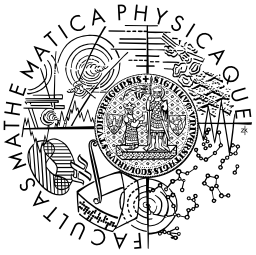


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Michal Kudlík

Testy nezávislosti pro mnohorozměrná data

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Ing. Marek Omelka, PhD.
Studijní program:	Matematika
Studijní obor:	Finanční a pojistná matematika

Praha 2016

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 27.7.2016 Michal Kudlík

Názov práce: Testy nezávislosti pre mnohorozmerné dáta

Autor: Bc. Michal Kudlík

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci diplomovej práce: Ing. Marek Omelka, PhD., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto diplomová práca je prehľadom testov nezávislosti pre mnohorozmerné dáta. Práca obsahuje testy nezávislosti kategoriálnych a spojitých náhodných veličín, testy predpokladajúce normálne rozdelenie dát, neparametrické asymptotické testy a permutačné testy s potrebným teoretickým aparátom s aplikáciou Monte Carlo metódy. Využitím vhodného štatistického softwaru s vhodne zvolenými skutočnými dátami ukáže vhodnosť jednotlých testov a pomocou simulačnej štúdie skontroluje dodržiavanie hladiny testov a porovná silu zvolených testov. Na základe simulačnej štúdie taktiež diskutuje o vhodnosti použitia jednotlivých testov pre rôzne situácie.

Klúčové slová: nezávislosť, permutačné a asymptotické testy nezávislosti, Monte Carlo metóda, simulačná štúdia

Title: Tests of independence for multivariate data

Author: Bc. Michal Kudlík

Department: Department of Probability and Mathematical Statistics

Supervisor: Ing. Marek Omelka, PhD., Department of Probability and Mathematical Statistics

Abstract: This thesis is an overview of tests of independence for multidimensional data. The report includes tests on independence of categorical and continuous random variables, tests assuming normal distribution of data, asymptotic nonparametric tests and permutation tests with application of the Monte Carlo method. This thesis shows the suitability of tests with properly chosen real data, checks significance level and compares the strength of the selected tests by simulation study while using appropriate statistical software. Based on the simulation study the thesis discusses an appropriateness of the use of different tests for different situations.

Keywords: independence, permutation and asymptotic tests of independence, Monte Carlo method, simulation study

Touto cestou by som sa rád poďakoval vedúcemu mojej diplomovej práce p. Ing. Marekovi Omelkovi, PhD. za odbornú pomoc, usmernenie a čas, ktorý mi venoval počas konzultácií. Za trpezlivosť, pochopenie a podporu počas celého môjho štúdia ďakujem priateľke Katke a celej rodine.

Obsah

Úvod

Už v skorých počiatkoch štatistiky sa pred vedcami otvárala téma ako nájsť vzťah medzi dvomi alebo viacerými súbormi pozorovaných veličín. S neustálym zlepšovaním výpočtovej techniky sa v posledných desaťročiach tento problém rozšíril takmer do všetkých oblastí vedy. Táto práca je podnietená týmto praktickým problémom a zaoberá sa detekciou ľubovoľnej miery závislosti medzi pozorovanými dátami. Práca je inšpirovaná najmä dielami Kendall [1], Omelka a Hudecová [7], Legendre a Legendre [5] a Pesarin a Salmaso [14].

Cieľom teoretickej časti tejto práce je prehľadne popísať testy nezávislosti, ich predpoklady, hypotézy a alternatívy, na ktoré sa jednotlivé testy zameriavajú. Ďalšie ciele práce sú: vhodne ukázať praktické použitie jednotlivých testov s reálnymi dátami, v simulačnej štúdii ilustrovať vlastnosti vybraných testov, porovnať ich a vyvodiť vhodnosť jednotlivých testov pre rôzne situácie a dáta.

Prvá kapitola charakterizuje základné testy hypotéz nezávislosti pre kategoriálne veličiny. Zaoberá sa testom hypotézy nezávislosti dvoch a viacerých náhodných veličín na základe kontingenčnej tabuľky. Táto kapitola je určená ako základný prehľad štatistických testov nezávislosti kategoriálnych veličín, testu dobrej zhody a testu pomerom vierohodností pre multinomické rozdelenie.

Druhá kapitola nadväzuje na prvú, no hypotézu a alternatívu skúmame pre dve spojité náhodné veličiny. Táto kapitola obsahuje prehľad základných štatistických nástrojov na zistenie (ne)závislosti medzi náhodnými veličinami: Pearsonov, Kendalov a Spearmanov korelačný koeficient. Spomenuté korelačné koeficienty sú odvodené ako jednotlivé prípady všeobecného korelačného koeficientu definovaného v Kendall [1]. Záver druhej kapitoly je venovaný testu nezávislosti medzi náhodnou veličinou a náhodným vektorom pomocou koeficientu mnohonásobnej korelácie alebo koeficientu determinácie.

V tretej kapitole sú uvedené testy nezávislosti dvoch a viacerých podvektorov mnohorozmerného normálneho rozdelenia. Testy sú založené na teórii maximálnej vierohodnosti, na základe ktorej sa porovnávajú determinanty výberových kovariančných matíc. V nasledujúcej kapitole sú uvedené asymptotické neparametrické testy nezávislosti dvoch vektorov podľa Arthur a Gyorfi [3], ktoré porovnávajú odhad pravdepodobnostnej miery združeného rozdelenia a súčinu marginálnych rozdelení dvoch podvektorov na disjunktnom rozdelení príslušného pravdepodobnostného priestoru.

V piatej kapitole je charakterizovaný základný princíp permutačných testov uvedený v Pesarin a Salmaso [14], ktorý je konkrétne prispôsobený testom nezávislosti. Následne je ukázaný postup podmienenej Monte Carlo metódy, ktorá pri K rôznych náhodných permutáciách náhodného výberu umožní odhad distribúcie testovej štatistiky, posúdenie jej významnosti a teda aj získanie odhadu p-hodnoty permutačných testov pri vhodne zvolenej permutačnej testovej štatistike. Príklady testových štatistík z predchádzajúcich kapitol použité na základe uvedeného princípu permutačných testov sú uvedené v závere piatej kapitoly.

V poslednej teoretickej kapitole uvádzame príklady koeficientov podobností či vzdialeností a konkrétne mnohorozmerné permutačné testy nezávislosti dvoch vektorov, ktoré sú definované využitím vzdialeností či podobností. Sú to napr. Mantlov test (Mantel [6]) alebo DCOV test (Omelka a Hudecová [7]). Posled-

nou uvedenou možnosťou permutačného testu nezávislosti dvoch vektorov bude Protest, ktorý je založený na pôvodných pozorovaniach, nie na maticiach vzdialeností.

Praktickú časť začneme príkladmi testov aplikovaných so skutočnými dátami, kde porovnáme jednotlivé testy, napr. asymptotické a permutačné testy s rovnakou testovou štatistikou, alebo mnohorozmerné permutačné testy uvedené v šiestej kapitole. Prácu zakončíme simulačnou štúdiou pre rôzne modelové situácie. V štúdii skontrolujeme hladinu testov, porovnáme realizovanú silu zvolených testov a zhrnieme vhodnosť použitia testov na jednotlivých príkladoch.

1. Kategoriálne veličiny

Základným štatistickým nástrojom pri teste hypotézy H_0 : {náhodné veličiny sú nezávislé} oproti alternatíve H_1 : {H_0 neplatí} pre kategoriálne náhodné veličiny je Pearsonova χ^2 testová štatistika. Pre jej zavedenie musíme ozrejmiť pojem kontingenčnej tabuľky a na základe multinomického rozdelenia prezentujeme test dobrej zhody založený na Pearsonovej χ^2 testovej štatistike a následne jeho alternatívu, test pomerom vierohodností. Najskôr sa budeme zaoberať testami H_0 pre dve kategoriálne náhodné veličiny, potom pre tri a na záver kapitoly aj medzi náhodnou veličnou a dvojicou náhodných veličín. Uvažované testy môžeme jednoduchým spôsobom rozšíriť aj pre viac náhodných veličín, no test hypotézy H_0 pre viac ako tri kategoriálne náhodné veličiny sa v praxi zriedkakedy využíva.

1.1 Dvojrozmerné kontingenčné tabuľky

Nech X, Z sú dve kategoriálne veličiny nadobúdajúce hodnoty $\{1,\dots,J\}$, resp. $\{1,\dots,K\}$. Nech N je pevné prirodzené číslo a $(X_1,Z_1)^\top,\dots,(X_N,Z_N)^\top$ je náhodný výber s rozsahom N. Pozorované hodnoty náhodných veličín môžeme reprezentovať napr. ako kategórie. V rámci štatistickej analýzy budeme sledovať, aké hodnoty nadobúdajú jednotlivé dvojice z náhodného výberu, resp. do ktorej dvojrozmernej kategórie patrí daná dvojica $(X,Z)^\top$. Označme počet dvojíc klasifikovaných do j-tej kategórie veličiny X a k-tej kategórie veličiny Z ako $n_{j,k} = \sum_{i=1}^{N} \mathbb{I}(X_i = j, Z_i = k)$, ktorý budeme považovať za náhodnú veličinu, ktorá sa v literatúre zvykne označovať ako absolútna združená frekvencia[1] príslušnej kategórie. Pozorované frekvencie môžeme následne vhodne usporiadať do tabuľky, ktorá sa nazýva kontingenčná tabuľka (pozri Tabuľka 1.1).

Tabuľka 1.1: Kontingenčná tabuľka obsahujúca $J \geq 2$ riadkov a $K \geq 2$ stĺpcov.

	$Z=1$ … $Z=K$	
$X=1$	$n_{1,1}$ … $n_{1,K}$	$n_{1,*}$
$X=2$	$n_{2,1}$ … $n_{2,K}$	$n_{2,*}$
⋮	⋮ ⋮ ⋮	⋮
$X=J$	$n_{J,1}$ … $n_{J,K}$	$n_{J,*}$
	$n_{*,1}$ … $n_{*,K}$	N

V kontingenčnej tabuľke sú riadkové absolútne frekvencie $n_{j,*}$ a stĺpcové absolútne frekvencie $n_{*,k}$ definované vťahmi

$$n_{j,*} = \sum_{k=1}^{K} n_{j,k},\ n_{*,k} = \sum_{j=1}^{J} n_{j,k}, \tag{1.1}$$

pričom platí

$$\sum_{j=1}^{J}\sum_{k=1}^{K} n_{j,k} = \sum_{k=1}^{K} n_{*,k} = \sum_{j=1}^{J} n_{j,*} = N. \tag{1.2}$$

[1] česky pozorovaná četnost

Označme pravdepodobnosť výskytu dvojice (X,Z) v j-tej kategórii veličiny X a k-tej kategórii veličiny Z ako $p_{j,k} = \mathsf{P}(X = j,Z = k)$. Označme náhodný vektor $\mathsf{n} = (n_{1,1}, \ldots, n_{J,K})^\top$ a vektor pravdepodobností $\mathsf{p} = (p_{1,1}, \ldots, p_{J,K})^\top$. Takto zavedený náhodný vektor n má multinomické rozdelenie s parametrami N a p (ozn. $\mathrm{Mult}_{JK}(N, \mathsf{p})$). Pozri napr. Anděl [2], kapitola 11. Z vlastností multinomického rozdelenia plynie $\mathsf{E}\, n_{j,k} = Np_{j,k}$, čo označíme ako $m_{j,k}$. Pre pravdepodobnosti a očakávané frekvencie potom platí

$$p_{j,*} = \sum_{k=1}^{K} p_{j,k},\ p_{*,k} = \sum_{j=1}^{J} p_{j,k},\ p_{*,*} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{j,k} = 1, \tag{1.3}$$

$$m_{j,*} = \sum_{k=1}^{K} m_{j,k},\ m_{*,k} = \sum_{j=1}^{J} m_{j,k},\ m_{*,*} = \sum_{j=1}^{J}\sum_{k=1}^{K} m_{j,k} = N. \tag{1.4}$$

Do kontingenčnej tabuľky (pozri Tabuľka 1.2) môžeme usporiadať podobne ako frekvencie tak aj pravdepodobnosti v multinomickom rozdelení.

Tabuľka 1.2: Tabuľka pravdepodobností popisujúca združené a marginálne rozdelenie vektora $(X,Z)^\top$.

	$Z = 1 \ldots Z = K$	
$X = 1$	$p_{1,1} \ \ldots \ p_{1,K}$	$p_{1,*}$
$X = 2$	$p_{2,1} \ \ldots \ p_{2,K}$	$p_{2,*}$
$\vdots$	$\vdots \ \ \vdots \ \ \vdots$	$\vdots$
$X = J$	$p_{J,1} \ \ldots \ p_{J,K}$	$p_{J,*}$
	$p_{*,1} \ \ldots \ p_{*,K}$	1

Poznamenajme, že združené rozdelenie vektora $(X,Z)^\top$ je určené pravdepodobnosťami $p_{j,k}$. Marginálne rozdelenie veličiny X určujú pravdepodobnosti $p_{j,*} = \mathsf{P}(X = j)$ a marginálne rozdelenie veličiny Z je dané $p_{*,k} = \mathsf{P}(Z = k)$.

V tejto časti kapitoly nás bude zaujímať spôsob, ako testovať hypotézu

$$H_0 : \{X \text{ a } Z \text{ sú nezávislé náhodné veličiny}\} \tag{1.5}$$

pre dve kategoriálne veličiny X a Z oproti alternatíve

$$H_1 : \{X \text{ a } Z \text{ nie sú nezávislé náhodné veličiny}\}. \tag{1.6}$$

Predpokladajme, že kategoriálne náhodné veličiny X a Z sú nezávislé, t. j. platí H_0. Potom pre každé $j = 1, \ldots, J$ a $k = 1, \ldots, K$ musí platiť

$$\mathsf{P}(X = j,Z = k) = \mathsf{P}(X = j)\,\mathsf{P}(Z = k), \quad \text{resp.} \quad p_{j,k} = p_{j,*}p_{*,k}. \tag{1.7}$$

1.2 Test dobrej zhody

Jednou z možností, ako skúmať nezávislosť medzi dvojicou kategoriálnych veličín X a Z, je χ^2 test dobrej zhody pre multinomické rozdelenie. Predpokladajme existenciu vyššie zavedeného vektora $\mathsf{n} \sim \mathrm{Mult}_{JK}(N, \mathsf{p})$ a definujme podľa Anděl [2]

(kapitola 11) Pearsonovu testovú štatistiku testu dobrej zhody v multinomickom rozdelení vzťahom

$$\chi_n^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{j,k} - \hat{m}_{j,k})^2}{\hat{m}_{j,k}}. \tag{1.8}$$

V testovej štatistike (1.8) porovnávame napozorované frekvencie $n_{j,k}$ s odhadmi očakávaných frekvencií $\hat{m}_{j,k}$ multinomického rozdelenia. Na výpočet χ_n^2 je navyše treba odhadnúť vektor pravdepodobností p, pretože pre odhad očakávaných frekvencií použijeme vzťah $\hat{m}_{j,k} = N p_{j,*} p_{*,k}$. Odhady pravdepodobností $p_{j,*}$ a $p_{*,k}$ môžeme odvodiť pomocou teórie maximálnej vierohodnosti. Pri platnosti hypotézy H_0 vyplýva z (1.7), že vektor pravdepodobností $\mathsf{p} = (p_{1,1}, \ldots, p_{J,K})^\top$ je funkciou iba $d = J - 1 + K - 1$ parametrov $p_{1,*}, \ldots, p_{J-1,*}$ a $p_{*,1}, \ldots, p_{*,K-1}$. Maximálne vierohodný odhad zložiek vektora p je rovný

$$\hat{p}_{j,k} = \hat{p}_{j,*} \hat{p}_{*,k} = \frac{n_{j,*} n_{*,k}}{N^2}, \tag{1.9}$$

pre všetky $j = 1, \ldots, J$ a $k = 1, \ldots, K$ a odhady očakávaných frekvencií $m_{j,k}$ sú pri platnosti H_0 rovné $\hat{m}_{j,k} = N \hat{p}_{j,k} = N \hat{p}_{j,*} \hat{p}_{*,k} = \frac{n_{j,*} n_{*,k}}{N}$.

Potom testovú štatistiku pri odhadnutých očakávaných frekvenciách môžeme vyjadriť v tvare

$$\chi_n^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{j,k} - \frac{n_{j,*} n_{*,k}}{N}\right)^2}{\frac{n_{j,*} n_{*,k}}{N}}. \tag{1.10}$$

Tvrdenie 1. *Pri predpoklade platnosti hypotézy H_0 má testová štatistika χ_n^2 (1.10) asymptotické χ^2 rozdelenie s $(J-1)(K-1)^2$ stupňami voľnosti, ozn. $\chi^2_{(J-1)(K-1)}$.*

Odvodenie a dôkaz je k dispozícii v Anděl [2], v kapitole 12. Hypotézu H_0 budeme zamietať na asymptotickej hladine α, ak $\chi_n^2 \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$, kde $\chi^2_{(J-1)(K-1)}(1-\alpha)$ je $(1-\alpha)$ kvantil $\chi^2_{(J-1)(K-1)}$ rozdelenia. Asymptotika v tomto prípade vyžaduje dostatočne veľký celkový počet pozorovaní. K záverom, kde všetky očakávané frekvencie neprekročili hodnotu aspoň 5, je treba pristúpiť opatrne, pretože aproximácia rozdelením χ^2 môže byť nepresná. V tomto prípade je nutné zlúčiť niektoré kategórie a predísť malým očakávaným frekvenciám.

1.3 Test pomerom vierohodností

Ukážme alternatívny spôsob ako testovať hypotézu (1.5) oproti alternatíve (1.6) uvedený vo Wilks [4]. Test H_0 môžeme previesť na otázku, či klasifikácia prvkov v kontingenčnej tabuľke podľa riadkov je nezávislá od klasifikácie podľa stĺpcov. Pravdepodobnosť, že náhodný výber zložený z N prvkov je rozdelený podľa multinomického rozdelenia tak, že $n_{j,k}$ je počet prvkov prislúchajúcich do j-teho riadku a k-teho stĺpca pre všetky j a k, je daná vzťahom

$$\frac{N!}{n_{1,1} n_{1,2} \cdots n_{J,K}} \prod_{j=1}^{J} \prod_{k=1}^{K} p_{j,k}^{n_{j,k}}. \tag{1.11}$$

Uvažujme podobne ako vo Wilks [4] pomer (1.11) pri predpoklade platnosti H_0, t. j. pri platnosti (1.7), a (1.11) pri jedinej reštrikcii $\sum_{j=1}^{J} \sum_{k=1}^{K} p_{j,k} = 1$,

t. j. bez predpokladu platnosti H_0. Tento pomer budeme nazývať vierohodnostný pomer a počítať nasledujúcim vzťahom

$$\Lambda = \frac{\prod_{j=1}^{J}\prod_{k=1}^{K}(n_{j,*}n_{*,k})^{n_{j,k}}}{\prod_{j=1}^{J}\prod_{k=1}^{K}(Nn_{j,k})^{n_{j,k}}}, \tag{1.12}$$

kde pre odhad pravdepodobností $p_{j,k}$ v (1.11) použijeme vzťah (1.9). Následne uveďme tvrdenie, na základe ktorého budeme môcť zamietať H_0.

Tvrdenie 2. *Zaveďme testovú štatistiku*

$$\lambda_n = -2\log(\Lambda) = 2\sum_{j=1}^{J}\sum_{k=1}^{K} n_{j,k}\log\frac{n_{j,k}}{\hat{m}_{j,k}}. \tag{1.13}$$

Testová štatistika λ_n definovaná v (1.13) *má v prípade platnosti H_0 asymptotické χ^2 rozdelenie s počtom stupňov voľnosti, ktorý je rovný $(J-1)(K-1)$.*

Dôkaz môžeme nájsť vo Wilks [4]. Testová štatistika λ_n sa občas nazýva deviancia a značí G^2. Hypotézu H_0 budeme zamietať na asymptotickej hladine α pre $\lambda_n \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$, kde $\chi^2_{(J-1)(K-1)}(1-\alpha)$ je $(1-\alpha)$ kvantil rozdelenia χ^2 s $(J-1)(K-1)$ stupňami voľnosti.

1.4 Viacrozmerné kontingenčné tabuľky

Uvedené testy hypotézy nezávislosti medzi dvoma kategoriálnymi veličinami môžeme rozšíriť aj pre test hypotézy vzájomnej nezávislosti viacerých kategoriálnych veličín. Zovšeobecnenie pre viacrozmerné kontingenčné tabuľky uvedieme na príklade trojrozmernej kontingenčnej tabuľky, ktorá obsahuje J riadkov, K stĺpcov a L úrovní. Uvažujme tri kategoriálne náhodné veličiny X, Z a Y nadobúdajúce hodnoty $\{1,\ldots,J\}$, resp. $\{1,\ldots,K\}$, resp. $\{1,\ldots,L\}$. Pre viac ako tri kategoriálne veličiny bude postup analogický.

Označme $(X_1,Z_1,Y_1)^\top,\ldots,(X_N,Z_N,Y_N)^\top$ náhodný výber z realizácií náhodného vektora $(X,Z,Y)^\top$. Nech $n_{j,k,l} = \sum_{i=1}^{N}\mathbb{I}(X_i = j,\ Z_i = k,\ Y_i = l)$ je náhodná veličina s pravdepodobnosťou $p_{j,k,l} = \mathsf{P}(X = j,\ Z = k,\ Y = l)$. Potom náhodný vektor $\mathsf{n} = \{n_{j,k,l};\ \forall j = 1,\ldots,J;\ k = 1,\ldots,K;\ l = 1,\ldots,L\}$ má multinomické rozdelenie s parametrami N a p.

1.4.1 Vzájomná nezávislosť kategoriálnych veličín

Predpokladajme, že náhodné veličiny X, Z a Y sú vzájomne nezávislé. Potom pre každé $j = 1,\ldots,J$, $k = 1,\ldots,K$ a $l = 1,\ldots,L$ platí:

$$\mathsf{P}(X = j,Z = k,Y = l) = \mathsf{P}(X = j)\,\mathsf{P}(Z = k)\,\mathsf{P}(Y = l), \tag{1.14}$$

čo môžeme zapísať aj ako

$$p_{j,k,l} = p_{j,*,*}\cdot p_{*,k,*}\cdot p_{*,*,l}, \tag{1.15}$$

kde $p_{j,*,*},\ p_{*,k,*}$ a $p_{*,*,l}$ sú definované podobne ako (1.3) v dvojrozmernom prípade.

V tomto prípade nás bude zaujímať test nulovej hypotézy

$$H_0 : \{X,\ Y \text{ a } Z \text{ sú vzájomne nezávislé náhodné veličiny}\} \tag{1.16}$$

oproti alternatíve

$$H_1 : \{X,\ Y \text{ a } Z \text{ nie sú vzájomne nezávislé náhodné veličiny}\}. \tag{1.17}$$

Pre tento účel opäť využijeme teóriu maximálnej vierohodnosti pre multinomické rozdelenie a odvodíme maximálne vierohodný odhad zložiek vektora p, ktorý je pri platnosti H_0 rovný

$$\hat{p}_{j,k,l} = \hat{p}_{j,*,*}\hat{p}_{*,k,*}\hat{p}_{*,*,l} = \frac{n_{j,*,*}n_{*,k,*}n_{*,*,l}}{N^3}, \tag{1.18}$$

pre $j = 1, \ldots, J$, $k = 1, \ldots, K$ a $l = 1, \ldots, L$. Potom pri platnosti H_0 platí $\hat{m}_{j,k,l} = N\hat{p}_{j,k,l} = N\hat{p}_{j,*,*}\hat{p}_{*,k,*}\hat{p}_{*,*,l} = \frac{n_{j,*,*}n_{*,k,*}n_{j,*,*}}{N^2}$.

