

**SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
STAVEBNÁ FAKULTA**

Evidenčné číslo: SvF-5343-50366

**MULTIVARIAČNÁ ANALÝZA ROZPTYLU A
KOVARIANCIE
DIPLOMOVÁ PRÁCA**

Študijný program: Matematicko-počítačové modelovanie

Číslo študijného odboru: 1114

Názov študijného odboru: 9.1.9 aplikovaná matematika

Školiace pracovisko: Katedra matematiky a deskriptívnej geometrie

Vedúci záverečnej práce/školiťel': RNDr. Jana Kalická, PhD.

Bratislava 2013

Bc. Ivona Fojtíková

Čestné prehlásenie

Vyhlasujem, že som diplomovú prácu vypracovala samostatne za použitia uvedenej odbornej literatúry a odbornej pomoci vedúcej práce.

Bratislave 24. 05. 2013

.....

vlastnoručný podpis

Pod'akovanie

Ďakujem vedúcej práce RNDr. Jane Kalickej, PhD., za jej pomoc, odbornú radu a pripomienky pri vypracovaní predkladanej práce.

Abstrakt

Predkladaná práca sa zaoberá testami viacrozmernej analýzy rozptylu. Samotná multivariačná analýza rozptylu (MANOVA) je rozšírením analýzy rozptylu (ANOVA), ktorá skúma vzťah medzi závislou intervalovou premennou Y a jednou alebo viacerými nezávislými nominálnymi premennými X (faktormi). MANOVA porovnáva vektory priemerov v skupinách podľa úrovné jednotlivých faktorov. Testy jednofaktorovej a dvojfaktorovej analýzy sú implementované v systéme Mathematica 8.0. Práca dopĺňa tieto testy o testy pre viacrozmernú analýzu rozptylu. Vhodnosť použitia testov je overená v simulačných štúdiách. Takisto pomocou týchto testov práca spracováva reálne hydrometeorologické dáta. Podobne ako sa ANOVA pridaním intervalového faktora rozšíri na ANCOVA, možno MANOVA rozšíriť na MANCOVA (multivariačná analýza kovariancie). MANOVA a MANCOVA sa používajú tam, kde má skúmaná vlastnosť viacrozmerný charakter – napr. v psychológii a sociológii.

Kľúčové slová: testy hypotéz, viacrozmerná analýza rozptylu, viacrozmerná analýza kovariancie

Abstract

The work deals with multivariate analysis of variance tests. Multivariate analysis of variance (MANOVA) is an extension of univariate analysis of variance, which describes the dependence Y of one or more nominal variables (factors). MANOVA compares vectors of means in groups according to levels of each factor. One-factor and two-factor ANOVA tests are implemented in software Mathematica 8.0. The work completes these tests with tests for multivariate analysis of variance. Applicability of these tests is verified in simulation studies. In the end of the work, MANOVA tests are used for testing real hydrometeorological data. Similarly as ANOVA after adding interval factor is extended to ANCOVA, MANOVA can be extended to MANCOVA (multivariate analysis of covariance). MANOVA and MANCOVA are used where a studied feature has multidimensional character – e.g. in psychology, sociology.

Key words: hypothesis testing, multivariate analysis of variance, multivariate analysis of covariance

Obsah

1. Úvod	1
2. Jednofaktorová analýza rozptylu	3
3. Viacrozmerné normálne rozdelenie	6
4. Testovanie rovnosti kovariančných matíc	9
5. Viacrozmerné porovnanie dvoch výberov	12
6. Viacrozmerná analýza rozptylu	14
6.1. Testovanie rovnosti dvojíc vektorov stredných hodnôt	18
6.2. Testovanie veličín, ktoré zapríčiňujú rozdielnosť stredných hodnôt	20
7. Simulačné štúdie pre Hotellingov T^2 test	22
8. Simulačné štúdie pre viacrozmernú analýzu rozptylu	23
9. Testovanie reálnych hydrometeorologických dát	25
10. Viacrozmerná analýza kovariancie	38
11. Záver	42
12. Použitá literatúra	43

1. Úvod

Pojem štatistika pochádza z latinského výrazu „ statisticum collegium “ a talianskeho „ statista “. Nemecký výraz „statistik“, ktorý zaviedol roku 1749 Gottfried Achenwall, sa týkal výhradne analýzy dát o štáte a predstavoval vedu o štáte. Termín štatistika získal všeobecný význam „ zber a analýza údajov “ až začiatkom 19. storočia.

Moderná štatistika má induktívny charakter. Je to vedná disciplína, ktorej úlohou je určiť postupy zberu dát a na základe zozbieraných hodnôt urobiť závery. Pri určení postupov zberu dát, ako aj pri ich vyhodnocovaní, sa opiera o teóriu pravdepodobnosti [5].

Napriek dlhej histórii štatistiky bola väčšina jej metód a analytických postupov popísaných až koncom 19. storočia a začiatkom 20. storočia. Štatistické metódy pomáhajú identifikovať, študovať ale aj riešiť mnohé závažné problémy vo všetkých oblastiach života [5].

Predkladaná diplomová práca sa zaoberá viacrozmernými štatistickými metódami, konkrétne multivariačnou analýzou rozptylu a kovariance (MANOVA, MANCOVA). Svojím obsahom nadväzuje na bakalársku prácu, v ktorej sme sa zaoberali analýzou rozptylu a kovariance a testami, ktoré nahrádzajú analýzu rozptylu v prípade, ak nie je splnený jej predpoklad o rovnosti disperzií. Diplomová práca je rozdelená na dve hlavné časti. V prvej časti popisujeme testy multivariačnej analýzy rozptylu. Tieto testy potom otestujeme na simulačných štúdiách a následne pomocou nich testujeme hydrometeorologické dáta. V druhej časti popisujeme multivariačnú analýzu kovariancie.

Viacrozmerné štatistické metódy majú na rozdiel od dvojrozmerných niekoľko výhod. Prvým dôvodom je mnohorozmernosť reality. Takmer nikdy nemožno jednu premennú dokonale vysvetliť pomocou druhej premennej. Napríklad úspešnosť liečby (závislá premenná) nikdy nezávisí iba od druhu podaného lieku, ale aj od mnohých iných faktorov akými sú: genetika, životospráva, vek.

Druhým dôvodom prečo sa má uprednostniť využitie viacrozmernej metódy je možnosť výskytu falošných korelácií. Falošná korelácia je neexistujúca korelácia medzi premennými X a Y, ktorá sa dôsledkom iných nezohľadnených premenných zdá ako silná.

Tretím dôvodom uprednostňovania viacrozmerných metód je možnosť skúmania interakcií medzi faktormi (nezávislými premennými) [5].

Samotná multivariačná analýza rozptylu (MANOVA) je rozšírením analýzy rozptylu (ANOVA), ktorá skúma vzťah medzi závislou intervalovou premennou Y a jednou alebo viacerými nezávislými nominálnymi premennými X (faktormi). ANOVA porovnáva priemery závislej premennej v skupinách podľa úrovne jednotlivých faktorov (napr. priemernú výšku u mužov a žien) a overuje, či zistené rozdiely môžu byť iba náhodné (medzi premennou a faktorom nie je vzťah), alebo sú štatisticky významné (medzi premennou a faktorom je vzťah). MANOVA porovnáva vektory priemerov v skupinách podľa úrovne jednotlivých faktorov. Ak sa porovnávajú iba dve skupiny sú výsledky MANOVA identické s Hotellingovým T^2 testom, ktorý je rozšírením Studentovho t -testu. Okrem Wilksovej Lambda, ktorá je multivariačným rozšírením F -testu, existujú pri MANOVA ďalšie tri testovacie štatistiky: Lawley-Hotellingova stopa, Pillaiova stopa, Royov najväčší koreň.

MANOVA (tak ako ANOVA) umožňuje popri samotných faktoroch testovať interakcie medzi faktormi, plánované porovnávanie a tiež rozlišuje medziskupinové a vnútroskupinové faktory [7].

Podobne ako sa ANOVA pridaním intervalového faktora rozšíri na ANCOVA, možno MANOVA rozšíriť na MANCOVA (multivariačná analýza kovariancie). MANOVA a MANCOVA sa používajú tam, kde má skúmaná vlastnosť viacrozmerný charakter – napr. v psychológii a sociológii. Typickým príkladom môže byť vyhodnocovanie liečby v psychológii, kde faktormi sú napr.: typ lieku – placebo vs. aktívne antidepresíva a typ psychoterapie – klinická vs. kognitívna. Na porovnanie úspešnosti liečby je meraných a porovnávaných viacero druhov psychologických indexov pri každom pacientovi. Multivariačnou analýzou rozptylu potom overujeme, či vektory stredných hodnôt týchto indexov sú rovnaké a teda, či záleží na type lieku a type psychoterapie.

2. Jednofaktorová analýza rozptylu

Jednofaktorová (One-Way ANOVA) je najjednoduchšou formou analýzy rozptylu, ktorá skúma vzťah medzi intervalovou a nominálnou premennou. Vo svojej parametrickej podobe predpokladá normalitu rozdelenia a tzv. homoskedasticitu (rovnosť rozptylov).

Nech $X_{i1}, X_{i2}, \dots, X_{in_i}$ je náhodný výber z rozdelenia $N(\mu_i, \sigma^2)$, kde $i = 1, \dots, k$. Označme X_{ij} j-te meranie v i-tom výbere a n_i počet pozorovaní v i-tom náhodnom výbere. Je potrebné testovať hypotézu:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu.$$

Pred samotnou realizáciou testu treba spočítať nasledujúce štatistiky

$$X_{i.} = \sum_{j=1}^{n_i} X_{ij}, \quad (2.1)$$

$$X_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}, \quad (2.2)$$

ako aj priemery v jednotlivých triedach a celkový priemer

$$X_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \frac{X_{i.}}{n_i}, i = 1, 2, \dots, k \quad (2.3)$$

$$x_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} = \frac{X_{..}}{N} \quad (2.4)$$

$$N = \sum_{i=1}^k n_i. \quad (2.5)$$

Ďalej definujeme:

celkový súčet štvorcov:

$$S_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - x_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}^2 - \frac{X_{..}^2}{N} \quad (2.6)$$

riadkový súčet štvorcov:

$$S_A = \sum_{i=1}^k n_i (x_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^k \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} \quad (2.7)$$

reziduálny súčet štvorcov:

$$S_e = S_T - S_A \quad (2.8)$$

Výsledky môžeme zapísať do tabuľky (tab. 2. 1):

Variabilita	Súčet štvorcov	Stupne voľnosti	Podiel	F
riadková	S_A	$k-1$	$\frac{S_A}{k-1} = (1)$	$\frac{(1)}{(2)}$
reziduálna	S_e	$N-k$	$\frac{S_e}{N-k} = (2)$	
celková	S_T	$N-1$		

tab. 2. 1 Súhrnná tabuľka jednofaktorovej analýzy rozptylu

F štatistika má Fisher-Snedecorovo rozdelenie s $(k-1)$ a $(N-k)$ stupňami voľnosti. Hypotézu H_0 o rovnosti stredných hodnôt zamietame, ak F štatistika je väčšia ako príslušný kvantil Fisher-Snedecorovho rozdelenia s danými stupňami voľnosti [1].

Príklad 2.1

Úlohou je porovnať úspešnosť absolventov gymnázií z troch rôznych krajov Slovenska pri prijímacej skúške z matematiky na vysokú školu. Dosiahnuté výsledky 15 náhodne vybraných študentov sú uvedené v tabuľke (tab. 2. 2).

Kraj 1	Kraj 2	Kraj 3
60	59	52
59	55	58
63	56	54
66	56	55
57	54	51

tab. 2. 2 Dáta k príkladu 2.1

Príklad bol vypočítaný pomocou softvéru Mathematica 8.0, ktorá má v sebe zabudovanú procedúru na výpočet jednofaktorovej analýzy rozptylu. Výstupom tejto procedúry je nasledujúca tabuľka (tab. 2. 3).

		DF	SumOfSq	MeanSq	FRatio	PValue
{ANOVA →	Model	2	130.	65.	8.29787	0.00546109
	Error	12	94.	7.83333		
	Total	14	224.			
CellMeans →	All		57.			
	Model[Kraj1.]		61.			
	Model[Kraj2]		56.			
	Model[Kraj3]		54.			
PostTests →	Model → Bonferroni	{ {Kraj1., Kraj2}, {Kraj1., Kraj3} }				
	Duncan	{ {Kraj1., Kraj2}, {Kraj1., Kraj3} }				

tab. 2. 3. Výsledná tabuľka programu Mathematica 8.0 pre príklad 2.1

Výsledok testovacej štatistiky pre daný príklad je 8.29787, ktorý musíme porovnať s kvantilom Fisher-Snedecorovho rozdelenia $F_{0.95}(2, 12)$. Na základe teórie analýzy rozptylu hypotézu H_0 zamietame, t.j. úspešnosť študentov na prijímacej skúške z matematiky závisí na kraji, v ktorom študovali. Záver o zamietnutí, resp. nezamietnutí H_0 možno urobiť aj pomocou Phodnoty (PValue). Platí, ak je PValue menšia ako hladina významnosti α , hypotézu H_0 zamietame. PValue je pre dané dáta rovná 0.00546109, zatiaľ čo hladina významnosti α bola nastavená na hodnotu 0.05. Hypotézu H_0 teda zamietame na príslušnej hladine významnosti.

