom

$$\begin{aligned} OS(x_i^*) &= A^* OS(x_i) \\ OR(x_i^*) &= A^* OR(x_i), \end{aligned} \tag{1}$$

kde $A^* = abs(det(A))(A^{-1})^T$.

Nato, aby sme dosiahli affínnu invariantnosť našich testov, musíme na naše dáta použiť štandardizáciu. Pretože sú už naše funkcie affínne ekvivariantné, nepotrebujeme ich štandardizovať vnútornou štandardizáciou, ale stačí vonkajšia štandardizácia. Vonkajšia štandardizácia v prípade pôvodných testov v kapitole 2.2 a 2.3 by zaručila len rotačnú invariantnosť, lebo klasické priestorové znamienko $S(x)$ je len rotačne invariantné, no nie afínne ekvivariantné. V prípade affínne ekvivariantných testov zaručí vonkajšia štandardizácia affínnu invariantnosť. Toto tvrdenie si v ďalšej časti aj dokážeme. Vonkajšia štandardizácia je vysvetlená v nasledujúcej definícií.

Definícia 3.2. *Nech $x_1, x_2, ..., x_n \in \mathbb{R}^p$, potom vonkajšia štandardizácia vektorov je definovaná ako*

$$K^{-1/2}(x_i - \hat{\mu}) \qquad \forall i,$$

kde $\hat{\mu}$ je Ojov medián a $K = \frac{1}{n}\sum_i (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$ je kovariančná matica vypočítaná z dát.

Ďalej uvedieme pomocné tvrdenia, ktoré nám pomôžu v ďalších dôkazoch.

Lema 3.3. *Nech $A_1, ...A_n$ sú $k \times m$ matice, a $tr(A)$ je stopa matice A. Potom stopa spĺňa nasledujúcu vlastnosť:*

$$tr(\sum_i A_i) = \sum_i tr(A_i)$$

Lema 3.4. *Nech A je $k \times m$ matica a B je $m \times k$ matica a $l \in \mathbb{R}$ je skalár. Potom platí:*

$$\begin{aligned} tr(AB) &= tr(BA) \\ tr(lA) &= l \cdot tr(A) \end{aligned}$$

Tieto lemy nebudeme dokazovať, lebo ich pokladáme za všeobecne známe. Dôkaz si možno vyhľadať v [7]. Všimnime si, že prvá časť lemy 3.4 sa dá zovšeobecniť do všeobecného tvaru, kde môžme vziať ľubovoľný počet matíc, poprehadzovať ich tak, aby nám sedeli dimenzie, stopa bude stále rovnaká.

Veta 3.5. *Nech* $r_1, r_2, ..., r_n \in \mathbb{R}^d$ *a* $K = \frac{1}{n}\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top$, *potom*

$$\frac{1}{n}\sum_i (K^{-1/2}(r_i - \hat{\mu})))^\top K^{-1/2}(r_i - \hat{\mu}) = d$$

Dôkaz. Využijeme vlastnosť, že ak k je skalár, potom $tr(k) = k$.

$$\begin{aligned}
\frac{1}{n}\sum_i (K^{-1/2}(r_i - \hat{\mu}))^\top K^{-1/2}(r_i - \hat{\mu}) &= \frac{1}{n}\sum_i (r_i - \hat{\mu})^\top (K^{-1/2})^\top K^{-1/2}(r_i - \hat{\mu}) \\
&= \frac{1}{n}\sum_i (r_i - \hat{\mu})^\top K^{-1}(r_i - \hat{\mu}) \\
&= \frac{1}{n}\sum_i (r_i - \hat{\mu})^\top (\frac{1}{n}\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}(r_i - \hat{\mu}) \\
&= \frac{1}{n}\sum_i (r_i - \hat{\mu})^\top (\frac{1}{n}\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}(r_i - \hat{\mu}) \\
&= \frac{1}{n}\sum_i (r_i - \hat{\mu})^\top n(\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}(r_i - \hat{\mu}) \\
&= tr(\sum_i (r_i - \hat{\mu})^\top (\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}(r_i - \hat{\mu})) \\
&\overset{lema\ 3.3}{=} \sum_i tr((r_i - \hat{\mu})^\top (\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}(r_i - \hat{\mu})) \\
&\overset{lema\ 3.4}{=} \sum_i tr((r_i - \hat{\mu})(r_i - \hat{\mu})^\top (\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}) \\
&\overset{lema\ 3.3}{=} tr(\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top (\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top)^{-1}) \\
&= tr(I_d) \\
&= d
\end{aligned}$$

□

Dôsledok 3.6. *Nech* $r_1, r_2, ..., r_n \in \mathbb{R}^d$,$A$ *je* $d \times d$ *regulárna matica,*
$K = \frac{1}{n}\sum_i (r_i - \hat{\mu})(r_i - \hat{\mu})^\top$, $K_A = \frac{1}{n}\sum_i A(r_i - \hat{\mu})(A(r_i - \hat{\mu}))^\top$. *Potom platí*

$$\frac{1}{n}\sum_i (K^{-1/2}(r_i - \hat{\mu}))^\top K^{-1/2}(r_i - \hat{\mu}) = \frac{1}{n}\sum_i (K_A^{-1/2}A(r_i - \hat{\mu}))^\top K_A^{-1/2}A(r_i - \hat{\mu}) = d$$

Dôkaz. Použijeme označenie $Ar_i = y_i$ a $A\hat{\mu} = \hat{\mu_A}$, kde $\hat{\mu_A}$ je Ojov medián z dát y_i, čo platí vďaka afínnej invariantnosti Ojovho mediánu. Potom $K_A = \frac{1}{n}\sum_i (y_i - \hat{\mu}_A)(y_i - \hat{\mu}_A)^\top$.

$$\frac{1}{n}\sum_i (K_A^{-1/2}A(r_i - \hat{\mu}))^\top K_A^{-1/2}A(r_i - \hat{\mu}) = \frac{1}{n}\sum_i (K_A^{-1/2}(y_i - \hat{\mu}_A))^\top K_A^{-1/2}(y_i - \hat{\mu}_A)$$
$$\overset{\text{Veta 3.5}}{=} d$$

□

Veta 3.7. *Nech* $x_1, x_2, ..., x_n \in \mathbb{R}^p$, $y_1, y_2, ..., y_n \in \mathbb{R}^q$. *Nech* A *je regulárna* $p \times p$ *matica a* B *je regulárna* $q \times q$ *matica. Ďalej nech* $K_x = \frac{1}{n}\sum_i (x_i - \hat{\mu}_x)(x_i - \hat{\mu}_x)^\top$, $K_y = \frac{1}{n}\sum_i (y_i - \hat{\mu}_y)(y_i - \hat{\mu}_y)^\top$, $K_{Ax} = \frac{1}{n}\sum_i (Ax_i - \hat{\mu}_{Ax})(Ax_i - \hat{\mu}_{Ax})^\top$ *a* $K_{By} = \frac{1}{n}\sum_i (By_i - \hat{\mu}_{By})(By_i - \hat{\mu}_{By})^\top$. *Označme si* $x_i' = (x_i - \hat{\mu}_x)$ *a* $y_i' = (y_i - \hat{\mu}_y)$. *Ďalej si označme* $X = \begin{pmatrix} | & | & | \\ x_1' & \dots & x_n' \\ | & | & | \end{pmatrix}$ *a* $Y = \begin{pmatrix} | & | & | \\ y_1' & \dots & y_n' \\ | & | & | \end{pmatrix}$. *Potom platí*

$$tr(((K_{Ax}^{-1/2}AX)(K_{By}^{-1/2}BY)^\top)^\top((K_{Ax}^{-1/2}AX)(K_{By}^{-1/2}BY))^\top)$$

sa rovná

$$tr(((K_x^{-1/2}X)(K_y^{-1/2}Y)^\top)^\top((K_x^{-1/2}X)(K_y^{-1/2}Y))^\top).$$

Dôkaz. Ako prvé si treba uvedomiť, že platí $K_{Ax} = AK_xA^\top$, čo možno jednoducho vidieť z definície K_{Ax}. Z toho vyplýva, že $K_{Ax}^{-1} = (A^\top)^{-1}K_x^{-1}A^{-1}$. Podobne to platí pre K_{By}. Pomocou tohto:

$$\begin{aligned}
&tr(((K_{Ax}^{-1/2}AX)(K_{By}^{-1/2}BY)^\top)^\top((K_{Ax}^{-1/2}AX)(K_{By}^{-1/2}BY)^\top)) \\
&\qquad = tr(K_{By}^{-1/2}BYX^\top A^\top K_{Ax}^{-1/2}K_{Ax}^{-1/2}AXY^\top B^\top K_{By}^{-1/2}) \\
&\qquad \overset{Lema\ 3.4}{=} tr(X^\top A^\top K_{Ax}^{-1/2}K_{Ax}^{-1/2}AXY^\top B^\top K_{By}^{-1/2}K_{By}^{-1/2}BY) \\
&\qquad = tr(X^\top A^\top K_{Ax}^{-1}AXY^\top B^\top K_{By}^{-1}BY) \\
&\qquad = tr(X^\top A^T(A^\top)^{-1}K_x^{-1}A^{-1}AXY^\top B^\top(B^\top)^{-1}K_y^{-1}B^{-1}BY) \\
&\qquad = tr(X^\top K_x^{-1}XY^\top K_y^{-1}Y) \\
&\qquad = tr(X^\top K_x^{-1/2}K_x^{-1/2}XY^\top K_y^{-1/2}K_y^{-1/2}Y) \\
&\qquad = tr((K_x^{-1/2}X)^\top K_x^{-1/2}X(K_y^{-1/2}Y)^\top K_y^{-1/2}Y) \\
&\qquad \overset{Lema\ 3.4}{=} tr(((K_x^{-1/2}X)(K_y^{-1/2}Y)^\top)^\top((K_x^{-1/2}X)(K_y^{-1/2}Y)^\top)).
\end{aligned}$$

□

V ďalšom dôkaze využijeme vlastnosť

$$\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)^\top = XY^\top, \tag{2}$$

kde $X = \begin{pmatrix} | & | & | \\ x_1' & \dots & x_n' \\ | & | & | \end{pmatrix}$, $x_1, x_2, ..., x_n \in \mathbb{R}^p$ a $Y = \begin{pmatrix} | & | & | \\ y_1' & \dots & y_n' \\ | & | & | \end{pmatrix}$, $y_1, y_2, ..., y_n \in \mathbb{R}^q$.