Pre test hypotézy (1.16) uvedieme zovšeobecnenie testových štatistík χ^2 a λ_n pre trojrozmernú kontingenčnú tabuľku a analogické tvrdenie k tvrdeniam 1 a 2. Testovú štatistiku χ^2_n môžeme vyjadriť v tvare

$$\chi^2_n = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(n_{j,k,l} - \hat{m}_{j,k,l})^2}{\hat{m}_{j,k,l}} \tag{1.19}$$

a λ_n v tvare

$$\lambda_n = 2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{j,k,l}\log\frac{n_{j,k,l}}{\hat{m}_{j,k,l}} = 2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{j,k,l}\log(n_{j,k,l}) + 4N\log N$$

$$- 2\sum_{j=1}^{J} n_{j,*,*}\log(n_{j,*,*}) - 2\sum_{k=1}^{K} n_{*,k,*}\log(n_{*,k,*}) - 2\sum_{l=1}^{L} n_{*,*,l}\log(n_{*,*,l}). \tag{1.20}$$

Tvrdenie 3. *Testové štatistiky χ^2_n (1.19) a λ_n (1.20) majú pri platnosti H_0 asymptotické χ^2 rozdelenie s $JKL - J - K - L + 2$ stupňami voľnosti.*

Dôkaz tvrdenia je podobný ako v predchádzajúcich prípadoch pre 2 veličiny. Tvar (1.20) je uvedený aj vo Wilks [4] (výraz č. 17). Hypotézu (1.16) budeme zamietať na asymptotickej hladine α na základe tvrdenia 3 v prípade, že testové štatistiky χ^2_n, resp. λ_n, sú väčie alebo rovné ako $(1 - \alpha)$ kvantil χ^2 rozdelenia s $(JKL - J - K - L + 2)$ stupňami voľnosti.

1.4.2 Nezávislosť náhodnej veličiny a náhodného vektora

Za zaujímavý môžeme považovať aj test hypotézy

$$H_0^* : \{\text{náhodná veličina } X \text{ je nezávislá od náhodného vektora } (Y,Z)^\top\} \tag{1.21}$$

oproti alternatíve

$$H_1^* : \{X \text{ nie je nezávislá od náhodného vektora } (Y,Z)^\top\}, \tag{1.22}$$

pričom vzťah medzi Y a Z je ľubovoľný. Pri platnosti H_0^* je pravdepodobnosť $p_{j,k,l}$ daná vzťahom

$$\mathsf{P}\left(X=j,Z=k,Y=l\right)=\mathsf{P}\left(X=j\right)\mathsf{P}\left(Z=k,Y=l\right)=p_{j,*,*}p_{*,k,l}, \tag{1.23}$$

kde naviac platí $\sum_{j=1}^{J} p_{j,*,*}=1$ a $\sum_{k=1}^{K}\sum_{l=1}^{L} p_{*,j,k}=1$. Počet neznámych parametrov je v tomto prípade $d=J-1+JK-1$ a maximálne vierohodné odhady sú rovné $\hat{p}_{*,k,l}=\frac{n_{j,*,*}}{N}$ a $\hat{p}_{j,*,*}=\frac{n_{*,k,l}}{N}$. Ak platí H_0^*, potom odhadnuté očakávané frekvencie vyjadríme ako

$$\hat{m}_{j,k,l}=\frac{n_{j,*,*}n_{*,k,l}}{N}. \tag{1.24}$$

Tvrdenie 4. *Nech platí H_0^* a $\hat{m}_{j,k,l}$ je dané vzťahom* (1.24)*. Potom χ_n^2 definované v* (1.19) *a*

$$\lambda_n=2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{j,k,l}\log\frac{n_{j,k,l}}{\hat{m}_{j,k,l}} \tag{1.25}$$

majú asymptotické χ^2 rozdelenie s $IJK-1-d=(J-1)(KL-1)$ stupňami voľnosti.

Hypotézu H_0^* zamietame na asymptotickej hladine α na základe tvrdenia 4, ak testové štatistiky χ_n^2, resp. λ_n, sú väčie alebo rovné ako $(1-\alpha)$ kvantil rozdelenia χ^2 s $(J-1)(KL-1)$ stupňami voľnosti. Všimnime si, že tento problém je ekvivalentný s testom hypotézy nezávislosti medzi dvomi náhodnými veličinami X a W, kde $W=Y\times Z$ je kategoriálna veličina, ktorá nadobúda hodnoty $\{1,\ldots,KL\}$. Dôkaz tvrdenia 4 je preto totožný postupne s dôkazmi tvrdení 1 a 2.

2. Korelačné koeficienty

V druhej kapitole práce nás budú opäť zaujímať možnosti testov hypotézy

$$H_0 : \{X \text{ a } Z \text{ sú nezávislé náhodné veličiny}\} \tag{2.1}$$

oproti alternatíve

$$H_1 : \{X \text{ a } Z \text{ nie sú nezávislé náhodné veličiny}\}. \tag{2.2}$$

V prvej časti práce sme testovali H_0 pre dve alebo tri kategoriálne veličiny, čo môžeme podobne rozšíriť aj pre viac kategoriálnych náhodných veličín. Pozornosť v tejto časti práce budeme venovať spojitým náhodným veličinám s konečnými rozptylmi.

V tejto súvislosti je jednou z najčastejšie používaných mier lineárnej závislosti medzi dvoma náhodnými veličinami Pearsonov korelačný koeficient. Ako ďalšie miery závislostí môžeme použiť koeficienty založené na poradí realizácie náhodného výberu zo združeného rozdelenia dvoch náhodných veličín, napr. Kendallov či Spearmanov koeficient poradovej korelácie.

V tomto kontexte spomenieme aj koeficient mnohonásobnej korelácie a koeficient determinácie, ktoré merajú lineárnu závislosť náhodnej veličiny na náhodnom vektore ľubovoľného rozmeru. Štatistickú analýzu teda môžeme založiť na skutočných hodnotách pozorovaného výberu alebo na poradí náhodných veličín v náhodnom výbere.

Na úvod zaveďme definíciu všeobecného korelačného koeficientu, ktorý je zovšeobecnením špecifických typov koeficientov korelácie: výberového korelačného koeficientu, Kendallovho aj Spearmanovho koeficientu poradovej korelácie. Definujme pre každú dvojicu náhodných veličín X_i a X_j z náhodného výberu (X_1,Y_1), $\ldots$, (X_n,Y_n) skóre a_{ij}, špecifické pre náhodnú veličinu X, ktoré spĺňa jedinú podmienku $a_{ij} = -a_{ji}$. Podobne definujme skóre špecifické pre náhodné veličiny Y_i a Y_j, ktoré označíme b_{ij} a podobne ako a_{ij} bude spĺňať $b_{ij} = -b_{ji}$. Poznamenajme, že a_{ij} a b_{ij} definujeme pre $i = j$ rovné 0. Zaveďme podľa Kendall [1], kapitola 2, všeobecnený korelačný koeficient[1]

$$\mathcal{T} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2}}. \tag{2.3}$$

2.1 Koeficienty poradovej korelácie

Ako prvé konkrétne príklady (2.3) uvedieme dva najpoužívanejšie príklady koeficientov poradovej korelácie, Kendallovo τ a Spearmanovo ρ. Koeficienty poradovej korelácie slúžia ako prostriedok pri meraní miery závislosti medzi poradiami dvoch náhodných výberov alebo na meranie intenzity poradovej korelácie v náhodnom výbere. Predpokladajme, že máme k dispozícii náhodný výber (X_1,Y_1), $\ldots$, (X_n,Y_n) zo združeného rozdelenia spojitých veličín X a Y. Poradím R_i^X náhodnej veličiny X_i, resp. poradím R_i^Y náhodnej veličiny Y_i, nazývame počet

[1] angl. general correlation coefficient

náhodných veličín $X_1, \ldots, X_n$, resp. $Y_1, \ldots, Y_n$, ktoré sú menšie alebo rovné X_i, resp. Y_i.

Ak by sme X_i, resp. Y_i, reprezentovali ako čísla, môže nastať situácia, že niektoré z nich sú rovnaké a vytvárajú tzv. zhody. Pri predpoklade spojitých náhodných veličiny X a Y nastávajú zhody s pravdepodobnosťou 0. Štatistická teória, ktorá sa zaoberá testovaním nezávislosti v prípade zhôd je spracovaná napr. v Kendall [1] od kapitoly 3. Teóriou korelačných koeficientov v prípade zhôd sa v tejto práci nebudeme zaoberať.

V Kendall [1] (kapitola 1) sa vyžaduje, aby koeficient poradovej korelácie spĺňal nasledujúce tri vlastnosti:

(i) ak zhoda medzi poradiami je perfektná, t. j. každá dvojica náhodných veličín má rovnaké poradie v oboch výberoch, potom $\mathcal{T}$ nadobúda hodnoty 1 a vyjadruje perfektnú pozitívnu závislosť;

(ii) ak nastane presný opak, t. j. perfektná nezhoda alebo opačné poradie, potom $\mathcal{T}$ je rovné - 1 a vyjadruje perfektnú negatívnu závislosť;

(iii) ak nastane iný prípad než (i) alebo (ii), musí byť tento koeficient ohraničený zhora 1 a zdola - 1, t. j. $\mathcal{T} \in (-1,1)$. Rastúca hodnota koeficientu by mala v rozumných prípadoch predstavovať rastúcu zhodu medzi náhodnými veličinami, resp. ich poradiami.

2.1.1 Kendallov koeficient poradovej korelácie

Predpokladajme, že na základe náhodného výberu veličín X a Y vieme zostrojiť poradia náhodných veličín $\{(R_1^X,R_1^Y), \ldots (R_n^X,R_n^Y)\}$. Ako sme vyššie naznačili, množinu poradí náhodnej veličiny budeme rozumieť ako postupnosť prirodzených čísel od 1 do n. Na základe poradí budeme chcieť zisiť, či medzi náhodnými veličinami X a Y existuje nejaká závislosť. Každej dvojici poradí R_i^X,R_j^X, resp. R_i^Y,R_j^Y, priradíme skóre a_{ij}, resp. b_{ij}. Ak i-tý prvok náhodného výberu bude mať menšie poradie ako j-tý, priradíme tejto dvojici skóre +1 a v opačnom prípade -1. Teda pre X definujme skóre ako

$$a_{ij} = +1 \text{ ak } R_i^X < R_j^X; \;\; a_{ij} = -1 \text{ ak } R_i^X > R_j^X. \tag{2.4}$$

Podobne definujeme $b_{i,j}$ pre Y. Pre takto definované skóre dvoch charakteristík budeme (2.3) značiť τ a nazývať Kendallovo τ alebo Kendallov koeficient poradovej korelácie. Bez vplyvu na všeobecnosť môžeme skóre (2.4) definovať aj na základe realizácie náhodných výberov. Porovnávať poradia veličín v náhodnom výbere je totiž ekvivalentné porovnávaniu samostatných veličín v náhodnom výbere.

Tento pomer indikuje vzťah medzi dvoma náhodnými veličinami X a Y. V skutočnosti hodnota $\tau = 0$ indikuje nezávislosť medzi veličinami, hodnoty 1 a -1 zase perfektnú pozitívnu, resp. negatívnu závislosť. Ak sú všetky poradia v oboch výberoch rovnaké, potom obidve skóre sú rovné 1 a teda menovateľ a čitateľ zlomku (2.3) nadobúdajú svoje maximum.

Vo všeobecnosti ak porovnávame poradia v n-prvkovom náhodnom výbere náhodných veličín X a Y, počet porovnávaných dvojíc je rovný kombinačnému

číslu $\binom{n}{2}=\frac{1}{2}n(n-1)$, pretože vyberáme dva prvky z n-prvkovej množiny. Menovateľ (2.3) pri skóre definovanom (2.4) je rovný $\frac{1}{2}n(n-1)$.

Nech P reprezentuje súčet kladných súčinov $a_{ij}b_{ij}$ a Q absolútnu hodnotu súčtu záporných súčinov $a_{ij}b_{ij}$, čo môžeme zapísať ako

$$P = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}\ \mathbb{I}_{(a_{ij}b_{ij}>0)} \ \text{ a } \ Q = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}\ \mathbb{I}_{(a_{ij}b_{ij}<0)}. \tag{2.5}$$

Na prvý pohľad je zrejmé, že aj $P+Q$ je rovné $\frac{1}{2}n(n-1)$. Pre takto definované P a Q sa dostávame k známej formule pre τ (pozri napr. Kendall [1]), ktorá sa dá vyjadriť pomocou P a Q definovaných v (2.5) ako

$$\tau_n = \frac{P-Q}{\frac{1}{2}n(n-1)} = 1 - \frac{2Q}{\frac{1}{2}n(n-1)} = \frac{2P}{\frac{1}{2}n(n-1)} - 1. \tag{2.6}$$

Koeficient, ktorý sme práve zaviedli podľa Kendall [1], poskytuje určitý druh pohľadu na priemernú mieru zhody medzi dvojicami poradí medzi náhodnými veličinami, pretože vyjadruje rozdiel medzi hodnotami poradí v rovnakom a v protichodnom smere. Z tohto dôvodu môže τ slúžiť ako meradlo súladu medzi dvoma poradiami. V praxi však môžeme naraziť na problém, kedy nastávajú zhody medzi realizáciami veličín. Ako sme zmienili na začiatku kapitoly, teóriu zaoberajúcu sa zhodami môžeme nájsť v Kendall [1].

Predpokladajme, že nás zaujíma test hypotézy H_0 (2.1) oproti alternatíve H_1 (2.2). K tomuto účelu definujme $S = P$ - Q. V Kendall [1] je ukázané, že rozdelenie S je pri platnosti H_0 pre všetky n symetrické a asymptoticky normálne pre rastúce $n \geq 10$, so strednou hodnotou 0 a rozptylom $\frac{n(n-1)(2n+5)}{18}$ (ak nenastávajú zhody). Pre počet pozorovaní n menší ako 10 Kendall [1] vypracoval presné tabuľky kritických hodnôt pre τ, podľa ktorých je možné zamietať hypotézu H_0. Nájdeme ich v Appendixe Kendall [1], v tabuľke B. Pre $n \geq 10$ definujme testovú štatistiku z_n ako

$$z_n = \tau_n \sqrt{\frac{9n(n-1)}{2(2n+5)}}. \tag{2.7}$$

Potom pri platnosti H_0 a $n \geq 10$ má testová štatistika z_n asymptoticky normálne rozdelenie $N(0,1)$ a H_0 zamietneme na asymptotickej hladine α, ak platí $|z_n| \geq \Phi^{-1}(1-\alpha/2)$, kde $\Phi^{-1}(1-\alpha/2)$ je $(1-\alpha/2)$ kvantil normovaného normálneho rozdelenia.

2.1.2 Spearmanov koeficient poradovej korelácie

Ďalším príkladom koeficientu poradovej korelácie je Spearmanovo ρ. V tomto prípade budeme opäť pracovať s poradiami R_i^X a R_i^Y náhodných veličín X_i a Y_i v jednotlivých náhodných výberoch. Tentokrát sa nebudeme zaoberať smerom zmeny, ale konkrétnou veľkosťou zmeny poradia. Na rozdiel od (2.4) budeme pre náhodné veličiny X a Y definovať skóre

$$a_{ij} = R_j^X - R_i^X \quad \text{analogicky} \quad b_{ij} = R_j^Y - R_i^Y. \tag{2.8}$$

Všeobecnený korelačný koeficient (2.3) pri definícii skóre (2.8) označíme ako ρ_n a má tvar

$$\rho_n = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(R_j^X - R_i^X)(R_j^Y - R_i^Y)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(R_j^X - R_i^X)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}(R_j^Y - R_i^Y)^2}}. \tag{2.9}$$

Keďže R_i^X a R_i^Y nadobúdajú všetky hodnoty od 1 do n, potom sú obidva činitele v menovateli (2.9) rovnaké. Po roznásobení čitateľa a využití $\sum_{j=1}^{n} R_j^X = \sum_{j=1}^{n} R_j^Y = \frac{1}{2}n(n-1)$ môžeme Spearmanov koeficient ρ zjednodušiť do tvaru

$$\rho_n = \frac{2n\sum_{i=1}^{n} R_i^X R_i^Y - \frac{1}{2}n^2(n-1)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}(R_j^X - R_i^X)^2}. \tag{2.10}$$

Označme $S(d^2)$ ako súčet kvadratických diferencií poradí X a Y. Potom

$$S(d^2) = \sum_{i=1}^{n}(R_i^X - R_i^Y)^2 = 2\sum_{i=1}^{n}(R_i^X)^2 - 2\sum_{i=1}^{n} R_i^X R_i^Y. \tag{2.11}$$

Ak využijeme rozklad $S(d^2)$, môžeme čitateľ (2.10) upraviť do tvaru

$$2n\sum_{i=1}^{n}(R_i^X)^2 - \frac{1}{2}n^2(n-1)^2 - nS(d^2) = \frac{1}{6}n^2(n^2-1) - nS(d^2). \tag{2.12}$$

Podobne môžeme upraviť aj menovateľa zlomku (2.10)

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(R_j^X - R_i^X)^2 = 2n\sum_{i=1}^{n}(R_i^X)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} R_i^X R_i^Y = \frac{1}{6}n^2(n^2-1) \tag{2.13}$$

a napokon sa dostávame k známej formule pre Spearmanovo ρ v tvare

$$\rho_n = 1 - \frac{6S(d^2)}{n^3 - n}. \tag{2.14}$$

Podobne ako pre τ_n platí podľa Kendall [1], kapitola 5, že rozdelenie ρ_n je pri predpoklade platnosti H_0 symetrické, stredná hodnota ρ_n je rovná 0, rozptyl $\frac{1}{n-1}$ a pre rastúce n má ρ_n asymptoticky (približne) normálne rozdelenie. Skutočné rozdelenie ρ_n sa však blíži k normálnemu rozdeleniu pomalšie ako pre τ_n.

Hypotéza H_0 je ekvivalentná hypotéze $H_0^+ : \{\rho = 0\}$ oproti alternatíve $H_1^+ : \{\rho \neq 0\}$. Pre $n \leq 30$ musíme hodnoty testovej štatistiky porovnávať s vypracovanými tabuľkovými hodnotami kritických hodnôt Spearmanovho poradového koeficientu. Kritické hodnoty $S(d^2)$ môžeme nájsť napr. v Appendixe Kendall [1], Tabuľka č. 2, alebo priamo kritické hodnoty ρ_n sú dostupné v Anděl [2], v tabuľke 18.6. Pre $n > 30$ využijeme podľa Anděl [2], str. 235, že pri platnosti H_0 je rozdelenie náhodnej veličiny $\sqrt{n-1}\ \rho_n \xrightarrow[n\to\infty]{\mathcal{D}} N(0,1)$. Hypotézu H_0 budeme v tomto prípade zamietať na asymptotickej hladine α v prípade, že platí $|\sqrt{n-1}\ \rho_n| \geq \Phi^{-1}(\alpha/2)$.

2.2 Výberový korelačný koeficient

Závislosť medzi dvoma náhodnými veličinami X a Y s konečnými druhými momentmi meriame vo väčšine prípadov pomocou Pearsonovho korelačného koeficientu ρ, ktorého definíciu a vlastnosti môžeme nájsť v Anděl [2], kapitola 2.5. Často však máme k dispozícii iba náhodný výber $(X_1,Y_1), \ldots, (X_n,Y_n)$ zo združeného rozdelenia veličín X a Y, no nie priamo združené rozdelenie týchto dvoch veličín, na základe ktorého môžeme vyjadriť ρ.

Preto sa zavádza výberový korelačný koeficient[2] r_n, ktorý budeme definovať podľa Anděl [2] ako podiel výberovej kovariancie a druhej odmocniny súčinu výberových rozptylov, ak sú kladné. K praktickému výpočtu výberového korelačného koeficientu sa využíva vzorec

$$r_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\overline{X}\,\overline{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2)}}, \tag{2.15}$$

kde $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, resp. $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, sú výberové priemery marginálnych rozdelení X, resp. Y.

Koeficienty poradovej korelácie sme založili na množine poradí náhodných veličín v náhodnom výbere. Na test H_0 môžeme takisto využiť aj samotný náhodný výber, nie iba množinu poradí. Podobne ako v predchádzajúcich prípadoch špecifikujeme konkrétny tvar skóre pre všeobecný korelačný koeficient (2.3). Definujme skóre

$$a_{ij} = X_j - X_i \quad \text{analogicky} \quad b_{ij} = Y_j - Y_i \tag{2.16}$$

a dosadením (2.16) do (2.3) dostávame

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_j - X_i)(Y_j - Y_i)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_j - X_i)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}(Y_j - Y_i)^2}}, \tag{2.17}$$

čo jednoduchým roznásobením výrazov môžeme upraviť na tvar (2.15). Ďalej uvedieme tvrdenie, pomocou ktorého môžeme zamietať hypotézu H_0.

Tvrdenie 5. *Nech $(X_1,Y_1), \ldots, (X_n,Y_n)$ je náhodný výber z dvojrozmerného normálneho rozdelenia, pre ktorý platí $Var(X) > 0$, $Var(Y) > 0$ a $\rho = 0$. Definujme pre $n \geq 3$ testovú štatistiku*

$$T = \sqrt{(n-2)}\frac{r_n}{\sqrt{1 - r_n^2}}. \tag{2.18}$$

Potom testová štatistika T (2.18) má študentovo t rozdelenie s $n-2$ stupňami voľnosti.

Dôkaz tvrdenia 5 môžeme nájsť v Anděl [2] na str. 117. Poznamenajme, že test hypotézy H_0 je pri platnosti predpokladov z tvrdenia 5 ekvivalentný testu hypotézy

[2] angl. product-moment correlation coefficient

$$H_0^* : \{\rho = 0\} \text{ oproti alternatíve } H_1^* : \{\rho \neq 0\}. \tag{2.19}$$

Na základe tvrdenia 5 budeme zamietať hypotézu H_0 na hladine α, ak $|T| \geq t_{n-2}(1-\alpha/2)$, kde $t_{n-2}(1-\alpha/2)$ je $(1-\alpha/2)$ kvantil rozdelenia t_{n-2}. V princípe tento test vyžaduje, aby náhodný výber pochádzal z dvojrozmerného normálneho rozdelenia. Ak náhodný výber nespĺňa predpoklad normality, môžeme využiť princíp permutačných testov, ktorým je venovaná samostatná kapitola práce.

Ukázali sme, že zo zovšeobecneného koeficientu (2.3) môžeme odvodiť všetky tri známe korelačné koeficienty τ, ρ a ρ v závislosti na voľbe skórovacej metódy medzi dvomi náhodnými veličinami. V prípade τ to bolo najjednoduchšie možné priradenie skóre, ktoré priradzovalo kladné alebo záporné jednotky podľa poradia vo výbere. Toto priradzovanie nezáviselo na tom, ako ďaleko od seba boli jednotlivé poradia. Skórovacia metóda pre ρ je viac sofistikovanejšia ako τ, pretože priradzuje väčšiu váhu rozdielom medzi poradiami, ktoré majú medzi sebou viacero iných prvkov. Avšak na základe τ, na rozdiel od ρ, môžeme usúdiť, akým smerom je porušená hypotéza.