Po zamietnutí nulovej hypotézy môžeme tzv. Post testami zistiť, ktoré skupiny sa od seba výrazne odlišujú. Softvér Mathematica 8.0 obsahuje napr. Bonferroniho test, Duncanov test alebo Tukeyho test. Pre náš príklad sme použili dva z nich a to Bonferroniho a Duncanov test. Oba testy ukázali, že výraznejšie sa od seba odlišujú kraj 1 a kraj 2, kraj 1 a kraj 3. Odlišnosť kraja 2 a kraja 3 nebola preukázaná.

Ako sme už spomenuli, jedným z predpokladov ANOVA je rovnosť disperzií dát. Tento predpoklad však nie je v praxi vždy splnený. Vtedy používame namiesto analýzy rozptylu iné testy, ktoré takisto ako ANOVA testujú rovnosť stredných hodnôt, no nepredpokladajú rovnosť rozptylov. Medzi takéto testy patria napr. Cochranov test, Brown-Forsytheov test, Welchov test či modifikovaný Welchov test.

3. Viacrozmerné normálne rozdelenie

Použitelnosť mnohých klasických štatistických induktívnych metód častokrát súvisí s predpokladom o normálnom rozdelení sledovaných veličín. Častá podmienka normality vyplýva nielen z toho, že odvodené riešenia majú dobré vlastnosti a dobre sa s nimi pracuje, ale aj preto, že nie je ťažké túto podmienku splniť. Metódy popísané v práci vychádzajú z predpokladu, že dáta spĺňajú podmienku normality, teda boli vybrané z viacrozmerného normálneho rozdelenia.

Viacrozmerné normálne rozdelenie je rozšírením jednorozmerného normálneho rozdelenia pre prípad $p \geq 2$ náhodných veličín.

Hustota mnohorozmerného normálneho rozdelenia má tvar:

$$f(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}, \quad (3.1)$$

kde $\mathbf{x} \in \mathbb{R}^p$, vektor $\boldsymbol{\mu}$ je vektor p stredných hodnôt veličín $X_1, X_2, \dots, X_p$. $\Sigma_{(p \times p)}$ je symetrická a pozitívne semidefinitná kovariančná matica. Hodnosti matice Σ , t.j. $h(\Sigma)$, sa hovorí *rád rozdelenia*. Ak $h(\Sigma) = p$, resp. $h(\Sigma) < p$, tak hovoríme, že mnohorozmerné normálne rozdelenie je regulárne, resp. singulárne.

Ak má náhodný vektor $\mathbf{x}$ viacrozmerné normálne rozdelenie, potom platí, že:

- lineárne kombinácie prvkov z $\mathbf{x}$ majú normálne rozdelenie
- všetky podmnožiny $\mathbf{x}$ majú normálne rozdelenie
- nekorelovanosť náhodných veličín z $\mathbf{x}$ znamená aj ich nezávislosť
- všetky podmienené rozdelenia sú normálne

Z uvedených tvrdení vyplývajú nasledujúce vety:

1. Nech náhodný vektor $\mathbf{x}$ má p -rozmerné normálne rozdelenie $N_p(\boldsymbol{\mu}, \Sigma)$, potom lineárna kombinácia $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$, kde $\mathbf{A}$ je nenulová matica konštant typu $m + p$ a $\mathbf{b}$ je m -rozmerný vektor konštant, má m -rozmerné normálne rozdelenie s vektorom stredných hodnôt $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ a s kovariančnou maticou $\mathbf{A}\Sigma\mathbf{A}^T$.

2. Ak má náhodný vektor $\mathbf{x}$ p-rozmerné normálne rozdelenie $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, potom náhodný vektor $\mathbf{u} = \mathbf{P}(\mathbf{x} - \boldsymbol{\mu})$ má p-rozmerné normálne rozdelenie s nulovým vektorom stredných hodnôt a s jednotkovou kovariančnou maticou iba vtedy, ak je $\mathbf{P}$ regulárna matica, taká že $\mathbf{P}^T \mathbf{P} = \boldsymbol{\Sigma}^{-1}$.

3. Rozdeľme náhodný vektor $\mathbf{x}$ s p-rozmerným normálnym rozdelením $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ na také dve podmnožiny x_1 a x_2 , že

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ a } \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

potom marginálne rozdelenie $\mathbf{x}_1$ je normálne s parametrami $\boldsymbol{\mu}_1$ a $\boldsymbol{\Sigma}_{11}$ a rovnako marginálne rozdelenie $\mathbf{x}_2$ je normálne s parametrami $\boldsymbol{\mu}_2$ a $\boldsymbol{\Sigma}_{22}$. Rovnako podmienené rozdelenie $\mathbf{x}_2$ vzhľadom k $\mathbf{x}_1$ je normálne s parametrami

$$\boldsymbol{\mu}_2 + \Sigma_{22} \Sigma_{11}^{-1} (x_1 - \mu_1) \text{ a } \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}, \quad (3.2)$$

ako aj rozdelenie $\mathbf{x}_1$ vzhľadom k $\mathbf{x}_2$ je normálne s parametrami

$$\boldsymbol{\mu}_1 + \Sigma_{11} \Sigma_{22}^{-1} (x_2 - \mu_2) \text{ a } \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}. \quad (3.3)$$

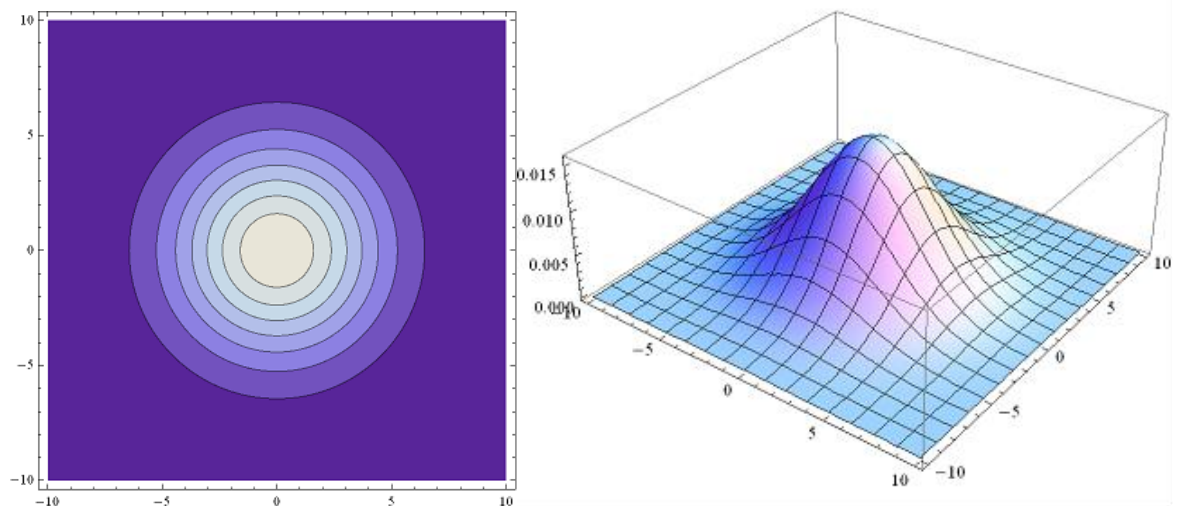
Vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ sú teda nezávislé v prípade, že matice Σ_{12} a Σ_{21} sú nulové.

4. Ak má náhodný vektor $\mathbf{x}$ p-rozmerné normálne rozdelenie $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, potom zložky vektora $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2} \mathbf{x}$ majú normálne rozdelenie a sú nezávislé.

Z geometrického hľadiska predstavuje výraz $\mathbf{c}^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ kvadratickú formu vzdialenosti $\mathbf{x}$ od $\boldsymbol{\mu}$. Sú to kontúry elipsoidu so stredom v $\boldsymbol{\mu}$ a s poloosami tvaru $c\sqrt{\lambda_j} \mathbf{v}_j$, kde λ_j sú vlastné čísla a $\mathbf{v}_j$ sú ortonormálne vektory matice $\boldsymbol{\Sigma}$. Hustota $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dosahuje svoje maximum v strednej hodnote. Obrisy elipsoidu možno použiť na vizualizáciu dvojrozmerných dát, t.j. pre dvojrozmerné normálne rozdelenie.

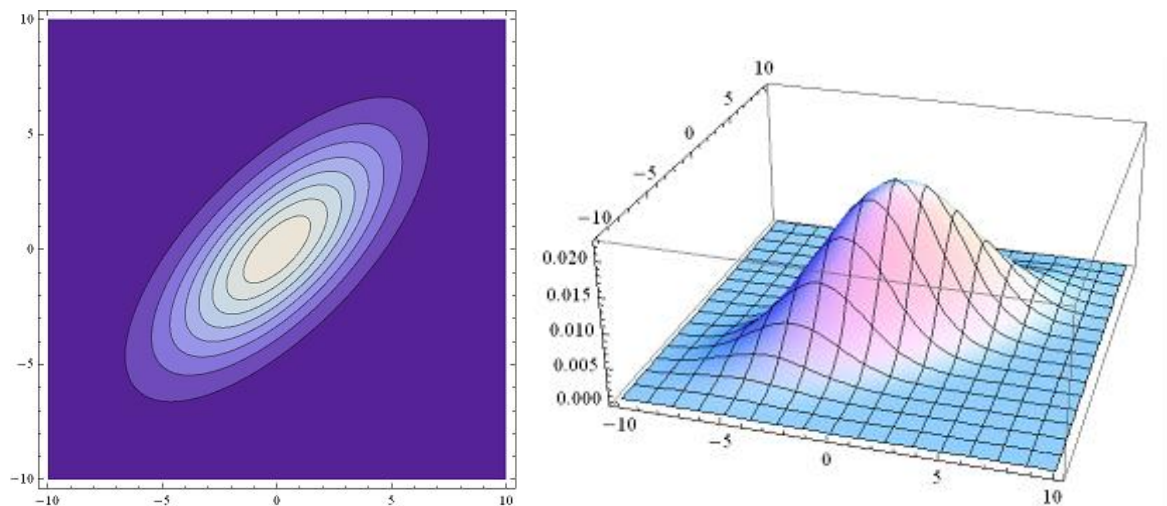
V prípade dvoch veličín X_1 a X_2 so strednými hodnotami μ_1 a μ_2 , s rozptylmi σ_1^2 , σ_2^2 a kovarianciou σ_{12} je determinant kovariančnej matice $\boldsymbol{\Sigma}$ možné vyjadriť ako $\sigma_1^2 \sigma_2^2 (1 - \rho^2)$, kde ρ je korelačný koeficient definovaný $\sigma_{12}/\sigma_1 \sigma_2$. Uhol medzi regresnými priamkami charakterizuje stupeň lineárnej závislosti medzi X_1 a X_2 . V prípade nezávislosti sú na seba obe priamky kolmé, zatiaľ čo v prípade funkčnej závislosti obe priamky

splývajú. Tento rozdiel vidíme aj na nasledujúcich obrázkoch. Obrázok 3.1 zobrazuje hustotu a izočiaru dvojrozmerného normálneho rozdelenia pre dáta, ktorých stredná hodnota sa rovná 0 a disperzia je 10. Korelačný koeficient je rovný 0, t.j. dáta sú nezávislé a regresné priamky sú na seba kolmé. Na rozdiel od obr. 3.2, ktorý zobrazuje hustotu a izočiaru dvojrozmerného normálneho rozdelenia pre dáta, ktorých korelačný koeficient je 0.7, t.j. dáta nie sú nezávislé a regresné priamky nie sú na seba kolmé [2].



obr. 3. 1 Hustota dvojrozmerného normálneho rozdelenia s parametrami

$$\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 10, \rho = 0$$



obr. 3. 2 Hustota dvojrozmerného normálneho rozdelenia s parametrami

$$\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 10, \rho = 0.7$$

4. Testovanie rovnosti kovariančných matíc

Vo viacrozmerných úlohách je často potrebné overiť nielen hypotézu zaoberajúcu sa rozptylmi všetkých premenných, ale aj všetkých kovariancií. Rovnosť kovariančných matíc vo všetkých skupinách je podmienkou na uplatnenie niektorých testov zameraných na vektory stredných hodnôt. Navyše, overenie hypotézy o rovnosti kovariančných matíc je často prvým krokom viacrozmernej štatistickej analýzy.

Testujeme hypotézu $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_K$ oproti alternatívnej hypotéze, ktorá jej platnosť popiera.

Testovací postup vyžaduje spočítať štatistiku M

$$M = \frac{\prod_{k=1}^K |S_k(\mathbf{x})|^{\frac{n_k-1}{2}}}{|\bar{S}|^{\frac{n-K}{2}}}, \quad (4.1)$$

v ktorej

$$\bar{S} = \frac{\sum_{k=1}^K S_k(\mathbf{x})(n_k - 1)}{n - K} = \frac{\sum_{k=1}^K Q_k(\mathbf{x})}{n - K} \quad (4.2)$$

a $S_k(\mathbf{x})$, $k = 1, 2, \dots, K$ je kovariančná matica pochádzajúca z k -tej skupiny, pre ktorú platí

$$S_k = \frac{1}{n - 1} Q. \quad (4.3)$$

Matica Q sa nazýva aj Wishartova matica. Môžeme ju zapísať v maticovom tvare

$$Q = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \quad (4.4)$$

Užitočné sú dve asymptotické aproximácie rozdelenia M :

1. Veličina

$$V = -2(1 - c_1) \ln M \quad (4.5)$$

má chí-kvadrát rozdelenie s $p(p + 1)(K - 1)/2$ stupňami voľnosti, kde

$$c_1 = \frac{2p^2 + 3p - 1}{6(p + 1)(K - 1)} \left(\sum_{k=1}^K \frac{1}{n_k - 1} - \frac{1}{n - K} \right). \quad (4.6)$$

Aproximácia chí-kvadrát rozdelením môže byť použitá v prípade, že počet premenných p ani počet skupín K nie je väčší ako 5 a všetky výberové rozsahy n_k sú aspoň 20.