Veta 3.8. *Nech $x_1, x_2, ..., x_n \in \mathbb{R}^p$, $y_1, y_2, ..., y_n \in \mathbb{R}^q$. Nech A je regulárna $p \times p$ matica a B je regulárna $q \times q$ matica. Ďalej, nech $K_{fx} = \frac{1}{n}\sum_i (f(x_i) - \hat{\mu}_{fx})(f(x_i) - \hat{\mu}_{fx})^\top$, $K_{fy} = \frac{1}{n}\sum_i (f(y_i) - \hat{\mu}_{fy})(f(y_i) - \hat{\mu}_{fy})^\top$, $K_{fAx} = \frac{1}{n}\sum_i (f(Ax_i) - \hat{\mu}_{fAx})(f(Ax_i) - \hat{\mu}_{fAx})^\top$, $K_{fBy} = \frac{1}{n}\sum_i (f(By_i) - \hat{\mu}_{fBy})(f(By_i) - \hat{\mu}_{fBy})^\top$ a $f(x) : R^d \to R^d$ je afínne eqvivariantná funkcia v zmysle 1 a $\hat{\mu_{fz}}$ označuje Ojov medián z dát $f(z_1), f(z_2), ..., f(z_n)$. Potom platí*

$$\begin{aligned}
&\|ave_i\{K_{fAx}^{-1/2}(f(Ax_i) - \hat{\mu}_{fAx})(K_{fBy}^{-1/2}(f(By_i) - \hat{\mu}_{fBy}))^\top\}\|^2 \\
&\qquad = \|ave_i\{K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx})(K_{fy}^{-1/2}(f(y_i) - \hat{\mu}_{fy}))^\top\}\|^2
\end{aligned}$$

Dôkaz. Teraz platí trochu iná vec ako v predošlej vete. $K_{fAx} = A^* K_{fx}(A^*)^T$, kde A^* je matica definovaná v afínnej eqvivariantnosti Ojovej znamienkovej funkcie a centrovanej

hodnosti vo vzorci (1). Z toho vyplýva, že $K_{fAx}^{-1} = ((A^*)^T)^{-1}K_{fx}^{-1}(A^*)^{-1}$. Podobne to platí pre K_{fBy}. Budeme to dokazovať pre normu bez priemeru, keďže počet dát je pre Ax_i a x_i konštantný. Podobne pre y_i a By_i. Prenásobením matice sa počet dát nemení. Označme si $X_f = \begin{pmatrix} | & | & | \\ (f(x_1)-\hat{\mu}_{fx}) & \dots & (f(x_n)-\hat{\mu}_{fx}) \\ | & | & | \end{pmatrix}$ a $Y_f = \begin{pmatrix} | & | & | \\ (f(y_1)-\hat{\mu}_{fy}) & \dots & (f(y_n)-\hat{\mu}_{fy}) \\ | & | & | \end{pmatrix}$

$$
\begin{aligned}
&\| \sum_i K_{fAx}^{-1/2}(f(Ax_i) - \hat{\mu}_{fAx})(K_{fBy}^{-1/2}(f(By_i) - \hat{\mu}_{fBy}))^\top \|^2 \\
&\qquad = \| \sum_i K_{fAx}^{-1/2}A^*(f(x_i) - \hat{\mu}_{fx})(K_{fBy}^{-1/2}B^*(f(y_i) - \hat{\mu}_{fy}))^\top \|^2 \\
&\qquad \overset{(2)}{=} \|K_{fAx}^{-1/2}A^*X_f(K_{fBy}^{-1/2}B^*Y_f)^T\|^2 \\
&\qquad = tr((K_{fAx}^{-1/2}A^*X_f(K_{fBy}^{-1/2}B^*Y_f)^T)^T(K_{fAx}^{-1/2}A^*X_f(K_{fBy}^{-1/2}B^*Y_f)^T)) \\
&\qquad \overset{Veta3.7}{=} tr((K_{fx}^{-1/2}X_f(K_{fy}^{-1/2}Y_f)^T)^T(K_{fx}^{-1/2}X_f(K_{fy}^{-1/2}Y_f)^T)) \\
&\qquad = \|K_{fx}^{-1/2}X_f(K_{fy}^{-1/2}Y_f)^T\|^2 \\
&\qquad \overset{(2)}{=} \| \sum_i K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx})(K_{fy}^{-1/2}(f(y_i) - \hat{\mu}_{fy}))^\top \|^2
\end{aligned}
$$

□

Ako neskôr uvidíme veta 3.8 a dôsledok 3.6 nám budú garantovať, že naše nové testy budú po transformácii affínne invariantné. Navyše z dôsledku 3.6 platí, že ak počítame

$$ave_i\{(K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx}))^\top(K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx}))\},$$

kde $f(x)$ je ľubovoľná affínne eqvivariantná funkcia, v zmysle 1, potom platí

$$
\begin{aligned}
&ave_i\{(K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx}))^\top(K_{fx}^{-1/2}(f(x_i) - \hat{\mu}_{fx}))\} = \\
&ave_i\{(K_{Afx}^{-1/2}(f(Ax_i) - \hat{\mu}_{Afx}))^\top(K_{Afx}^{-1/2}(f(Ax_i) - \hat{\mu}_{Afx}))\} = d,
\end{aligned}
\tag{3}
$$

čo nám značne zjednoduší testy.

3.1 Nové Spearmanovo rho

Ojovo štandardizované centrované poradie má tvar $OR_i^x = K_{ORx}^{-1/2}(OR(x_i))$, kde $K_{ORx} = \frac{1}{n}\sum_i OR_i^x OR_i^{x^\top}$ predstavuje výberovú kovariančnú maticu.

Nové Spearmanovo rho ρ^2_{nove} má tvar

$$\rho^2_{nove} = \frac{npq}{(oc_x oc_y)^2} \|ave\{OR^x_i OR^{y^T}_i\}\|^2,$$

kde

$$oc^2_x = ave\{OR^{x^T}_i OR^x_i\},$$

a

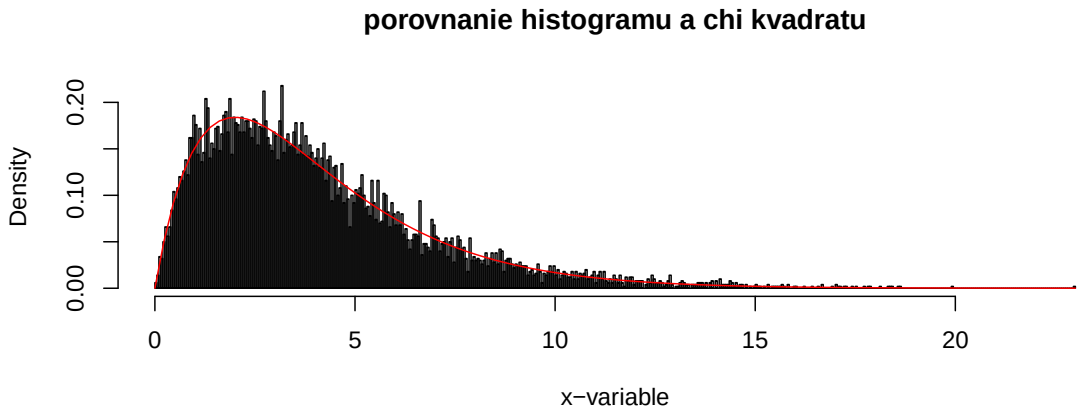
$$oc^2_y = ave\{OR^{y^T}_i OR^y_i\}.$$

Ak v rovnosti (3) nahradíme funkciu f Ojovým centrovaným poradím, tak dostaneme, že $oc^2_x = p$ a $oc^2_y = q$, a náš nový test bude mať tvar

$$\rho^2_{nove} = n\|ave\{OR^x_i OR^{y^T}_i\}\|^2.$$

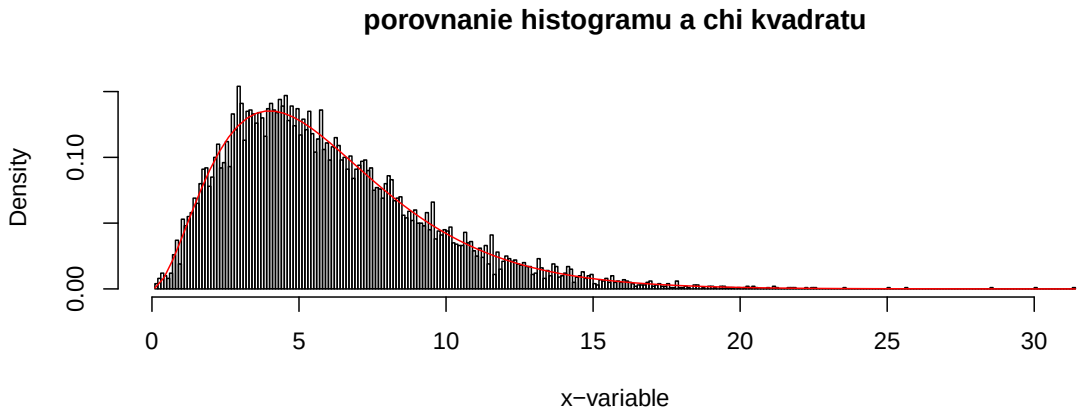
Podľa vety 3.8 (kde $f = OR$ a $\hat{\mu}$ je nulový vektor, keďže netreba centrovať) je ρ^2_{nove} affínne invariantné. Rozdelenie tejto testovej štatistiky by malo byť rovnaké, ako rozdelenie pôvodnej testovej štatistiky, pretože sme len bežné priestorové znamienko $S(x)$ nahradili Ojovým priestorovým znamienkom $OS(x)$, no definícia centrovaného poradia sa formálne nezmenila, rovnako ako definícia ρ^2. Ďalšou zmenou bolo nahradenie vnútornej štandardizácie vonkajšou, čo by výsledné asymptotické rozdelenie tiež nemalo ovplyvniť. Tento predpoklad sme si následne overili na 100000 nezávislých umelo vygenerovaných sadách $n = 100$ dát.

Ako prvé sme si zvolili $p = q = 2$. Tieto dáta boli aj vnútorne nezávislé. Nasimulované ρ^2_{nove} sme následne porovnávali s chí-kvadrát rozdelením so 4 stupňami voľnosti ($p \cdot q = 4$). Výsledok tohto môžme vidieť na obrázku 1, kde sme si nakreslili histogram z 100000 vypočítaných ρ^2_{nove} , spolu s chí-kvadrát rozdelením. Chyba 1. druhu nám v tomto teste vyšla 4.9% oproti pôvodnému testu, ktorý mal nasimulovanú chybu 1. druhu rovnú 5.2%.



Obr. 1: Porovnanie histogramu nasimulovaných ρ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = q = 2$, $n = 100$

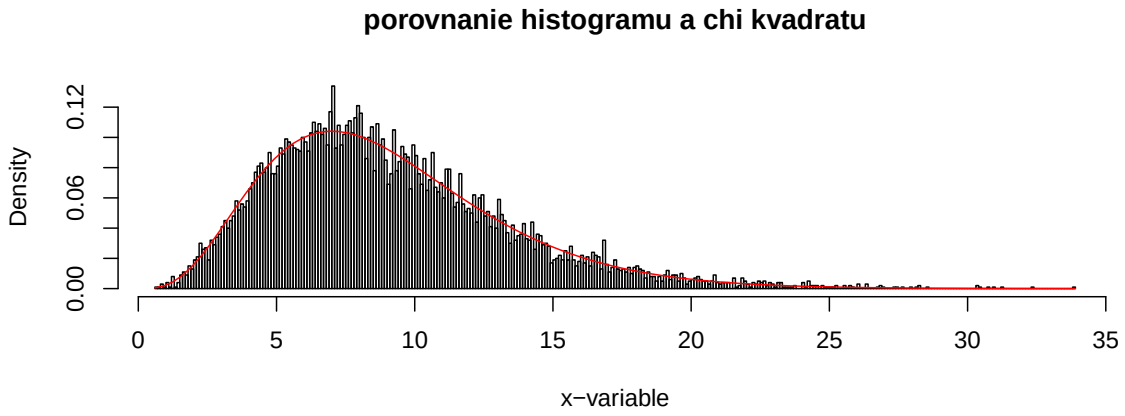
Ako druhé sme porovnávali dve sady dát, kde prvá bola 2−rozmerná s vnútornými závislosťami, s kovariančnou maticou $K = \left(\begin{smallmatrix} 2 & 3 \\ 3 & 7 \end{smallmatrix}\right)$. Druhá sada dát bola 3−rozmerná bez vnútorných závislostí. Nasimulované ρ^2_{nove} sme následne porovnali s chí-kvadrát rozdelením s 6 stupňami voľnosti. Chyba 1. druhu nám v tomto teste vyšla 4.8% oproti pôvodnému testu, ktorý mal nasimulovanú chybu 1. druhu rovnú 5.1%.