Metóda výpočtu ρ zachytí hodnotu rozdielu medzi napozorovanými hodnotami náhodných veličín. Hlavnou nevýhodou korelačného koeficientu je ale linearita, t. j. korelačný koeficient dokáže zachytiť iba lineárny vzťah medzi náhodnými veličinami X a Y.

2.3 Koeficient mnohonásobnej (lineárnej) korelácie

V predchádzajúcom texte sme sa zaoberali testom nezávislosti medzi dvoma náhodnými veličinami. Teraz uvedieme test hypotézy

$$H_0^{'} : \{\text{náhodná veličina } Y \text{ je nezávislá od náhodného vektora } \mathbf{X}_{(p\times 1)}\} \tag{2.20}$$

oproti alternatíve $H_1^{'} : \{H_0^{'} \text{ neplatí }\}$.

Test založíme na koeficiente mnohonásobnej korelácie $\rho_{Y,\mathbf{X}}$, ktorého definíciu a vlastnosti nájdeme napr. v Anděl [2] na str. 40. $\rho_{Y,\mathbf{X}}$ je vlastne najväčší korelačný koeficient medzi Y a veličinou $\alpha + \boldsymbol{\beta}^\top \mathbf{X}$, kde $\alpha \in \mathbb{R}$ a $\boldsymbol{\beta} \in \mathbb{R}^p$. Koeficient mnohonásobnej korelácie nezapadá do kontextu všeobecného korelačného koeficientu, no môžeme ho využiť pre test hypotézy nezávislosti medzi náhodnou veličinou Y a celým vektorom $\mathbf{X}$.

Nech Y je náhodná veličina, $p \geq 1$ a $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ náhodný vektor s konečnými druhými momentmi, podľa ktorých máme k dispozícii náhodný výber

$$\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \end{pmatrix}, \tag{2.21}$$

s regulárnou výberovou korelačnou maticou $\mathbb{R}_{\mathbf{XX}}$. Potom výberový koeficient mnohonásobnej korelácie $r_{Y,\mathbf{X}}$ definujeme (podľa Anděl [2], str.123) ako nezáporné číslo spĺňajúce

$$r^2_{Y,\mathbf{X}} = \mathbb{R}_{Y\mathbf{X}} \mathbb{R}^{-1}_{\mathbf{XX}} \mathbb{R}_{\mathbf{X}Y}, \tag{2.22}$$

kde $\mathbb{R}_{Y\mathbf{X}}$, resp. $\mathbb{R}_{\mathbf{X}Y}$, sú výberové korelačné matice medzi Y a $\mathbf{X}$ zostrojené na základe konkrétnej realizácie náhodného výberu (2.21). Nasledujúce tvrdenie nám umožní test hypotézy $H_0^{'}$ a jeho dôkaz môžeme nájsť v Anděl [2] na str. 125.

Tvrdenie 6. *Predpokladajme, že vektory v* (2.21) *sú výberom z normálneho rozdelenia s regulárnou variančnou maticou a* $n > p + 1$. *Pri platnosti* $\rho_{Y,\mathbf{X}} = 0$ *má testová štatistika*

$$F = \frac{r^2_{Y,\mathbf{X}}}{1 - r^2_{Y,\mathbf{X}}} \frac{n-p-1}{p} \tag{2.23}$$

F *rozdelenie s* p *a* $n - p - 1$ *stupňami voľnosti, ktoré značíme* $F_{p,n-p-1}$.

Na základe tvrdenia 6 budeme zamietať hypotézu $H_0^{'}$ na hladine α, ak platí $F \geq F_{p,n-p-1}(1-\alpha)$, kde $F_{p,n-p-1}(1-\alpha)$ je $(1-\alpha)$ kvantil rozdelenia $F_{p,n-p-1}$.

Výberový koeficient mnohonásobnej korelácie je v tesnom vzťahu s koeficientom determinácie R^2 definovanom v kontexte lineárnej regresie, pozri Zvára [15], kapitola 3.5. Koeficient determinácie R^2 vyjadruje, aká časť rozptylu σ^2_Y náhodnej veličiny Y je vysvetlená lineárnou kombináciou $p \geq 1$ vysvetľujúcich veličín X_1, X_2, ..., X_p.

V prípade lineárneho modelu $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, kde ϵ je náhodná veličina so strednou hodnotou 0 a kladným rozptylom σ^2, je koeficient determinácie R^2 zhodný s druhou mocninou výberového koeficientu mnohonásobnej korelácie $r^2_{Y,\mathbf{X}}$ medzi náhodnou veličinou Y a náhodným vektorom $\mathbf{X}$. Potom pre test hypotézy $H_0^{'}$ môžeme pri predpoklade $\epsilon \sim N(0,\sigma^2)$ opäť využiť tvrdenie 6, kde $r^2_{Y,\mathbf{X}} = R^2$.

Všimnime si, že predpoklady sú v tomto prípade mierne rozličné od predpokladov v tvrdení 6, no testové štatistiky sú v oboch prípadoch totožné. Tvrdenie 6 vyžaduje normalitu celého náhodného výberu (2.21), zatiaľ čo pri využití koeficientu determinácie je postačujúca normalita náhodnej veličiny Y, ktorú modelujeme pomocou lineárnej kombinácie prvkov vektora $\mathbf{X}$.

Uvažujme inú situáciu. Nech $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_q)^\top$ je q-rozmerný náhodný vektor vytvorený ľubovoľnou transformáciou vektora $\mathbf{X}$ a $Z_1 = 1$, napr. $Z_2 = X_1$, $Z_3 = X_1^2$. Zaveďme lineárny model v nelineárnych prvkoch vektora $\mathbf{X}$, ktorý môžeme zapísať v tvare

$$Y = \boldsymbol{\beta}^\top \mathbf{Z} + \epsilon, \tag{2.24}$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^\top$ je vektor regresných koeficientov a ϵ je náhodná veličina s rozdelením $N(0,\sigma^2)$. Potom môžeme testovať hypotézu $H_0^{'}$ medzi náhodnou veličinou Y a vektorom $(Z_2, \ldots, Z_q)^\top$ pomocou testovej štatistiky

$$F = \frac{r^2_{Y,\mathbf{Z}}}{1 - r^2_{Y,\mathbf{Z}}} \frac{n-q}{q-1} = \frac{R^2}{1-R^2} \frac{n-q}{q-1}, \tag{2.25}$$

kde R^2 je koeficient determinácie pre model 2.24. Pri platnosti H_0 má testová štatistika 2.25 $F_{q-1,n-q}$ rozdelenie s $q-1$ a $n-q$ stupňami voľnosti. Uvažujme pre názornú ukážku model (2.24) v tvare $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, pre náhodné veličiny Y a X s konečnými rozptylmi. Na základe testovej štatistiky (2.25) môžeme zistiť, či medzi náhodnými veličinami Y a X existuje lineárny alebo kvadratický vzťah.

3. Mnohorozmerné normálne rozdelenie

Uvažujme p-rozmerný náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)^\top$, $\mathbf{Y} = (Y_1, \ldots, Y_q)^\top$ nech je q-rozmerný náhodný vektor a $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$ je $(p+q)$-rozmerný náhodný vektor, ktorý má $(p+q)$-rozmerné normálne rozdelenie so strednou hodnotou $\boldsymbol{\mu}$ a rozptylovou[1] maticou $\mathrm{Var}(\mathbf{Z})$. V tejto časti práce nás bude zaujímať test hypotézy

$$H_0 : \{\text{vektor } \mathbf{X} \text{ je nezávislý od vektora } \mathbf{Y}\} \tag{3.1}$$

oproti alternatíve

$$H_1 : \{\text{vektory } \mathbf{X} \text{ a } \mathbf{Y} \text{ nie sú nezávislé}\}. \tag{3.2}$$

Označme variančnú maticu vektora $\mathbf{Z}$ ako

$$\mathrm{Var}(\mathbf{Z}) = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_Y \end{bmatrix}, \tag{3.3}$$

kde $\boldsymbol{\Sigma}_{YX} = \boldsymbol{\Sigma}_{YX}^\top$, podmatica $\boldsymbol{\Sigma}_X$ je kovariančná matica vektora $\mathbf{X}$ s rozmermi $p \times p$ a podobne podmatica $\boldsymbol{\Sigma}_Y$ je kovariančná matica vektora $\mathbf{Y}$ s rozmermi $q \times q$. Matica $\boldsymbol{\Sigma}_{XY}$ s rozmermi $p \times q$ označuje kovariančnú maticu medzi vektormi $\mathbf{X}$ a $\mathbf{Y}$, t. j. $\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \boldsymbol{\Sigma}_{XY}$. Ak platí hypotéza H_0, potom pre kovariančnú maticu platí

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0, \text{ kde } \boldsymbol{\Sigma}_0 = \begin{bmatrix} \boldsymbol{\Sigma}_X & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_Y \end{bmatrix}, \quad \text{resp. } \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^\top = \mathbf{0}, \tag{3.4}$$

kde $\boldsymbol{\Sigma}_X$, resp. $\boldsymbol{\Sigma}_Y$, sú ľubovoľné nenulové pozitívne definitné matice s rozmermi $p \times p$, resp. $q \times q$, a $\mathbf{0}$ reprezentuje matice s príslušnými rozmermi obsahujúce nulové prvky.

3.1 Wilksova štatistika

Skôr ako prejdeme k samotnému odvodeniu testov hypotézy H_0 pri predpokladaní mnohorozmernej normality, uveďme definíciu Wilksovho rozdelenia Λ, ktoré využijeme na test H_0. Teoretické zavedenie Wilksovho rozdelenia, odvodenie hustoty a jeho aplikácie sú dostupné v článku Pham-Gia [11]. Zároveň pre lepšiu predstavu o Wilksovom rozdelení uvedieme Hotelingovo i Wishartovo rozdelenie a ich vzťahy s Wilksovým rozdelením.

Ako prvé definujme Wishartovo rozdelenie $W_p(\boldsymbol{\Sigma},n)$. Nech náhodný vektor $\mathbf{Y}_i \sim N_p(\mathbf{0},\Sigma)$ pre všetky $i = 1, \ldots, n$ a $\mathbb{Y}_{n\times p} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\top$. Potom definujeme Wishartovo rozdelenie ako

$$\mathbf{M}_{(p\times p)} = \mathbb{Y}^\top \mathbb{Y} \tag{3.5}$$

a označíme $\mathbf{M} \sim W_p(\boldsymbol{\Sigma},n)$. Ďalej uvažujme $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbb{I})$ a $\mathbf{M} \sim W_p(\mathbb{I},n)$. Potom hovoríme, že náhodná veličina

$$T = n\, \mathbf{Z}^\top \mathbf{M}^{-1} \mathbf{Z} \tag{3.6}$$

[1] inak aj variančná, kovariančná alebo variančno-kovariačná

má Hotellingovo T^2 rozdelenie s parametrami $p \in \mathbb{N}$ a $n \in \mathbb{N}$, čo označíme ako $T^2(n,p)$. Poznamenajme, že Hottelingovo rozdelenie je zovšeobecnením študentovho t rozdelenia do viacrozmerného priestoru.

Pomocou Wishartovho rozdelenia (3.5) budeme definovať aj Wilksovo Λ rozdelenie a označíme ako $\Lambda_{p,q,n}$. Uvažujme dvojicu nezávislých náhodných Wishartových matíc $\mathbf{A} \sim W_p(\mathbf{\Sigma},n)$ a $\mathbf{B} \sim W_p(\mathbf{\Sigma},q)$, pričom $n > p$. Potom rozdelenie podielu dvoch determinantov

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A}+\mathbf{B}|} \tag{3.7}$$

sa nazýva Wilksovo rozdelenie a značí $\Lambda_{p,q,n}$. V praxi je bežné aproximovať Wilksovu štatistiku pomocou Chi-kvadrát rozdelenia Bartlettovou transformáciou v tvare

$$-\left(n - \frac{p-q+1}{2}\right) \log \Lambda_{p,q,n} \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{pq}. \tag{3.8}$$

Podľa Pham-Gia [11], časť 6.1, je možné vyjadriť vzťah aj medzi $T \sim T^2(n,p)$ a (3.7)

$$\Lambda^{\frac{2}{n}} = \left(1 + \frac{T^2}{n-1}\right)^{-1}. \tag{3.9}$$

Vzťah medzi $\Lambda_{p,q,n}$ a $T^2(n,p)$, resp. χ^2_{pq}, môžeme využiť napríklad na odvodenie kritických hodnôt Wilksovho rozdelenia alebo na aproximáciu rozdelenia testovej štatistiky Λ, ktorú definujeme v nasledujúcej časti.

3.2 Test pomerom vierohodností

Ďalej predpokladajme, že máme k dispozícii náhodný výber $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ náhodných vektorov $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top \sim N_{p+q}(\boldsymbol{\mu},\mathbf{\Sigma})$. V tejto časti práce využijeme teóriu maximálnej vierohodnosti a na nej založený test pomerom vierohodností, ktorého odvodenie môžeme nájsť v Anděl [2], kapitola 15.6.

Aplikovaním teórie maximálnej vierohodosti získame na základe náhodného výberu z mnohorozmerného normálneho rozdelenia maximálne vierohodný odhad kovariančnej matice $\mathbf{\Sigma}$ vektora $\mathbf{Z}$ v tvare

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i - \bar{\mathbf{Z}}_n)(\mathbf{Z}_i - \bar{\mathbf{Z}}_n)^\top = \frac{n-1}{n}\mathbb{S}, \tag{3.10}$$

kde $\bar{\mathbf{Z}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i$ je maximálne vierohodný odhad vektora $\boldsymbol{\mu} = \mathsf{E}\,\mathbf{Z}$ a $\mathbb{S}$ je výberová kovariančná matica vektora $\mathbf{Z}$. Z dôvodu ďalšieho postupu uvedieme aj tvar logaritmickej vierohodnostnej funkcie $\ell_n(\boldsymbol{\mu}, \mathbf{\Sigma})$ pre mnohorozmerné normálne rozdelenie s parametrami $\boldsymbol{\mu}$ a $\mathbf{\Sigma}$, ktorá má tvar

$$\ell_n(\boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{n}{2}\log(|\mathbf{\Sigma}|) - \sum_{i=1}^{n}\frac{1}{2}(\mathbf{Z}_i - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{Z}_i - \boldsymbol{\mu}) - \frac{n}{2}\log(2\pi). \tag{3.11}$$

Pre ďalšie odvodenie upravíme prostredný člen výrazu (3.11) do tvaru (Tr značí stopu štvorcovej matice)

$$\sum_{i=1}^{n}\frac{1}{2}(\mathbf{Z}_i - \boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{Z}_i - \boldsymbol{\mu}) = \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Z}_i - \boldsymbol{\mu})(\mathbf{Z}_i - \boldsymbol{\mu})^\top\right). \tag{3.12}$$

Označme $\ell_n(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ logaritmickú vierohodnostnú funkciu $(p+q)$-rozmerného normálneho rozdelenia so strednou hodnotou $\boldsymbol{\mu}$ a variančnou maticou $\mathrm{Var}(\mathbf{Z})$ v maximálne vierohodnom odhade $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\Sigma}}$. Podobne označme $\ell_n(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}_0)$ logaritmickú vierohodnostnú funkciu v maximálne vierohodnom odhade $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\Sigma}}_0$ spočítanú za platnosti nulovej hypotézy H_0, t. j. $\hat{\boldsymbol{\Sigma}}_0$ je v blokovo diagonálnom tvare

$$\hat{\boldsymbol{\Sigma}}_0 = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_X & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_Y \end{bmatrix} \text{ a } \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_X & \hat{\boldsymbol{\Sigma}}_{XY} \\ \hat{\boldsymbol{\Sigma}}_{YX} & \hat{\boldsymbol{\Sigma}}_Y \end{bmatrix}. \tag{3.13}$$

Na vektor $\boldsymbol{\mu}$ nekladieme v oboch prípadoch žiadne požiadavky. Pri predpoklade splnenia podmienok regularity metódy maximálnej vierohodnosti môžeme uviesť nasledujúce tvrdenie.

Tvrdenie 7. *Ak platí hypotéza H_0, potom testová štatistika*

$$2\log\lambda_n = 2(\ell_n(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) - \ell_n(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}_0)) \tag{3.14}$$

má asymptotické χ^2_{pq} rozdelenie s pq stupňami voľnosti.

Počet stupňov voľnosti χ^2 rozdelenia pre test pomerom vierohodností je daný ako rozdiel počtu parametrov alternatívy a hypotézy. V tomto prípade určíme počet stupňov voľnosti ako rozdiel počtu prvkov v dolnej trojuholníkovej matici s rozmerom $p+q$ a súčtu počtu prvkov v dolných trojuholníkových maticiach s rozmermi p a q, t. j.

$$\frac{(p+q)(p+q+1)}{2} - \frac{q(q+1)}{2} - \frac{p(p+1)}{2} = pq. \tag{3.15}$$

Tvrdenie 7 nám umožňuje testovať H_0 pre dostatočný počet pozorovaní. Hypotézu budeme zamietať na asymptotickej hladine spoľahlivosti α, ak $2\log\lambda_n \geq \chi^2_{pq}(1-\alpha)$, kde $\chi^2_{pq}(1-\alpha)$ je $1-\alpha$ kvantil χ^2_{pq} rozdelenia.

Testovú štatistiku $2\log\lambda_n$ môžeme využitím (3.12) upraviť do tvaru

$$2\log\lambda_n = -n\log\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_X||\hat{\boldsymbol{\Sigma}}_Y|} - \sum_{i=1}^{n} \mathrm{Tr}\left((\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}_0^{-1})(\mathbf{Z}_i - \hat{\boldsymbol{\mu}})(\mathbf{Z}_i - \hat{\boldsymbol{\mu}})^\top\right), \tag{3.16}$$

kde druhú časť výrazu zjednodušíme nasledujúcim spôsobom

$$n\mathrm{Tr}\left((\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}_0^{-1})\sum_{i=1}^{n}\frac{1}{n}(\mathbf{Z}_i - \hat{\boldsymbol{\mu}})(\mathbf{Z}_i - \hat{\boldsymbol{\mu}})^\top\right) = n\mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}\right). \tag{3.17}$$

Je zjavné, že $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}} = \mathbb{I}_{(p+q)}$, kde $\mathbb{I}$ je jednotková matica a $\mathrm{Tr}(\mathbb{I}_{(p+q)}) = p+q$. Ďalej upravme aj maticový šúčin $\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}$ v (3.17) následovne

$$\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_X^{-1} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_Y^{-1} \end{bmatrix} \times \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_X & \hat{\boldsymbol{\Sigma}}_{XY} \\ \hat{\boldsymbol{\Sigma}}_{YX} & \hat{\boldsymbol{\Sigma}}_Y \end{bmatrix} = \begin{bmatrix} \mathbb{I}_p & \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\Sigma}}_{XY} \\ \hat{\boldsymbol{\Sigma}}_Y^{-1}\hat{\boldsymbol{\Sigma}}_{YX} & \mathbb{I}_q \end{bmatrix}, \tag{3.18}$$

ktorého stopa je rovná opäť $p+q$ a teda výraz (3.17) je rovný 0. Odvodili sme výslednú formulu testovej štatistiky (3.14) testu pomerom vierohodností do tvaru

$$2\log\lambda_n = -n\log\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_X||\hat{\boldsymbol{\Sigma}}_Y|}. \tag{3.19}$$

V Rencher [10], v kapitole 7.4, je uvedená alternatívna testová štatistika k testu hypotézy H_0 označovaná Λ a definovaná vzťahom

$$\Lambda = \frac{|\mathbb{S}|}{|\mathbb{S}_X||\mathbb{S}_Y|}, \tag{3.20}$$

kde $\mathbb{S}$ je bloková výberová kovariančná matica v tvare

$$\mathbb{S} = \begin{bmatrix} \mathbb{S}_X & \mathbb{S}_{XY} \\ \mathbb{S}_{YX} & \mathbb{S}_Y \end{bmatrix}. \tag{3.21}$$

Všimnime si, že zlomok vo výraze (3.19) a zlomok (3.20) sú podobné, pretože medzi $\mathbb{S}$ a $\hat{\boldsymbol{\Sigma}}$ platí vzťah (3.10). S využitím (3.10) navyše zisťujeme, že tieto výrazy sú totožné, t. j.

$$\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_X||\hat{\boldsymbol{\Sigma}}_Y|} = \Lambda. \tag{3.22}$$

Poznamenajme, že menovateľ (3.20) je konzistentným odhadom $|\boldsymbol{\Sigma}_X||\boldsymbol{\Sigma}_Y|$, teda determinantu $\boldsymbol{\Sigma}$, ak platí H_0, t. j. ak $\boldsymbol{\Sigma}_{XY} = \mathbf{0}$. Wilksova testová štatistika Λ porovnáva, ako blízko je odhad $\boldsymbol{\Sigma}$ s odhadom variančnej matice za platnosti hypotézy H_0. Λ môžeme podľa Rencher [10] (časť 7.4) vyjadriť aj pomocou vlastných čísel matice $\mathbb{K} = \mathbb{S}_Y^{-1}\mathbb{S}_{YX}\mathbb{S}_X^{-1}\mathbb{S}_{XY}$, ktoré označíme ako $\lambda_i \neq 0$, pre $i = 1, \ldots, k$, kde $k = \min(p,q)$ je celkový počet týchto čísel. Potom $\Lambda = \sum_{i=1}^{k}(1 - \lambda_i)$.

Za platnosti H_0 je podľa Rencher [10] Λ rozdelená podľa Wilksovho rozdelenia $\Lambda_{p,q,n-1-q}$ (3.7). Hypotézu H_0 budeme zamietať, ak hodnota testovej štatistiky Λ bude menšia alebo rovná α kvantilu $\Lambda_{p,q,n-1-q}$ rozdelenia, ozn. $\Lambda_{\alpha,p,q,n-1-q}$. V Rencher [10], na str. 566 - 573, môžeme nájsť tabuľky kritických hodnôt Wilksovho rozdelenia $\Lambda_{p,q,n-1-q}$. Pri platnosti H_0 môžeme využiť aj Bartlettovu aproximáciu (3.8) testovej štatistiky Λ a zamietať H_0 pre hodnoty testovej štatistiky väčšie ako $1-\alpha$ kvantil χ^2_{pq} rozdelenia. Navyše pre veľké hodnoty n sú testové štatistiky (3.14) a (3.8) až na multiplikatívnu konštantu totožné.

3.3 Nezávislosť viacerých podvektorov

Nech

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} \quad \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

kde

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i1} \\ \mathbf{Y}_{i2} \\ \vdots \\ \mathbf{Y}_{ip_i} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_k \end{bmatrix} \quad a \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2k} \\ \vdots & \vdots & & \\ \boldsymbol{\Sigma}_{k1} & \boldsymbol{\Sigma}_{k2} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix},$$

pričom $\sum_{i=1}^{k} p_i = p$ a $\mathbf{Y}_i$ reprezentujú rozdelenie vektora $\mathbf{Z}$, nie samostatné výbery z nezávislých vektorov. Predpokladajme, že máme k dispozícii náhodný výber $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ pozostávajúci z p-rozmerných vektorov $\mathbf{Z} \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

V tejto časti práce sa budeme na základe realizácie náhodného výberu zaoberať testovaním hypotézy o nezávislosti viacerých podvektorov mnohorozmerného vektora pri predpoklade mnohorozmerného normálneho rozdelenia. Formálne túto hypotézu môžeme zapísať

$$H_0 : \{\mathbf{Y}_1, \ldots, \mathbf{Y}_k \text{ sú navzájom nezávislé náhodné vektory}\} \tag{3.23}$$

oproti alternatíve $H_1 : \{H_0 \text{ neplatí}\}$. V prípade platnosti hypotézy H_0 platí pre bloky kovariančnej matice $\mathbf{\Sigma}_{ij} = 0$ pre všetky $i \neq j$, čo môžeme vyjadriť aj ako $\mathbf{\Sigma} = \mathbf{\Sigma}_0$, kde

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{kk} \end{bmatrix}, \tag{3.24}$$

pričom $\mathbf{\Sigma}_{ii}$ sú ľubovoľné štvorcové $(p_i \times p_i)$ pozitívne definitné matice $\forall i = 1, \ldots k$.