2. Veličina

$$F = (-2b) \ln M \quad (4.7)$$

má F rozdelenie s d_1 a d_2 stupňami voľnosti, kde

$$d_1 = \frac{p(p + 1)(K - 1)}{2}, \quad (4.8)$$

$$d_2 = \frac{d_1 + 2}{|c_2 - c_1^2|}, \quad (4.9)$$

$$b = \frac{1 - c_1 - \frac{d_1}{d_2}}{d_1}, \quad (4.10)$$

c_1 sa počíta rovnako ako v (4.6) a

$$c_2 = \frac{(p - 1)(p + 2)}{6(K - 1)} \left(\sum_{k=1}^K \frac{1}{(n_k - 1)^2} - \frac{1}{(n - K)^2} \right). \quad (4.11)$$

Ak $c_2 - c_1^2 < 0$, tak

$$- \frac{b_1 d_2 \ln M}{d_1 + 2b_1 d_1 \ln M} \quad (4.12)$$

má približne F rozdelenie s d_1 a d_2 stupňami voľnosti, kde

$$b_1 = \left(1 - c_1 - \left(\frac{2}{d_2}\right)\right) d_2. \quad (4.13)$$

Aproximácia F rozdelením býva lepšia ako aproximácia chí-kvadrát rozdelením [2].

Príklad 4.1

Na 45 vzorkách ropy pochádzajúcich z 3 rôznych ložísk boli namerané hodnoty týchto premenných:

X_1 podiel vanadia

X_2 podiel železa

X_3 podiel nasýtených uhlíkovodíkov

X_4 podiel aromatických uhlíkovodíkov

Úlohou je testovať hypotézu o rovnosti kovariančných matíc náhodného vektora $\mathbf{x}^T = (X_1, X_2, X_3, X_4)$. Z 3 skúmaných ložísk máme k dispozícii nezávislé náhodné výbery s rozsahmi $n_1 = 7$, $n_2 = 8$, $n_3 = 30$ [2].

Na overenie hypotézy o rovnosti kovariančných matíc bol použitý program vytvorený v softvéri Mathematica 8.0. Test bol počítaný na hladine významnosti $\alpha = 0.05$.

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Test Kovariančných Matíc	2.12508×10^{-9}	1.50138	20.	1070.37
Príslušný kvantil test.štatistiku: 1.58036				
nezamietam H_0				

tab. 4.1. Výsledky k príkladu 4.1

Keďže štatistika F je menšia ako príslušný kvantil, nulovú hypotézu o rovnosti kovariančných matíc nezamietame.

5. Viacrozmerné porovnanie dvoch výberov

Predpokladajme, že máme k dispozícii dva nezávislé náhodné výbery z viacrozmerného normálneho rozdelenia s rozsahmi n_1, n_2 . Súčet n_1, n_2 označme ako n . Pridajme ešte predpoklad rovnosti kovariančných matíc $\Sigma_1 = \Sigma_2 = \Sigma$. Test rovnosti kovariančných matíc bol popísaný v kapitole 4 tejto práce.

K posúdeniu rovnosti vektora stredných hodnôt vypočítame veličiny $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1, \mathbf{S}_2$ (index udáva číslo výberu - skupiny) podľa vzťahov (4. 3), (4. 4).

Vypočítame spoločnú výberovú kovariančnú maticu

$$\bar{\mathbf{S}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}, \quad (5. 1)$$

ktorá je bodovým odhadom Σ .

Nech $\delta = \mu_1 - \mu_2$ (najčastejšie volíme $\delta = 0$). Potom štatistika

$$T^2 = \frac{n_1 n_2}{n} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta)^T \bar{\mathbf{S}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta) \quad (5. 2)$$

má Hotellingovo rozdelenie s parametrami $p, (n - p - 1)$ a

$$F = \frac{n - p - 1}{p} \frac{T^2}{n - 2} \quad (5. 3)$$

má F rozdelenie s rovnakými parametrami. Podobne ako v prípade jedného výberu, oblasť spoľahlivosti pre rozdiel medzi strednými hodnotami obsahuje všetky δ , pre ktoré je hodnota veličiny $F < F_{1-\alpha}(p, n - p - 1)$. Zodpovedá jej vnútro p -rozmerného elipsoidu so stredom v $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$.

Pre overenie hypotézy H_0 sa vypočíta hodnota testovacieho kritéria uvedená v rovnici (5. 3). Ak táto hodnota je väčšia ako kvantil $F_{1-\alpha}(p, n - p - 1)$, zamietame hypotézu $H_0: \mu_1 = \mu_2$ o zhode dvoch vektorov stredných hodnôt na hladine významnosti α [2].

Príklad 5. 1

V náhodnom agrokultúrnom experimente bolo porovnávaných osem rôznych polí. Na štyroch bolo použité nízkodusičnanové hnojivo a na zvyšných štyroch vysokodusičnanové hnojivo. Na všetkých poliach bol meraný výnos (počet semien na rastlinu) v gramoch a váha semien (váha 100 náhodne vybraných semien na rastlinu) v miligramoch. Zaujímame sa o vplyv obsahu dusičnanov v hnojive na výnos a hmotnosti semien [8].

Na overenie hypotézy o rovnosti stredných hodnôt bol použitý program vytvorený v softvéri Mathematica 8.0. Jeho výstupom je tabuľka (tab. 5. 1):

VÝSLEDNÁ TABUĽKA				
	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Hotellingov T^2 test	56.544	23.56	2	5
Príslušný kvantil pre test.štatistiku: 5.78614				
zamietam H_0				
T-test pre závislu premennú: 1				
{ Statistic P-Value , The null hypothesis that the mean difference is 0				
T -1.46385 0.19357				
is not rejected at the 5. percent level based on the T test., 0.19357 }				
nezamietam H_0				
T-test pre závislu premennú: 2				
{ Statistic P-Value , The null hypothesis that the mean difference is 0				
T -1.21268 0.270829				
is not rejected at the 5. percent level based on the T test., 0.270829 }				
nezamietam H_0				

tab. 5. 1 Výsledky k príkladu 5.1

Keďže hodnota štatistiky je väčšia ako príslušný kvantil, nulovú hypotézu na hladine významnosti $\alpha = 0.05$ zamietame, t.j. výnosy aj hmotnosti semien závisia od obsahu dusičnanov v hnojive. Pretože nulová hypotéza bola zamietnutá, môžeme pokračovať v testovaní jednotlivých premenných pomocou T-testu zabudovaného v softvéri Mathematica 8.0. Pre náš príklad, prekvapujúco, T-test ani pre jednu premennú, nulovú hypotézu nezamietol.

6. Viacrozmerná analýza rozptylu

Namiesto jedného pozorovania budeme teraz uvažovať vektor p pozorovaní. Pre viacrozmernú analýzu rozptylu usporiadame pozorovania do dátovej matice $\mathbf{Y}$ typu $n \times p$ tak, že riadky matice $\mathbf{y}_{hi}^T = (y_{hi1}, y_{hi2}, \dots, y_{hip})$ zodpovedajú meraniam na i -tej jednotke v h -tej skupine. Skupiny môžu mať vo všeobecnosti rôzny rozsah.

Rovnako ako pri jednorozmernej analýze rozptylu použijeme model

$$\mathbf{y}_{hi} = \boldsymbol{\mu}_h + \boldsymbol{\varepsilon}_{hi} \quad (6.1)$$

alebo

$$\mathbf{y}_{hi} = \boldsymbol{\mu} + \boldsymbol{\alpha}_h + \boldsymbol{\varepsilon}_{hi} \quad (6.2)$$

$$i = 1, 2, \dots, n_h, h = 1, 2, \dots, H,$$

kde $\mathbf{y}_{hi}$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_h$, $\boldsymbol{\alpha}_h$ a $\boldsymbol{\varepsilon}_{hi}$ sú p -rozmerné vektory. $\mathbf{y}_{hi}$ je vektor pozorovaní na i -tej jednotke v h -tej skupine, $\boldsymbol{\mu}$ všeobecná konštanta, $\boldsymbol{\mu}_h$ vektor stredných hodnôt v h -tej skupine, $\boldsymbol{\alpha}_h$ vektor efektov h -tej úrovne skúmaného faktora a $\boldsymbol{\varepsilon}_{hi}$ vektor náhodných zložiek. Predpokladáme nezávislosť $\boldsymbol{\varepsilon}_{hi} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, rovnosť kovariančných matic. Vektor $\mathbf{y}_{hi}$ má potom rozdelenie $N(\boldsymbol{\mu}_h, \boldsymbol{\Sigma})$.

Označme

$$\bar{\mathbf{y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{y}_{hi} \quad (6.3)$$

a

$$\mathbf{S}_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\mathbf{y}_{hi} - \bar{\mathbf{y}}_h)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_h)^T \quad (6.4)$$

vektor priemerov a výberovú kovariančnú maticu v h -tej skupine a

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{h=1}^H \bar{\mathbf{y}}_h n_h \quad (6.5)$$

a

$$\mathbf{S} = \frac{1}{n - H} \sum_{h=1}^H \mathbf{S}_h (n_h - 1) \quad (6.6)$$

celkový vektor priemerov a spoločnú výberovú kovariančnú maticu.

Na overenie hypotézy $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_H$ o rovnosti vektorov stredných hodnôt v H skupinách alebo hypotézu $H_0: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \dots = \boldsymbol{\alpha}_H = 0$ môžeme použiť niekoľko testovacích štatistík. Označme

$$\mathbf{T} = \sum_{h=1}^H \sum_{i=1}^{n_h} (\mathbf{y}_{hi} - \bar{\mathbf{y}})(\mathbf{y}_{hi} - \bar{\mathbf{y}})^T \quad (6.7)$$

maticu popisujúcu celkovú variabilitu a

$$\mathbf{E} = \sum_{h=1}^H \sum_{i=1}^{n_h} (\mathbf{y}_{hi} - \bar{\mathbf{y}}_h)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_h)^T = (n - H)\mathbf{S} \quad (6.8)$$

maticu vnútroskupinovej variability. Medziskupinovú variabilitu vyjadríme pomocou matice

$$\mathbf{B} = \sum_{h=1}^H n_h (\bar{\mathbf{y}}_h - \bar{\mathbf{y}})(\bar{\mathbf{y}}_h - \bar{\mathbf{y}})^T \quad (6.9)$$

pričom platí

$$\mathbf{T} = \mathbf{B} + \mathbf{E}. \quad (6.10)$$

Označme λ_j , $j = 1, 2, \dots, p$, charakteristické čísla matice $\mathbf{B}\mathbf{E}^{-1}$ a $(1 + \lambda_j)$ sú charakteristické čísla matice $\mathbf{T}\mathbf{E}^{-1} = \mathbf{B}\mathbf{E}^{-1} + \mathbf{I}$. Ďalej zoradíme charakteristické čísla tak, že $\lambda_{(1)} \geq \lambda_{(2)} \geq \dots \geq \lambda_{(p)}$ a označme $s = h(\mathbf{B})$. Na s nenulových charakteristických číslach matice $\mathbf{B}\mathbf{E}^{-1}$ sú založené nasledujúce štyri najpoužívanéjšie štatistiky vo viacrozmernej analýze rozptylu:

$$W = \prod_{l=1}^s \frac{1}{1 + \lambda_l} \quad (6.11)$$

$$P = \sum_{l=1}^s \frac{\lambda_l}{1 + \lambda_l} \quad (6.12)$$

$$LH = \sum_{l=1}^s \lambda_l \quad (6.13)$$

$$R = \lambda_{max} \quad (6.14)$$

Štatistiky je možné transformovať na veličiny, ktoré majú približne F-rozdelenie. Výpočet transformovaných testovacích štatistík a ich stupňov voľnosti je uvedený v nasledujúcej tabuľke (tab. 6. 1).

Štatistika	F_{approx}	v_1	v_2
Pillai	$\frac{v_2 P}{v_1 (s - P)}$	$s(2m_p + s + 1)$	$s(2n_p + s + 1)$
Wilks	$\frac{v_2}{v_1} \frac{1 - W^{1/t}}{W^{1/t}}$	pv_H	$ft - \frac{pv_H}{2} + 1$
Lawley-Hotelling	$\frac{v_2}{sv_1} LH$	pv_H	$s(v_E - p - 1) + 2$
Roy	$\frac{v_2}{v_1} R$	$2m_p + s + 1$	$2n_p + s + 1$

tab. 6. 1 Súhrnná tabuľka MANOVA

Pri výpočte transformovaných hodnôt F_{approx} a stupňov voľnosti v_1 a v_2 využijeme vzťahy

$$s = \min(p, v_H), \quad (6.15)$$

$$m_p = \frac{|p - v_H| - 1}{2}, \quad (6.16)$$

$$n_p = \frac{v_E - p - 1}{2}, \quad (6.17)$$

$$f = v_E - \frac{p + 1 - v_H}{2}, \quad (6.18)$$

$$v_H = H - 1, \quad (6.19)$$

$$v_E = n - H, \quad (6.20)$$

$$t = \begin{cases} 1, & \text{inak} \\ \sqrt{\frac{p^2 v_H^2 - 4}{p^2 + v_H^2 - 5}}, & \text{ak } p^2 + v_H^2 - 5 > 0 \end{cases} \quad (6.21)$$

Testovanú hypotézu zamietneme na hladine významnosti α , ak hodnota testovanej štatistiky F je väčšia ako kvantil $F_{1-\alpha}(v_1, v_2)$.