Obr. 2: Porovnanie histogramu nasimulovaných ρ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 2$ a $q = 3$, $n = 100$

Ako tretiu testovaciu sadu sme zvolili obe dáta 3−rozmerné s vnútornými závislosťami, kde kovariančná matica bola $K_1 = \left(\begin{smallmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{smallmatrix}\right)$ pre rozdelenie iksov, respektíve $K_2 =$

$\begin{pmatrix} 3 & 0 & 1 \\ 0 & 5 & 2 \\ 1 & 2 & 9 \end{pmatrix}$ pre rozdelenie ypsilonov. Výsledok možno vidieť na nasledujúcom obrázku. Chyba 1. druhu nám v tomto teste vyšla 4.7% oproti pôvodnému testu, ktorý mal nasimulovanú chybu 1. druhu rovnú 5.1%.



Obr. 3: Porovnanie histogramu nasimulovaných ρ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 3$, $n = 100$

Ako si môžme všimnúť, na všetkých troch obrázkoch sa naše histogramy zhodujú s príslušným chí kvadrát rozdelením, a teda je vysoko pravdepodobné, že naše nové testové štatistiky majú za platnosti nezávislosti rovnaké asymptotické rozdelenie ako ich predchodcovia. Overili sme to aj na ďalších experimentoch, no tie tu pre prehľadnosť nebudeme pripájať.

3.2 Nové Kendallovo tau

Podobne ako v predošlej časti, aj tu budeme používať Ojovú centrovanú hodnosť, ku ktorej sa pridá Ojová znamienková funkcia. Po štandardizácii bude mať táto funkcia nasledovný tvar $OS^x_{ij} = K^{-1/2}_{OSx} OS(x_i - x_j)$, kde K_{OSx} je kovariančná matica vypočítaná z dát $OS(x_i - x_j)$ pre všetky i, j kde $i \neq j$. Centrovať netreba. K hodnote $OS(x_i - x_j)$ je rovnako pravdepodobná hodnota $OS(-(x_i - x_j)) = -OS(x_i - x_j)$, čiže OS_{ij} je symetricky rozdelený okolo nuly, a preto ho netreba centrovať.

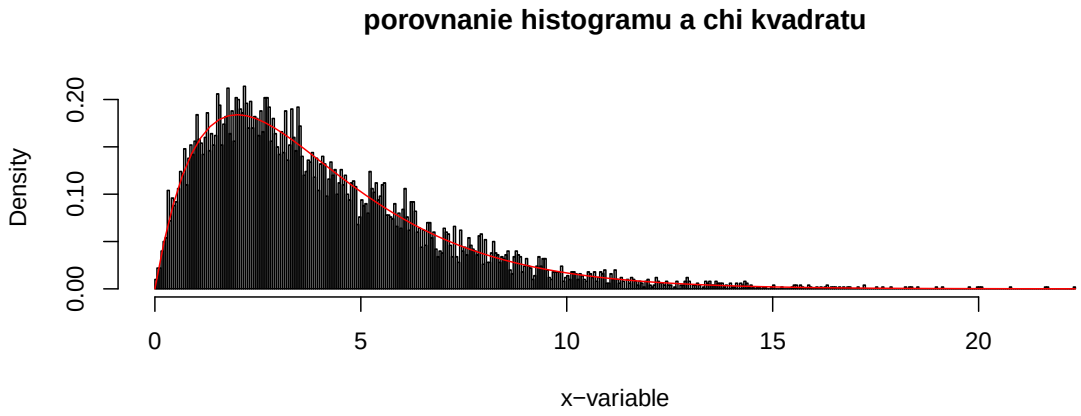
Testová štatistika má nasledovný tvar

$$\tau^2_{nove} = \frac{npq}{(oc_x oc_y 2)^2} \| ave\{OS^x_{ij} OS^{y^\top}_{ij}\} \|^2,$$

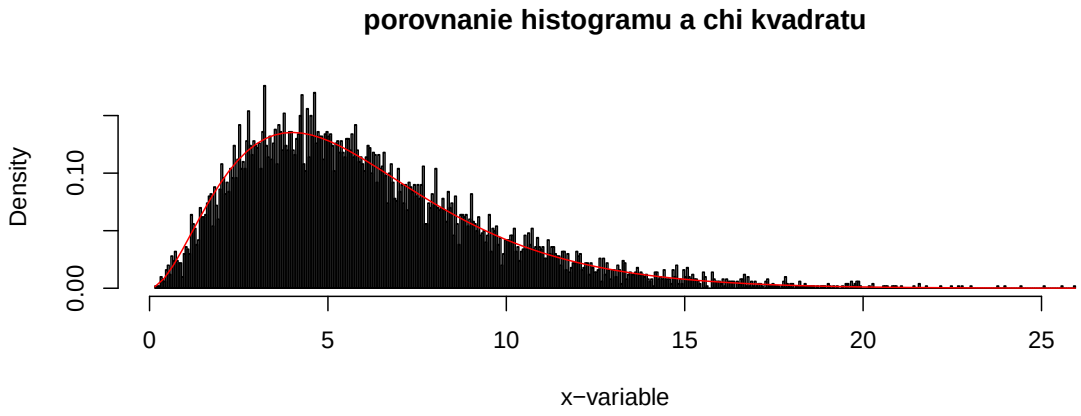
kde oc_x^2 a oc_y^2 sú počítané ako v novom Spermanovhom rho. Priemer je tu počítaný cez všetky hodnoty i, j. Z rovnosti (3) nám však opäť vyplýva, že $oc_x^2 = p$ a $oc_y^2 = q$, a teda náš nový test bude mať tvar

$$\tau^2_{nove} = \frac{n}{4} \| ave\{OS^x_{ij} OS^{y^\top}_{ij}\} \|^2.$$

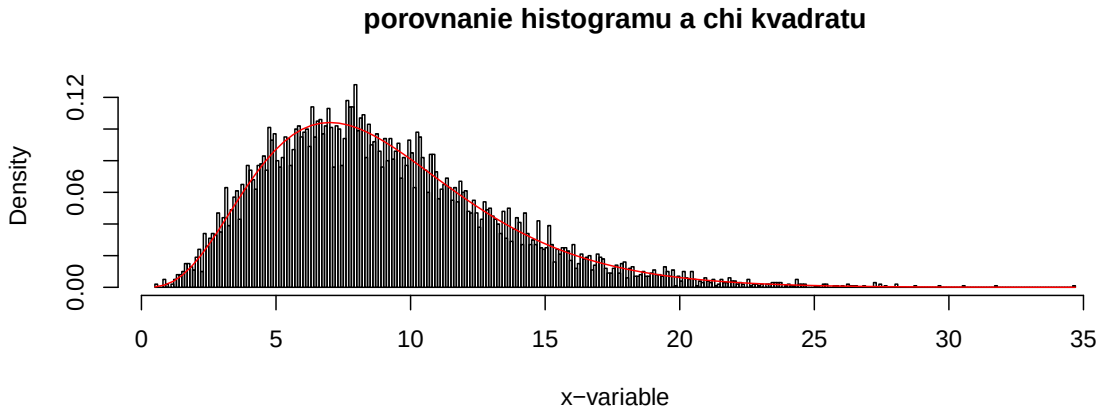
Podľa vety 3.8 (kde $f = OS$ a $\hat{\mu}$ je nulový vektor, keďže netreba centrovať) je τ^2_{nove} affínne invariantné. Podobne ako v predošlej časti sme z nami vygenerovaných sád dát počítali tentokrát nové Kendalovo tau τ^2_{nove}, čoho výsledky môžme vidieť na obrázkoch 4, 5, 6. Zvolili sme rovnaký počet simulácií a rovnaké kovariančné matice ako v predošlej časti. Rozmer dát bol $n = 50$. Chyba 1. druhu nám v tomto teste vyšla 5%, 4.9% a 5% oproti pôvodnému testu, ktorý mal nasimulovanú chybu 1. druhu rovnú 5.3%, 5.2% a 5.4%.



Obr. 4: Porovnanie histogramu nasimulovaných τ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = q = 2$, $n = 50$



Obr. 5: Porovnanie histogramu nasimulovaných τ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 2$, $n = 50$



Obr. 6: Porovnanie histogramu nasimulovaných τ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 3$, $n = 50$

Vidíme, že sa nám opäť zhoduje histogram s príslušným chi-kvadrát rozdelením, a je veľmi pravdepodobné, že náš predpoklad o asymptotickom rozdelení χ^2_{pq} je správny.

Aj túto testovú štatistiku sme podrobili ďalším experimentom, ktoré nám opäť dali pozitívny výsledok, ale pre prehľadnosť ich neuvádzame.

3.3 Nový Blomquistov test

Ako sa uvádza v [4], všeobecný tvar testu s príslušným asymptotickým chí-kvadrát rozdelením za platnosti nezávislosti má tvar

$$Q^2 = \frac{npq}{ave\{\hat{T}(x_i)^\top \hat{T}(x_i)\} \cdot ave\{\hat{T}(y_i)^\top \hat{T}(y_i)\}} \|ave\{\hat{T}(x_i)\hat{T}(y_i)^\top\}\|^2, \tag{4}$$

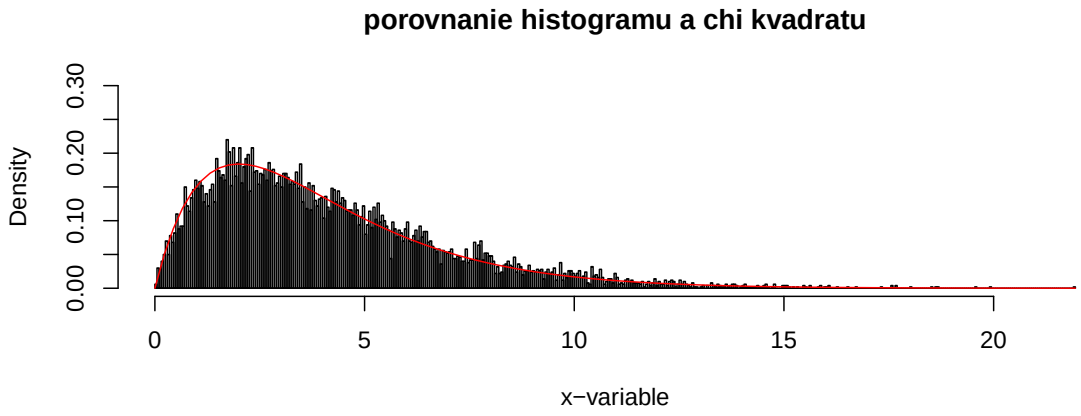
kde $\hat{T}(x)$ je ľubovoľná štandardizovaná funkcia skóre $T(x)$, pričom v [4] sa používala výhradne vnútorná štandardizácia, keďže $T(x)$ nebola affínne ekvivariantná. Ak bude naša funkcia skóre affínne ekvivariantná, môžeme použiť vonkajšiu štandardizáciu, vďaka čomu sa testová štatistika zjednoduší na tvar