Využitím vierohodnostnej funkcie môžeme odvodiť rozdelenie testovej štatistiky pre test H_0. Znovu označíme $\ell_n(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$ ako logaritmickú vierohodnostnú funkciu p-rozmerného normálneho rozdelenia so strednou hodnotou $\boldsymbol{\mu}$ a variančnou maticou $\mathbf{\Sigma}$ v maximálne vierohodnom odhade $\hat{\boldsymbol{\mu}}$ a $\hat{\mathbf{\Sigma}}$. Podobne označme $\ell_n(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}_0)$ logaritmickú vierohodnostnú funkciu v maximálne vierohodnom odhade $\hat{\boldsymbol{\mu}}$ a $\hat{\mathbf{\Sigma}}_0$ spočítanú pri platnosti nulovej hypotézy H_0, t. j. variančná matica $\hat{\mathbf{\Sigma}}_0$ má tvar

$$\hat{\mathbf{\Sigma}}_0 = \begin{bmatrix} \hat{\mathbf{\Sigma}}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Sigma}}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\mathbf{\Sigma}}_{kk} \end{bmatrix} \text{ a } \hat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{\mathbf{\Sigma}}_{11} & \hat{\mathbf{\Sigma}}_{12} & \cdots & \hat{\mathbf{\Sigma}}_{1k} \\ \hat{\mathbf{\Sigma}}_{21} & \hat{\mathbf{\Sigma}}_{22} & \cdots & \hat{\mathbf{\Sigma}}_{2k} \\ \vdots & \vdots & & \vdots \\ \hat{\mathbf{\Sigma}}_{k1} & \hat{\mathbf{\Sigma}}_{k2} & \cdots & \hat{\mathbf{\Sigma}}_{kk} \end{bmatrix}, \tag{3.25}$$

pričom na vektor $\boldsymbol{\mu}$ znovu nekladieme žiadne podmienky. Opäť predpokladajme splnenie podmienok regularity. Za platnosti H_0 potom platí analogické tvrdenie k tvrdeniu 7.

Tvrdenie 8. *Testová štatistika*

$$2\log\lambda_n = 2(\ell_n(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) - \ell_n(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}_0)) \tag{3.26}$$

má v prípade platnosti H_0 asymptoticky χ^2_{df} rozdelenie s $df = \frac{1}{2}(p^2 - \sum_{i=1}^k p_i^2)$ stupňami voľnosti.

Počet stupňov voľnosti odvodíme podobne ako (3.15). Počet parametrov v matici $\mathbf{\Sigma}$ je $\frac{1}{2}p(p+1)$. Ak platí H_0, potom je počet parametrov v matici $\mathbf{\Sigma}_0$ rovný súčtu parametrov v blokových maticiach $\mathbf{\Sigma}_{ii}$, t. j. $\frac{1}{2}\sum_{i=1}^k p_i(p_i+1)$. Výsledný počet stupňov voľnosti χ^2 rozdelenia testovej štatistiky (3.26) dostaneme ako rozdiel počtu parametrov v matici $\mathbf{\Sigma}$ a v matici $\mathbf{\Sigma}_0$. Po úprave dostávame $\frac{1}{2}(p^2 - \sum_{i=1}^k p_i^2)$.

Hypotézu H_0 zamietame na základe tvrdenia 8 na hladine spoľahlivosti α, ak $2\log\lambda_n \geq \chi^2_{df}(1-\alpha)$, kde $\chi^2_{df}(1-\alpha)$ je $1-\alpha$ kvantil χ^2_{df} rozdelenia. Testovú štatistiku (3.26) môžeme analogicky (3.16) upraviť do tvaru

$$2\log\lambda_n = -n \log \frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_{11}||\hat{\mathbf{\Sigma}}_{22}| \cdots |\hat{\mathbf{\Sigma}}_{kk}|}. \tag{3.27}$$

Podobne ako v prípade dvoch podvektorov ďalšou z možných testových štatistík (označme U) je podľa Rencher [10] (pokračovanie časti 7.4) testová štatistika definovaná vzťahom

$$U = \frac{|\mathbb{S}|}{|\mathbb{S}_{11}||\mathbb{S}_{22}|\cdots|\mathbb{S}_{kk}|}, \tag{3.28}$$

kde $\mathbb{S}$ je výberová kovariančná matica odhadnutá z náhodného výberu pozostávajúceho z n pozorovaní a je rozdelená podobne ako $\mathbf{\Sigma}$. Rozdiel oproti testovej štatistike založenej na metóde maximálnej vierohodnosti (3.27) je znovu v použití odhadu kovariančnej matice, no pri praktickom výpočte je zlomok v (3.27) opäť totožný s testovou štatistikou U. Znovu poznamenajme, že menovateľ zlomku (3.28) je pri platnosti H_0 rovný determinantu matice $\mathbf{S}$, keďže $\mathbf{\Sigma}_{ij} = 0$ pre všetky $i \neq j$.

Pri platnosti H_0 má testová štatistika U Wilksovo Λ rozdelenie iba v prípade $k = 2$, no pre $k \geq 2$ môžeme podľa Rencher [10] testovú štatistiku U aproximovať pomocou χ^2 rozdelenia využitím transformácie

$$U^* = -(n-1)c \log U, \tag{3.29}$$

kde

$$c = 1 - \frac{1}{12(df)(n-1)}(2a + 3b), \tag{3.30}$$

$$a = p^3 - \sum_{i=1}^{k} p_i^3, \tag{3.31}$$

$$b = p^2 - \sum_{i=1}^{k} p_i^2. \tag{3.32}$$

Počet stupňov voľnosti χ^2 rozdelenia testovej štatistiky U^* je rovnaký ako pre (3.26), t. j. $df = \frac{1}{2}(p^2 - \sum_{i=1}^{k} p_i^2)$. Pri platnosti H_0 teda platí $U^* \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{df}$. Všimnime si, že asymptotické testové štatistiky U^* (3.29) a $2\log\lambda_n$ (3.27) sú asymptoticky rovnako rozdelené a pre veľké hodnoty n prakticky totožné.

Hypotézu H_0 budeme pomocou testovej štatistiky U^* zamietať na hladine spoľahlivosti α, ak $U^* \geq \chi^2_{df}(1-\alpha)$, kde $\chi^2_{df}(1-\alpha)$ je $1-\alpha$ kvantil χ^2_{df} rozdelenia s df stupňami voľnosti.

Pre $p_i = 1 \ \ \forall\ i = 1, \ldots, k$ dostávame špeciálny prípad, v ktorom testujeme hypotézu

$$H_0 : \{\text{zložky } \mathbf{Z} = (Y_1, \ldots, Y_k)^\top \text{ sú navzájom nezávislé náhodné veličiny}\} \tag{3.33}$$

oproti alternatíve $H_1 : \{H_0 \text{ neplatí}\}$ a ak je splnená hypotéza H_0, potom platí $\sigma_{jk} = 0 \ \ \forall\ k \neq j$.

4. Neparametrické testy nezávislosti

Zatiaľ sme sa venovali testom nezávislosti dvoch náhodných veličín, či už kategoriálnych alebo nominálnych, alebo testom nezávislosti dvoch a viacerých podvektorov pri predpoklade normality. V tejto časti práce sa sústredíme na rozšírenie tejto problematiky, a to na testy nezávislosti dvoch viacrozmerných náhodných vektorov bez predpokladu na ich rozdelenie, ktoré sú inšpirované prácou Arthur a Gyorfi [3]. Uvažujme preto dve prirodzené čísla p a q, ktoré budú reprezentovať rozmer náhodných vektorov $\mathbf{X}$, resp. $\mathbf{Y}$. Nech $(p+q)$-rozmerné rovnako rozdelené a nezávislé (*i.i.d.*) náhodné vektory $(\mathbf{X}_1^\top,\mathbf{Y}_1^\top)^\top, \ldots, (\mathbf{X}_n^\top,\mathbf{Y}_n^\top)^\top$ na $\mathbb{R}^p \times \mathbb{R}^q$ tvoria náhodný výber dvojíc náhodných vektorov definovaných na rovnakom pravdepodobnostnom priestore. Pre takto definovaný pravdepodobnostný model označíme pravdepodobnostné rozdelenie náhodného vektora $(\mathbf{X}^\top,\mathbf{Y}^\top)^\top$ ako ν, kým rozdelenia náhodných vektorov $\mathbf{X}$, resp. $\mathbf{Y}$, budeme označovať ako μ_1, resp. μ_2.

Ako sme už viackrát spomenuli, aj v tejto časti nás bude zaujímať test hypotézy $H_0 : \{\mathbf{X}$ a $\mathbf{Y}$ sú nezávislé$\}$, čo môžeme zapísať pomocou pravdepodobnostných rozdelení ako

$$H_0 : \{\nu = \mu_1 \times \mu_2\}, \tag{4.1}$$

oproti alternatíve $H_1 : \{\mathbf{X}$ a $\mathbf{Y}$ nie sú nezávislé$\}$, pričom nebudeme špecifikovať konkrétnu združenú či marginálnu distribúciu náhodných vektorov. Uvažujme ďalej jednoznačný a konečný rozklad $P_n = \{A_{n,1}, \ldots, A_{n,m_{1,n}}\}$ priestoru $\mathbb{R}^p$, kde $A_{n,i}$ a $A_{n,j}$ sú po dvoch disjunktné množiny pre $i \neq j$ a zjednotenie všetkých $A_{n,i}$ tvorí $\mathbb{R}^p$. Podobne definujme jednoznačný a konečný rozklad $Q_n = \{B_{n,1}, \ldots, B_{n,m_{2,n}}\}$ priestoru $\mathbb{R}^q$.

Nasledujúce testy založíme na vyššie uvedenom rozklade P_n a Q_n daného pravdepodobnostného priestoru. Následným vyhodnotením vhodnej testovej štatistiky na zvolených diskrétnych podpriestoroch uskutočníme štatistické rozhodnutie. Základom tejto metódy bude otázka: Ako empiricky odhadnúť združené rozdelenie náhodného vektora $(\mathbf{X}^\top,\mathbf{Y}^\top)^\top$ a jeho marginálne rozdelenia vektorov $\mathbf{X}$, resp. $\mathbf{Y}$. Z tohto dôvodu označme ν_n, $\mu_{n,1}$ a $\mu_{n,2}$ ako empirické odhady mier ν, μ_1, resp. μ_2, založené na náhodnom výbere $\{(\mathbf{X}_1^\top,\mathbf{Y}_1^\top)^\top, \ldots, (\mathbf{X}_n^\top,\mathbf{Y}_n^\top)^\top\}$, ktoré pre ľubovoľné dve Borelovské podmnožiny $A \subset \mathbb{R}^p$ a $B \subset \mathbb{R}^q$ definujeme

$$\begin{aligned}
\nu_n(A \times B) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{(\mathbf{X}_i^\top,\mathbf{Y}_i^\top)^\top \in A\times B\}}, \\
\mu_{n,1}(A) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in A\}}, \\
\mu_{n,2}(B) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{Y}_i \in B\}}.
\end{aligned} \tag{4.2}$$

4.1 Test založený na testovej štatistike L

Pre test hypotézy H_0 definujme prvú testovú štatistiku L_n podľa Arthur a Gyorfi [3] v tvare

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) = \sum_{A\in P_n}\sum_{B\in Q_n}|\nu_n(A\times B)-\mu_{n,1}(A)\cdot\mu_{n,2}(B)|. \tag{4.3}$$

Testová štatistika L_n je jedným zo spôsobov, ktorými môžeme navzájom porovnávať miery ν a $\mu_1\times\mu_2$, ktoré sú pri platnosti H_0 totožné. V Arthur a Gyorfi [3] je uvedené nasledujúce trdenie o rozdelení testovej štatistiky L_n, na základe ktorého je odvodený asymptotický test H_0 na hladine spoľahlivosti α. Nasledujúce a aj ostatné tvrdenia v tejto kapitole uvedieme bez dôkazov. Tie môžeme nájsť práve v Arthur a Gyorfi [3].

Tvrdenie 9. *Nech sú splnené podmienky*

$$\lim_{n\to\infty}\frac{m_{1,n}\cdot m_{2,n}}{n}=0, \tag{4.4}$$

$$\lim_{n\to\infty}\max_{A\in P_n}\mu_1(A)=0,\quad \lim_{n\to\infty}\max_{B\in Q_n}\mu_2(B)=0. \tag{4.5}$$

Potom za platnosti hypotézy H_0 existuje centrovacia postupnosť $(C_n)_{n\geq 1}$, ktorá závisí na pravdepodobnostnom rozdelení ν tak, že

$$\frac{\sqrt{n}}{\sigma}(L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-C_n)\xrightarrow[n\to\infty]{\mathcal{D}}N(0,1), \tag{4.6}$$

kde $\sigma = 1-\frac{2}{\pi}$.

V prípade platnosti H_0 udáva tvrdenie 9 asymptotické rozdelenie testovej štatistiky (4.3) pre test nezávislosti dvoch viacrozmerných vektorov, ktorý zamieta H_0, ak $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ nadobúda veľké hodnoty. Uveďme podľa Arthur a Gyorfi [3] spôsob, ako zamietať hypotézu H_0 na základe tvrdenia 9. Predpokladajme platnosť H_0 a podmienok v tvrdení 9. Potom z dôkazu tvrdenia 9 v Arthur a Gyorfi [3] vyplýva, že centrovaciu postupnosť C_n môžeme zhora ohraničiť konštantou $\sqrt{2/\pi}\sqrt{\frac{m_{1,n}\cdot m_{2,n}}{n}}$. Preto hypotézu H_0 zamietame na asymptotickej hladine α, ak

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) > \sqrt{2/\pi}\sqrt{\frac{m_{1,n}\cdot m_{2,n}}{n}}+\frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha), \tag{4.7}$$

kde $\sigma = 1-2/\pi$ a Φ^{-1} je kvantilová funkcia štandardného normálneho rozdelenia.

4.2 Test založený na Log-likelihood štatistike

Hypotézu H_0 môžeme testovať aj ďalšou testovou štatistikou navrhnutou v Arthur a Gyorfi [3]. Definujme podľa Arthur a Gyorfi [3] príslušnú logaritmicko-vierohodnostnú[1] testovú štatistiku ako

$$I_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) = \sum_{A\in P_n}\sum_{B\in Q_n}\nu_n(A\times B)\log\frac{\nu_n(A\times B)}{\mu_{n,1}(A)\cdot\mu_{n,2}(B)}. \tag{4.8}$$

[1] angl. Log-likelihood

Log-likelihood testová štatistika je založená na rozšírení I-divergentnej[2] testovej štatistiky (pozri Arthur a Gyorfi [3]). Nasledujúce tvrdenie udáva možnosť, ako zamietať H_0.

Tvrdenie 10. *Nech sú splnené podmienky* (4.4) *a* (4.5). *Potom pri platnosti* H_0 *platí pre ľubovoľné reálne* x

$$P\{\frac{2nI_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-m_{1,n}m_{2,n}}{\sqrt{2m_{1,n}m_{2,n}}}\geq x\}\xrightarrow[n\to\infty]{}1-\Phi(x). \tag{4.9}$$

Na základe tvrdenia 10 budeme zamietať H_0 v prípade, že

$$\frac{2nI_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-m_{1,n}m_{2,n}}{\sqrt{2m_{1,n}m_{2,n}}}\geq\Phi^{-1}(1-\alpha), \tag{4.10}$$

alebo ekvivalentne, ak

$$I_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\geq\frac{\Phi^{-1}(1-\alpha)\sqrt{2m_{1,n}m_{2,n}}+m_{1,n}m_{2,n}}{2n}. \tag{4.11}$$

4.3 Test založený na Pearsonovej štatistike

Ďalšou možnou testovou štatistikou pre test H_0, ktorá je uvedená v Arthur a Gyorfi [3], je Pearsonova χ^2_n testová štatistika v tvare

$$\chi^2_n(\nu_n,\mu_{n,1}\times\mu_{n,2})=\sum_{A\in P_n}\sum_{B\in Q_n}\frac{(\nu_n(A\times B)-\mu_{n,1}(A)\cdot\mu_{n,2}(B))^2}{\mu_{n,1}(A)\cdot\mu_{n,2}(B)}. \tag{4.12}$$

V Arthur a Gyorfi [3] sa domnievajú, že pre nasledujúcu transformáciu testovej štatistiky χ^2_n platí pri platnosti H_0 a podmienok (4.4) a (4.5) nasledujúce

$$\frac{1}{\sqrt{2m_{1,n}\cdot m_{2,n}}}(n\cdot\chi^2_n(\nu_n,\mu_{n,1}\times\mu_{n,2}))-m_{1,n}\cdot m_{2,n})\xrightarrow[n\to\infty]{\mathcal{D}}N(0,1). \tag{4.13}$$

Na základe tejto domnienky môžeme zamietať hypotézu H_0 na asymptotickej hladine α, ak

$$\chi^2_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\geq\frac{\Phi^{-1}(1-\alpha)\sqrt{2m_{1,n}\cdot m_{2,n}}+m_{1,n}\cdot m_{2,n}}{n}. \tag{4.14}$$

Poznamenajme, že rozdelenia uvedených testových štatistík sú citlivé na voľbu P_n a Q_n, a teda aj $m_{n,1}$, resp. $m_{n,2}$. Pri neopatrnej voľbe rozdelenia pravdepodobnostného priestoru nemusia byť podmienky (4.4) a (4.5) splnené, a preto môžu byť prezentované asymptotické testy nepresné.

V nasledujúcej kapitole uvedieme spôsob, ako využiť testové štatistiky uvedené v tejto kapitole v kontexte permutačných testov. Na rozdiel od asymptotických neparametrických testov sú permutačné testy presné testy a ich použitím môžeme posúdiť významnosť napozorovanej testovej štatistiky. Navyše permutačné testy poskytujú spoľahlivý spôsob posúdenia významnosti χ^2_n, zatiaľ čo domnienka o rozdelení χ^2_n nebola dokázaná.

[2] angl. Kullback-Leibler divergence

5. Permutačné testy

Táto kapitola je venovaná základnému princípu permutačných testov a ich využitiu na testovanie nezávislosti. Podrobné teoretické spracovanie všeobecných permutačných testov a ich použitie nielen v kontexte testov nezávislosti je dostupné v Pesarin a Salmaso [14]. Základnou myšlienkou permutačných testov je ich nezávislosť od pravdepodobnostného rozdelenia, z ktorého náhodný výber pochádza.

5.1 Definícia permutačných testov

Uvažujme náhodné veličiny X_1 a X_2 definované na pravdepodobnostnom priestore $(\mathcal{X},\mathcal{F},P_1)$, resp. $(\mathcal{X},\mathcal{F},P_2)$, kde P_1 a P_2 patria do rodiny neparametrických distribúcii $\mathcal{P}$. Nech $\mathbf{X}_j = (X_{j1},\ldots,X_{jn})$ sú náhodné výbery veličín X_j na pravdepodobnostnom priestore $(\mathcal{X},\mathcal{F},P_j)$ pre $j = 1,2$ a P je združené rozdelenie vektora $(X_1,X_2)^\top$. V Pesarin a Salmaso [14] sa zaoberajú testom všeobecnej hypotézy $H_0^* : \{P_1 = P_2\}$ oproti jednostrannej alternatíve, že jedno z rozdelení je stochasticky dominantnejšie, alebo oproti alternatíve $H_1^* : \{P_1 \neq P_2\}$. V rámci tejto práce nás bude v kontexte permutačných testov zaujímať test hypotézy

$$H_0 : \{P = P_1 \times P_2\} \tag{5.1}$$

oproti alternatíve

$$H_1 : \{P \neq P_1 \times P_2\}. \tag{5.2}$$

Bez vplyvu na všeobecnosť môžeme H_0 a H_1 rozšíriť aj na dve viacrozmerné rozdelenia. Konkrétny spôsob uvedieme na konci podkapitoly.

Označme $\mathbf{X} = ((X_{11},X_{21})^\top,\ldots,(X_{1n},X_{2n})^\top)$ náhodný výber zo združeného rozdelenia $(X_1,X_2)^\top$ a $\mathbf{X}_i = (X_{1i},X_{2i})^\top$ pre $i = 1,\ldots,n$. Definujme podľa Pesarin a Salmaso [14], kapitola 2, podmienený[1] priestor hodnôt $\mathcal{X}_{|\mathbf{X}}$ založený na $\mathbf{X}$ pri platnosti H_0, ako množinu všetkých $\mathbf{X}^*$ z $\mathcal{X}\times\mathcal{X}$, ktoré sú pravdepodobnostne ekvivalentné k $\mathbf{X}$, t. j. pravdepodobnostný pomer $\frac{dP(\mathbf{X}_i)}{dP(\mathbf{X}_i^*)}$ nezávisí na rozdelení P pre všetky $i = 1,\ldots,n$. Takto definovaný $\mathcal{X}_{|\mathbf{X}}$ obsahuje podľa Pesarin a Salmaso [14] množinu všetkých permutácií prvkov $\mathbf{X}$. Permutáciu prvkov $\mathbf{X}$ vysvetlíme v nasledujúcej podkapitole.

Potom môžeme podľa Pesarin a Salmaso [14] podmienenú pravdepodobnosť náhodného javu $A \in \mathcal{F}$ pri danej množine permutácií $\mathcal{X}_{|\mathbf{X}}$ vyjadriť nezávisle na rozdelení P, t. j.

$$\mathsf{P}\left[A \in \mathcal{F}; P|\mathcal{X}_{|\mathbf{X}}\right] = \mathsf{P}\left[A \in \mathcal{F}|\mathcal{X}_{|\mathbf{X}}\right]. \tag{5.3}$$

Z toho vyplýva, že permutačné rozdelenie ľubovoľne vybranej testovej štatistiky $T : \mathcal{X}_{|\mathbf{X}} \to \mathbb{R}^1$ definované $P_T(t|\mathcal{X}_{|\mathbf{X}}) = \mathsf{P}\left[T \leq t|\mathcal{X}_{|\mathbf{X}}\right]$ je P-invariantné. Navyše pre každé konečné výbery je počet prvkov, ozn. $M^{(n)}$, priestora $\mathcal{X}_{|\mathbf{X}}$ konečný a ak sa v $\mathcal{X}_{|\mathbf{X}}$ nenachádzajú viacnásobné prvky, potom permutácie $\mathbf{X}^*$ sú podmienene $\mathcal{X}_{|\mathbf{X}}$ rovnako pravdepodobné s pravdepodobnosťou $\frac{1}{M^{(n)}}$ ak $\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}$,

[1] v Pesarin a Salmaso [14] angl. conditional reference space

inak je pravdepodobnosť rovná 0. Následne môžeme definovať a počítať pravdepodobnosť (5.3) ako

$$\mathsf{P}\left[\mathbf{X}^* \in \mathcal{F}|\mathcal{X}_{|\mathbf{X}}\right] = \frac{\sum_{\mathbf{X}^* \in \mathcal{F}} dP(\mathbf{X}^*)}{\sum_{\mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}} dP(\mathbf{X}^*)} = \frac{\sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}_{(\mathbf{X}^* \in \mathcal{F})}}{M^{(n)}}. \tag{5.4}$$

Predpokladajme, že $T : \mathcal{X}_{|\mathbf{X}} \to \mathbb{R}^1$ je vhodná testová štatistika, pre ktorú môžeme bez vplyvu na všeobecnosť predpokladať, že odľahlé hodnoty napozorované pre všetky permutácie $\mathbf{X}^*$ svedčia proti hypotéze H_0. Pre príklady a vlastnosti vhodnej testovej štatistiky pre všeobecné permutačné testy pozri napr. Pesarin a Salmaso [14], kapitola 2.5.