Ak viacrozmerný test zamietne testovanú hypotézu H_0 , to znamená, že existujú rozdiely medzi skupinami, a teda môžeme pokračovať v podrobnejšej analýze [2].

Príklad 6.1

Testovali sme zloženie lieku rovnakého názvu vyrábaného v troch rôznych farmaceutických závodoch. Zisťoval sa obsah niacínu (X_1 , mg), obsah kyseliny DHA (X_2 , mg), obsah biotínu (X_3 , μg), obsah kyseliny listovej (X_4 , μg) a obsah pantoténu (X_5 , mg). K dispozícii máme 60 vzoriek – 15 vzoriek zo závodu 1, 20 vzoriek zo závodu 2 a 25 vzoriek zo závodu 3. Treba overiť na hladine významnosti $\alpha = 0.05$, či je štatisticky významný rozdiel v lieku z uvedených závodov.

Na overenie hypotézy o rovnosti vektora stredných hodnôt viacrozmerných dát bol použitý program vytvorený v softvéri Mathematica 8.0.

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Pillai	1.29399	19.7943	10	108
Hotelling	6.202	32.2504	10	104
Roy	5.389	58.2012	5	54
Wilks	0.0862159	25.5004	10	106

Príslušný kvantil pre Pillaiovu test.štatistiku: 1.91947

zamietam H_0

Príslušný kvantil pre Hotellingovu test.štatistiku: 1.92294

zamietam H_0

Príslušný kvantil pre Royovu test.štatistiku: 2.38607

zamietam H_0

Príslušný kvantil pre Wilksovu test.štatistiku: 1.92117

zamietam H_0

tab. 6. 2 Výsledná tabuľka k príkladu 6.1

Po porovnaní štatistík s príslušnými kvantilmi F rozdelenia zamietame nulovú hypotézu, t.j. medzi liekmi z uvedených závodov je štatisticky významný rozdiel. Ak zamietame hypotézu H_0 o rovnosti všetkých stredných hodnôt, je potrebné zistiť, ktoré dvojice stredných hodnôt sa líšia.

6.1. Testovanie rovnosti dvojíc vektorov stredných hodnôt

Po zamietnutí nulovej hypotézy v testoch viacrozmernej analýzy rozptylu, treba testovať, ktoré dvojice skupín sa od seba výrazne odlišujú.

Testujeme teda hypotézu:

$$H_0: \mu_i = \mu_j,$$

že stredné hodnoty μ_i a μ_j (stredné hodnoty sú p -rozmerné vektory) sa nelíšia ($i, j = 1, 2, \dots, H, i < j$).

Testovacia štatistika má tvar:

$$F_{ij} = \frac{n - H - p + 1}{p(H - 1)} U_{ij}^2 \quad (6. 1. 1)$$

Hypotézu H_0 zamietame, ak testovacia štatistika je väčšia ako kvantil $F_{1-\alpha}(v_1, v_2)$.

Pre stupne voľnosti platí, že

$$v_1 = \frac{(H - 1)(n - H - p)p}{n - 2 - pH + p} \quad (6. 1. 2)$$

$$v_2 = n - H - p + 1. \quad (6. 1. 3)$$

Pre U_{ij} v testovacej štatistike (6. 1. 1) platí

$$U_{ij}^2 = \frac{n_i n_j}{n_i + n_j} (\bar{y}_i - \bar{y}_j)' E^{-1} (\bar{y}_i - \bar{y}_j) \quad (6. 1. 4)$$

[9].

Pokračovanie príkladu 6. 1

Keďže všetky štyri základné testovacie štatistiky multivariačnej analýzy rozptylu zamietli nulovú hypotézu o rovnosti vektorov stredných hodnôt, pokračujeme v testovaní jednotlivých dvojíc vektorov. Výsledky týchto testov nasledujúce:

TESTOVANIE DVOJÍC

Test pre dvojicu skupin: 1,2

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	4.35877	10.8333	10.8333

Príslušný kvantil pre test.štatistiku: 1.98079

zamietam H_0

Test pre dvojicu skupin: 1,3

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	20.7728	10.8333	10.8333

Príslušný kvantil pre test.štatistiku: 1.98079

zamietam H_0

Test pre dvojicu skupin: 2,3

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	20.6204	10.8333	10.8333

Príslušný kvantil pre test.štatistiku: 1.98079

zamietam H_0

Po porovnaní hodnôt výsledných testovacích štatistík s príslušnými kvantilmi, nulovú hypotézu zamietame, t.j. medzi jednotlivými dvojicami liekov z rôznych závodov je štatisticky významný rozdiel.

6.2. Testovanie veličín, ktoré zapríčiňujú rozdielnosť stredných hodnôt

V ďalšej časti úlohy sa rieši problém určenia veličín, ktoré zapríčiňujú rozdielnosť stredných hodnôt. Zaujímame sa teda o to, ktoré z p zložiek vektorov $\mu_1, \mu_2, \dots, \mu_H$ sa líšia.

Testujeme hypotézu:

H_0 : i -ta zložka ($i = 1, 2, \dots, p$) nespôsobuje rozdielnosť vektorov stredných hodnôt

$\mu_1, \mu_2, \dots, \mu_H$.

Testovacia štatistika má tvar:

$$\chi^2 = \left(1 - n + \frac{p + H}{2}\right) \ln \frac{e_{ii}}{t_{ii}}, \quad (6.2.1)$$

kde e_{ii} sú diagonálne prvky matice E (6. 8) a t_{ii} sú diagonálne prvky matice T (6. 10).
 Nulovú hypotézu zamietame, ak testovacia štatistika je väčšia ako kvantil $\chi^2_{1-\alpha}(pH - p)$ [9].

Pokračovanie príkladu 6. 1

Poslednou časťou príkladu, je zistiť, ktoré látky obsiahnuté v lieku zapríčiňujú rozdiel. Výsledky testov sú nasledujúce:

TESTOVANIE VELIČÍN

Test veličiny: 1

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	22.064	10.

Príslušný kvantil pre test.štatistiku: 18.307

zamietam H_0

Test veličiny: 2

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	61.9865	10.

Príslušný kvantil pre test.štatistiku: 18.307

zamietam H_0

Test veličiny: 3

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	1.64213	10.

Príslušný kvantil pre test.štatistiku: 18.307

nezamietam H_0

Test veličiny: 4

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	82.1486	10.

Príslušný kvantil pre test.štatistiku: 18.307

zamietam H_0

Test veličiny: 5

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	51.7498	10.

Príslušný kvantil pre test.štatistiku: 18.307

zamietam H_0

Po porovnaní výsledných štatistík s príslušnými kvantilmi rozdelenia, nulovú hypotézu o rovnosti stredných hodnôt jednotlivých veličín zamietame v štyroch prípadoch, t.j. práve tieto štyri látky obsiahnuté v liekoch zapríčínujú rozdiely.

7. Simulačné štúdie pre Hotellingov T^2 test

V tejto simulačnej štúdii chceme testovať úspešnosť Hotellingovho T^2 testu, ktorým testujeme rovnosť vektorov stredných hodnôt dvoch skupín.

Simulácie boli realizované 10 000 krát pre dáta z normálneho rozdelenia a parametre, ako vektory stredných hodnôt a kovariančné matice sa vyberali týmito spôsobmi:

- vektory stredných hodnôt aj kovariančné matice rovnaké
- vektory stredných hodnôt rovnaké, kovariančné matice rôzne
- vektory stredných hodnôt rôzne, kovariančné matice rovnaké
- vektory stredných hodnôt aj kovariančné matice rôzne

Všetky testy boli počítané na hladine významnosti $\alpha = 0.05$. Výsledky simulácií sú spracované v nasledujúcej tabuľke (tab.7. 1).

	$\mu_1 = \mu_2$ $\Sigma_1 = \Sigma_2$	$\mu_1 = \mu_2$ $\Sigma_1 \neq \Sigma_2$	$\mu_1 \neq \mu_2$ $\Sigma_1 = \Sigma_2$	$\mu_1 \neq \mu_2$ $\Sigma_1 \neq \Sigma_2$
Počet zamietnutí	1979	7458	10000	8714
Počet nezamietnutí	8021	2542	0	1286
Percentuálna úspešnosť	80.21	25.42	100	87.14

tab. 7. 1 Výsledky simulácií pre Hotellingov T^2 testu

Výsledky simulácií ukazujú, že test má vysokú úspešnosť v troch simulovaných prípadoch zo štyroch. Najnižšia percentuálna úspešnosť je, keď dáta boli nasimulované s rovnakými vektormi stredných hodnôt a rôznymi kovariančnými maticami – 25.42 %. Úspešnosť 80.21 % je v prípade, ak dáta mali rovnaké aj vektory stredných hodnôt aj kovariančné matice.

V ďalších dvoch prípadoch mali dáta rôzne vektory stredných hodnôt a raz bol splnený predpoklad rovnosti kovariančných matíc a raz nie. V týchto prípadoch bola úspešnosť Hotellingovho T^2 testu ešte vyššia. Ak dáta nemali rovnaké kovariančné matice ani vektory stredných hodnôt, percentuálna úspešnosť testu je 87.14 %. Najvyššia úspešnosť testu je až 100 % a to v prípade, keď dáta mali rôzne vektory stredných hodnôt ale predpoklad rovnosti kovariančných matíc bol splnený.

8. Simulačné štúdie pre viacrozmernú analýzu rozptylu

V nasledujúcej simulačnej štúdii budeme testovať silu multivariačnej analýzy rozptylu (MANOVA), pomocou ktorej testujeme rovnosť vektorov stredných hodnôt dvoch a viacerých skupín. Na testovanie boli použité štyri základné testovacie štatistiky MANOVA, ktorými sú: Pillaiova, Royova, Wilksova a Lawley-Hotellingova štatistika.

Simulácie boli realizované 10 000 krát pre dáta z normálneho rozdelenia a parametre, ako vektory stredných hodnôt a kovariančné matice sa vyberali týmito spôsobmi:

- vektory stredných hodnôt aj kovariančné matice rovnaké
- vektory stredných hodnôt rovnaké, kovariančné matice rôzne
- vektory stredných hodnôt rôzne, kovariančné matice rovnaké
- vektory stredných hodnôt aj kovariančné matice rôzne

Všetky testy boli počítané na hladine významnosti $\alpha = 0.05$. Výsledky simulácií sú spracované v nasledujúcich tabuľkách (tab. 8. 1 – tab. 8. 4).

	$\mu_1 = \mu_2$ $\Sigma_1 = \Sigma_2$			
	Pillai	Hotelling	Roy	Wilks
Počet zamietnutí	380	594	1850	520
Počet nezamietnutí	9620	9406	8150	9480
Percentuálna úspešnosť	96.2	94.06	81.5	94.8

tab. 8. 1 Výsledky testov pre dáta s rovnakými vektormi stredných hodnôt a rovnakou kovariančnou maticou

	$\mu_1 \neq \mu_2$ $\Sigma_1 = \Sigma_2$			
	Pillai	Hotelling	Roy	Wilks
Počet zamietnutí	10000	10000	10000	10000
Počet nezamietnutí	0	0	0	0
Percentuálna úspešnosť	100	100	100	100

tab. 8. 2 Výsledky testov pre dáta s nerovnakými vektormi stredných hodnôt a rovnakou kovariančnou maticou

	$\mu_1 \neq \mu_2$ $\Sigma_1 \neq \Sigma_2$			
	Pillai	Hotelling	Roy	Wilks
Počet zamietnutí	8922	9817	9983	9712
Počet nezamietnutí	1078	183	17	288
Percentuálna úspešnosť	89.22	98.17	99.83	97.12

tab. 8. 3 Výsledky testov pre dáta s nerovnakými vektormi stredných hodnôt aj kovariančnými maticami

	$\mu_1 = \mu_2$ $\Sigma_1 \neq \Sigma_2$			
	Pillai	Hotelling	Roy	Wilks
Počet zamietnutí	1005	1528	2832	1372
Počet nezamietnutí	8995	8472	7168	8628
Percentuálna úspešnosť	89.95	84.72	71.68	86.28

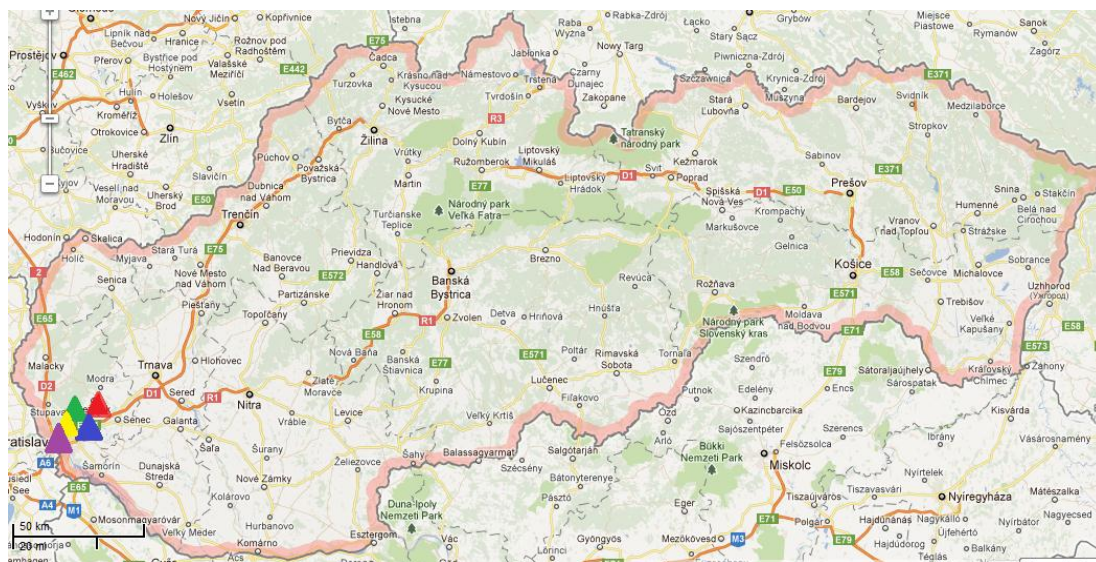
tab. 8. 4 Výsledky testov pre dáta s rovnakými vektormi stredných hodnôt a nerovnakými kovariančnými maticami

Výsledky simulácií ukazujú, že všetky štyri štatistiky majú vysokú úspešnosť v prípadoch, ak je splnený predpoklad rovnosti kovariančných matíc a dáta majú rovnakú strednú hodnotu. V tomto prípade je úspešnosť štatistík 81.5 % - 96.2 %. Ak nie je splnená rovnosť kovariančných matíc, ale dáta majú rovnakú strednú hodnotu, úspešnosť testov je v rozmedzí 71.68 % - 89.95 %. Pre takto nasimulované dáta mali štatistiky najnižšie úspešnosti vôbec.