$$Q^2 = n\|ave\{\hat{T}(x_i)\hat{T}(y_i)^\top\}\|^2,$$

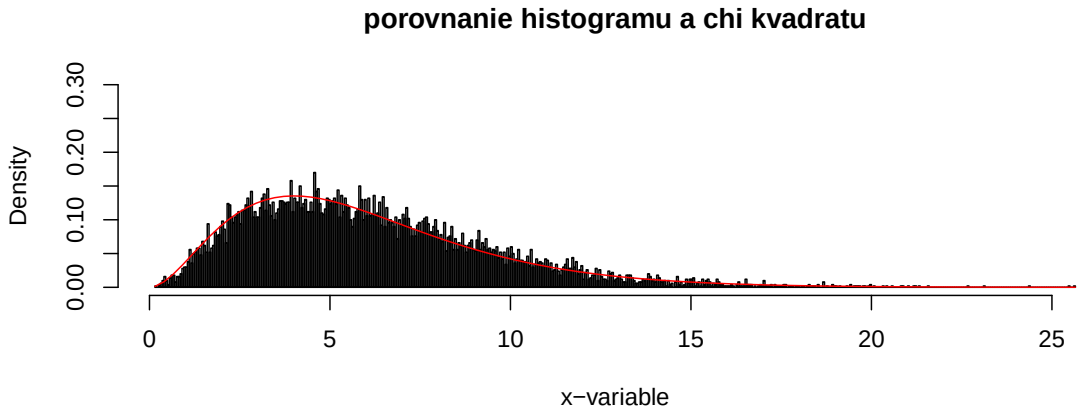
a budeme mať vlastnosť affínnej invariantnosti podľa vety 3.8 (kde $f = T$). V pôvodnom zovšeobecnení Blomquistoveho testu v článku [2] použil ako funkciu skóre vnútorne štandardizované priestorové znamienko $S(x)$, ktoré sme my nahradili vonkajšie štandardizovaným Ojovým priestorovým znamienkom $OS(x)$. Naša funkcia skóre má tvar $\hat{T}(x_i) = \hat{OS}(x_i) = K_{OSx}^{-1/2}((OS(x_i) - \hat{\mu}_{osx}))$, kde K_{OSx} je kovariančná matica vypočítaná z dát $OS(x_i)$ a $\hat{\mu}_{osx}$ je Ojov medián z dát $OS(x_i)$. Naša testová štatistika má tvar

$$B^2 = n\|ave\{\hat{OS}(x_i)\hat{OS}(x_i)^\top\}\|^2.$$

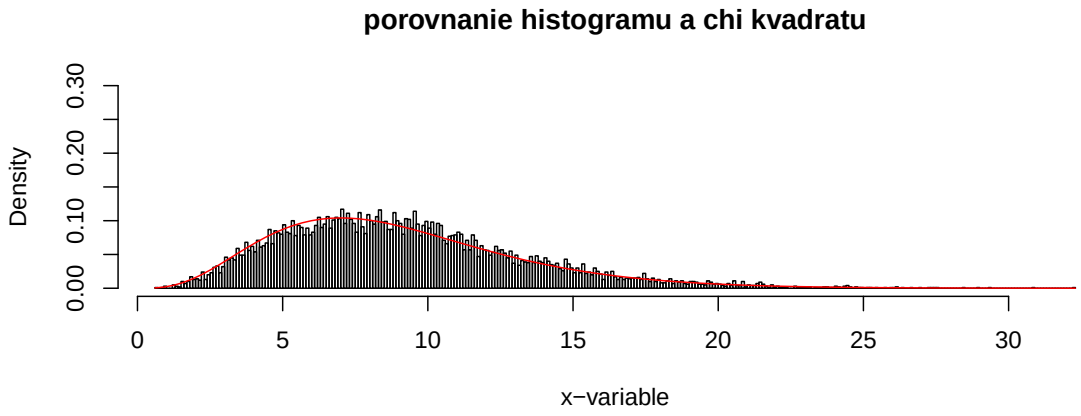
a vďaka vete 3.8 je affínne invariantná. Rovnako, ako v predošlej časti sme z nami vygenerovaných sád dát počítali tento krát novú Blomquistovu testovú štatistiku B^2, čoho výsledky môžme vidieť na obrázkoch 7, 8, 9. Zvolili sme rovnaký počet simulácií a rovnaké kovariančné matice ako v predošlej časti. Rozmer dát bol $n = 50$. Chyba 1. druhu nám v tomto teste vyšla 5%, 5.1% a 5% oproti pôvodnému testu, ktorý mal nasimulovanú chybu 1. druhu rovnú 5.3%, 5.4% a 5.4%.



Obr. 7: Porovnanie histogramu nasimulovaných B^2 s chí-kvadrát rozdelením χ^2_{pq} pre $p = q = 2$, $n = 50$



Obr. 8: Porovnanie histogramu nasimulovaných B^2 s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 2$, $n = 50$



Obr. 9: Porovnanie histogramu nasimulovaných B^2 s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 3$, $n = 50$

Opäť sa zhoduje histogram s príslušným chi-kvadrát rozdelením a je teda veľmi pravdepodobné, že náš predpoklad o asymptotickom rozdelení χ^2_{pq} je správny.

Aj túto testovú štatistiku sme podrobili ďalším experimentom, ktoré nám opäť dali pozitívny výsledok, ale pre prehľadnosť ich neuvádzame.

3.4 Nový test pomocou Ojovho znamienkového poradia

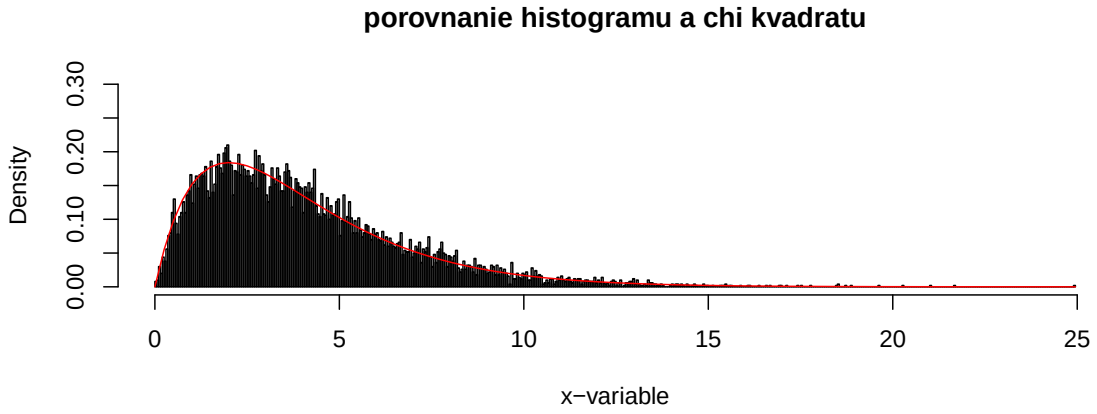
Ako posledné sme opäť vychádzali z všeobecného tvaru (4) testu definovaného v [4]. Tentokrát sme ako funkciu skóre použili Ojovo znamienkové poradie, ktoré je v [5] definované ako klasické Ojovo poradie, ktoré však nie je počítané z n dát $x_1, x_2, ..., x_n$, ale z $2n$ dát $x_1, x_2, ..., x_n, -x_1, -x_2, ..., -x_n$. Toto Ojovo znamienkové poradie budeme označovať $OSR(x)$.

Naša funkcia skóre má tvar $\hat{T}(x_i) = \hat{OSR}(x_i) = K^{-1/2}_{OSRx}((OSR(x_i) - \hat{\mu}_{OSRx}))$, kde K_{OSRx} je kovariančná matica vypočítaná z dát $OSR(x_i)$ a $\hat{\mu}_{OSRx}$ je Ojov medián z dát $OS(x_i)$. Naša testová štatistika má vďaka platnosti 3 tvar

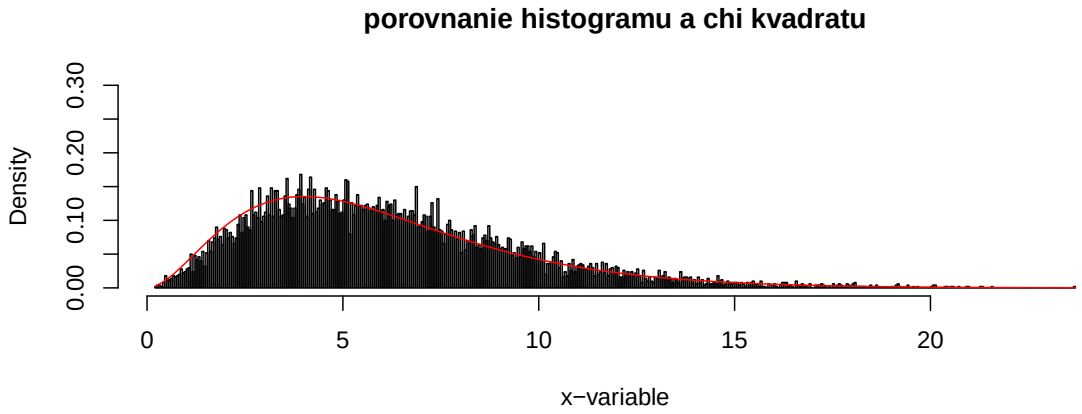
$$Q^2_{osr} = n\|ave\{\hat{OSR}(x_i)\hat{OSR}(x_i)^\top\}\|^2.$$

a vďaka vete 3.8 je afínne invariantná. Aj v tejto časti sme z nami vygenerovaných sád dát počítali tentokrát novú testovú štatistiku pomocou Ojovho znamienkového poradia

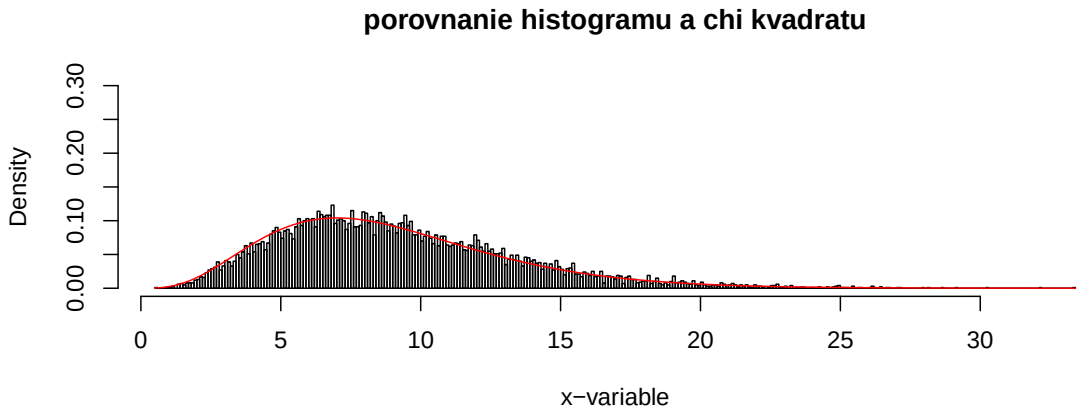
Q^2_{osr}, čoho výsledky môžme vidieť na obrázkoch 10, 11, 12. Zvolili sme rovnaký počet simulácií a rovnaké kovariančné matice ako v predošlej časti. Rozmer dát bol $n = 50$. Chyba 1. druhu nám v tomto teste vyšla 5.2%, 5.2% a 5.3%.