Definujme obor všetkých hodnôt permutačnej testovej štatistiky ako množinu $\mathcal{T}_{\mathbf{X}} = \{T^* = T(\mathbf{X}^*); \mathbf{X}^* \in \mathcal{X}_{|\mathbf{X}}\}$, ktorá obsahuje všetky možné hodnoty testovej štatistiky T pri hodnotách $\mathbf{X}^*$ z $\mathcal{X}_{|\mathbf{X}}$. Usporiadajme $M^{(n)}$ členov množiny $\mathcal{T}_{\mathbf{X}}$ do neklesajúceho radu tak, že $|T^*|_{(1)} \leq |T^*|_{(2)} \leq \cdots \leq |T^*|_{(M^{(n)})}$. Definujme pre každé $\alpha \in (0{,}1)$ permutačnú kritickú hodnotu $T_\alpha(\mathbf{X}^*) = T_\alpha = |T^*|_{(M_\alpha^{(n)})}$, kde $M_\alpha^{(n)} = \sum_{\mathcal{X}_{|\mathbf{X}}} \mathbb{I}_{\{|T(\mathbf{X}^*)| < T_\alpha\}}$ je počet permutačných testových štatistík, ktoré sú v absolútnej hodnote menšie ako hodnota T_α. Ako pri každom štatistickom teste, tak aj pre permutačný test uvedieme definíciu p-hodnoty. Pre $T : \mathcal{X}_{|\mathbf{X}} \to \mathbb{R}^1$, kedy značne odľahlé hodnoty napozorovanej testovej štatistiky hovoria proti platnosti H_0, definujeme p-hodnotu podobne ako v Pesarin a Salmaso [14]

$$p = p_T(\mathbf{X}) = L_T(T^0|\mathcal{X}_{|\mathbf{X}}) = \mathsf{P}\left[(|T| \geq |T^0|)|\mathcal{X}_{|\mathbf{X}}\right], \tag{5.5}$$

kde T^0 je napozorovaná hodnota testovej štatistiky T a L_T je tzv. funkcia prežitia[2] definovaná ako $L_T(t|\mathcal{X}_{|\mathbf{X}}) = \mathsf{P}\left[(|T| \geq |t|)|\mathcal{X}_{|\mathbf{X}}\right]$. Pre malé hodnoty $M^{(n)}$ (Pesarin a Salmaso [14] uvádzajú n do 25) môžeme $L_T(t|\mathcal{X}_{|\mathbf{X}})$ a jej doplnok do 1 (distribučnú funkciu testovej štatistiky, ozn. $F_T(t|\mathcal{X}_{|\mathbf{X}})$), určiť presne, a to vyhodnotením testových štatistík pre všetky prvky $\mathcal{X}_{|\mathbf{X}}$. V prípade veľkého $M^{(n)}$ ich môžeme odhadnúť pomocou podmienej metódy Monte Carlo, ktorej je venovaná nasledujúca podkapitola. Ak budeme uvažovať testovú štatistiku $T : \mathcal{X}_{|\mathbf{X}} \to \mathbb{R}_0^+$, potom sú absolútne hodnoty v uvádzaných výrazoch nepodstatné.

Ukázali sme ďalšiu možnosť, ako testovať nezávislosť dvoch náhodných veličín. Zamerajme sa na spôsob, ako odvodiť permutačný test nezávislosti pre dvojicu viacrozmerných vektorov. Predpokladajme, že máme k dispozícii realizáciu náhodného výberu $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, ktorá vznikne z náhodného vektora $\mathbf{X} = (X_1, \ldots, X_q)^\top$ pre $q \geq 2$. Nech $\mathbf{W} = (X_1, \ldots, X_p)^\top$ a $\mathbf{Z} = (X_{p+1}, \ldots, X_q)^\top$ pri $p \geq 1$. Zaujíma nás test hypotézy

$$H_0 : \{\text{podvektory } \mathbf{W} \text{ a } \mathbf{Z} \text{ sú nezávislé náhodné vektory}\} \tag{5.6}$$

oproti alternatíve

$$H_1 : \{\text{podvektory } \mathbf{W} \text{ a } \mathbf{Z} \text{ nie sú nezávislé náhodné vektory}\}, \tag{5.7}$$

pričom na distribúciu vektora $\mathbf{X}$ nekladieme žiadne požiadavky.

Ak permutačný test založíme na testovej štatistike $T : \mathbb{R}^q \to \mathbb{R}^1$, potom z hľadiska teórie a výpočtovej stránky prechádzame opäť k problému, ktorý je

[2] significance level (survival) function

ekvivaletný vyššie uvedenému permutačnému testu. Tento pricíp využijeme aj pri odvodení konkrétnych testov nezávislosti v nasledujúcej časti tejto práce.

Ak máme k dispozícii transformáciu dát $\psi : \mathbb{R}^q \to \mathbb{R}^2$, ktorá prevedie viacrozmerný problém na dvojrozmerný, môžeme opäť využiť tento permutačný test nezávislosti pre dve náhodné veličiny. Pre ďalšiu teóriu a aplikáciu viacrozmerných permutačných testov pozri Pesarin a Salmaso [14].

5.2 Podmienená Monte Carlo metóda (CMC)

Názov tejto metódy je preložený z Pesarin a Salmaso [14] z angl. *Conditional Monte Carlo (CMC)*. Termín podmienené Monte Carlo (CMC) sa používa na zdôraznenie, že ide o bežnú simuláciu Monte Carlo, ktorá je uskutočnená na náhodne zvolenej časti $\mathcal{X}_{|\mathbb{X}}$, množiny všetkých permutácií prvkov $\mathbb{X}$, kde budeme pokračovať opäť v značení náhodného výberu $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, ktorý vznikne z náhodného vektora $\mathbf{X} = (X_1, \ldots, X_q)^\top$ pre $q \geq 2$, kde $\mathbf{W} = (X_1, \ldots, X_p)^\top$ a $\mathbf{Z} = (X_{p+1}, \ldots, X_q)^\top$ pri $p \geq 1$.

Poznamenajme, že pôvodný náhodný výber pozorovaní $\mathbb{X}$ zostáva nemenný. Ak je množina $\mathcal{X}_{|\mathbb{X}}$ veľmi veľká do počtu všetkých prvkov $M^{(n)}$, potom môžeme podľa Pesarin a Salmaso [14] študovať množinu všetkých permutácií prvkov $\mathbb{X}$ na základe dostatočne veľkého množstva náhodne vybraných prvkov množiny $\mathcal{X}_{|\mathbb{X}}$. Priestor všetkých permutácií $\mathcal{X}_{|\mathbb{X}}$ je definovaný ako

$$\mathcal{X}_{|\mathbb{X}} = \{\mathbb{X}^* = \{(\mathbf{Z}_1, \mathbf{W}_{u_1^*}), \ldots, (\mathbf{Z}_n, \mathbf{W}_{u_n^*})\};\ \forall \mathbf{u}^* = (u_1^*, \ldots u_n^*)\}, \tag{5.8}$$

kde $\mathbf{u}^*$ je ľubovoľná permutácia prvkov množiny $\{1, \ldots, n\}$. Z toho vyplýva, že kardinalita tejto množiny je rovná počtu všetkých permutácií na n-prvkovej množine, t. j. $M^{(n)} = n!$. Poznamenajme, že pri platnosti H_0 je možné každému náhodnému vektoru $\mathbf{Z}_i$, priradiť s rovnakou pravdepodobnosťou ľubovoľný vektor $\mathbf{W}_j$ pre všetky $i,j = 1, \ldots, n$.

Túto myšlienku realizujeme pomocou opakovaného výpočtu testovej štatistiky, ktorá je spočítaná pri ľubovoľnej permutácii pozorovanej množiny dát. Opäť pripomeňme, že tento proces prebieha bez realizácie novej štúdie, v zmysle: raz napozorujeme, viackrát testujeme. Postup CMC sa podľa Pesarin a Salmaso [14] skladá z nasledujúcich krokov:

(a) Spočítame pozorovanú hodnotu T^0 testovej štatistiky T : $T^0 = T(\mathbb{X})$ na pôvodnej realizácii náhodného výberu $\mathbb{X}$.

(b) Náhodne permutujeme množinu dát $\mathbb{X}$. Permutáciu označíme ako $\mathbb{X}^* = (\mathbf{X}_{u_1^*}, \ldots, \mathbf{X}_{u_n^*})$, kde $\mathrm{u}^* = (u_1^*, \ldots, u_n^*)$ je náhodná permutácia množiny $\{1, \ldots, n\}$. Napríklad pre náhodné veličiny X_1 a X_2 dostaneme permutáciu $\mathbb{X}^*$ permutovaním druhej zložky vektorov $\mathbf{X}_i = (X_{1i}, X_{2i})^\top$, t. j. $\mathbb{X}^* = ((X_{11}, X_{2u_1^*})^\top, \ldots, (X_{1n}, X_{2u_n^*})^\top)$. Pre dva náhodné vektory realizujeme permutáciu $\mathbb{X}^*$ permutovaním druhej zložky vektorov $\mathbf{X}_i = (\mathbf{Z}_i^\top, \mathbf{W}_i^\top)^\top$ pre $i = 1, \ldots, n$. Výsledná permutácia je potom zložená z vektorov $\mathbf{X}_i^*$, kde prvý podvektor $\mathbf{Z}_i$ zostáva fixný a náhodne k nemu priradíme podvektor $\mathbf{W}_{u_i^*}$, t. j. $\mathbb{X}^* = ((\mathbf{Z}_1^\top, \mathbf{W}_{u_1^*}^\top)^\top, \ldots, (\mathbf{Z}_n^\top, \mathbf{W}_{u_n^*}^\top)^\top)$.

(c) Spočítame testovú štatistiku $T^* = T(\mathbb{X}^*)$ pri ľubovoľnej permutácii $\mathbb{X}^*$.

(d) Nezávisle opakujeme krok (b) a (c) K-krát, pričom každá permutácia sa objaví práve raz. Množina obsahujúca K rôznych permutácií $\mathbb{X}^*$ je náhodný výber z množiny všetkých permutácií $\mathcal{X}_{|\mathbb{X}}$, na ktorom založíme štatistickú analýzu.

Vo výsledku dostávame K príslušných hodnôt T^*, ktoré reprezentujú časť permutačného rozdelenia testovej štatistiky T. Na základe tejto množiny pozorovaní môžeme štatisticky odhadnúť permutačnú distribučnú funkciu $F_T(t|\mathcal{X}_{|\mathbb{X}})$ testovej štatistiky T a funkciu prežitia $L_T(t|\mathcal{X}_{|\mathbb{X}})$. Empirickú distribučnú funkciu testovej štatistiky T dostaneme podľa Pesarin a Salmaso [14] ako

$$\hat{F}_T^*(t) = \sum_{k=1}^{K} \frac{\mathbb{I}_{(|T_k^*| \leq |t|)}}{K} \tag{5.9}$$

a empirickú funkciu prežitia

$$\hat{L}_T^*(t) = \sum_{k=1}^{K} \frac{\mathbb{I}_{(|T_k^*| \geq |t|)}}{K}. \tag{5.10}$$

S rastúcim počtom náhodných permutácií $K \to n!$, resp. iterácií CMC, konvergujú odhady $\hat{L}_T^*(\cdot)$ a $\hat{F}_T^*(\cdot)$ s istotou[3] ku skutočným $L_T(\cdot|\mathcal{X}_{|\mathbb{X}})$ a $F_T(\cdot|\mathcal{X}_{|\mathbb{X}})$. Preto k vyhodnoteniu zhody pozorovaných dát s nulovou hypotézou H_0 môžeme pri dostatočnom množstve permutácií využiť odhady $\hat{L}_T^*(\cdot)$ a $\hat{F}_T^*(\cdot)$. V praxi je odhad p-hodnoty v prípade veľkých hodnôt svedčiacich proti platnosti H_0 daný vzťahom

$$\hat{p}(\mathbb{X}) = \hat{p} = \hat{L}_K^*(T^0) = \sum_{k=1}^{K} \frac{\mathbb{I}_{(|T_k^*| \geq |T^0|)}}{K}. \tag{5.11}$$

Nech α je predom stanovená hladina testu. Ak $\hat{p} \leq \alpha$, potom môžeme konštatovať, že empirické pozorovania svedčia proti platnosti H_0, ktorú týmto zamietame.

5.3 Príklady permutačných testov

Nech $\mathbb{X} = \{(\mathbf{Z}_1^\top, \mathbf{W}_1^\top)^\top, \ldots, (\mathbf{Z}_n^\top, \mathbf{W}_n^\top)^\top\}$ je náhodný výber z q-rozmerného rozdelenia náhodného vektora $\mathbf{X} = (\mathbf{Z}^\top, \mathbf{W}^\top)^\top$ s neznámym rozdelením so združenou distribučnou funkciou F na $\mathbb{R}^q$. Teraz využijeme vyššie formulovanú teóriu permutačných testov a budeme testovať nezávislosť dvoch náhodných vektorov $\mathbf{Z}_{(p_1 \times 1)}$ a $\mathbf{W}_{(p_2 \times 1)}$, pre $p_1 + p_2 = q$. Nulovú hypotézu H_0 tohto problému môžeme formulovať ako

$$H_0 : \{F(\mathbf{z},\mathbf{w}) = F_1(\mathbf{z}) \cdot F_2(\mathbf{w});\ \forall (\mathbf{z}^\top, \mathbf{w}^\top)^\top \in \mathbb{R}^q\}, \tag{5.12}$$

kde F_1 je marginálna distribučná funkcia náhodného vektora $\mathbf{Z}$ a F_2 je marginálna distribučná funkcia náhodného vektora $\mathbf{W}$. Alternatívu budeme formulovať ako $H_1 : \{H_0 \text{ neplatí}\}$, teda náhodné vektory $\mathbf{Z}$ a $\mathbf{W}$ nie sú nezávislé.

[3] česky skoro jiste

Uvažujme $p_1 = p_2 = 1$. Potom na skúmanie lineárneho vzťahu medzi náhodnými veličinami Z a W môžeme použiť testovú štatistiku $T = \sum_{i=1}^{n} Z_i W_i$, ktorej permutačné rozdelenie dostaneme podmienenou Monte Carlo metódou založenou na permutačnej testovej štatistike $T^* = \sum_{i=1}^{n} Z_i W_{u_n^*}$, kde $\mathrm{u}^* = (u_1^*, \ldots, u_n^*)$ je náhodná permutácia množiny $\{1, \ldots, n\}$. Všimnime si podobnosť T s výberovým korelačným koeficientom (2.15), kde druhý zlomok vo výraze (2.15) závisí na zvolenej permutácii u^* iba výrazom $\sum_{i=1}^{n} Z_i W_i$, ktorý je zhodný s testovou štatistikou T.

Týmto spôsobom je možné testovať hypotézu H_0 pre náhodné veličiny Z a W alternatívne k výberovému korelačnému koeficientu. Navyše v predpokladoch tvrdenia 5 je predpoklad normálne rozdeleného náhodného výberu, ktorý pri permutačnom teste nepotrebujeme. Výhodou permutačného testu teda je, že nevyžaduje predpoklad na rozdelenie náhodného výberu. Permutačné testy môžeme takisto použiť aj pri voľbe testových štatistík vo forme poradových korelačných koeficientov (2.6), resp. (2.14).

Jednou z ďalších výhod permutačných testov oproti asymptotickým testom uvedeným v druhej kapitole je, že nevyžadujú veľké množstvo pozorovaní pre splnenie asymptotického rozdelenia testovej štatistiky. Zatiaľ čo asymptotické testy korelačných koeficientov majú asymptotickú hladinu spoľahlivosti, permutačné testy sú presné testy, a preto pri dostatočnom množstve permutácií majú presnú, nie asymptotickú, hladinu spoľahlivosti α.

V Pesarin a Salmaso [14](kapitola 2.6, príklad 2) uvádzajú ďalšiu testovú štatistiku v tvare $T = \phi(P,P_1,P_2)$, kde ϕ je ľubovoľná metrika na neparametrickej rodine distribúcií $\mathcal{P}$, ktorá udáva vzdialenosť medzi pravdepodobnostnými distribúciami P a $P_1 \times P_2$. Takisto uvádzajú, že takto volená T je vhodnejšia permutačná testová štatistika k testu H_0 než testové štatistiky založené na korelačných koeficientoch. Ako príklad je uvedená testová štatistika v tvare

$$T = \sqrt{n} \cdot \sup_{A=A_1\times A_2\in\mathcal{B}} |\hat{P}(A) - \hat{P}_1(A_1) \times \hat{P}_2(A_2)|, \tag{5.13}$$

kde $\mathcal{B}$ je vhodná neprázdna množina javov, $\hat{P}(\cdot)$, $\hat{P}_1(\cdot)$ a $\hat{P}_2(\cdot)$ sú vhodné výberové odhady pravdepodobnostných distribúcií P, P_1 a P_2 (pozri (4.2)).

Ako ďalšie možnosti testových štatistík, nielen pre dvojrozmerné rozdelenie P, môžeme využiť príklady testových štatistík z predchádzajúcej kapitoly (4.3), (4.8) alebo (4.12), pretože svedčia veľkými hodnotami proti platnosti H_0.

Podobne ako testy v štvrtej kapitole, aj permutačné testy sú nezávislé na pravdepodobnostnom rozdelení náhodného výberu vektorov, no väčšina záverov v Arthur a Gyorfi [3] je podobne ako testy korelačných koeficientov založená na asymptotikých vlastnostiach testových štatistík. Z toho dôvodu môžeme permutačné testy považovať za spresnenie asymptotických testov. Navyše pre testovú štatistiku (4.12) v Arthur a Gyorfi [3] iba predpokladajú pri splnení uvedených podmienok asymptoticky normálne rozdelenie.

Pomocou permutačných testov však môžeme určiť významnosť testových štatistík a na základe testovej štatistiky (4.12) môžeme realizovať presný test H_0 aj bez exaktného dôkazu asymptotického rozdelenia testovej štatistiky (4.12) pri splnení podmienok tvrdenia 10.

6. Mnohorozmerné testy nezávislosti

V tejto kapitole rozvinieme myšlienku z predchádzajúcich častí a budeme prezentovať postupy a možnosti ako testovať hypotézu

$$H_0 : \{\text{náhodné vektory } \mathbf{X} \text{ a } \mathbf{Y} \text{ sú nezávislé}\} \tag{6.1}$$

voči alternatíve

$$H_1 : \{\text{náhodné vektory } \mathbf{X} \text{ a } \mathbf{Y} \text{ nie sú nezávislé}\} \tag{6.2}$$

pomocou permutačných testov. Testy založíme ako obvykle na dvoch náhodných výberoch náhodných vektorov $\mathbf{X}$ a $\mathbf{Y}$, medzi ktorými chceme testovať H_0.

Prezentujeme test s názvom Protest, ktorý založeníme na skutočných hodnotách realizácie náhodných výberov a testy založené na maticiach vzdialeností alebo podobností, Mantlov test a DCOV test. Na začiatku kapitoly uvedieme základné charakteristiky koeficientov podobností a vzdialeností, a potom budeme formulovať konkrétne testy a ich predpoklady.

6.1 Koeficienty podobnosti S

Koeficienty podobnosti, vo všeobecnosti označované písmenom S alebo v krátkosti podobnosti, sa využívajú na meranie vzťahu medzi dvojma objektmi. Na rozdiel od väčšiny koeficientov vzdialeností, miery podobnosti nemajú nikdy vlastnosti metrík (pozri koeficienty vzdialeností), pretože je vždy možné nájsť dva objekty A a B, ktoré sú viac podobné než súčet ich podobností s iným, vzdialenejším objektom C.

Z toho vyplýva, že podobnosti nemožno použiť priamo na popis umiestnenia objektov v metrických priestoroch, preto musia byť transformované na vzdialenosti. Príklady koeficientov podobností môžeme nájsť v Legendre a Legendre [5], v kapitole 7.3. Delia sa na symetrické a nesymetrické a všetky koeficienty podobností, ktoré majú obor hodnôt v intervale od 0 do 1, môžu byť transformované na koeficienty vzdialeností nasledujúcimi spôsobmi:

$$D = 1 - S, \qquad D = \sqrt{1 - S}, \qquad D = \sqrt{1 - S^2}. \tag{6.3}$$

6.2 Koeficienty vzdialenosti D

Koeficienty vzdialeností sú funkcie, ktoré nadobúdajú svoju maximálnu hodnotu (často 1) pre dva objekty, ktoré sú úplne odlišné alebo hodnotu 0 medzi dvoma objektmi, ktoré sú identické vo všetkých zložkách. Koeficienty vzdialeností môžeme podľa Legendre a Legendre [5], kapitoly 7.4, rozdeliť do troch skupín:

(i) Prvá skupina sa skladá z metrík, ktoré majú spoločné nasledujúce štyri vlastnosti:

1.) minimum je rovné 0: ak $a = b$, potom $D(a, b) = 0$;

2.) pozitivita: ak $a \neq b$, potom $D(a, b) > 0$;

3.) symetria: $D(a, b) = D(b, a)$;

4.) trojuholníková nerovnosť: $D(a, b) + D(b, c) \geq D(a, c)$.

Pre ilustráciu uveďme zopár príkladov. Nech $\mathbf{x} = (x_1, \ldots, x_p)^\top$ a $\mathbf{y} = (y_1, \ldots, y_p)^\top$ sú dva body patriace do p-rozmerného priestoru (Euklidovského). Euklidovská vzdialenosť, ktorá je definovaná (podľa Legendre a Legendre [5]) medzi dvoma bodmi $\mathbf{x}$ a $\mathbf{y}$ vzťahom

$$D_E(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{6.4}$$

je najčastejšie používanou metrikou. Euklidovská vzdialenosť nie je zhora ohraničená, rastie s počtom dimenzií priestoru a závisí na merítku každého smeru v priestore. Tento problém môžeme obísť použitím štandardizovaných premenných na základe množiny bodov $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, resp. $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, pre ktoré nahradíme x_i v (6.4) výrazmi $\frac{x_i - \overline{x}_i}{s_{x_i}}$, kde $\overline{x}_i$ je výberový priemer a s_{x_i} je výberová smerodajná odchýlka i-tej zložky bodu $\mathbf{x}$. Podobne by se nahradili y_i. Euklidovská vzdialenosť je konkrétnym prípadom Minkowskiho metriky (pre $r = 2$):

$$D_M(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^r\right]^{1/r}. \tag{6.5}$$

Pre $r > 2$ dáva (6.5) dôležitosť veľkým absolútnym rozdielom, a preto sa v praxi vyskytuje použitie tejto metriky pre $r > 2$ veľmi zriedkavo. Naopak, veľmi často sa využíva manhattanská vzdialenosť[1] (6.5) s $r = 1$ a samozrejme aj Euklidovská vzdialenosť pre $r = 2$. Ak (6.5) vynásobíme zlomkom $\frac{1}{p}$ dostávame vzdialenosť, ktorá pre $r = 1$ nutne nerastie s rastúcim počtom dimenzií p.