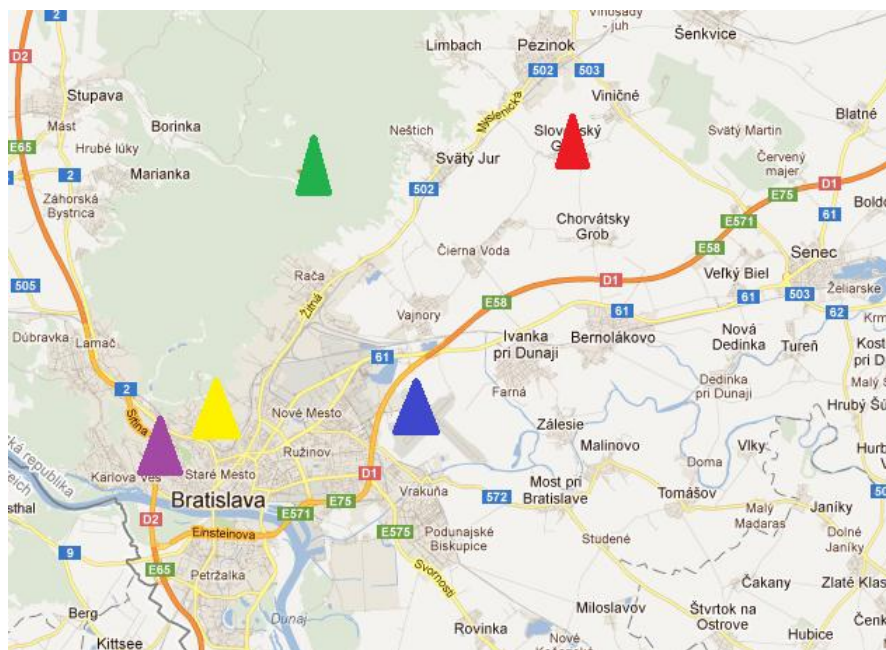
Naopak najvyššie úspešnosti všetkých štyroch štatistík sú v prípade, keď dáta nemajú rovnaké vektory stredných hodnôt, ale predpoklad rovnosti kovariančných matíc je splnený. Úspešnosť všetkých štyroch štatistík je až 100 %. V poslednom prípade, keď dáta nemajú rovnaké ani vektory stredných hodnôt ani kovariančné matice, úspešnosť testov je 89.22 % - 99.83 %.

9. Testovanie reálnych hydrometeorologických dát

V nasledujúcej kapitole budeme testovať reálne hydrometeorologické dáta poskytnuté Slovenským hydrometeorologickým ústavom (SHMÚ) testami popisovanými v kapitole 6 tejto práce. Testované reálne dáta sú z hydrometeorologických staníc: Bratislava - Mlynská Dolina, Malý Javorník, Bratislava - Koliba, Slovenský Grob, Bratislava - letisko. Umiestnenie hydrometeorologických staníc v rámci celého Slovenska zobrazuje obr. 9. 1 a bližšie umiestnenie staníc obr. 9. 2.



obr. 9. 1. Umiestnenie testovaných hydrometeorologických staníc v rámci SR



obr. 9. 2. Bližšie umiestnenie testovaných hydrometeorologických staníc

Legenda:

-  Slovenský Grob
-  Malý Javorník
-  Bratislava – letisko
-  Bratislava - Koliba
-  Bratislava – Mlynská dolina

Poskytnuté dáta sú z rokov 1982 až 2010 a popisujú maximálne denné teploty, minimálne denné teploty, priemerné denné teploty, rýchlosť vetra, úhrn zrážok, výšku snehovej pokrývky a tlak vzduchu. V nasledujúcej tabuľke (tab. 9. 1) uvádzame časť dát spolu s vysvetlivkami.

IndKl	Datum	T_MAX_1	T_MIN_1	T_DEN_PR	VIE_RYCH	ZRA_UHRN	SNEH_PO	TLAK_PR
11810	1.2.1982	2,1	-1,7	-0,8	6	0	0	1004,4
11810	2.2.1982	1,1	-3,9	-2,3	4,7	0	0	1008,3
11810	3.2.1982	0,6	-6,8	-4,4	2	0	0	1010,1
11810	4.2.1982	-0,4	-5,8	-2,8	4	0	0	1014,7
11810	5.2.1982	2	-5,4	-2	4,7	0	0	1013,8
11810	6.2.1982	0,2	-5,2	-4,1	1,3	0,2	0	1005,4
11810	7.2.1982	5,6	-4,9	0,6	1	0	0	1001,8
11810	8.2.1982	0,6	-0,8	-0,1	2,7	0,3	0	996,6
11810	9.2.1982	6	0,3	3	4	0	0	1004,9
11810	10.2.1982	6	-1,8	2,4	2,7	0	0	1011,2
11810	11.2.1982	3,6	-0,7	0,5	4	0	0	1009,6
11810	12.2.1982	1,1	-2,3	-1,6	2,7	0	0	1006,4
11810	13.2.1982	-1,5	-3,1	-1,9	4,3	0	0	1003,1
11810	14.2.1982	-0,4	-2,4	-1,4	5,3	0	0	1000,4
11810	15.2.1982	0	-2,6	-1,4	4,7	0	0	998,4
11810	16.2.1982	4,6	-2,4	1,3	3,3	0	0	1000,5
11810	17.2.1982	5,3	-0,7	1,9	4	0	0	1002,4
11810	18.2.1982	1,8	-1,1	0,6	4	0	0	1002,1
11810	19.2.1982	1,1	-4,9	-2,9	2	0	0	1003,9

tab. 9. 1. Ukážka poskytnutých reálnych hydrometeorologických dát

Vysvetlivky:

IndKl – kód hydrometeorologickej stanice

T_MAX_1 – maximálna denná teplota vzduchu

T_MIN_1 – minimálna denná teplota vzduchu

T_DEN_PR – priemerná denná teplota vzduchu

VIE_RYCH_PR – priemerná denná rýchlosť vetra

ZRA_UHRN_1 – denný úhrn zrážok

SNEH_POKR – výška snehovej pokrývky

TLAK_PR – priemerný denný tlak vzduchu

Poskytnuté dáta predstavovali denné hodnoty spomínaných faktorov, no pre naše testovanie boli tieto dáta upravené na mesačné priemery. Z poskytnutých dát sme využili dáta, ktoré zaznamenávali max. dennú teplotu, min. dennú teplotu, priemernú dennú teplotu, denný úhrn zrážok a výšku snehovej pokrývky.

Keďže programy na viacrozmernú analýzu rozptylu boli vytvorené v softvéri Mathematica 8.0, vstupné dáta musia mať nasledujúci tvar:

```
{{{26.57419355,15.63225806,20.49677419,3.05483871},
{23.01612903,14.06451613,18.12903226,3.476666667},
```

```
{25.82580645,15.98064516,20.5,3.003225806},
{26.75806452,15.80322581,21.1,7.607692308},
{27.2,15.15806452,20.83870968,7.173333333}},
{{24.33548387,13.25483871,18.41612903,1.706451613},
{21.18387097,12.04516129,15.69333333,2.719354839},
{23.54516129,13.50967742,18.18387097,2.341935484},
{24.72580645,13.36774194,18.19032258,2.25483871},
{24.57741935,13.21290323,18.63225806,1.506451613}},
{{28.74193548,17.06774194,22.17741935,3.629032258},
{24.74193548,16.20645161,20.3,2.929032258},
{28.50322581,17.60645161,22.52580645,3.358064516},
{28.85806452,16.96129032,22.63548387,4.1},
{28.69032258,17.11290323,23.16774194,2.977419355}}}
```

Zobrazené dáta sú priemerné hodnoty menovaných hydrometeorologických faktorov za mesiac júl v rokoch 1982 (1. skupina), 1996 (2. skupina) a 2010 (3. skupina). Prvý údaj vo vektore je priemerná maximálna denná teplota, druhý predstavuje priemernú minimálnu dennú teplotu, tretí priemernú dennú teplotu a štvrtý priemerný mesačný úhrn zrážok. Ak by sme testovali niektorý zo zimných mesiacov, piaty údaj by predstavoval priemernú mesačnú výšku snehovej pokrývky.

V tejto kapitole ďalej zobrazíme výstupy programov testov viacrozmernej analýzy rozptylu pre vyššie uvedené dáta.

Pred samotnými testami viacrozmernej analýzy rozptylu musíme najskôr testovať, či sa rovnajú kovariančné matice týchto dát. Rovnosť kovariančných matíc je totiž jeden z predpokladov viacrozmernej analýzy rozptylu. Výsledok testu (tab. 9. 2) je:

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Test Kovariančných Matic	4.33972×10^{-14}	1.48774	20.	516.896
Príslušný kvantil test.štatistiku: 1.59091				
nezamietam H_0				

tab. 9. 2 Výsledok testu pre test rovnosti kovariančných matíc

Test na rovnosť kovariančných matíc overuje hypotézu H_0 , ktorá tvrdí, že kovariančné matice sú zhodné. Podľa výstupu programu tohto testu, hypotézu H_0 nezamietame, t.j. kovariančné matice sú zhodné.

Keďže základný predpoklad analýzy rozptylu je splnený, môžeme prejsť k testom samotnej viacrozmernej analýzy rozptylu. Uvádzame výstupy programov (tab. 9. 3.) spolu s rozhodnutím o zamietnutí alebo nezamietnutí nulovej hypotézy.

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Pillai	1.33192	4.9841	8	20
Hotelling	14.104	14.104	8	16
Roy	13.434	33.585	4	10
Wilks	0.0414855	8.79674	8	18

Príslušný kvantil pre Pillaiovu test.štatistiku: 2.44706

zamietam H0

Príslušný kvantil pre Hotellingovu test.štatistiku: 2.5911

zamietam H0

Príslušný kvantil pre Royovu test.štatistiku: 3.47805

zamietam H0

Príslušný kvantil pre Wilksovu test.štatistiku: 2.51016

zamietam H0

tab. 9. 3. Kompletná tabuľka výsledkov testov viacrozmernej analýzy rozptylu

TESTOVANIE DVOJÍC

Test pre dvojicu skupin: 1,2

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	6.97705	12.8	12.8

Príslušný kvantil pre test.štatistiku: 3.05234

zamietam H0

Test pre dvojicu skupin: 1,3

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	2.24948	12.8	12.8

Príslušný kvantil pre test.štatistiku: 3.05234

nezamietam H0

Test pre dvojicu skupin: 2,3

	hodnota štatistiky	stupeň voľnosti v1	stupeň voľnosti v2
Test dvojíc	14.5758	12.8	12.8

Príslušný kvantil pre test.štatistiku: 3.05234

zamietam H0

TESTOVANIE VELIČÍN

Test veličiny: 1

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	9.12837	8.

Príslušný kvantil pre test.štatistiku: 15.5073

nezamietam H0

Test veličiny: 2

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	23.1653	8.

Príslušný kvantil pre test.štatistiku: 15.5073

zamietam H0

Test veličiny: 3

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	14.2671	8.

Príslušný kvantil pre test.štatistiku: 15.5073

nezamietam H0

Test veličiny: 4

	hodnota štatistiky	stupeň voľnosti 1
Test veličín	6.24018	8.