Obr. 10: Porovnanie histogramu nasimulovaných Q^2_{osr} s chí-kvadrát rozdelením χ^2_{pq} pre $p = q = 2$, $n = 50$



Obr. 11: Porovnanie histogramu nasimulovaných Q^2_{osr} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 2$, $n = 50$



Obr. 12: Porovnanie histogramu nasimulovaných Q^2_{osr} s chí-kvadrát rozdelením χ^2_{pq} pre $p = 3$ a $q = 3$, $n = 50$

Opäť sa zhoduje histogram s príslušným chi-kvadrát rozdelením, a je veľmi pravdepodobné, že náš predpoklad o asymptotickom rozdelení χ^2_{pq} je správny.

Aj túto testovú štatistiku sme podrobili ďalším experimentom, ktoré nám opäť dali pozitívny výsledok, ale pre prehľadnosť ich neuvádzame.

4 Porovnanie jednotlivých metód

V tejto časti sa budeme venovať porovnaniu jednotlivých testov vzhľadom na rôzne sady dát. V celej tejto kapitole budeme dáta generovať len z normálneho rozdelenia, avšak všetky tieto testy sme spravili aj pre mnohorozmerné Studentove t-rozdelenia, pre rôzne stupne voľnosti, ale keďže výsledky sú takmer totožné, pre prehľadnosť ich neuvádzame.

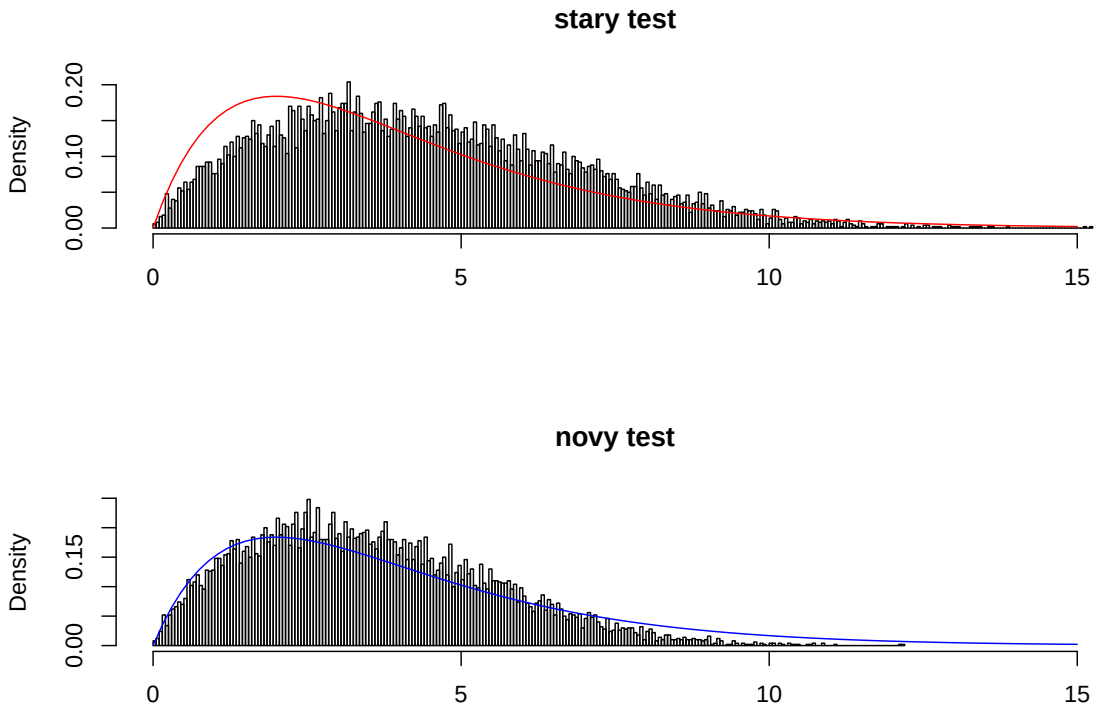
4.1 Vlastnosti testov pre malé sady dát

Nie vždy máme veľké sady dát na testovanie, rozhodli sme sa najprv zistiť, ako sa testy správajú pre malé sady dát. Chceme zistiť, či sa skutočné rozdelenie testových štatistík v prípade nezávislosti podobá na príslušný chí-kvadrát, keďže dokázané je to len asymptomaticky, t.j. len pre $n \to \infty$. Na tieto testy sme si zvolili $n = 10$ pre rôzne p a q. Na porovnanie budeme používať:

1. porovnanie histogramu s hustotou príslušného chí-kvadrát rozdelenia,
2. strednú hodnotu vypočítanú z vygenerovaných testových štatistík v porovnaní s očakávanou strednou hodnotou,
3. koľko vygenerovaných testových štatistík je za 5% kritickou hodnotou príslušného chí kvadrátu.

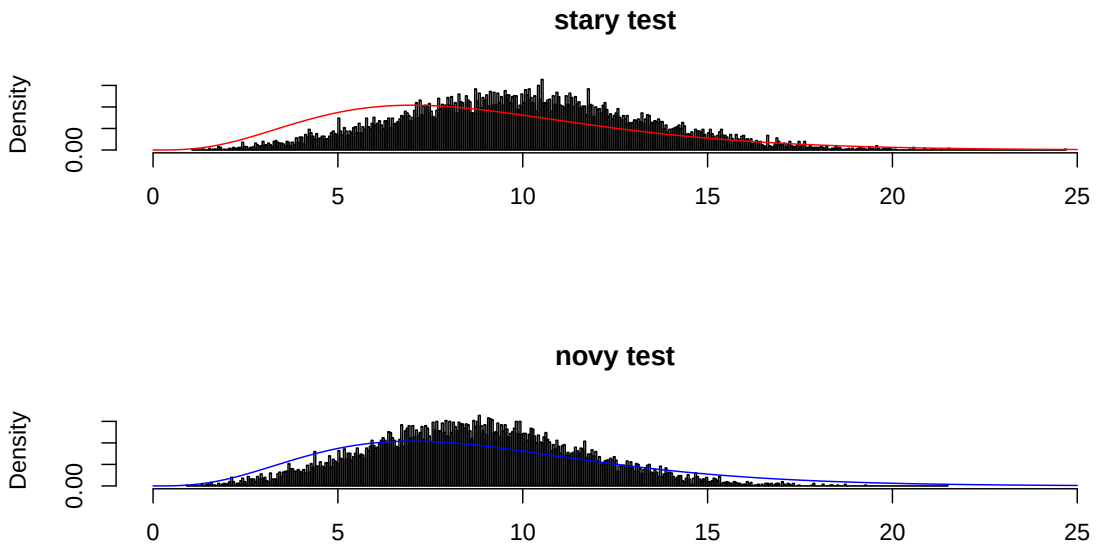
4.1.1 Spearmanovo rho

Ako prvé sme sa zamerali na test založený na Spearmanovom rho. Najprv sme si zvolili $p = q = 2$, kde boli naše dáta bez vnútorných závislostí a boli navzájom nezávislé. Tieto dáta sme si vygenerovali 10000 krát a porovnali s príslušným chí-kvadrát rozdelením. Výsledok môžme vidieť na obrázku 13.



Obr. 13: Porovnanie histogramu nasimulovaných ρ^2 a ρ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq}

Ako môžme vidieť, oba testy sa stále pomerne presne riadia príslušným chí-kvadrát rozdelením, aj keď je n malé. Overili sme si to aj na výpočte strednej hodnoty, ktorá nám v prípade pôvodného testu vyšla 4.63 a v prípade nového testu 4.35, a keďže skutočná stredná hodnota chí-kvadrátu s 4 stupňami voľnosti je 4, tak náš nový test vyšiel v tomto porovnaní lepšie. Keď sa zameriame na 5% kritickú hodnotu, v pôvodnom teste nám za 5% kritickou hodnotou rozdelenia χ^2_4, ktorá je rovná 9.488, skončilo 6.2% dát, pričom v novom teste len 5.7% dát. Teda aj v tomto porovnaní vyšiel nový test o čosi lepšie. Následne sme si zvolili $p = 3$ a $q = 4$. Tu sme si už generovali $3-$rozmerné dáta x_i, ktoré mali vnútorné závislosti dané kovariančnou maticou $K_1 = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}$ a $4-$rozmerné dáta y_i bez vnútorných závislostí, teda ich kovariančná matica bola rovná $K_2 = I_4$. Výsledok môžme vidieť na nasledujúcom obrázku.



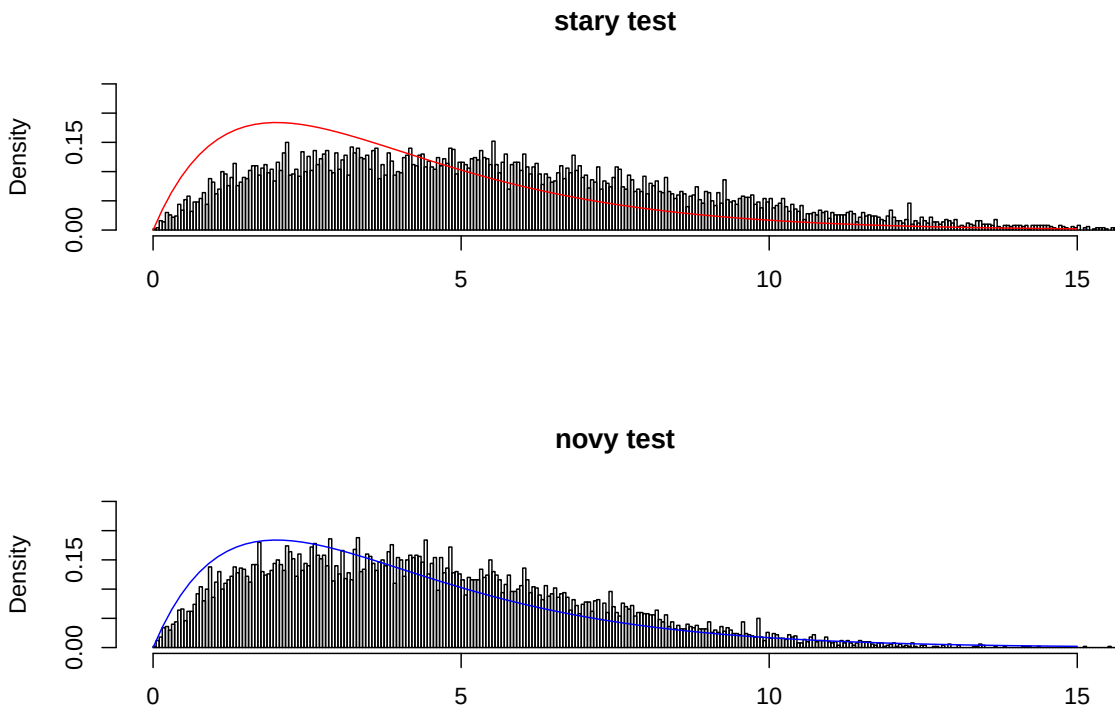
Obr. 14: Porovnanie histogramu nasimulovaných ρ^2 a ρ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq}

Aj v tomto prípade sa oba testy riadia približne chí-kvadrát rozdelením, ale keď sa pozrieme na príslušne stredné hodnoty, stredná hodnota pôvodného testu nám vyšla 13.72 a nového testu vyšla 13.64. Čo sa týka počtu dát, ktoré prekročili 5% kritickú hodnotu, ktorá je rovná 21.026, tak v prípade pôvodného testu to bolo 7.5% a v prípade nového testu 7.1%.