(ii) Druhou skupinou sú pseudometriky alebo semimetriky, pre ktoré neplatí axióm trojuholníkovej nerovnosti. Jedným z uvedených príkladov semimetrík v Legendre a Legendre [5] je D_{sm} vzdialenosť, ktorá je definovaná pre dva body z p-rozmerného priestoru $\mathbf{x} = (x_1, \ldots, x_p)^\top$ a $\mathbf{y} = (y_1, \ldots, y_p)^\top$ vzťahom

$$D_{sm}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{p} |x_i - y_i|}{\sum_{i=1}^{p}(x_i + y_i)}. \tag{6.6}$$

(iii) Tretia skupina sú nemetriky[2]. Tieto koeficienty môžu nadobúdať záporné hodnoty, čím je porušená jedna z vlastností metrík, pozitivita. V Legendre a Legendre [5] uvádzajú ako príklad nemetrickej vzdialenosti $D_n = 1 - S$, kde S reprezentuje binárny Kulczynskiho koeficient podobnosti. Negatívne vzdialenosti sú intuitívne nezmyselné a problematicky interpretovateľné. Z toho dôvodu by sme sa všeobecne ich použitiu mali vyhýbať a použiť ich v prípade, ak existuje veľmi jasný dôvod ich použitia.

[1] angl. manhattan distance

[2] angl. nonmetrics

6.3 Mantlov test

Hlavným uplatnením Mantlovho testu publikovaného v Mantel [6] je možnosť porovnávať dve matice podobností, resp. dve matice vzdialeností, $\mathbb{X}^D$ a $\mathbb{Y}^D$. Bez vplyvu na všeobecnosť budeme matice značiť horným indexom D bez ohľadu na to, či sa jedná o maticu vzdialeností alebo podobností. Tento test sa ale v praxi využíva hlavne pri testovaní nezávislosti medzi dvoma viacrozmernými vektormi.

Nech $\mathbf{X} = (X_1, \ldots, X_p)$ je p-rozmerný náhodný vektor a $\mathbf{Y} = (Y_1, \ldots, Y_q)$ je q-rozmerný náhodný vektor. V praxi štatistici často majú k dispozícii namiesto realizácie náhodného výberu iba matice vzdialeností či podobností. Predpokladajme, že už máme k dispozícii zostrojené štvorcové matice podobností alebo vzdialeností. Nech $\mathbb{X}_{n\times n}$, ozn. $\mathbb{X}^D$, je matica podobností alebo vzdialeností, ktorá vznikne z náhodného výberu rovnako rozdelených p-rozmerných vektorov $\mathbf{X}_1$, $\ldots$, $\mathbf{X}_n$ a $\mathbb{Y}_{n\times n}$, ozn. $\mathbb{Y}^D$, je matica podobností alebo vzdialeností zostrojená z náhodného výberu rovnako rozdelených q-rozmerných vektorov $\mathbf{Y}_1$, $\ldots$, $\mathbf{Y}_n$, pričom pozorovania na jednotlivých subjektoch sú uvedené v rovnakom poradí, t. j. pozorovania náhodných vektorov $\mathbf{X}_i$ a $\mathbf{Y}_i$ sú uskutočnené na rovnakých subjektoch. Poznamenajme, že z definície koeficientov vzdialeností, resp. podobností, sú diagonálne prvky matíc $\mathbb{X}^D$ a $\mathbb{Y}^D$ nulové, resp. jednotkové.

Naviažme na začiatok kapitoly a prejdime k Mantlovej štatistike pre test hypotézy H_0. V Mantel [6], Omelka a Hudecová [7] alebo aj v Legendre a Legendre [5], v kapitole 10.5, je uvedená základná forma Mantlovej testovej štatistiky z_M, ktorá je definovaná ako

$$Z_M = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} x_{ij} y_{ij}, \tag{6.7}$$

kde x_{ij}, resp. y_{ij}, sú hodnoty koeficientov podobností alebo vzdialeností v maticiach $\mathbb{X}^D$, resp. $\mathbb{Y}^D$. Inou testovou štatistikou môže byť podľa Legendre a Legendre [5] štandardizovaná forma testovej štatistiky (6.7) v tvare

$$r_M = \frac{1}{n(n-1)/2 - 1} \sum_{i=1}^{n-1} \sum_{j=i+1, j\neq i}^{n} \frac{x_{ij} - \overline{x}}{S_x} \frac{y_{ij} - \overline{y}}{S_y}, \tag{6.8}$$

kde $n(n-1)$ je počet nenulových vzdialeností v maticiach $\mathbb{X}^D$ alebo $\mathbb{Y}^D$, resp. počet prvkov v matici $\mathbb{X}^D$ alebo $\mathbb{Y}^D$, okrem diagonálnych prvkov a

$$\overline{x} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} x_{ij}, \tag{6.9}$$

$$S_x = \sqrt{\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (x_{ij} - \overline{x})^2}. \tag{6.10}$$

Podobne by sme definovali $\overline{y}$ a S_y. Poznamenajme, že testová štatistika (6.8) je obmedzená zhora +1 a zdola -1 ako korelačný koeficient, čo viedlo k myšlienke použiť alternatívne k vzdialenostiam v testovej štatistike (6.7) aj poradia koeficientov vzdialeností či podobností.

Hypotézu H_0 budeme testovať na základe testovej štatistiky Z_M, resp. r_M podľa Mantel [6] permutačným testom. Pozri kapitolu o permutačných testoch.

V Mantel [6] je odvodenie rozdelenia ilustrované na príklade pozorovaní prípadov rakoviny, kde náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$ reprezentuje miesta výskytu rakoviny a náhodný výber $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ reprezentuje čas výskytu rakoviny na osobe $i = 1, \ldots, n$, kde $\mathbf{X}$ a $\mathbf{Y}$ sú v konečnom dôsledku náhodné veličiny.

Za platnosti H_0 máme k dispozíci pozorovania n subjektov (osôb), ktoré môžeme pri predpoklade náhodného párovania miesta výskytu rakoviny s jej výskytom v čase ľubovoľne permutovať. Vo všeobecnosti môžeme povedať, že napozorovaný vektor hodnôt pre ľubovoľný subjekt mohol byť napozorovaný pre ľubovoľný iný subjekt. Celkovo môžeme získať $n!$ rovnako pravdepodobných množín párov, t. j. ako sme v predchádzajúcej kapitole konštatovali, kardinalita množiny všetkých permutácii $M^{(n)} = n!$.

Realizácia konkrétnych permutácií na zostrojenie permutačnej testovej štatistiky je dosiahnutá permutáciou riadkov a stĺpcov v jednej z matíc vzdialeností alebo využitím bodu (b) v CMC metóde, t. j. je nutné permutovať náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$, resp. $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, a potom znovu zostrojiť maticu vzdialeností. Tieto dva postupy sú ekvivalentné, pretože prehodenie i-tého a j-tého riadku a stĺpca matice $\mathbb{Y}^D$ zodpovedá prehodeniu i-tého a j-tého prvku vo výbere $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$.

Permutáciu náhodného výberu realizujeme bez vplyvu na všeobecné riešenie permutovaním $\mathbf{Y}_{u_1^*}, \ldots, \mathbf{Y}_{u_n^*}$, kde $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)^\top$ je náhodná permutácia prvkov množiny $\{1, \ldots, n\}$ alebo môžeme náhodne permutovať riadky a stĺpce matice $\mathbb{Y}^D$, pričom náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$, resp. matica $\mathbb{X}^D$, zostávajú fixné. Týmto spôsobom ostávajú diagonálne prvky matice $\mathbb{Y}^D$ nulové a jej mimodiagonálne prvky sa s rovnakou pravdepodobnosťou objavia na jednej z $n(n-1)/2$ mimodiagonálnych pozícií príslušnej trojuholníkovej matice. Následne aplikujeme bod (c) v CMC metóde a prepočítame permutačnú testovú štatistiku Z_M^*, resp. r_M^*. Opakovaním bodu (b) a (c) pri dostatočnom množstve rôznych permutácií dostávame množinu hodnôt permutačných testových štatistík Z_M^* alebo r_M^*.

Takto dostávame odhad permutačného rozdelenia Mantlovej testovej štatistiky v prípade platnosti H_0. V prípade, že skutočná hodnota Mantlovej testovej štatistiky bola v porovnaní s oborom hodnôt permutačnej testovej štatistiky signifikantne príliš extrémna na to, aby bola považovaná za pravdepodobnú pri platnosti H_0, potom H_0 zamietame. Na hladine spoľahlivosti α môžeme na štatistické rozhodnutie využiť aj p-hodnotu testu, ktorú spočítame podľa (5.5).

6.4 DCOV test

V nasledujúcich riadkoch ukážeme inú alternatívu ako testovať hypotézu H_0. Test, ktorý predstavíme budeme nazývať DCOV test, podobne ako v Omelka a Hudecová [7]. Test je založený na teoretickej štúdii kovariancií vzdialeností[3] v Szekely [8], podľa ktorej vznikla skratka DCOV test. Podľa Szekely [8] korelácia vzdialeností R predstavuje nový prístup ako testovať hypotézu H_0 a zovšeobecňuje myšlienku korelácie pre náhodné vektory $\mathbf{X} \in \mathbb{R}^p$ a $\mathbf{Y} \in \mathbb{R}^q$ s konečnými prvými momentmi v dvoch smeroch:

- $\mathsf{R}(\mathbf{X}, \mathbf{Y})$ je definovaný pre $\mathbf{X}$ a $\mathbf{Y}$ ľubovoľnej dimenzie,
- $\mathsf{R}(\mathbf{X}, \mathbf{Y}) = 0$ charakterizuje nezávislosť $\mathbf{X}$ a $\mathbf{Y}$ ľubovoľnej dimenzie.

[3] angl. distance covariance

Korelácia vzdialeností R má vlastnosti skutočnej miery závislosti, t. j. dokáže rozoznať závislostnú štruktúru medzi dvoma vektormi $\mathbf{X}$ a $\mathbf{Y}$. Korelácia vzdialeností nadobúda minimum $\mathsf{R} = 0$ len ak $\mathbf{X}$ a $\mathbf{Y}$ sú nezávislé, maximum $\mathsf{R} = 1$ len ak $\mathbf{X}$ a $\mathbf{Y}$ sú úplne závislé, t. j. ak existuje nenulové číslo $b \in \mathbb{R}$, vektor $\mathbf{d} \in \mathbb{R}^q$ a ortogonálna matica $\mathbb{C}$, potom platí $\mathbf{Y} = b\mathbb{C}\mathbf{X} + \mathbf{d}$ a v ostatných prípadoch $0 < \mathsf{R} < 1$. Korelácia vzdialeností by mala dokázať rozoznať aj nelineárnu závislosť, a preto je vhodnejšia než Mantlov test, ktorý bol v poslednej dobe kritizovaný.

Znovu predpokladajme náhodný model, kde $\mathbf{X} = (X_1, \ldots, X_p)$ je p-rozmerný náhodný vektor a $\mathbf{Y} = (Y_1, \ldots, Y_q)$ je q-rozmerný náhodný vektor. Opäť nás bude zaujímať test hypotézy H_0 medzi vektormi $\mathbf{X}$ a $\mathbf{Y}$ oproti alternatíve, kde hypotéza H_0 neplatí. Označme matice napozorovaných dát podľa teoretického modelu ako $\mathbb{X}_{n\times p}$ a $\mathbb{Y}_{n\times q}$, kde i-tý riadok matíc $\mathbb{X}$ a $\mathbb{Y}$ popisuje rovnaký subjekt. Označme matice euklidovských vzdialeností $\mathbb{X}^D$ a $\mathbb{Y}^D$, ktoré získame z pozorovaných dát $\mathbb{X}$ a $\mathbb{Y}$ podobne ako v predchádzajúcom prípade.

Ďalej zaveďme podľa Szekely [8], resp. Omelka a Hudecová [7], štvorcové matice Δ_X, resp. Δ_Y, s rozmermi $n \times n$, ktoré nazveme Gowerove centrovacie matice. Tieto matice budú obsahovať v i-tom riadku a j-tom stĺpci prvky $\delta^X_{i,j}$, resp. $\delta^Y_{i,j}$, kde:

$$\delta^X_{i,j} = x_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x}_{**}, \tag{6.11}$$

pričom

$$\overline{x}_{i*} = \frac{1}{n-1}\sum_{j=1}^{n} x_{ij}, \quad \overline{x}_{*j} = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}, \quad \overline{x}_{**} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}, \tag{6.12}$$

kde x_{ij} sú prvky matice vzdialeností $\mathbb{X}^D$ a $\delta^Y_{i,j}$ zavedieme podobne. Omelka a Hudecová [7] odporúčajú pracovať s testovou štatistikou R_{d_E}, ktorú prezentovali Szekely [8], je špecifickým prípadom R (pozri Szekely [8]) a má tvar:

$$R_{d_E} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta^X_{ij}\delta^Y_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta^X_{ij})^2 \sum_{i=1}^{n}\sum_{j=1}^{n}(\delta^Y_{ij})^2}}. \tag{6.13}$$

Všimnime si podobnosť (6.13) so všeobecnými korelačnými koeficientmi definovanými v 2. kapitole. Na rozdiel od Mantlovho testu, kde testová štatistika je vypočítaná z pôvodných matíc vzdialeností, R_{d_E} prezentuje iný spôsob výpočtu testovej štatistiky pomocou konštrukcie Gowerových centrovaných vzdialeností v maticiach Δ_X a Δ_Y. V princípe ako pri Mantlovej testovej štatistike, tak aj pri R_{d_E} ide o snahu merať, ako sú malé či veľké hodnoty x_{ij} v porovnaní s hodnotami y_{ij}. Rozdiel je iba v spôsobe, na základe ktorého budeme robiť závery. Pri počítaní r_M porovnávame x_{ij}, resp. y_{ij}, s celkovým priemerom vzdialeností $\overline{x}$, resp. $\overline{y}$. Na druhej strane pri zostrojení koeficientu δ^X_{ij}, resp. δ^Y_{ij}, sa navyše berú do úvahy priemerné vzdialenosti pre i-té a j-té pozorovanie.

V Omelka a Hudecová [7] autori uvádzajú spôsob, ako testovať H_0 pomocou testovej štatistiky R_{d_E}. Ako pri Mantlovom teste, tak aj v tomto prípade využijeme permutačný test. Permutačnú testovú štatistiku $R^*_{d_E}$ zostrojíme permutáciou riadkov a stĺpcov v jednej z pôvodných matíc vzdialeností. Opäť bez vplyvu na všeobecné riešenie budeme náhodne permutovať riadky a stĺpce matice $\mathbb{Y}^D$

alebo podobne ako pri Mantlovom teste môžeme využiť bod (b) v CMC metóde a permutovať náhodný výber $\mathbf{Y}_{u_1^*}, \ldots, \mathbf{Y}_{u_n^*}$ podľa permutácie $\mathbf{u}^* = u_1^*, \ldots, u_n^*$.

Potom zostrojíme maticu vzdialeností $\mathbf{Y}^{D*}$ a z nej Δ_Y^*. Následne aplikujeme bod (c) v CMC metóde a prepočítame príslušnú permutačnú testovú štatistiku R_d^*. Opakovaním bodu (b) a (c) pri rôznych permutáciách dostávame množinu hodnôt testových štatistík $R_{d_E}^*$. V prípade, že skutočná hodnota testovej štatistiky je veľmi odľahlá v porovnaní s oborom permutačnej testovej štatistiky, potom H_0 zamietame.

Podľa Omelka a Hudecová [7] je možné použiť k výpočtu (6.11) aj iné matice vzdialeností než euklidovské vzdialenosti. V Omelka a Hudecová [7], časť 2.2, preto navrhujú použiť na test H_0 testovú štatistiku (6.13) (ozn. podľa Omelka a Hudecová [7] $R_{d_E^2}$) založenú na Gowerových centrovacích maticiach Δ_X, resp. Δ_Y, kde prvky (6.11) sú vyjadrené na základe matíc vzdialeností $\mathbb{X}^D$, resp. $\mathbb{Y}^D$, s druhými mocninami euklidovskej vzdialenosti. Navyše Omelka a Hudecová [7] poukázali na zhodu medzi $R_{d_E^2}$ a koeficientom RV predstaveným v Escoufier [9], ktorý spĺňa nasledujúce vlastnosti

(i) $0 \leq RV \leq 1$.

(ii) Pre $p = q = 1$ je RV zhodné s druhou mocninou Pearsonovho korelačného koeficientu.

(iii) $RV = 0$ ak empirická kovariančná matica medzi $\mathbb{X}$ a $\mathbb{Y}$ je nulová matica.

(iv) $RV = 1$ ak existuje matica $\mathbb{B}_{p\times q}$ a vektor $\mathbf{a}_{q\times 1}$ tak, že $\mathbb{X} = \mathbb{Y}\mathbb{B} + \mathbf{1}_n\mathbf{a}^\top$, kde $\mathbb{B}\mathbb{B}^\top$ je $k \neq 0$ násobok $p \times p$-rozmernej jednotkovej matice.

V Omelka a Hudecová [7] sa poukazuje na kritiku Mantlovho testu pri detekcii nelineárneho vzťahu medzi náhodnými vektormi. Naopak pri detekcii iba lineárneho vzťahu zdôrazňujú vhodnosť testovej štatistiky $R_{d_E^2}$. Na druhej strane upozornili na prácu Szekely [8], Corollary 2, kde ukázali konzistentnosť voči ľubovoľnej alternatíve pre test založený na testovej štatistike

$$nV_{d_E}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}^X\delta_{ij}^Y, \tag{6.14}$$

t. j. ukázali konvergenciu $nV_{d_E}^2 \xrightarrow[n\to\infty]{\mathcal{P}} \infty$ pre každé dva závislé vektory a pri platnosti H_0 navyše ukázali asymptotické rozdelenie (6.14).

Na základe simulačnej štúdie v Szekely [8] takisto ukázali rast sily permutačného testu s (6.14) k 1 a aj dodržiavanie hladiny tohto permutačného testu v rôznych modelovaných situáciách. Z toho a z vlastností R_{d_E} poukazujú Omelka a Hudecová [7] na konvergenciu sily testu aj s testovou štatistikou R_{d_E} s rastúcim počtom pozorovaní pri každom porušení nezávislosti k 1, zatiaľ čo pri platnosti H_0 simulačnou štúdiou ukázali, že DCOV test dodržiava stanovenú hladinu spoľahlivosti.

6.5 PROTEST

Ďalšia metóda, ktorou môžeme testovať H_0, je metóda založená na Procrustes[4] analýze navrhnutej v Gower [12], ktorá sa využíva najmä v analýze tvaru. Znovu vychádzame z rovnakých predpokladov a značení ako v celej kapitole. Chceme teda určiť nezávislosť medzi p-rozmerným náhodným vektorom $\mathbf{X}$ a príslušným q-rozmerným náhodným vektorom $\mathbf{Y}$. Opäť uvažujme príslušné náhodné výbery z $\mathbf{X}$ a $\mathbf{Y}$ a označme matice $\mathbb{X}_{n\times p}$ a $\mathbb{Y}_{n\times p}$ obsahujúce v riadkoch príslušné realizácie náhodného výberu vektorov $\mathbf{X}$ a $\mathbf{Y}$. Poznamenajme, že tak ako pri Mantlovom a DCOV teste, i tu predpokadáme realizáciu náhodných vektorov $\mathbf{X}_i$ a $\mathbf{Y}_i$ na rovnakom subjekte pre všetky $i = 1, \ldots, n$.

Následne sa pre danú dvojicu bodov hľadá najlepšia (najbližšia) poloha bodu $\mathbf{X}_i$, označme $\mathbf{X}_i^{'}$, ktorá bude čo najbližšou k príslušnému bodu $\mathbf{Y}_i$ využitím rotácie ($\mathbf{X}_i^{'} = \mathbf{R}\mathbf{X}_i$, kde $\mathbf{R}_{p\times p}$ je matica rotácie), posunutia ($\mathbf{X}_i^{'} = \mathbf{X}_i + \mathbf{d}$, kde $\mathbf{d}$ je p-rozmerný bod v priestore), zväčšenia, resp. zmenšenia ($\mathbf{X}_i^{'} = \mathbf{X}_i d$, kde $d \in \mathbb{R}$) a špeciálnej ortogonálnej rotácie ($\mathbf{X}_i^{'} = \mathbf{R}\mathbf{X}_i + \mathbf{d}$, kde $|R| = 1$). V tomto kontexte sa ako kritérium používa minimalizácia reziduálneho súčtu štvorcov medzi pozorovanými bodmi označovaná

$$m^2_{\mathbf{X}_i\mathbf{Y}_i} = \Delta^2(\mathbf{X}_i^{'}, \mathbf{Y}_i), \tag{6.15}$$

kde $\Delta(\mathbf{X}_i^{'}, \mathbf{Y}_i)$ je euklidovská vzdialenosť medzi bodmi $\mathbf{X}_i^{'}$ a $\mathbf{Y}_i$.

Na tomto základe pokračuje Jackson [13] a navrhuje použiť permutačný test k posúdeniu významnosti testovej štatistiky $m^2_{\mathbf{XY}} = \sum_{i=1}^{n} m^2_{\mathbf{X}_i\mathbf{Y}_i}$, ktorá slúži na určenie stupňa zhody medzi dvoma maticami a pôvodne bola navrhnutá v ortogonálnej Procrustes analýze v Gower [12]. V kapitole 7.5 v Legendre a Legendre [5] je pre testovú štatistiku uvedený nasledujúci vzťah

$$m^2_{\mathbf{XY}} = m^2_{\mathbf{YX}} = 1 - \mathrm{Tr}(\mathbb{W}), \tag{6.16}$$

kde maticu $\mathbb{W}$ dostaneme nasledujúcim spôsobom. Predpokladajme, že máme k dispozícii matice $\mathbb{Y}$ a $\mathbb{X}$, kde x_{ij}, resp. y_{ij}, sú pozorované hodnoty. Tieto matice následne vycentrujeme a znormujeme, t. j. výsledné matice budú mať prvky

$$x^*_{ij} = \frac{x_{ij} - \overline{x}_j}{S_x}, \quad \text{resp.} \quad y^*_{ij} = \frac{y_{ij} - \overline{y}_j}{S_y}, \quad \forall i,j = 1, \ldots, n, \tag{6.17}$$

kde $\overline{x}_j$ je priemer v j-tom stĺpci matice $\mathbb{X}$ a S_x je daný ako

$$S_x = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij})^2}. \tag{6.18}$$

Podobne by sme definovali $\overline{y}_j$ a S_y.