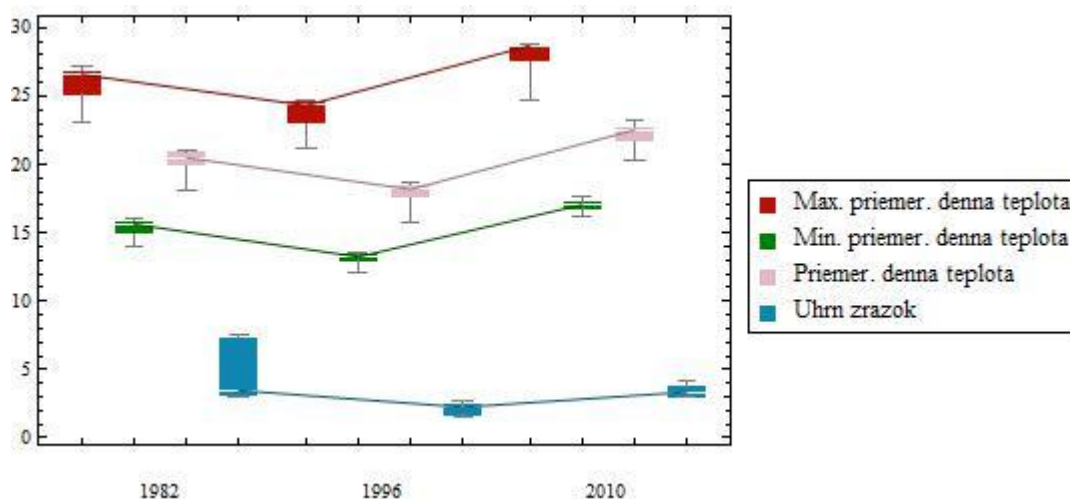
Príslušný kvantil pre test.štatistiku: 15.5073

nezamietam H0

Štyri testy viacrozmernej analýzy rozptylu zamietajú nulovú hypotézu, t.j. vektory stredných hodnôt sa nerovnajú, preto pokračujeme ďalším testovaním. V ďalšom testovaní zisťujeme, ktoré skupiny sa od seba výrazne odlišujú, a ktoré veličiny spôsobujú nerovnosť vektorov stredných hodnôt. Najskôr overujeme platnosť nulovej hypotézy, ktorá tvrdí, že medzi skupinami nie je štatisticky významný rozdiel. Na základe výstupov programu štatisticky významný rozdiel je medzi skupinami 1 a 2 a medzi 2 a 3.

Poslednou časťou testovania je zistiť, ktoré veličiny spôsobujú nerovnosť vektorov stredných hodnôt. Hypotézu H_0 zamietame iba v jednom prípade a to pre veličinu: priemerná denná minimálna teplota.

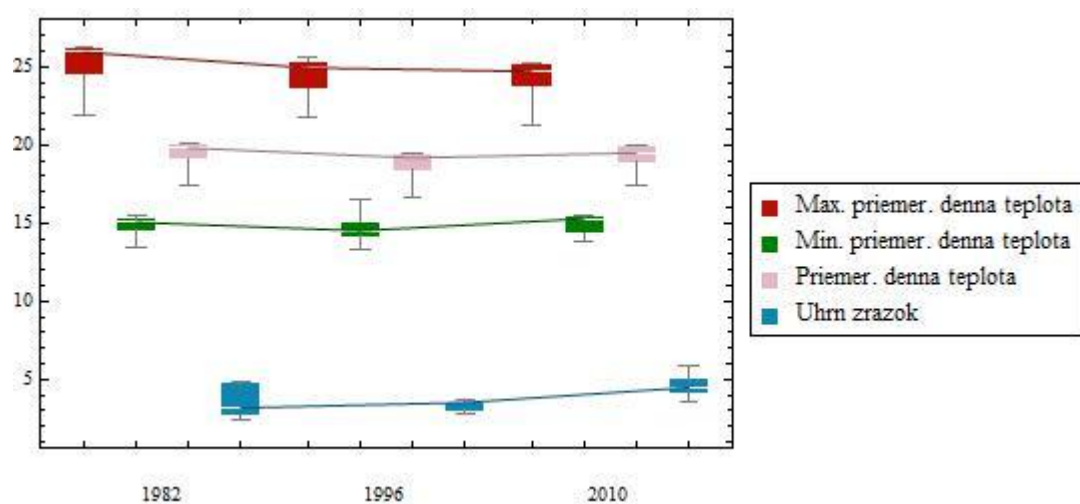
Kvôli názornosti uvádzame Boxove grafy testovaných dát (obr. 9. 3). Z grafu je vidieť, že Boxove grafy neležia na jednej úrovni pre jednotlivé zložky, čo svedčí v neprospech hypotézy H_0 .



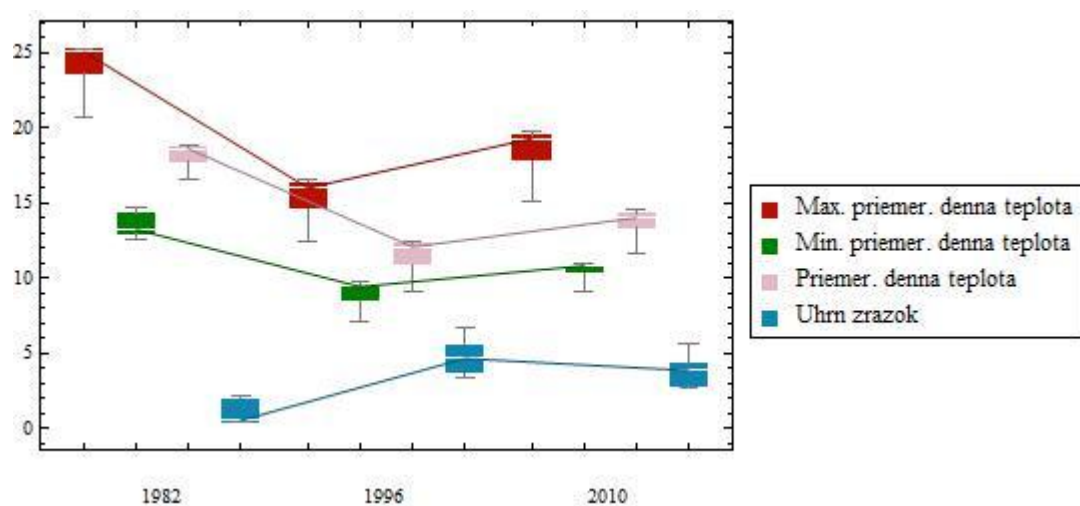
obr. 9. 3 Graf testovaných dát – mesiac júl

Podobne sme testovali aj ostatné mesiace. Nie však pri všetkých môžeme použiť testy MANOVA, pretože test rovnosti kovariančných matic nulovú hypotézu zamietol a teda predpoklad MANOVA nie je splnený. Výsledky testov zobrazuje tabuľka (tab. 9. 4).

Kvôli lepšej názornosti nasledujúce grafy (obr. 9. 4 – 9. 5) zobrazujú vybrané dáta, pre ktoré sa nulová hypotéza viacrozmernej analýzy rozptylu prevažne nezamietla (obr. 9.4) a dáta, pre ktoré sa H_0 prevažne zamietla (obr. 9. 5).



obr. 9. 4. Hydrometeorologické dáta pre mesiac august



obr. 9. 5. Hydrometeorologické dáta pre mesiac september

tab. 9. 4. Výsledky testov pre reálne hydrometeorologické dáta

Mesiac	Test Σ	Manova testy				Testy dvojíc skupín			Testy jednotlivých faktorov				
		Pillai	Wilks	Hotelling	Roy	1 a 2	1 a 3	2 a 3	1	2	3	4	5
1.	zamietam H_0	-	-	-	-	-	-	-	-	-	-	-	-
2.	zamietam H_0	-	-	-	-	-	-	-	-	-	-	-	-
3.	zamietam H_0	-	-	-	-	-	-	-	-	-	-	-	-
4.	nedá sa overiť H_0	-	-	-	-	-	-	-	-	-	-	-	-
5.	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	zamietam H_0	-
6.	nezamietam H_0	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	-
7.	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	nezamietam H_0	nezamietam H_0	-
8.	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	zamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	nezamietam H_0	-
9.	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	-
10.	nezamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	zamietam H_0	nezamietam H_0	zamietam H_0	nezamietam H_0	nezamietam H_0	-
11.	zamietam H_0	-	-	-	-	-	-	-	-	-	-	-	-
12.	zamietam H_0	-	-	-	-	-	-	-	-	-	-	-	-

Vysvetlivky:

Test Σ = Test rovnosti kovariančných matic

faktor 1 = priemerná max. denná teplota

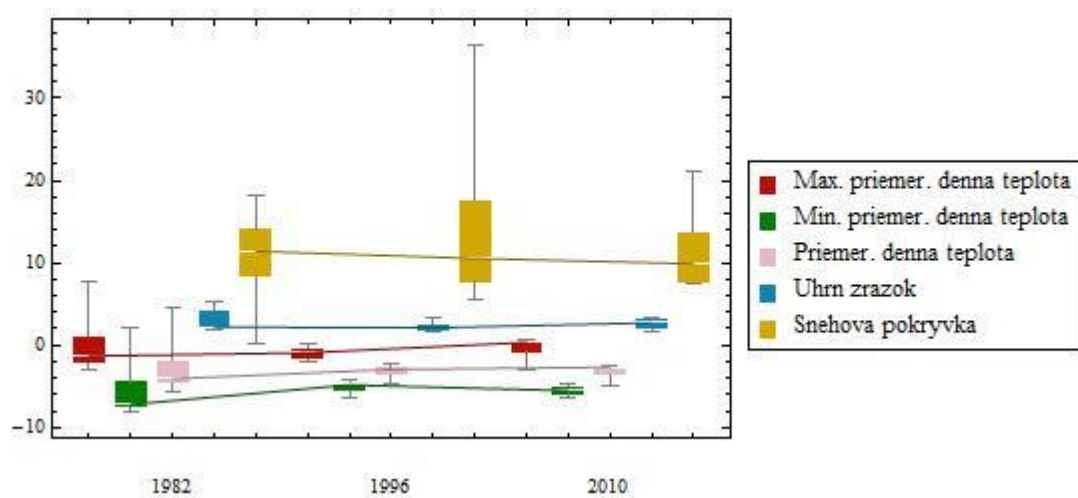
faktor 2 = priemerná min. denná teplota

faktor 3 = priemerná denná teplota

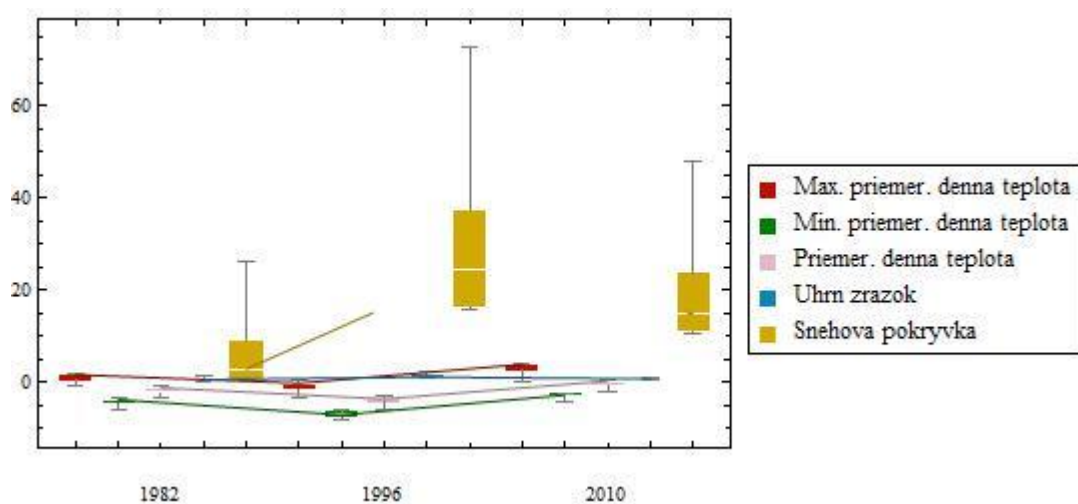
faktor 4 = úhrn zrážok

faktor 5 = výška snehovej pokrývky

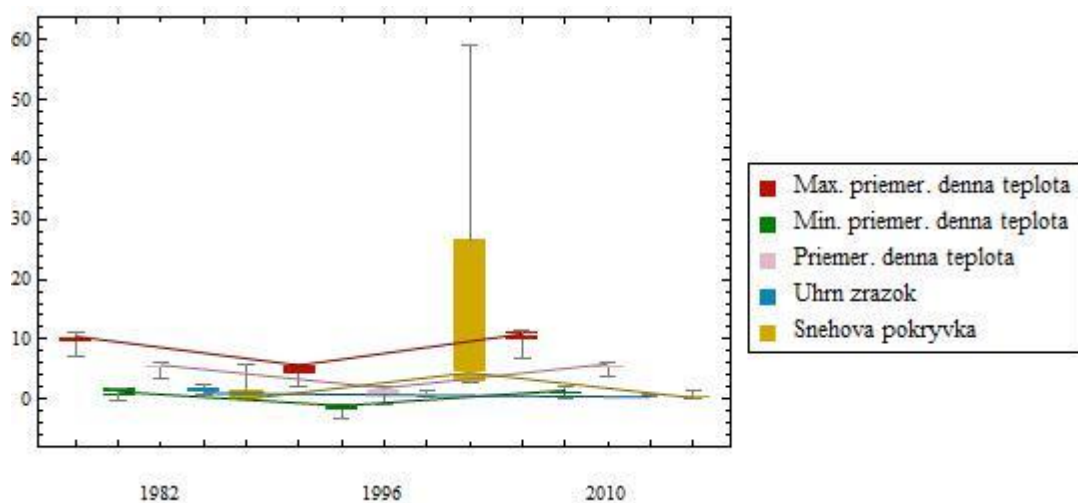
Nasledujúce grafy (obr. 9. 6 – obr. 9. 17) zobrazujú dáta vo všetkých testovaných mesiacoch.