4.1.2 Kendalovo tau

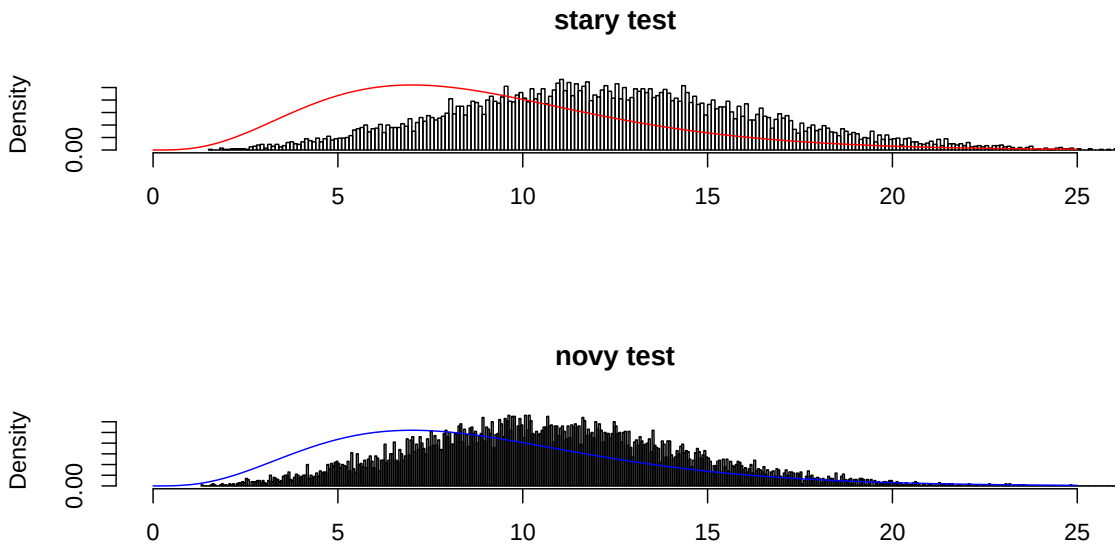
Rovnaký postup ako pre Spearmanovo rho sme zvolili aj pre Kendalovo tau. Náš počet dát bol $n = 10$, kde sme dáta generovali 10000−krát. Na porovnávanie jednotlivých testov sme zvolili rovnakú stratégiu ako v predošlej podkapitole, na ohodnotenie testov použijeme porovnanie histogramu s príslušným chí-kvadrátom, strednú hodnotu a percento testových štatistík za 5% kritickou hodnotou príslušného chí-kvadrátu.

Ako prvé sme zvolili $p = q = 2$, kde obe sady dát boli bez vnútorných závislostí. Výsledok tohto porovnania možno vidieť na obrázku 15.



Obr. 15: Porovnanie histogramu nasimulovaných τ^2 a τ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq}

Je vidieť, že pôvodný test sa vychýlil viac ako nový test, čo si môžme všimnúť aj na nasledujúcich charakteristikách. Konkrétne to môžme vidieť na porovnaní stredných hodnôt, kde očakávaná stredná hodnota chí-kvadrátu so 4 stupňami voľnosti je 4, pričom stredná hodnota dát pôvodného testu je 5.52 a stredná hodnota dát nového testu je 3.88. Keď sa pozrieme na 5% kritickú hodnotu, ktorá je rovná 9.488, tak v prípade pôvodného testu ju prekročilo 12.4%. V prípade nového testu 4.7%. Ďalej sme si vzali $p = 3$ a $q = 4$. Tu sme si už generovali $3-$rozmerné dáta x_i, ktoré mali vnútorné závislosti dané kovariančnou maticou $K_1 = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}$ a $4-$rozmerné dáta y_i bez vnútorných závislostí, ich kovariančná matica bola rovná $K_2 = I_4$. Výsledok môžme vidieť na nasledujúcom obrázku.

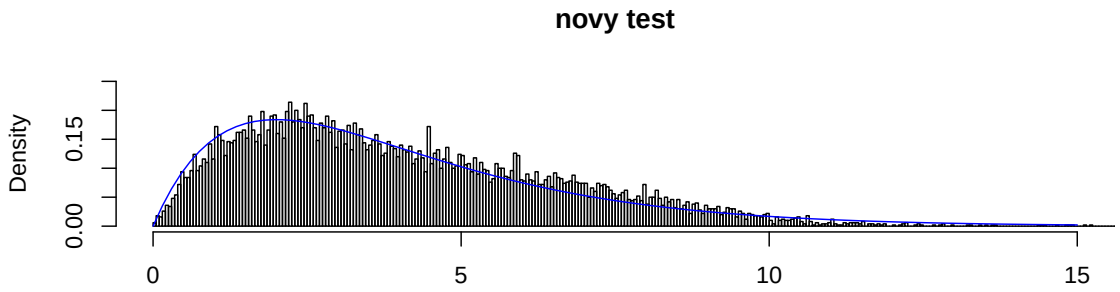


Obr. 16: Porovnanie histogramu nasimulovaných τ^2 a τ^2_{nove} s chí-kvadrát rozdelením χ^2_{pq}

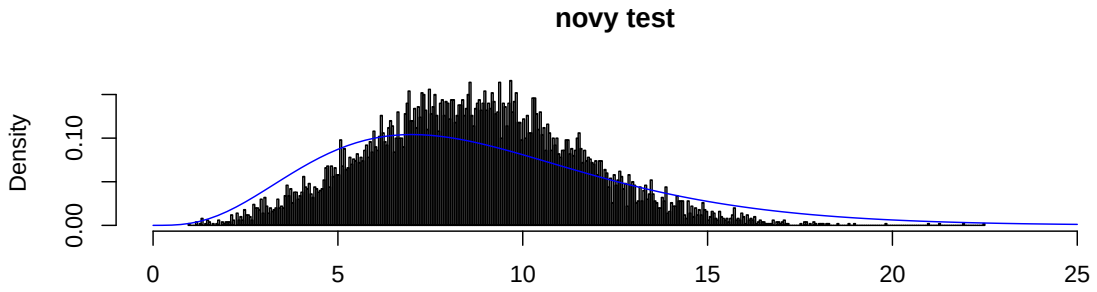
Vidíme, že opäť sú obe skutočné rozdelenia posunuté doprava, a preto budú častejšie zamietať nezávislosť dát. Pôvodný test je viac vychýlený, pričom ma strednú hodnotu vypočítanú z dát rovnú 16.43, oproti novému testu, ktorého stredná hodnota vygenerovaných testov je 13.94. Skutočná hodnota strednej hodnoty chí kvadrátu s 12 stupňami voľnosti je ekvivalentná počtu stupňov voľnosti a teda 12. Keď sa pozrieme na 5% kritickú hodnotu,ktorá je rovná 21.026, tak v prípade pôvodného testu ju prekročilo až 14.1% a v prípade nového testu 6.7%.

4.1.3 Blomquistov test

Aj v tejto podkapitole sme overovali vlastnosti našej novej testovej štatistiky v porovnaní s pôvodnou testovou štatistikou pre rovnaký typ dát ako v predošlých podkapitolách, v prvom prípade $p = q = 2$ bez vnútorných závislostí a v druhom prípade sme si vzali $p = 3$ a $q = 4$. Tu sme si už generovali $3-$rozmerné dáta x_i, ktoré mali vnútorné závislosti dané kovariančnou maticou $K_1 = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}$ a $4-$rozmerné dáta y_i bez vnútorných závislostí, teda ich kovariančná matica bola rovná $K_2 = I_4$. Výsledky nájdeme na nasledujúcich obrázkoch.



Obr. 17: Porovnanie histogramu nasimulovaných B^2 s chí-kvadrát rozdelením χ^2_{pq} s 4 stupňami voľnosti



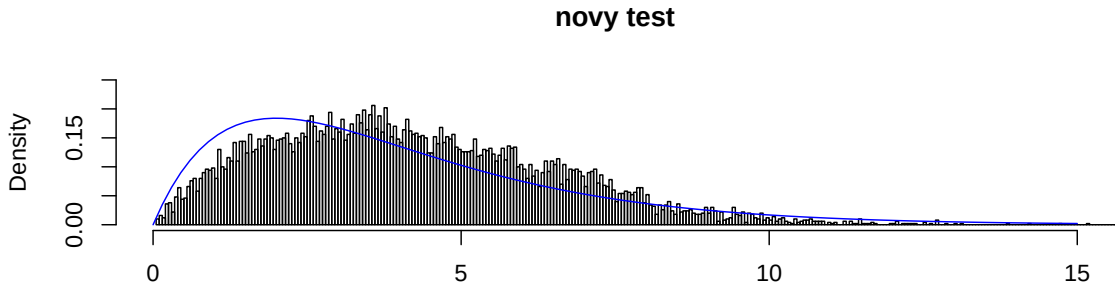
Obr. 18: Porovnanie histogramu nasimulovaných B^2 s chí-kvadrát rozdelením χ^2_{pq} s 12 stupňami voľnosti

Ako môžeme vidieť v tomto prípade neuvádzame starý test, keďže vnútorná štandardizácia tohto testu pre mnohé prípady nekonverguje, a preto by bolo nasimulovať veľa testov zdĺhavé. Pre $p = q = 2$ nám stredná hodnota vyšla 4.42 a počet dát, ktoré prekročili 5%-kritickú hodnotu bolo 6.1%. Pre $p = 3$ a $q = 4$ nám stredná hodnota vyšla 13.33 a počet dát, ktoré prekročili 5%-kritickú hodnotu bolo 6.8%.

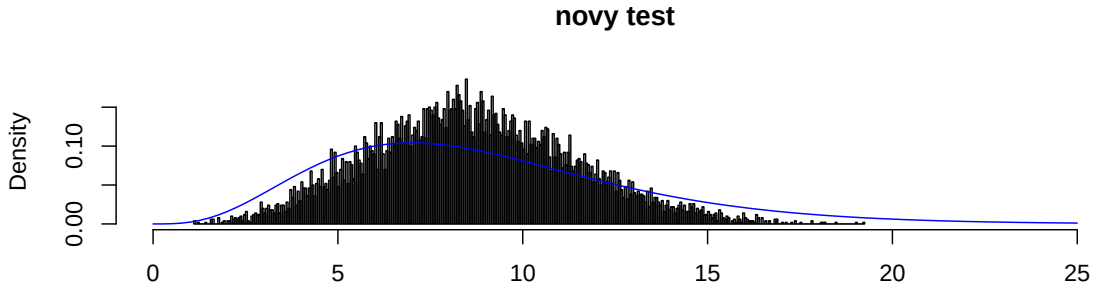
4.1.4 Test pomocou Ojovho znamienkového poradia

Tu už nemáme s čím porovnávať, keďže pôvodne takýto test neexistoval, takže si len overíme vlastnosti nášho nového. Overovali sme ich pre rovnaký typ dát ako v

predošlých podkapitolách, v prvom prípade $p = q = 2$ bez vnútorných závislostí, v druhom prípade pre $p = 3$ a $q = 4$, kde prvé dáta boli vnútorne závislé. Výsledok môžeme vidieť na nasledujúcich obrázkoch.