Označme novovzniknuté matice symbolmi $\mathbb{Y}^*$ a $\mathbb{X}^*$. Maticu $\mathbb{W}$ zo vzťahu (6.16) dostaneme singulárnym rozkladom matice $\mathbb{X}^{*T}\mathbb{Y}^*$, ktorý je daný vzťahom

$$\mathbb{X}^{*T}\mathbb{Y}^* = \mathbb{V}\,\mathbb{W}\,\mathbb{U}^\top, \tag{6.19}$$

kde $\mathbb{U}$ a $\mathbb{V}$ sú ortogonálne matice a $\mathbb{W}$ je diagonálna matica, ktorá obsahuje singulárne čísla súčinu matíc $\mathbb{X}^{*T}\mathbb{Y}^*$. Jackson [13] poukazuje, že v tejto súvislosti

[4] prebraté z angl.

je vhodné používať matice rovnakých dimenziií, t. j. $p = q$. V opačnom prípade môžu nastať problémy pri stanovení symetrickej testovej štatistiky (6.16).

Testová štatistika (6.16) vypovedá o miere zhody medzi dvoma pozorovanými maticami. V Jackson [13] je na posúdenie významosti testovej štatistiky navrhnuté použiť permutačný test s podobným postupom ako pri Mantlovom alebo DCOV teste. Týmto spôsobom odhadneme pravdepodobnosť výskytu pozorovanej testovej štatistiky oproti veľkému množstvu štatistík, ktoré dostaneme náhodným permutovaním pôvodných dát. Permutačné testové štatistiky získame náhodným permutovaním riadkov v jednej z matíc $\mathbb{X}$, resp. matice $\mathbb{Y}$.

Bez vplyvu na všeobecné riešenie budeme opäť náhodne permutovať riadky matice $\mathbb{Y}$, kedy opäť využijeme bod (b) v CMC metóde a permutujme náhodný výber $\mathbf{Y}_{u_1^*}, \ldots, \mathbf{Y}_{u_n^*}$. Matica $\mathbb{X}$ ostáva fixná a náhodne meníme riadky matice $\mathbb{Y}$ podľa permutácie $\mathbf{u}^* = u_1^*, \ldots, u_n^*$. Poznamenajme, že týmto permutačným postupom nemeníme hodnoty medzi rôznymi zložkami vektora $\mathbf{Y}$, teda vnútromaticová kovariančná štruktúra zostáva nezmenená, t. j. pre každú permutáciu $\mathbf{Y}_{u_1^*}, \ldots, \mathbf{Y}_{u_n^*}$ je odhad kovariančnej matice vektora $\mathbf{Y}^*$ rovnaký ako pre prvotné pozorovanie náhodného výberu.

Opakovaním bodu (b) a (c) pri rôznych permutáciách dostávame množinu hodnôt testových štatistík $m^2_{\mathbf{XY}}$. Pri veľkom počte pozorovaní n opäť využijeme podmienenú Monte Carlo metódu a realizujeme dostatočné množstvo náhodných permutácií, na základe ktorých môžeme odhadnúť p-hodnotu permutačného testu. V prípade, že skutočná hodnota testovej štatistiky bola príliš extrémna v porovnaní s oborom permutačnej testovej štatistiky na to, aby bola považovaná za pravdepodobnú za platnosti H_0, potom H_0 zamietame.

7. Praktické použitie testov s reálnymi dátami

V tejto kapitole poukážeme na možnosti praktického využitia testov predstavených v teoretickej časti práce, a to už s reálnymi dátami. Testy budeme pomocou softwaru R (R Core Team [17]) aplikovať na finančné dáta (výnosy akcií svetových spoločností) získané z Yahoo!Finance [16] a na dáta o pacientoch so zhubnými kožnými nádormi s názvom melanoma dostupné v balíčku s názvom boot (Canty a Ripley [18]). Všetky štatistické testy budeme realizovať na hladine spoľahlivosti $\alpha = 0{,}05$.

7.1 Dáta melanoma

Dátový súbor melanoma obsahuje 205 pozorovaní pacientov so zhubným kožným nádorom (malígnym melanómom). U každého pacienta sú uvedené nasledujúce údaje:

- počet dní (tim), ktorých sa dožil od operácie zhubného nádoru, ozn. X,
- vek pri prvej operácii (age), ozn. Y,
- kalendárny rok prvej operácie (year), ozn. Z,
- šírka nádoru v mm (thickness), ozn. W.

Na tomto mieste je nutné poznamenať, že ide o cenzorované dáta, t. j. ukončenie sledovania pacientov nastalo v dvoch prípadoch: smrť alebo ukončenie štúdie. Touto oblasťou sa podrobnejšie zaoberá analýza prežitia. Tieto dáta však slúžia iba ako ukážka aplikácie prezentovaných testov.

V praxi môže padnúť obvyklá otázka: Závisí počet prežitých dní po operácii na šírke nádoru? Uvažujme preto hypotézu

$$H_0 : \{\text{počet prežitých dní od operácie } (X) \text{ nezávisí na šírke nádoru } (Y)\} \quad (7.1)$$

oproti alternatíve

$$H_1 : \{\text{počet prežitých dní od operácie } (X) \text{ závisí ne šírke nádoru } (Y)\}. \quad (7.2)$$

Podobne môžeme kombinovať všetky štyri pozorované veličiny. V kontexte mnohorozmerných testov nezávilosti môžeme považovať za zaujímavý test hypotézy na jednej strane medzi počtom prežitých dní od operácie a rokom operácie (XY) a na druhej strane medzi šírkou nádoru a vekom pacienta (ZW):

$$H_0^* : \{XY \text{ je nezávislé od } ZW\} \quad (7.3)$$

oproti alternatíve

$$H_1^* : \{XY \text{ nie je nezávislé na } ZW\} \quad (7.4)$$

alebo ostatné dva možné varianty.

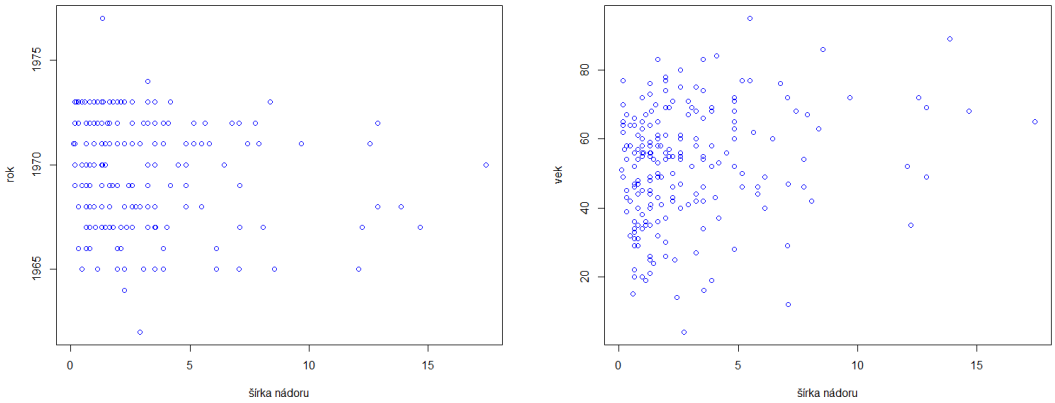
Na test H_0 a H_0^* využijeme neparametrické asymptotické testy prezentované v štvrtej kapitole, ktoré sú založené na testových štatistikách L_n (4.3), I_n (4.8), resp. χ_n^2, (4.12) alebo ich permutačné verzie χ_p^2, I_p a L_p, ktoré implementujeme v softwari R. Ďalej využijeme permutačné testy uvedené v predchádzajúcej kapitole: Mantlov test s euklidovskou (ozn. Mant_E) a manhattanskou vzdialenoťou (Mant_M), Protest (Prot), DCOV test založený na euklidovskej vzdialenosti (DCOV_E) a DCOV test založený na druhej mocnine euklidovskej vzdialenosti (DCOV_{E^2}). Mantlov test (funkcia mantel) i Protest (protest) sú k dispozícii v balíčku vegan (Oksanen a ost. [20]) a DCOV test (dcov.test) v balíčku energy (Rizzo a Szekely [19]). Pre test hypotézy H_0 môžeme využiť i asymptotické testy založené na korelačných koeficientoch a aj presné permutačné testy s korelačnými koeficientmi ako testovými štatistikami. Testy korelačných koeficientov (cor.test) sú štandardnou výbavou softwaru R a permutačné testy zavedieme v R podobne ako permutačné neparametrické testy. Všetky permutačné testy realizujeme pri množine náhodných permutácií s kardinalitou 10 000.

Pre všetky spomenuté štatistické testy na hladine $\alpha = 0{,}05$ zamietame hypotézu H_0 pri p-hodnote veľmi blízkej 0, čím je štatisticky preukázateľná závislosť počtu prežitých dní od operácie X a šírky nádoru W (pozri Tabuľka 7.1). Výsledky testov hypotéz nezávislosti medzi ostatnými kombináciami pozorovaných veličín sú zhrnuté v tabuľke 7.1. Podobný záver môžeme vyvodiť aj pre test nezávislosti medzi rokom operácie Z a počtom dožitých dní W, resp. vekom pacienta v čase operácie Y a počtom dožitých dní W (túto hypotézu nemôžeme zamietať iba permutačným testom na základe testovej štatistiky L_n, avšak s p-hodnotou 0,052).

Tabuľka 7.1: P-hodnoty testov s reálnymi dátami melanoma

test/H_0	X a Y	X a Z	X a W	Y a Z	Y a W	Z a W
DCOV_E	<0,001	<0,001	<0,001	0,002	0,003	0,081
DCOV_{E^2}	<0,001	<0,001	<0,001	0,009	0,003	0,054
Mant_E	0,003	<0,001	0,006	0,069	0,129	0,453
Mant_M	0,002	<0,001	0,012	0,074	0,114	0,464
Prot	<0,001	<0,001	<0,001	0,008	0,003	0,055
r	<0,001	<0,001	<0,001	0,007	0,002	0,057
r_p	<0,001	<0,001	<0,001	0,007	0,002	0,067
τ	<0,001	<0,001	<0,001	0,006	0,005	0,047
τ_p	<0,001	<0,001	<0,001	0,015	0,006	0,051
ρ	<0,001	<0,001	<0,001	0,016	0,004	0,043
ρ_p	<0,001	<0,001	<0,001	0,013	0,005	0,042
χ_p^2	0,015	<0,001	0,006	0,166	0,599	0,460
I_p	0,003	<0,001	<0,001	0,223	0,485	0,603
L_p	0,052	<0,001	<0,001	0,105	0,609	0,231

Všimnime si, že hypotézu nezávislosti šírky nádoru W a roku operácie Z nemôžeme zamietať väčšinou testov, až na testy τ, ρ i ρ_p, avšak s p-hodnotami o niečo menšími ako 0,05. Na obrázku 7.1 vľavo môžeme pozorovať rovnomerné rozmiestnenie zhluku bodov dát roku operácie a šírku nádoru, na základe ktorých väčšina testov neodhalila nejakú závislosť medzi Z a W.



Obr. 7.1: Zobrazený rok operácie a vek pacienta oproti šírke nádoru

Pri porovnaní mohorozmerných testov použitých v tomto jednorozmernom príklade môžeme konštatovať, že závislosť medzi skúmanými veličinami odhalili vo väčšine prípadov DCOV_E, DCOV_{E^2} a Protest. Podobnosť výsledkov DCOV_{E^2} a Protestu bola ukázaná aj v Omelka a Hudecová [7], Apendix B, kde autori porovnávajú testové štatistiky spomenutých testov. Navyše, pre $p = q = 1$ sú tieto dva testy zhodné s testom výberového korelačného koeficientu, čo vidíme aj na realizovaných p-hodnotách v tabuľke 7.1. Výsledky Mantlovho testu s rôznymi vzdialenosťami a permutačných testov s permutačnými štatistikami χ^2_p, L_p a I_p pre Y a Z, Y a W, resp. Z a W, napovedajú, že testy nie sú schopné zachytiť každú závislostnú štruktúru dát (pozri Obrázok 7.1 vpravo a Tabuľka 7.1).

V prípade testov prezentovaných v kapitole 4 môže mať rôzne delenie disjunktných množín rôzny dopad na výsledok testu. Pre praktickú implementáciu neparametrických testov budeme uvažovať ekvidistantné delenie podpriestorov $\mathbb{R}^4$, ktoré určíme podľa rozsahu pozorovanej veličiny a voľby $m_{n,1}$ a $m_{n,2}$. Pri zvolení $m_{n,1} = m_{n,2} = 3$, resp. 4, nemôžeme zamietať hypotézy nezávislosti dvojíc veličín (X, Y, Z a W) na základe kritických hodnôt neparametrických asymptotických testov skoro v žiadnom prípade (okrem X a Z). V porovnaní s permutačnými testami s rovnakými testovými štatistikami neboli asymptotické testy schopné odhaliť závislosť v niektorých prípadoch, zatiaľ čo permutačné testy áno.

Na tomto mieste môžeme poznamenať, že na príklade dát melanoma permutačné testy r_p, τ_p a ρ_p objavili štatistickú závislosť dát vo väčšom počte prípadov ako testy χ^2_p, L_p a I_p. To odporuje poznámke v Pesarin a Salmaso [14], v kapitole 2.6, príklad 2. Tieto dva stavy sú pravdepodobne spôsobené nesplnením predpokladov asymptotických testov alebo nedostatočným množstvom pozorovaní vzhľadom k počtu disjunktných podmnožín.

V tabuľke 7.2 sme zhrnuli výsledky testov hypotézy H_0^* pre všetky tri možné prípady dvojíc pozorovaných veličín. Znovu poznamenajme schopnosť detekcie závislosti pomocou DCOV_E testu, a v tomto prípade aj Mantlovho testu s obomi použitými vzdialenosťami. Protest a DCOV_{E^2} neodhalili závislosť medzi prežitými dňami po operáci s rokom operácie XZ a vekom pacienta so šírkou nádoru YW. Aj v tomto prípade vidíme podobné výsledky týchto dvoch testov.

Na základe kritických hodnôt pre asymptotické neparametrické testy s $m_{n,1} = m_{n,2} = 4$ zamietame pomocou testových štatistík χ^2_n a L_n hypotézu nezávislosti

Tabuľka 7.2: P-hodnoty testov s reálnymi dátami melanoma

test/H_0	XY a ZW	XZ a YW	XW a YZ
DCOV_E	<0,001	<0,001	<0,001
DCOV_{E^2}	<0,001	0,627	<0,001
Mant_E	0,003	<0,001	<0,001
Mant_M	0,002	<0,001	<0,001
Prot	<0,001	0,619	<0,001
χ^2_p	<0,001	0,004	0,013
L_p	<0,001	0,001	0,066
I_p	<0,001	0,003	<0,001

iba pre XY a ZW a testovou štatistikou I_n zamietame aj hypotézu nezávislosti medzi XW a YZ. V ostatných prípadoch nebola asymptotickými testami na dátach preukázaná závislosť (pozri Tabuľka 7.3), a teda hypotézy nemôžeme zamietať. Asymptotické výsledky testov môžeme porovnať s výsledkami presných permutačných testov. Ako vidíme v tabuľke 7.2, hypotézu nezávislosti nemôžeme zamietať iba medzi XW a YZ permutačným testom na základe štatistiky L_p. Aj v tomto prípade sa ukazujú permutačné testy v porovnaní s asymptotickými testami ako lepší štatistický nástroj. Môže to byť opäť spôsobené nedostatočným počtom pozorovaní pre platnosť asymptotiky testových štatistík.

Tabuľka 7.3: Testové štatistiky a kritické hodnoty testov

test/H_0	XY a ZW	XZ a YW	XW a YZ	kr. h.
χ^2_n	0,888	0,117	0,096	0,123
L_n	0,736	0,211	0,196	0,265
I_n	0,469	0,058	0,066	0,062

Ako najvhodnejšia voľba testu nezávislosti sa ukazuje DCOV test, no tentoraz aj Mantlov test pri oboch definíciách vzdialeností objavil, na rozdiel od Protestu a DCOV_{E^2} testu, závislosť medzi vektormi XY a ZW. Na tomto príklade sme ukázali, že asymptotické testy χ^2_n, I_n a L_n neodhalili závislosť vo väčšine prípadov, na rozdiel od ostatných testov, a preto je rozumné ich nepoužívať a uprednostniť ich permutačné verzie alebo DCOV_E. Mantlov test, DCOV_{E^2} test alebo Protest neodporúčame na základe tohto príkladu používať samostatne a pri podobnej analýze je vhodné použiť aj iný štatistický nástroj.

7.2 Výnosy akcií

V Yahoo!Finance [16] sú na výskumné účely voľne dostupné historické ceny akcií top svetových spoločností obchodovaných na finančnej burze, ktoré využijeme na ďalšie praktické účely. Štatistickú analýzu ale nemôžeme založiť na cenách akcií, pretože evidentne nespĺňajú predpoklad nezávislosti v náhodnom výbere. Z tohto dôvodu uprednostníme výnosy akcií vybraných desiatich spoločností,

ktoré vzniknú rozdielmi dvoch po sebe nasledujúcich cien akcií v časových jednotkách, dňoch.

V Yahoo!Finance [16] sú dostupné historické ceny akcií zaznamenané na začiatku dňa v okamihu otvorenia trhu[1], na konci dňa pri zatvorení trhu[2], ďalej minimálna a maximálna hodnota počas daného dňa a cena upravená o prípadné dividendy alebo rozdelenia akcií, tzv. adjusted close price, ktorú použijeme pre výpočet výnosov.

Budeme pracovať s dátovým súborom zloženým z cien akcií desiatich svetových spoločností (Ford Motor Co., Tesla Motors Inc., General Motors Co., American Airlines Group Inc., Marathon Oil Co., Apple Inc., Facebook Inc., Microsoft Co., Citigroup Inc., Bank of America Co.) zaznamenaných v pracovných dňoch od 18.5.2010 do 21.6.2016. K štatistickej analýze použijeme v súčte 1029 pozorovaní desiatich cien akcií, t. j. 1028 pozorovaní výnosov na jednu spoločnosť.

Z podstaty skúmaných dát a finančných trhov je veľmi pravdepodobné, že výnosy akcií budú v rovnakých odvetviach závislé náhodné veličiny alebo závislé viacrozmerné vektory. Zaujímavou však v mnohorozmernom prípade môže byť otázka, či sú dva vektory výnosov spoločnosti z dvoch rôznych odvetví nezávislé alebo nie. Pre túto myšlienku použijeme na test hypotézy

$$H_0 : \{\text{dva náhodné vektory výnosov sú nezávislé}\} \tag{7.5}$$

oproti alternatíve $H_1 : \{H_0 \text{ neplatí}\}$ mnohorozmerné testy: Protest, Mantlov test (s euklidovskou a manhattanskou vzdialenosťou) a DCOV test (s euklidovskou vzdialenosťou a jej druhou mocninou) pri použití desaťtisíc náhodných permutácií. Vzľadom na náročnú technickú implementáciu nebudeme neparametrické testy používať pre viac ako dvojrozmerné podvektory. Aj Arthur a Gyorfi [3] implementovali asymptotické testy uvedené v kapitole 4 iba pre jedno až trojrozmerné vektory.

Uvažujme vektor výnosov spoločností $\mathbf{X} = (\text{Apple, Microsoft, Facebook})^\top$ a druhý vektor výnosov spoločností $\mathbf{Y} = (\text{Tesla, General Motors, Ford})^\top$. Pomocou všetkých vyššie uvedených permutačných testov zamietame hypotézu H_0 na hladine $\alpha = 0{,}05$ pri dosiahnutej p-hodnote $< 0{,}001$. Takisto môžeme zistiť závislosť alebo nezávislosť medzi výnosmi spoločností Bank of America, Citigroup oproti Marathon Oil s American Airlines. Výsledky testov opäť potvrdzujú našu domnienku a zamietajú hypotézu nezávislosti pre tieto dva vektory výnosov.

Nakoniec sa dostávame k testu hypotézy pre dva päťrozmerné vektory výnosov. Nech $\mathbf{X} = (\text{Apple, Microsoft, Facebook, Bank of America, Citigroup})^\top$ a $\mathbf{Y} = (\text{Tesla, General Motors, Ford, Marathon Oil, American Airlines})^\top$ sú vektory výnosov uvedených spoločností. Uvažujme hypotézu H_0 oproti alternatíve H_1. Využitím uvažovaných testov znovu zamietame hypotézu H_0 na hladine $\alpha = 0{,}05$ pri dosiahnutej p-hodnote $< 0{,}001$.

Súbor s cenami akcií využijeme aj v nasledujúcej kapitole pri simulovaní dát z určitej závislostnej štruktúry. Na základe takto vygenerovaných dát budeme sledovať dosiahnutú silu a dodržiavanie hladiny permutačných testov. Simulačná štúdia ale tematicky nenadväzuje na túto praktickú časť a nemá slúžiť ani ako doplnenie praktických testov.

[1] angl. open price
[2] angl. close price

8. Simulačná štúdia

V predchádzajúcej kapitole sme ukázali výsledky jednotlivých testov na dátach dostupných v Yahoo!Finance [16] a dátach melanoma v R balíčku boot (Canty a Ripley [18]). V tejto časti práce budeme porovnávať vlastnosti jednotlivých testov na základe simulácií náhodných výberov z normálneho rozdelenia a jeho následnej transformácie. Potom budeme sledovať v simulačnom príklade so zvoleným zavislostným modelom realizované hladiny a sily jednotlivých testov hypotézy

$$H_0 : \{\text{dva náhodné vektory } \mathbf{X} \text{ a } \mathbf{Y} \text{ sú nezávislé}\} \tag{8.1}$$

oproti alternatíve $H_1 : \{H_0 \text{ neplatí}\}$.

Pozornosť upriamime na porovnanie Mantlovho testu, DCOV testu, Protestu a testov predpokladajúcich mnohorozmerné normálne rozdelenie: test pomerom vierohodností λ_n (3.19), Λ (3.20) a následnú Bartletovu transformáciu (3.8), ktorú označíme U^*. Všetky testy realizujeme na teoretickej hladine významnosti $\alpha = 0{,}05$. K praktickej realizácii všetkých testov opäť využijeme software R. Permutačné testy budeme po zohľadnení časovo náročných simulácií vykonávať pre tisíc náhodne generovaných permutácií.

Keďže permutačné testy sú presné testy, odhad hladiny spoľahlivosti testu by mal po zohľadnení simulačnej chyby zodpovedať teoretickej hladine spoľahlivosti $\alpha = 0{,}05$. Pri testoch vyžadujúcich predpoklad normality budeme pri výbere z normálneho rozdelenia opäť očakávať hodnoty hladiny významnosti približne 0,05. Poznamenajme, že test založený na testovej štatistike Λ pre dva podvektory z normálneho rozdelenia je presný test a test pomerom vierohodností λ_n a U^* sú na rozdiel od Λ asymptotické testy. Navyše k určeniu významnosti Λ budeme potrebovať kritické hodnoty Wilksovho rozdelenia, ktoré sú k dispozícii v Rencher [10], na str. 566 - 573.