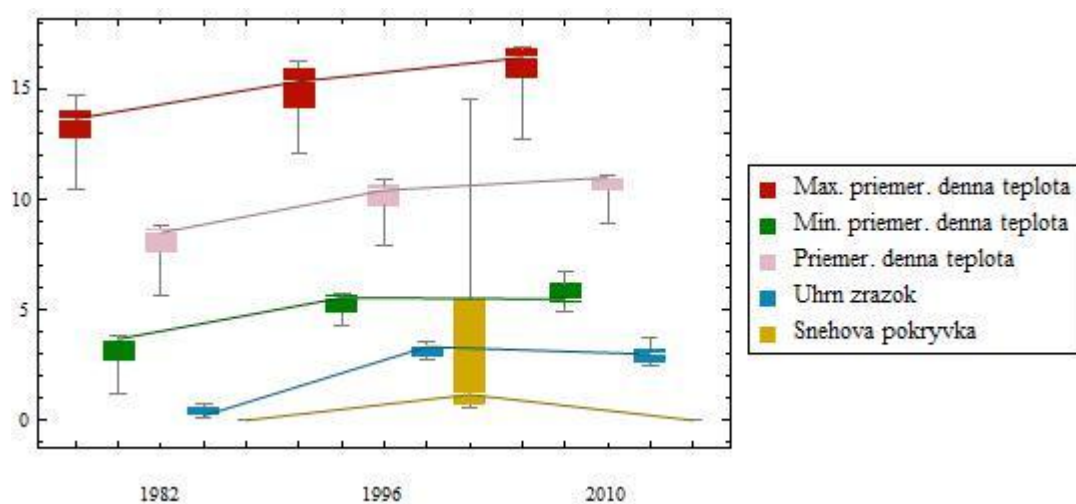
obr. 9. 6. Hydrometeorologické dáta pre mesiac január



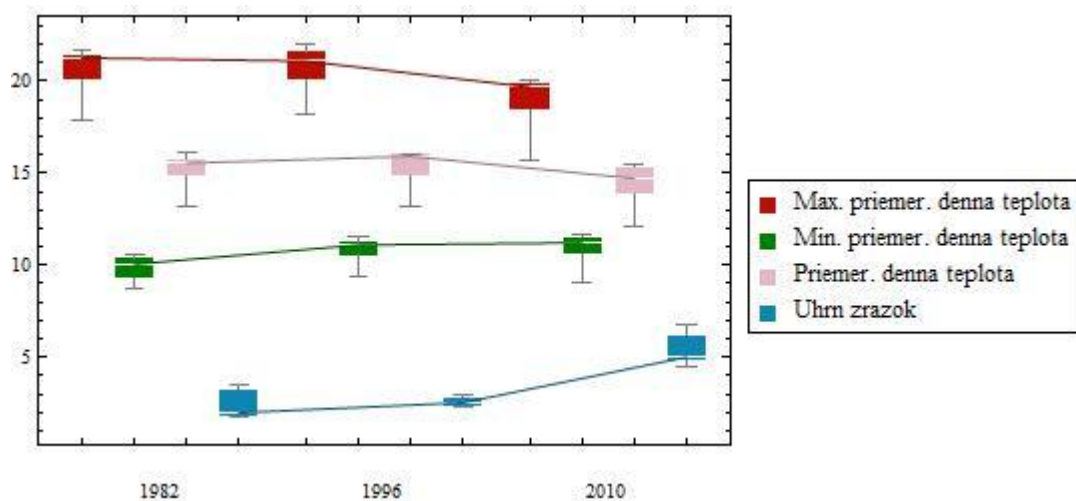
obr. 9. 7. Hydrometeorologické dáta pre mesiac február



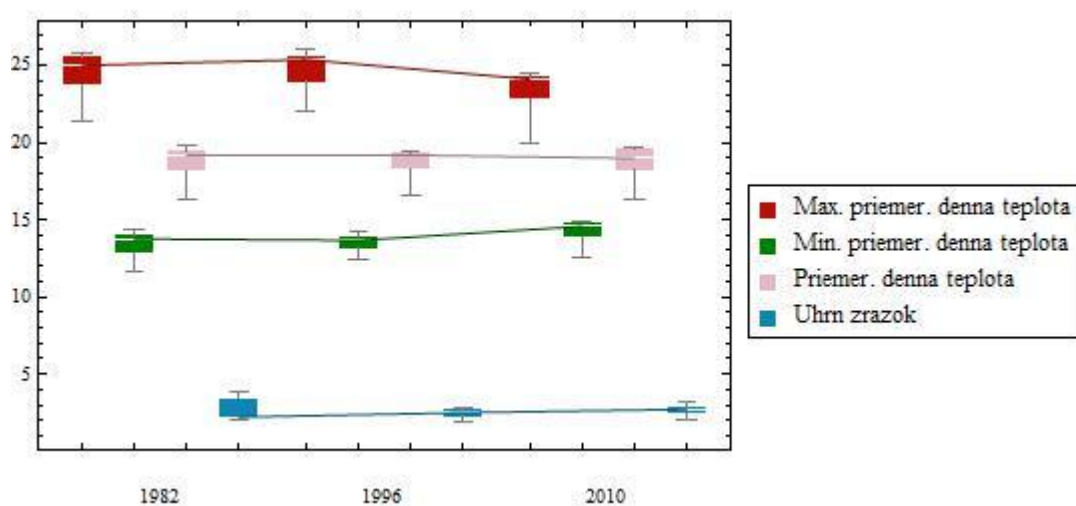
obr. 9. 8. Hydrometeorologické dáta pre mesiac marec



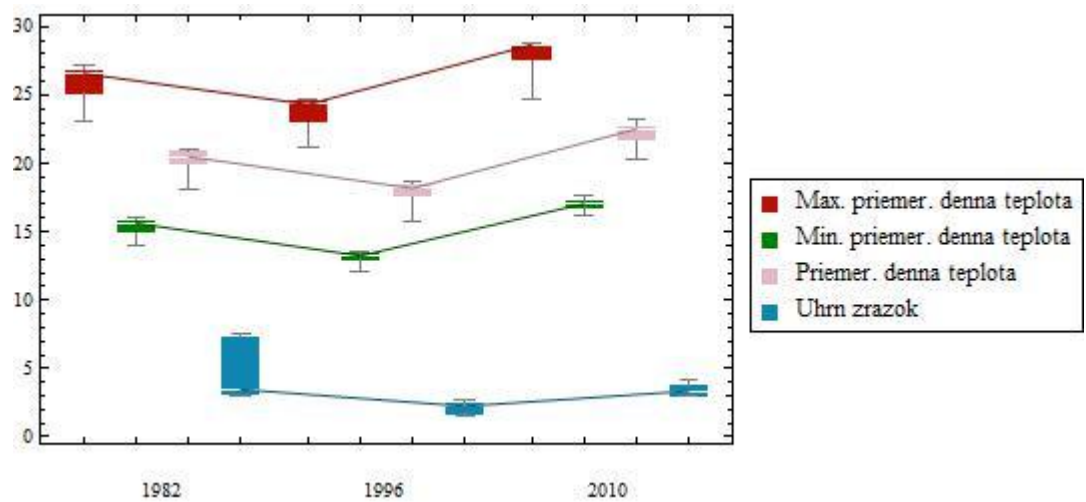
obr. 9. 9. Hydrometeorologické dáta pre mesiac apríl



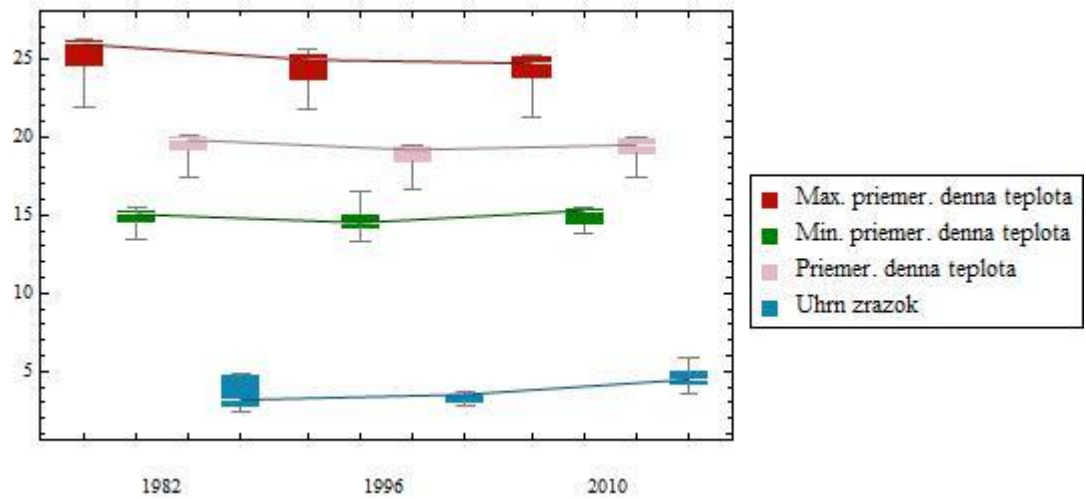
obr. 9. 10. Hydrometeorologické dáta pre mesiac máj



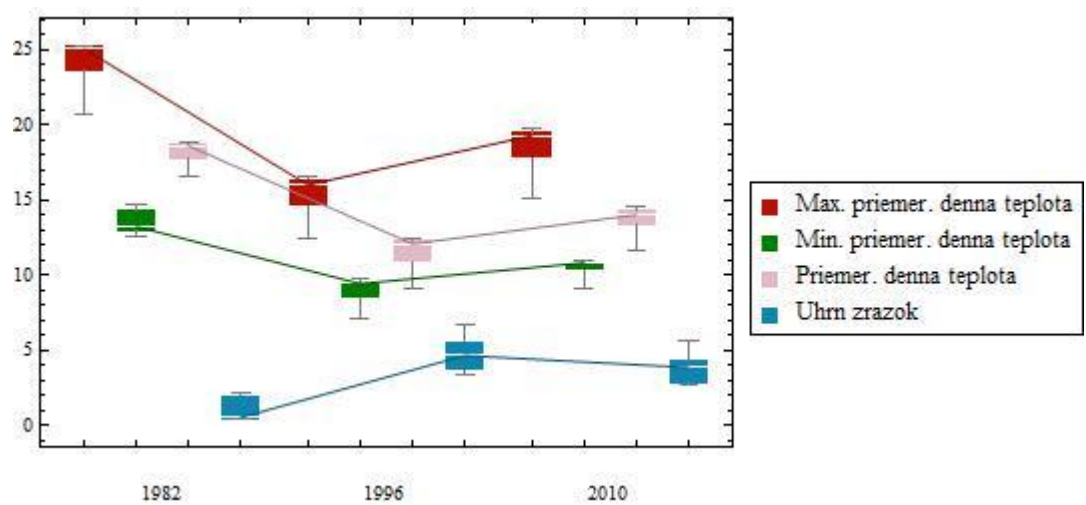
obr. 9. 11. Hydrometeorologické dáta pre mesiac jún



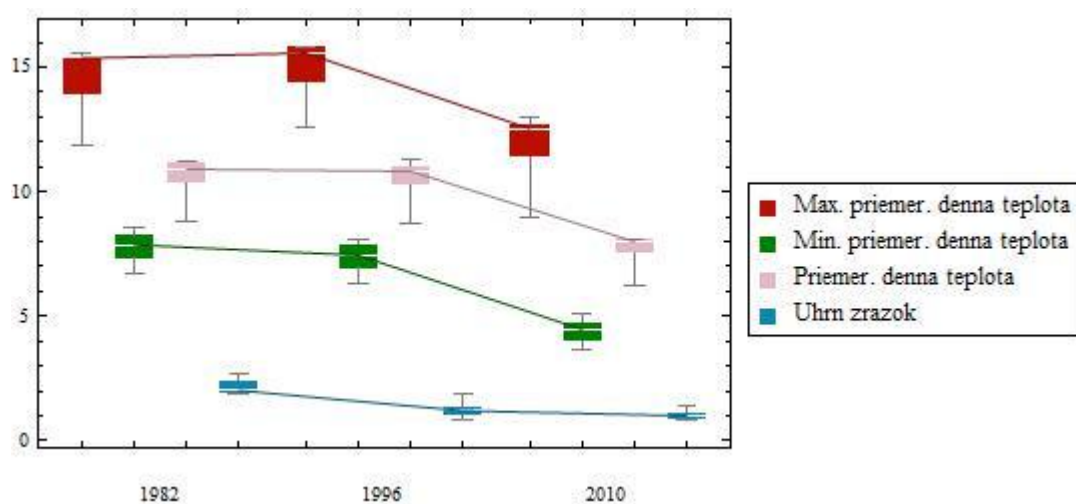
obr. 9. 12. Hydrometeorologické dáta pre mesiac júl