Obr. 19: Porovnanie histogramu nasimulovaných Q^2_{osr} s chí-kvadrát rozdelením χ^2_{pq} s 4 stupňami voľnosti



Obr. 20: Porovnanie histogramu nasimulovaných Q^2_{osr} s chí-kvadrát rozdelením χ^2_{pq} s 12 stupňami voľnosti

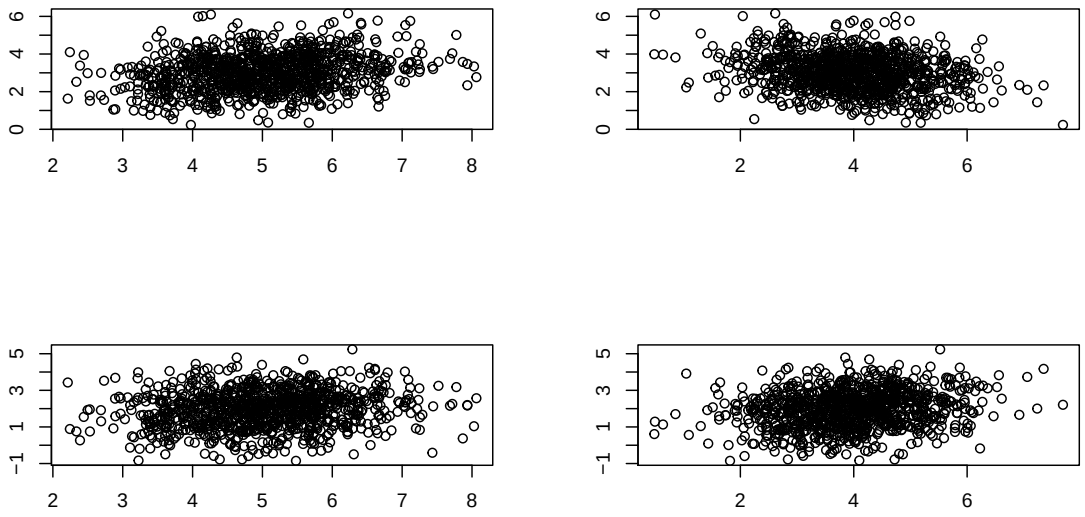
Pre $p = q = 2$ nám stredná hodnota vyšla 4.19 a počet dát, ktoré prekročili 5%-kritickú hodnotu bolo 4.5%. Pre $p = 3$ a $q = 4$ nám stredná hodnota vyšla 12.83 a počet dát, ktoré prekročili 5%-kritickú hodnotu bolo 4.6%.

4.2 Testy na závislých dátach pre $p = q = 2$ a $n = 30$

V tejto časti sa pozrieme na to, ako si testy poradia pri zamietaní nezávislosti dát, ktoré sú závislé. Budeme to skúmať pre fixné veľkosti dát a fixné dimenzie. Na vyjadrenie závislosti použijeme kovariančnú maticu $cov(x, y)$. Kovariančná matica bude mať nasledovný tvar $S = \begin{pmatrix} 1 & 0 & a & a \\ 0 & 1 & -a & a \\ a & -a & 1 & 0 \\ a & a & 0 & 1 \end{pmatrix}$, kde a je náš parameter vyjadrujúci závislosti medzi jednotlivými sadami dát. Tento parameter budeme postupne meniť, čo bude predstavovať narastajúce závislosti. Jeho hodnoty budú $a = 0.2, 0.3, 0.4, 0.5$.

4.2.1 Parameter $a = 0.2$

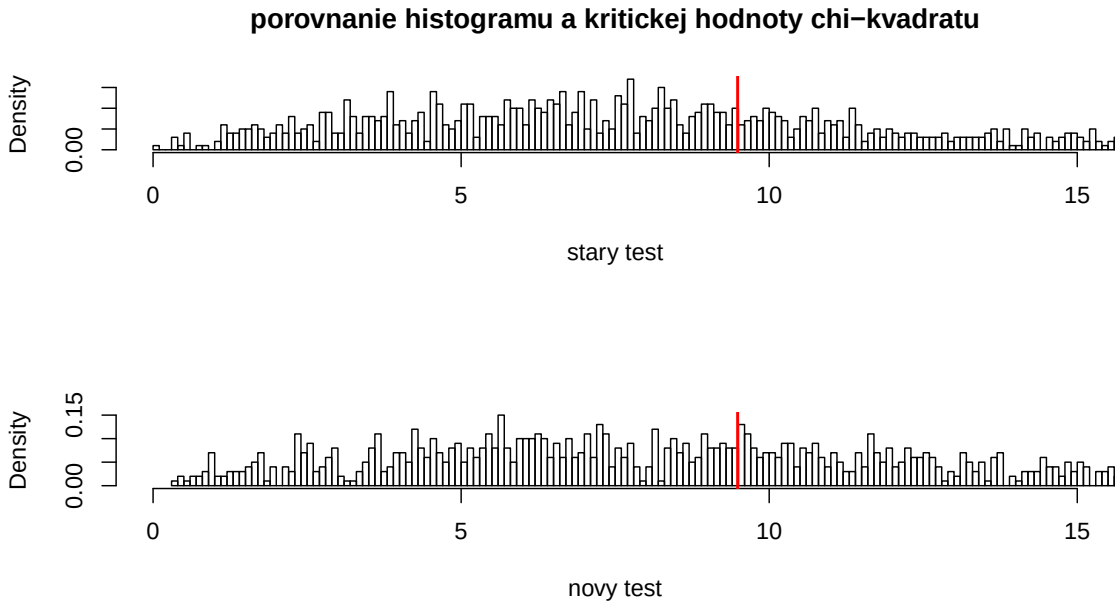
Ako prvé sme sa rozhodli sledovať vlastnosti v prípade málo závislých dát, čo môžeme vidieť aj na nasledujúcom obrázku.



Obr. 21: Jednotlivé závislosti medzi dátami

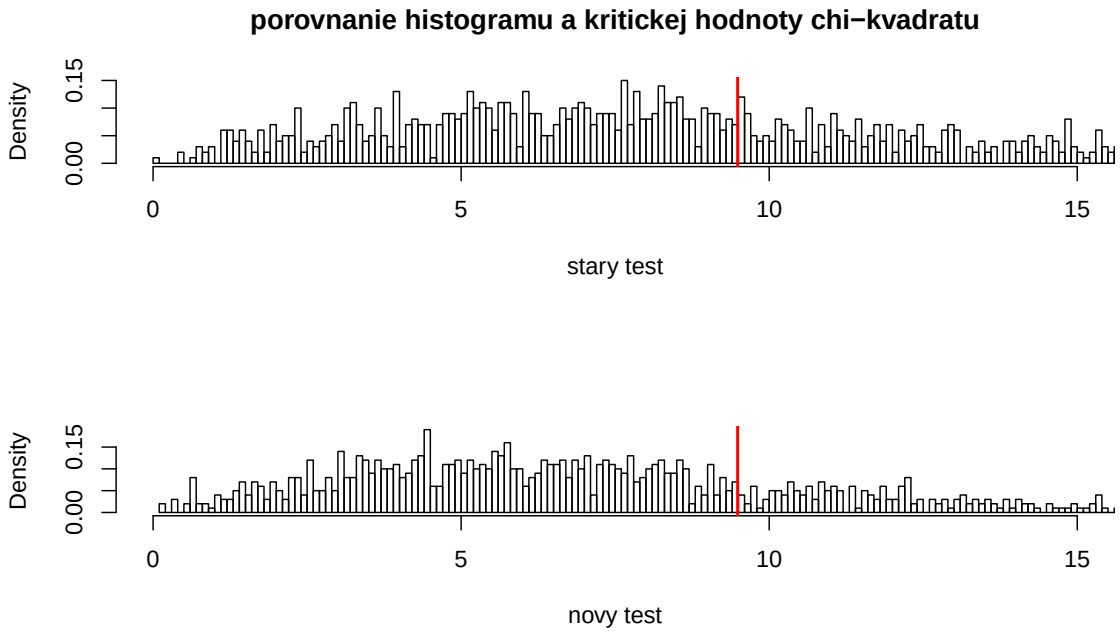
Teraz sme si z tohto rozdelenia generovali dáta a aplikovali na ne staré a nové testy, ktoré sme následne porovnávali. Na obrázku 22 vidíme ako si Spearmanov test poradil s testom nezávislosti pre tieto závislé dáta, ktoré sme generovali 1000-krát. Starý Spearmanov test zamietol hypotézu o nezávislosti dát 326-krát, pričom náš nový test ju zamietol 310-krát. Zvislá čiara na obrázku predstavuje 95% kritickú hodnotu

chí-kvadrátu.



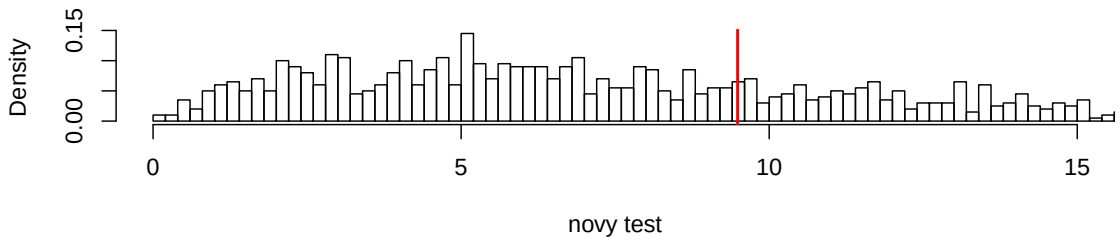
Obr. 22: histogram testových štatistík pre vygenerované dáta

Ďalej sme na tieto dáta aplikovali aj Kendalovo tau, výsledok je na ďalšom obrázku.