8.1 1. simulačný model

Zostavme pravdepodobnostný model pre vektor $\mathbf{Z}_{(10\times1)}$. Nech prvých päť náhodných veličín vektora $\mathbf{Z}$ tvorí podvektor $\mathbf{X} = (X_1, \ldots, X_5)^\top$ a ostatné veličiny vektora $\mathbf{Z}$ nech tvoria podvektor $\mathbf{Y} = (Y_1, \ldots, Y_5)^\top$. K uskutočneniu simulácií musíme špecifikovať pravdepodobnostné rozdelenie vektora $\mathbf{Z}$, na základe ktorého určíme závislosť, resp. nezávislosť podvektorov $\mathbf{X}$ a $\mathbf{Y}$. Nech $\mathbf{Z}$ je generovaný z desaťrozmerného normálneho rozdelenia so strednou hodnotou $\mathbf{0}_{(10\times1)}$ a rozptylovou maticou $\mathrm{Var}(\mathbf{Z})_{(10\times10)} = \boldsymbol{\Sigma}_c$

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \boldsymbol{\Sigma}_X & c\boldsymbol{\Sigma}_{XY} \\ c\boldsymbol{\Sigma}_{XY}^\top & \boldsymbol{\Sigma}_Y \end{bmatrix}, \tag{8.2}$$

kde $c = 0, \frac{5}{100}, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$ a 1. Počiatočná variančná matica $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^\top & \boldsymbol{\Sigma}_Y \end{bmatrix} \tag{8.3}$$

musí spĺňať podmienku pozitívne definitnej matice, t. j. $\forall\, \mathbf{x}_{(10\times1)} \neq \mathbf{0}_{(10\times1)}$ musí platiť $\mathbf{x}^\top\boldsymbol{\Sigma}\mathbf{x} > 0$. Určiť desaťrozmernú maticu $\boldsymbol{\Sigma}$ (8.3) tak, aby spĺňala podmienku pozitívnej definitnosti nie je vôbec jednoduché. Pre tento účel môžeme

využiť výberovú kovariančnú maticu odhadnutú z dát, ktorá tento predpoklad spĺňa. Preto k stanoveniu počiatočnej kovariančnej matice $\mathbf{\Sigma}$ využijeme dátový súbor cien, ktorý sme prezentovali v predchádzajúcej kapitole. $\mathbf{\Sigma}$ stanovíme ako odhad kovariančnej matice na základe kótovaných cien uvedených desiatich svetových spoločností na konci dňa upravených o prípadné dividendy a o štiepenie akcií, tzv. adjusted close prices. Pre lepšiu predstavu o simulovaných dátach uvedieme konkrétnu odhadnutú výberovú kovariančnú maticu (8.3) v číselnej podobe (v práci je zaokrúhlená na jedno desatinné miesto)

$$\begin{bmatrix} 462{,}8 & 150{,}8 & 488{,}8 & 31{,}9 & 63{,}6 & 1191{,}9 & 32{,}5 & 11{,}9 & -76{,}5 & 231{,}0 \\ 150{,}8 & 86{,}0 & 268{,}2 & 18{,}4 & 36{,}5 & 644{,}3 & 23{,}5 & 9{,}4 & -40{,}7 & 104{,}2 \\ 488{,}8 & 268{,}2 & 910{,}7 & 59{,}1 & 112{,}8 & 2112{,}5 & 75{,}1 & 30{,}2 & -143{,}3 & 336{,}7 \\ 31{,}9 & 18{,}4 & 59{,}1 & 8{,}3 & 20{,}7 & 210{,}3 & 11{,}6 & 5{,}1 & 1{,}0 & 31{,}0 \\ 63{,}6 & 36{,}5 & 112{,}8 & 20{,}7 & 59{,}4 & 468{,}4 & 31{,}0 & 14{,}0 & 11{,}2 & 65{,}8 \\ 1191{,}9 & 644{,}3 & 2112{,}5 & 210{,}3 & 468{,}4 & 6882{,}7 & 281{,}1 & 132{,}5 & -99{,}1 & 965{,}9 \\ 32{,}5 & 23{,}5 & 75{,}1 & 11{,}6 & 31{,}0 & 281{,}1 & 21{,}9 & 9{,}2 & 5{,}7 & 41{,}4 \\ 11{,}9 & 9{,}4 & 30{,}2 & 5{,}1 & 14{,}0 & 132{,}5 & 9{,}2 & 4{,}5 & 4{,}9 & 17{,}8 \\ -76{,}5 & -40{,}7 & -143{,}3 & 1{,}0 & 11{,}2 & -99{,}1 & 5{,}7 & 4{,}9 & 64{,}2 & -26{,}6 \\ 231{,}0 & 104{,}2 & 336{,}7 & 31{,}0 & 65{,}8 & 965{,}9 & 41{,}4 & 17{,}8 & -26{,}6 & 177{,}9 \end{bmatrix}$$

Poznamenajme, že táto simulačná štúdia nezávisí a ani nenadväzuje na predchádzajúcu praktickú časť a nebudeme ju ani aplikovať na dáta použité v predchádzajúcej kapitole. Navyše ceny akcií a ani výnosy cien akcií pravdepodobne nebudú pochádzať z normálneho rozdelenia. Tomu sa ale v tejto práci nebudeme venovať. Mnohorozmerné normálne rozdelenie sme volili vzhľadom na jeho jednoznačnosť závislostnej štruktúry.

Pre počiatočnú kovariančnú maticu (8.3) budeme pomocou pravdepodobnostného modelu pre vektor $\mathbf{Z} \sim N_{10}(\mathbf{0},\mathbf{\Sigma}_c)$ generovať náhodné výbery pre všetky hodnoty koeficientu c, pričom stále kontrolujeme závislostnú štruktúru generovaných dát. Nezávislosť dvoch podvektorov nastáva práve vtedy, keď $c = 0$. V ostatných prípadoch sa jedná o model dvoch závislých podvektorov, ktorého závislostnú štruktúru reprezentuje kovariančná matica $c\mathbf{\Sigma}_{XY}$.

8.2 2. simulačný model

Uvažujme ďalší možný simulačný postup pre náhodný vektor $\mathbf{Z}_{(10\times1)}$ s podvektormi $\mathbf{X} = (X_1, \ldots, X_5)^\top$ a $\mathbf{Y} = (Y_1, \ldots, Y_5)^\top$ s rovnakým pravdepodobnostným modelom pre vektor $\mathbf{Z}$ ako vyššie. Pripomeňme, že pre $c = 0$ sú podvektory $\mathbf{X}$ a $\mathbf{Y}$ nezávislé. Potom aj pre ľubovoľné funkcie $f : \mathbb{R}^5 \to \mathbb{R}^5$ a $g : \mathbb{R}^5 \to \mathbb{R}^5$ platí, že $f(\mathbf{X})$ a $g(\mathbf{Y})$ sú nezávislé náhodné vektory. Definujme preto transformácie získaných dát na základe modelu $\mathbf{Z} \sim N_{10}(\mathbf{0},\mathbf{\Sigma}_c)$ nasledovne: $\mathbf{X}' = (X_1^2, X_2^3, X_3, X_4, |X_5|)^\top$ a $\mathbf{Y}' = (\exp(Y_1), Y_2, Y_3^2, Y_4, Y_5)^\top$. Na základe takto definovaných transformácií budeme opäť ilustrovať testy hypotézy

$$H_0 : \{\mathbf{X}' \text{ a } \mathbf{Y}' \text{ sú nezávislé náhodné vektory}\} \tag{8.4}$$

oproti alternatíve $H_1 : \{H_0 \text{ neplatí}\}$ pre rôzne hodnoty koeficientu c.

Označme n ako rozsah náhodného výberu pre výbery z náhodneho vektora $\mathbf{Z} \sim N_{10}(\mathbf{0},\mathbf{\Sigma}_c)$. Pre každú hodnotu koeficientu c budeme simulovať $m = 1000$

realizácií náhodných výberov vektora Z s rozsahom n. Na základe takto postaveného pravdepodobnostného modelu budeme porovnávať mnohorozmerné testy hypotézy H_0 pre rozličné hodnoty n a c. Budeme sledovať vlastnosti testov hypotézy H_0, ich silu a hladinu, pre tri rozsahy náhodných výberov $n = 100$, 500 a 1000 pre všetky hodnoty parametra c, pričom nás bude zaujímať aj správanie sa testov s rastúcim n.

8.3 Zrhnutie výsledkov simulácií

V nasledujúcich tabuľkách zhrnieme výsledky simulačnej štúdie, kde budeme sledovať realizovanú hladinu (pre $c = 0$) a silu ($c > 0$) jednotlivých testov pri rôznych hodnotách n a c. Kritické hodnoty rozdelenia $\Lambda_{p,q,n}$ určíme približne na základe tabuľky v Rencher [10], na str. 566 - 573. Pretože v tabuľkách nie sú uvedené kritické hodnoty $\Lambda_{p,q,n}$ pre všetky hodnoty n, pre praktické účely využijeme nasledujúcu aproximáciu. Pri rozsahu výberu $n = 100$ použijeme namiesto $\Lambda_{0,05;5;5;94}$ hodnotu $\Lambda_{0,05;5;5;100} = 0{,}685$, pri $n = 500$ odhadneme $\Lambda_{0,05;5;5;494}$ váženým priemerom $\Lambda_{0,05;5;5;440} = 0{,}918$ a $\Lambda_{0,05;5;5;600} = 0{,}939$ s hodnotou 0,925 a pre $n = 1000$ použijeme kritickú hodnotu $\Lambda_{0,05;5;5;1000} = 0{,}963$.

Na základe simulačnej štúdie na dátach generovaných z normálneho rozdelenia zrhnutej v tabuľkách 8.1 až 8.3 môžeme konštatovať, že s rastúcim rozsahom výberov n sa odhad hladiny testu pohybuje stabilne okolo hodnoty 0,05, a teda simulačnou štúdiou sme potvrdili očakávané dodržiavanie hladiny pre všetky uvažované testy. Tento záver môžeme konštatovať aj pre permutačné testy použité na transformovaných dátach, ktoré pôvodne pochádzajú z normálneho rozdelenia (pozri Tabuľka 8.4 až Tabuľka 8.6).

V druhom prípade ale pozorujeme nedodržanie hladiny spoľahlivosti na úrovni 0,05 pre testy Λ, U^* a λ_n, čo je spôsobené transformáciou normálneho rozdelenia, t. j. porušením predpokladu normality. Napriek porušeniu tohto predpokladu existuje iná možnosť ako využiť testovú štatistiku U^* pre test H_0, a to použiť permutačný test s testovou štatistikou U^*, ktorý podľa definície permutačných testov bude dodržiavať hladinu spoľahlivosti na stanovenej úrovni α. Otázkou však ostáva akú silu bude mať tento permutačný test. V tabuľke 8.7 zhrnieme výsledky simulácií testu na dátach z druhého príkladu. V porovnaní s ostatnými permutačnými testami môžeme konštatovať, že tento test sa ukazuje ako ďalšia vhodná alternatíva aj pri porušení predpokladu normality.

Pri porovnaní sily permutačných testov v oboch modelových prípadoch zisťujeme nasledujúce. V prvom príklade usudzujeme na základe generovaných dát, že všetky testy dokázali odhaliť závislosť dvoch podvektorov $\mathbf{X}$ a $\mathbf{Y}$ s rastúcou silou testu pri rastúcom rozsahu výberov n a takisto pre rastúcu hodnotu koeficientu $c > 0$. Síce s rastúcim koeficientom c sledujeme nárast sily testov, no pozorujeme aj problém s detekciou závislosti pri alternatívach blízkych hypotéze, najevidentnejšie v Mantlovom teste.

Na prvý pohľad je z tabuliek 8.1 až 8.6 zjavné, že najvhodnejší test na detekciu závislosti medzi dvoma viacrozmernými vektormi je DCOV test a pri splnení predpokladu normality aj testy s testovými štatistikami Λ, U^* a λ_n. Pri transformácii dát pochádzajúcich z normálneho rozdelenia pozorujeme výrazne malú silu testov DCOV_{E^2} a Protest. Tieto testy dokonca vo väčšine prípadov nedokázali odhaliť závislosť podvektorov ani pri väčších hodnotách koeficientu c. Tento fakt

Tabuľka 8.1: Pomer počtu zamietnutí hypotézy H_0 k m v prvom simulovanom príklade pre $n = 100$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,064	0,047	0,060	0,053	0,042	0,042	0,045	0,045
0,05	0,096	0,064	0,105	0,074	0,054	0,052	0,055	0,057
0,1	0,147	0,106	0,174	0,139	0,125	0,059	0,057	0,131
0,2	0,429	0,374	0,460	0,492	0,331	0,173	0,179	0,330
0,3	0,832	0,790	0,849	0,864	0,625	0,423	0,415	0,636
0,4	0,995	0,991	0,997	0,988	0,879	0,767	0,757	0,881
0,5	1	1	1	1	0,973	0,976	0,960	0,973
0,6	1	1	1	1	1	1	1	1

Tabuľka 8.2: Pomer počtu zamietnutí hypotézy H_0 k m v prvom simulovanom príklade pre $n = 500$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,033	0,043	0,051	0,048	0,048	0,046	0,048	0,048
0,05	0,117	0,125	0,133	0,194	0,133	0,076	0,071	0,138
0,1	0,405	0,438	0,455	0,548	0,389	0,108	0,106	0,402
0,2	0,997	0,997	0,998	0,996	0,933	0,450	0,431	0,936
0,3	1	1	1	1	0,999	0,956	0,948	1
0,4	1	1	1	1	1	1	1	1

Tabuľka 8.3: Pomer počtu zamietnutí hypotézy H_0 k m v prvom simulovanom príklade pre $n = 1000$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,049	0,049	0,050	0,049	0,047	0,055	0,051	0,051
0,05	0,209	0,207	0,212	0,309	0,209	0,066	0,074	0,213
0,1	0,860	0,865	0,868	0,894	0,677	0,153	0,154	0,679
0,2	1	1	1	1	1	0,699	0679,	1
0,3	1	1	1	1	1	0,996	0,995	1

Tabuľka 8.4: Pomer počtu zamietnutí hypotézy H_0 k m v druhom simulovanom príklade pre $n = 100$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,081	0,068	0,098	0,040	0,060	0,052	0,047	0,059
0,05	0,108	0,084	0,125	0,059	0,047	0,053	0,051	0,044
0,1	0,127	0,101	0,147	0,112	0,050	0,054	0,053	0,044
0,2	0,252	0,220	0,272	0,289	0,059	0,080	0,097	0,058
0,3	0,528	0,473	0,557	0,609	0,069	0,146	0,166	0,063
0,4	0,881	0,844	0,892	0,917	0,076	0,281	0,325	0,075
0,5	0,989	0,986	0,990	0,991	0,088	0,483	0,543	0,088
0,6	1	1	1	1	0,092	0,738	0,828	0,089

Tabuľka 8.5: Pomer počtu zamietnutí hypotézy H_0 k m v druhom simulovanom príklade pre $n = 500$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,066	0,074	0,076	0,062	0,050	0,051	0,045	0,048
0,05	0,097	0,105	0,115	0,110	0,060	0,046	0,050	0,054
0,1	0,254	0,275	0,289	0,372	0,058	0,060	0,062	0,057
0,2	0,923	0,931	0,938	0,958	0,050	0,146	0,180	0,058
0,3	1	1	1	1	0,077	0,355	0,470	0,075
0,4	1	1	1	1	0,082	0,729	0,844	0,081
0,5	1	1	1	1	0,122	0,975	0,996	0,117

Tabuľka 8.6: Pomer počtu zamietnutí hypotézy H_0 k m v druhom simulovanom príklade pre $n = 1000$

c	Λ	U^*	λ_n	DCOV	DCOV_{E^2}	Mant_E	Mant_M	Prot
0	0,069	0,069	0,069	0,053	0,047	0,051	0,053	0,051
0,05	0,136	0,133	0,136	0,186	0,044	0,058	0,048	0,050
0,1	0,518	0,509	0,518	0,684	0,052	0,067	0,079	0,051
0,2	0,999	0,999	0,999	1	0,062	0,188	0,265	0,063
0,3	1	1	1	1	0,082	0,485	0,724	0,079
0,4	1	1	1	1	0,098	0,947	0,984	0,101
0,5	1	1	1	1	0,141	0,997	1	0,128

Tabuľka 8.7: Pomer počtu zamietnutí hypotézy H_0 k m pre permutačný test s testovou štatistikou U^*

c \ n	100	500	1000
0	0,040	0,055	0,054
0,05	0,059	0,078	0,098
0,1	0,056	0,209	0,436
0,2	0,141	0,875	0,995
0,3	0,342	1	1
0,4	0,751	1	1
0,5	0,970	1	1
0,5	1	1	1

môžeme vysvetliť voľbou uvažovanej transformácie, kde sme využili kvadratickú, kubickú či exponenciálnu transformáciu zložiek vektora.

Ako sme spomenuli pri teoretickom zavedení DCOV testu s testovou štatistikou $R_{d_E^2}$, tento test je vhodný k detekcii lineárneho vzťahu medzi náhodnými vektormi a preto zlyhal v odhalení závislosti takto zostrojených vektorov. Testová štatistika $R_{d_E^2}$ je podľa Omelka a Hudecová [7] založená na súčine druhých mocnín singulárnych čísel v diagonálnej matici W v testovej štatistike (6.16) Protestu, zatiaľ čo testová štatistika (6.16) je založená na súčte týchto singulárnych čísel.

Z tohto dôvodu môžeme očakávať podobné výsledky Protestu a DCOV$_{E^2}$ testu.

Ako najlepšia alternatíva medzi permutačnými testami sa na základe simulačnej štúdie javí DCOV test založený na korelácii euklidovských vzdialeností, podobne ako to bolo aj v praktickom príklade na dátach melanoma. Pre DCOV test s euklidovskou vzdialenosťou sme v teoretickej časti práce uviedli, že v Szekely [8] ukázali pre testovú štatistiku $nV^2_{d_E}$ konvergenciu sily testu k 1 pre dva závislé náhodné vektory a dodržiavanie hladiny testu v prípade platnosti H_0. Podobný záver konštatujú aj v Omelka a Hudecová [7] pre testovú štatistiku R_{d_E}. Poznamenajme, že DCOV test založený na testovej štatistike R_{d_E} zamieta H_0 pri jej veľkých hodnotách (blízkych 1), a teda pri dostatočne veľkom počte pozorovaní by mal aj DCOV test zamietať H_0 pre každé dva závislé vektory.

Simulačná štúdia skutočne ukázala, že pri porušení hypotézy H_0 rastie sila DCOV testu s testovou štatistikou R_{d_E} spolu s rastúcim počtom pozorovaní, pričom test stále dodržiava stanovenú hladinu spoľahlivosti. Menej vhodnou alternatívou je Mantlov test, ktorý pre malé hodnoty c preukázal slabú silu, no pre alternatívu vzdialenejšiu od hypotézy sa javí ako vhodný štatistický nástroj k odhaleniu závislosti dvoch podvektorov. Pre detekciu iba lineárnej závislosti sme na základe simulácií ukázali vhodnosť DCOV testu s testovou štatistikou $R_{d^2_E}$ a Protestu.

Záver

Ústredným motívom tejto diplomovej práce boli testy hypotéz nezávislosti, formulovanie predpokladov a alternatív jednotlivých testov. Jedným z cieľov práce bola prehľadná rešerš testov, ktorú sme štrukturovali do kapitol podľa študovaných hypotéz a alternatív. Následne sme využili teoretické poznatky a aplikovali ich na skutočných dátach i v simulačnej štúdii. Na základe výsledkov praktickej časti sme diskutovali o vhodnosti testov v jednotlivých situáciách, čím sme splnili zvyšné ciele.

V teoretickej časti práce sme uviedli podrobný prehľad testov rôznych hypotéz nezávislostí. Začali sme testami nezávislosti dvoch alebo viacerých kategoriálnych a spojitých veličín. Pri predpoklade normálneho rozdelenia sme ukázali možnosti, ako testovať vzájomnú nezávislosť $k \geq 2$ podvektorov normálne rozdeleného vektora. Ako ďalšie v poradí sme uviedli testové štatistiky založené na odhade pravdepodobnostných mier na základe disjunktného rozdelenia pravdepodobnostného priestoru.

Samostatnú kapitolu práce sme venovali i permutačným testom, ktoré sme formulovali na mieru testom nezávislosti pre vhodne volené testové štatistiky. Príklady permutačných testových štatistík sme čerpali aj z predchádzajúcich kapitol a aj z uvedenej literatúry. Samostatnú pozornosť sme venovali permutačným testom: Protest, Mantlov a DCOV test.

V praktickej časti práce ukazujeme výsledky jednotlivých testov, ktoré sme použili na pozorovaniach pacientov so zhubným kožným nádorom a finančných dátach. V simulačnej časti sme pomocou výberovej kovariančnej matice odhadnutej z cien akcií desiatich spoločností generovali náhodné výbery z normálneho rozdelenia, pomocou ktorých sme sledovali vlastnosti testov. Na základe praktickej časti sme porovnali jednotlivé testy a diskutovali o vhodnosti ich použitia.

Zoznam použitej literatúry

[1] Kendall, M. G., *Rank Corrlation Methods*. 4. vydanie. London: Griffin, 1970. ISBN 0-85264-199-0.

[2] Anděl, J., *Matematická Statistika*. 2. vydanie. Praha: SNTL Nakladatelství technické literatury, 1985 ISBN (nenašiel som v knihe, čo tu mam uviesť?).

[3] Gretton, A., Györfi, L., *Consistent Nonparametric Tests of Independence*. Journal of Machine Learning Research 11, 1391-1423, 2010.

[4] Wilks, S. S., *The Likelihood Test of Independence in Contingency Tables*. The Annals of Mathematical Statistics, 6(4), 190–196, 1935, Institute of Mathematical Statistics Stiahnuté z "`http://www.jstor.org/stable/2957689`".

[5] Legendre, P., Legendre, L., *Numerical ecology*. 2. anglické vydanie. Amsterdam: Elsevier Science B.V, 1998. ISBN 0-444-89249-4.

[6] Mantel, N., *The Detection of Disease Clustering and a Generalized Regression Approach*. Cancer Research 27 Part 1, 209-220, 1967.

[7] Omelka, M., Hudecová, Š., *A comparison of the Mantel test with a generalised distance covariance test*. Environmetrics; 24: 449–460, Wiley Online Library, 2013.

[8] Székely, J. G., Rizzo, L. M.,Bakirov, K. N., *Measuring and testing dependence by correlation of distances*. The Annals of Statistics, Vol. 35, No. 6, 2769–2794, 2007.

[9] Escoufier, Y., *Le traitement des variables vectorielles*. Biometrics 29: 751–760, 1973.

[10] Rencher, A C., *Methods of Multivariate Analysis*. Wiley-Interscience 2002, 2. vydanie, ISBN 0-471-41889-7.

[11] Pham-Gia, T., *Exact distribution of the generalized Wilks's statistic and applications*. Journal of Multivariate Analysis 99, 1698-1716, 2008.

[12] Gower, J.C., *Statistical methods of comparing different multivariate analyses of the same data.* Mathematics in the archaeological and historical sciences, 138-149, 1971.

[13] Jackson, D. A., *PROTEST: A PROcrustean Randomization TEST of community environment concordance*. Écoscience, 2(3), 297-303, 1995.

[14] Pesarin, F., Salmaso, L., *Permutation tests for complex data*. Wiley, 2010. ISBN: 978-0-470-51641-6.

[15] Zvára, K., *Regrese*. 1. vydanie. Praha: Matfyzpress, 2008. ISBN: 978-80-7379-041-8.

[16] Yahoo!Finance. *Ceny akcií spoločností.* Stiahnuté dňa 22.6.2016 z "http://finance.yahoo.com/stock-center//".

[17] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Viedeň, 2014. "http://www.R-project.org/".

[18] Canty, A., Ripley, B., *boot: Bootstrap R (S-Plus) Functions.* R package version 1.3-11.

[19] Rizzo, M. L., Szekely, G. J., *energy: E-statistics (energy statistics).* R package version 1.6.2. "http://CRAN.R-project.org/package=energy".

[20] Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., OHara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H., Wagner, H., *vegan: Community Ecology Package* R package version 2.3-5. "http://CRAN.R-project.org/package=vegan".