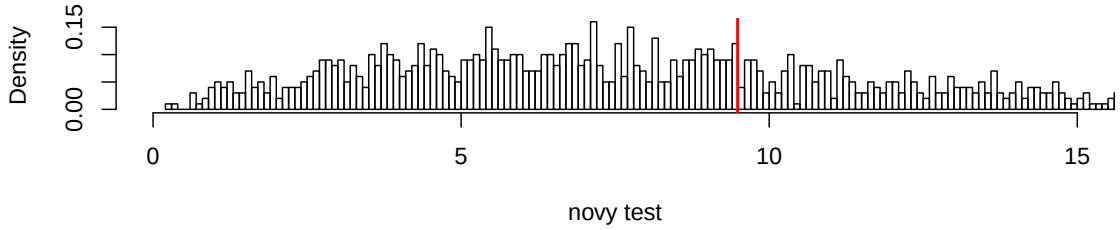


Obr. 23: histogram testových štatistík pre vygenerované dáta

V tomto prípade starý test zamietol hypotézu o nezávislosti 352-krát, pričom nový test 320-krát. Ako posledné sme na tieto dáta aplikovali Blomquistov test a test založený na Ojovom znamienkovom poradí, ktorých výsledky môžeme vidieť na obrázku 24, respektíve obrázku 25



Obr. 24: Blomquistov test

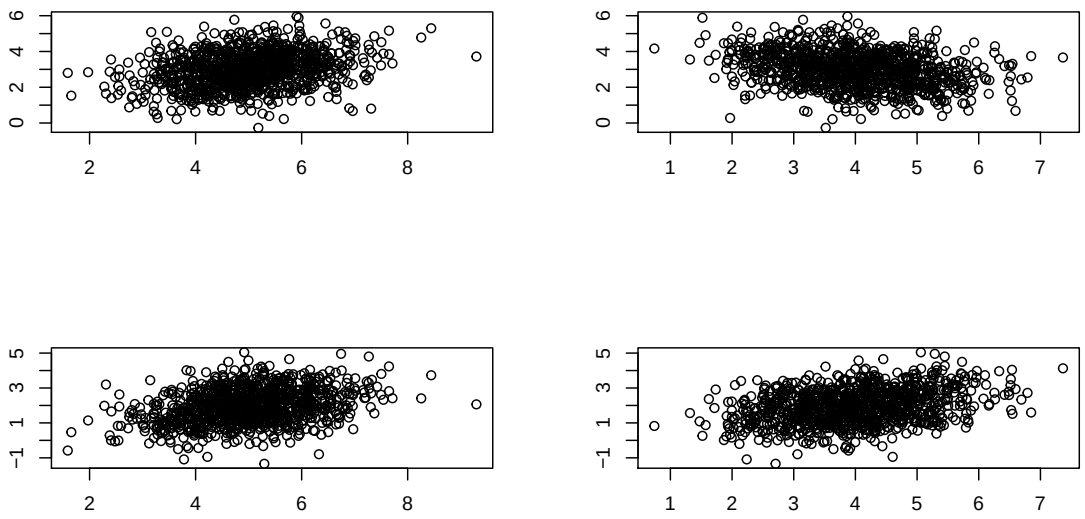


Obr. 25: test založený na Ojovom znamienkovom poradí

Počty zamietnutí boli 290 pre Blomquistov test, respektíve 322 pre test založený na Ojovom znamienkovom poradí. Ako vidíme, pre slabé závislosti mal najmenšiu chybu 2.stupňa pôvodný test založený na Kendalovom tau.

4.2.2 Parameter $a = 0.3$

Ďalej sme sa rozhodli sledovať vlastnosti v prípade viac závislých dát, čo môžno vidieť aj na nasledujúcom obrázku.



Obr. 26: Jednotlivé závislosti medzi dátami

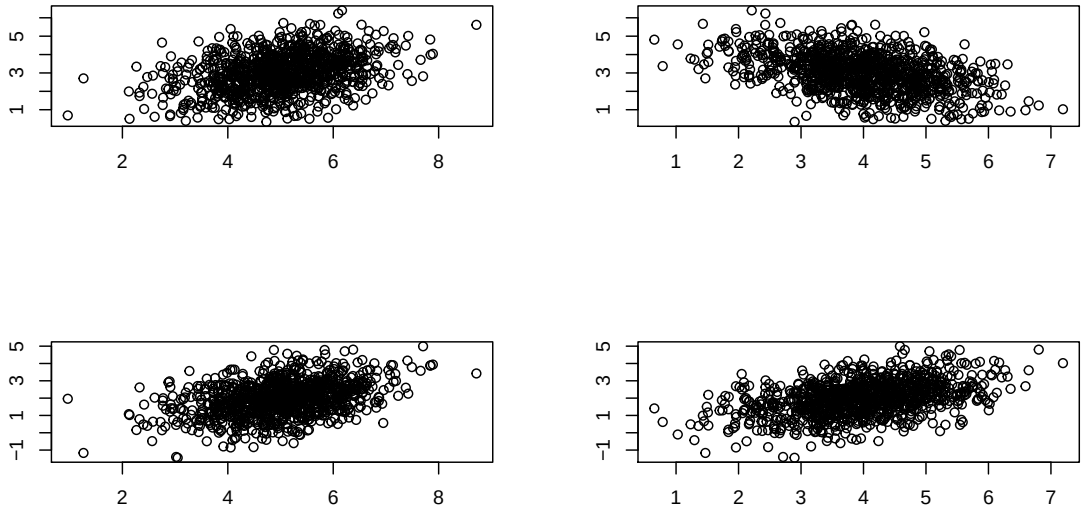
Opäť sme si z tohto rozdelenia generovali dáta, na ktoré sme aplikovali naše testy. Pre prehľadnosť uvádzame výsledky vo forme tabuľky.

a=0.3	ρ^2	τ^2	B^2	Q^2_{osr}
starý test	743	762	**x**	**x**
nový test	682	674	700	690

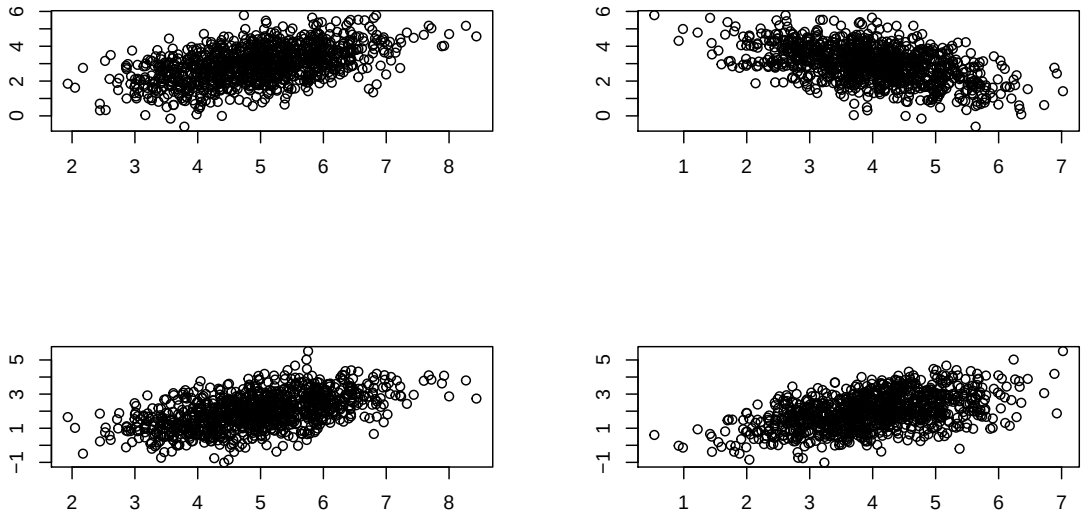
Vidíme, že opäť mal najmenšiu chybu 2.druhu pôvodný test založený na Kendalovom tau.

4.2.3 Parameter $a = 0.4$, $a = 0.5$

Ako posledné sme sa rozhodli sledovať vlastnosti v prípade silne závislých dát, čo môžno vidieť aj na nasledujúcich obrázkoch .



Obr. 27: Jednotlivé závislosti medzi dátami pre $a = 0.6$



Obr. 28: Jednotlivé závislosti medzi dátami pre $a = 0.7$

Opäť sme si z tohto rozdelenia generovali dáta, na ktoré sme aplikovali naše testy. Výsledky uvádzame v nasledujúcich tabuľkách.

a=0.4	ρ^2	τ^2	B^2	Q^2_{osr}
starý test	972	981	**x**	**x**
nový test	952	925	902	950

a=0.5	ρ^2	τ^2	B^2	Q^2_{osr}
starý test	1000	1000	**x**	**x**
nový test	1000	1000	982	1000

Pre takéto silne závislé dáta sa naše nové testy dokázali vyrovnať v chybe 2. druhu pôvodným testom, čo pri menších závislostiach neplatilo. V porovnaní na základe odhadnutej sily dopadli staré testy lepšie.

Záver

Cieľom našej diplomovej práce bolo v prvom rade pochopiť pôvodné testy nezávislosti, s ktorými sme následne pracovali. Jednoduché priestorové znamienko, ktoré je odvodené od priestorového mediánu, ktorý je založený na minimalizácií vzdialeností, sme nahradili zložitejšou funkciou. Konkrétne Ojovým priestorovým znamienkom, ktoré je odvodené z Ojovho mediánu, kde sa namiesto vzdialeností využívajú objemy. Ojovo priestorové znamienko si vyžaduje zložitejší výpočet, ale má vlastnosť affínnej invariantnosti, čo nám pomôže pri štandardizácii.

Ojovo priestorové znamienko sme následne použili na výpočet centrovaného poradia a tieto dve funkcie dosadili do pôvodných testov. Keďže pôvodné testy sú affínne invariantné, vyžadovali sme to aj od našich nových testov. Preto sme na Ojove priestorové znamienko, respektíve Ojovo centrované poradie, použili vonkajšiu štandardizáciu, ktorá stačila, keďže Ojovo priestorové znamienko je affínne ekvivariantné. To, že sú vzniknuté testy affínne invariantné, sme si následne pomocou zopár viet aj dokázali. Následne sme skúmali, či sa aj naše nové testy asymptoticky riadia príslušným chí-kvadrátom, ako to je v prípade ich predchodcov. Tento predpoklad sa nám potvrdil, keďže sme ho overili na mnohých simuláciách, kde sme generovali dáta viac ako 100000-krát. Ďalej sme testy porovnávali na malých sadách dát, kde sa nám ukázalo, že nový test založený na Spearmanovom rho je lepší ako všetky ostatné testy, čo sa týka chyby 1. druhu. V ďalšej časti sme porovnávali testy, keď boli naše dáta závislé. Pre malé závislosti mal najmenšiu chybu 2. druhu pôvodný test založený na Kendalovom tau. To, že staré testy zamietali viac závislých dát však môže byť spôsobené aj tým, že už pre nezávislé dáta sú tieto testy vychýlené doprava. Pri väčších závislostiach už mali testy zhruba rovnakú chybu 2. druhu. Čo sa týka odporúčaného používania našich testov, pre menšie sady dát odporúčame použiť naše testy. Taktiež, ak chceme počítať p-hodnotu na základe permutacií, tak pomocou nášho nového testu to vieme veľmi rýchlo vypočítať. V ostatných prípadoch má náš nový test omnoho jednoduchší tvar a porovnateľné výsledky, a preto ho môžme používať aj v takomto prípade.

V testovaní nezávislosti vidíme aj ďalší možný potenciál. Podľa našich zistení ešte nikto nemenil znamienkovú funkciu v testoch nezávislosti z článku [1]. Preto tam vidíme ďalšie možnosti na vylepšenie. Je tu možnosť napríklad použiť iné normy v priestoro-

vom znamienku. Tak isto je veľa ďalších testov, ktoré používajú klasické priestorové znamienko, a bolo by možno vhodné skúsiť spraviť túto výmenu aj tam a zisťovať, ako sa tieto testy budú správať.

Zoznam použitej literatúry

[1] Oja, H., Randles R.H.(2004): *Multivariete nonparametric tests,* Statistical Science , 598-605. 57-565-94

[2] Taskinen, S., Kankainen, A., Oja, H.(2003) :*Sign test of independence between two random vectors,* Statistics & Probability Letters 62, 9-21

[3] Hettmansperger, TP., Nyblom, J., Oja, H.(1994): *Affine Invariant Multivariete One-Sample Sign Tests,* Journal of the Royal Statistical Society. Series B(Methodological),vol.56,No.1,221-234

[4] Oja, H. (2010), Multivariate Nonparametric Methods with R, Springer

[5] Oja, H. (1999), *Affine invariant multivariate sign and rank tests and corresponding estimates: A review*, Scand. J. Statist., 26, 319–343.

[6] Taskinen, S., Oja, H., Randles, R.H.,(2005), *Multivariete nonparametric tests of independence,* J. Amer. Statist. Assoc. 100

[7] Greene, J. (2014),*Traces of Matrix Products*, Electronic Journal of Linear Algebra, Volume 27