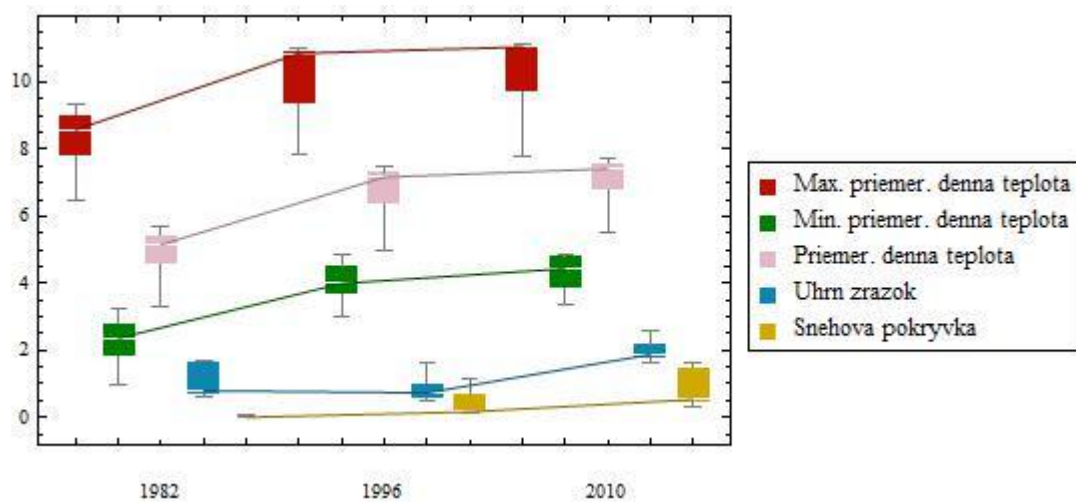
obr. 9. 13. Hydrometeorologické dáta pre mesiac august



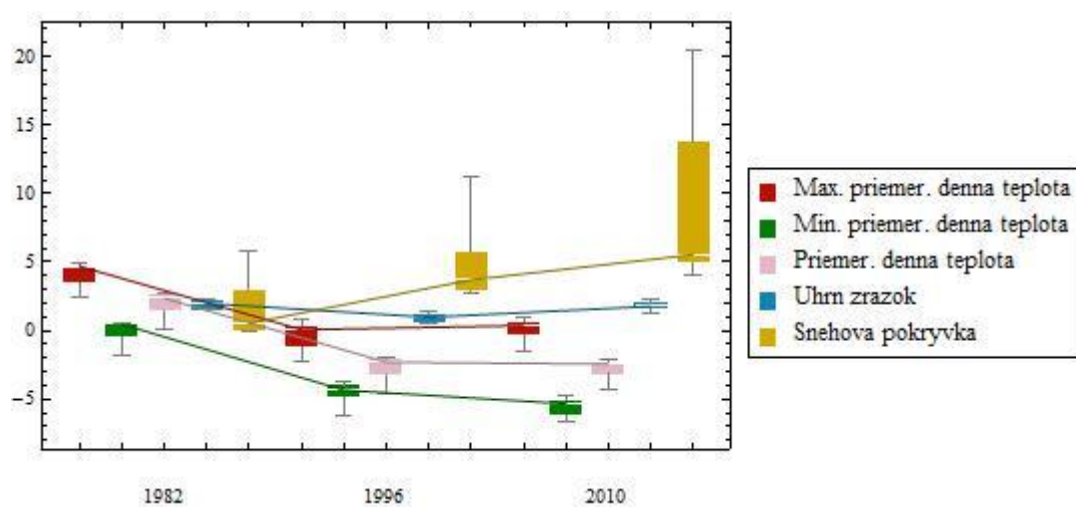
obr. 9. 14. Hydrometeorologické dáta pre mesiac september



obr. 9. 15. Hydrometeorologické dáta pre mesiac október



obr. 9. 16. Hydrometeorologické dáta pre mesiac november



obr. 9. 17. Hydrometeorologické dáta pre mesiac december

10. Viacrozmerná analýza kovariancie

Podobne ako v analýze rozptylu aj v analýze kovariancie existujú viacrozmerné úlohy. V tejto kapitole budeme uvažovať $p > 1$ vysvetľovaných premenných $Y_1, Y_2, \dots, Y_p$. Cieľom je určiť závislosť tejto množiny premenných na (pre jednoduchosť, uvažujme jednom) faktore A . Rovnako ako pri obyčajnej analýze kovariancie, študujeme túto závislosť očistenú od vplyvu množiny doprovodných premenných $X_1, X_2, \dots, X_q$. Testujeme teda hypotézu $H_0: \mu_1 = \mu_2 = \dots = \mu_H$ alebo $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_H = 0$.

Jednotlivé kroky viacrozmernej analýzy kovariancie (MANCOVA):

1. Dátovú maticu upravíme na tvar $[YX]$, t.j. naľavo hodnoty vysvetľovaných premenných a napravo hodnoty doprovodných premenných.
2. Pre takto upravenú maticu typu $n \times (p + q)$ vypočítame postupom obdobným ako v analýze rozptylu, teda podľa vzorcov (6. 7) a (6. 8), maticu T celkovej variability a maticu E reziduálnej variability.
3. Obe tieto matice typu $(p + q) \times (p + q)$ opäť rozdelíme na submatice zodpovedajúce skupine p vysvetľovaných a skupine q doprovodných premenných

$$T = \begin{bmatrix} T(y, y) & T(y, x) \\ T(x, y) & T(x, x) \end{bmatrix}, E = \begin{bmatrix} E(y, y) & E(y, x) \\ E(x, y) & E(x, x) \end{bmatrix} \quad (10. 1)$$

4. Vypočítame maticu všeobecných regresných koeficientov typu $p \times q$

$$B = T(y, x)[T(x, x)]^{-1} \quad (10. 2)$$

a maticu vnútroskupinových regresných koeficientov typu $p \times q$

$$\bar{B} = E(y, x)[E(x, x)]^{-1} \quad (10. 3)$$

5. Vypočítame opravené priemery vysvetľovaných premenných podľa

$$\bar{y}_{hj}^* = \bar{y}_{hj} - \sum_{g=1}^q b_{jg}(\bar{x}_{hg} - \bar{x}_g), \quad h=1,2,\dots,H, \quad j=1,2,\dots,p \quad (10. 4)$$

6. Submaticu $\mathbf{T}(\mathbf{y}, \mathbf{y})$ celkovej variability a submaticu $\mathbf{E}(\mathbf{y}, \mathbf{y})$ reziduálnej variability opravíme na vplyv doprovodných premenných

$$\mathbf{T}^* = \mathbf{T}(\mathbf{y}, \mathbf{y}) - \mathbf{B}\mathbf{T}(\mathbf{x}, \mathbf{y}), \mathbf{E}^* = \mathbf{E}(\mathbf{y}, \mathbf{y}) - \bar{\mathbf{B}}\mathbf{E}(\mathbf{x}, \mathbf{y}) \quad (10.5)$$

7. Určíme hodnotu Wilksovej štatistiky pre opravené submatice

$$\Lambda^* = \frac{|\mathbf{E}^*|}{|\mathbf{T}^*|} \quad (10.6)$$

Nízke hodnoty štatistiky Λ^* svedčia v prospech závislosti množiny premenných $\mathbf{y}$ na faktore A.

8. Rozdelenie štatistiky Λ^* aproximujeme F rozdelením. Podobne ako v kapitole 6 dostávame:

$$t = \sqrt{\frac{p^2(H-1)^2 - 4}{p^2 + (H-1)^2 - 5}} \quad (10.7)$$

$$m = n - \frac{p+H}{2} - q - 1, \quad (10.8)$$

$$v_1 = p(H-1), \quad (10.9)$$

$$v_2 = mt - \frac{p(H-1)}{2} + 1 \quad (10.10)$$

9. Vypočítame hodnotu testovacieho kritéria

$$F = \frac{v_2}{v_1} \frac{1 - \sqrt[t]{\Lambda^*}}{\sqrt[t]{\Lambda^*}} \quad (10.11)$$

Ak hodnota testovacieho kritéria prekročí kvantil $F_{1-\alpha}(v_1, v_2)$, pokladáme závislosť množiny premenných $\mathbf{y}$ na faktore A za dokázanú na danej hladine významnosti α [3].

Príklad 10. 1

Príklad bol inšpirovaný článkom [7]. Pre hypotetické ženské basketbalové družstvá sme simulovali množinu dát výsledkov výkonnostných testov, počas ktorých boli merané: šprint na 20 m, šprint 20 m – driblovanie, hod basketbalovou loptou, hod medicimbalom, 6 x 5 m šprint, 6 x 5 m šprint – driblovanie, obranný výskok, útočný výskok. Hráčky boli rozdelené do troch skupín podľa ich hráčskej pozície. Prvú skupinu tvorilo 35 obrankýň, druhú 14 útočníčok a tretiu skupinu 16 stredových hráčok. Za doprovodné premenné boli považované telesná výška a telesná hmotnosť. Chceli sme vedieť, či výsledky výkonnostných testov sú pre všetky typy hráčok rovnaké, alebo existuje závislosť medzi týmito výsledkami a hráčskou pozíciou.

Príklad bol počítaný pomocou programov na výpočet multivariačnej analýzy rozpylu a multivariačnej analýzy kovariancie vytvorených v softvéri Mathematica 8.0. Výsledky zobrazujú nasledujúce tabuľky (tab. 10. 1 – tab. 10. 2).

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Pillai	1.02303	5.65459	20	108
Hotelling	4.497	11.6922	20	104
Roy	4.224	22.8096	10	54
Wilks	0.150373	8.36758	20	106

Príslušný kvantil pre Pillaiovu test.štatistiku: 1.66854

zamietam H_0

Príslušný kvantil pre Hotellingovu test.štatistiku: 1.67233

zamietam H_0

Príslušný kvantil pre Royovu test.štatistiku: 2.01118

zamietam H_0

Príslušný kvantil pre Wilksovu test.štatistiku: 1.6704

zamietam H_0

tab. 10. 1 Výsledky MANOVA pre príklad 10.1

Po vypočítaní štatistík multivariačnej analýzy rozptylu a ich porovnaní s príslušnými kvantilmi, nulovú hypotézu zamietame. Výsledky výkonnostných testov závisia od hráčskej pozície.

V predchádzajúcom teste sme neuvažovali vplyv telesnej výšky a váhy hráčov na ich výsledky vo výkonnostných testoch. Predpokladá sa, že každá hráčska pozícia v basketbale si vyžaduje hráča s vhodnou výškou aj váhou. Predpokladajme teda, že medzi telesnou váhou a výškou a výsledkami výkonnostných testov existuje závislosť. Preto, na testovanie rovnosti vektorov skupinových priemerov použijeme multivariačnú analýzu kovariancie, ktorá odstráni vplyv doprovodných premenných, t.j. vplyv telesnej váhy a telesnej výšky.

VÝSLEDNÁ TABUĽKA

	hodnota štatistiky	F	stupeň voľnosti v1	stupeň voľnosti v2
Mancova	0.504676	9.32493	16.	366.
Príslušný kvantil pre test.štatistiku: 1.6712				
zamietam H0				

tab. 10. 2 Výsledky MANCOVA pre príklad 10.1

Po porovnaní štatistiky s príslušným kvantilom F rozdelenia zamietame nulovú hypotézu, t.j. aj po odstránení vplyvu doprovodných premenných sa vektory skupinových priemerov nerovnajú a teda výsledky výkonnostných testov závisia od hráčskej pozície.

11. Záver

Predkladaná práca popisuje viacrozmernú analýzu rozptylu (MANOVA), ktorá umožňuje porovnať stredné hodnoty viacerých viacrozmerných výberov. Podobne ako jednorozmerná analýza rozptylu má MANOVA predpoklady, ktoré musia testované dáta spĺňať. Jedným z týchto predpokladov je homoskedasticita, teda rovnosť rozptylov. Na overenie tohto predpokladu MANOVA používa test rovnosti kovariančných matíc, ktorý je bližšie v práci popísaný a ukázaný na príklade. Ak testované dáta pozostávajú z dvoch skupín, na testovanie rovnosti vektora stredných hodnôt používame Hotellingov T^2 test, ktorý je tiež v práci popísaný. Jeho úspešnosť bola testovaná simulačnými štúdiami. Simulačné štúdie ukázali, že úspešnosť testu je v troch prípadoch zo štyroch simulovaných prípadov pomerne vysoká. V prípade, že dáta pozostávajú z viac ako dvoch skupín na testovanie rovnosti stredných hodnôt používame testy viacrozmernej analýzy rozptylu. Úspešnosti týchto testov boli tiež testované simuláciami. Úspešnosť testov je vysoká, dokonca v prípade, ak dáta neboli simulované s rovnakými vektormi stredných hodnôt, no rovnosť kovariančných matíc bola splnená, percentuálna úspešnosť všetkých štyroch testov je 100 %. Testy MANOVA boli použité aj na reálnych hydrometeorologických dátach, ktoré boli poskytnuté SHMÚ v Bratislave. Za poskytnutie týchto dát ďakujeme. Testovaniu hydrometeorologických dát je venovaná 9. kapitola práce, v ktorej môžeme nájsť tabuľku všetkých výsledkov. V poslednej kapitole práce sme sa zaoberali testom multivariačnej analýzy kovariancie, ktorý sme prakticky ukázali aj na príklade z oblasti športu.

Všetky spomínané testy sú podrobnejšie popísané v teoretickej časti a prakticky vyskúšané na príkladoch, ktoré sú v záveroch každej kapitoly. Samotné sú naprogramované v softvéri Mathematica 8.0.

12. Použitá literatúra

1. FOJTÍKOVÁ, I.: Analýza roptylu a kovariancie, bakalárska práca. Bratislava, STU 2010
2. HEBÁK, P. a kol.: Vícerozměrné statistické metody [1]. Praha, Informatorium 2004
3. HEBÁK, P. a kol.: Vícerozměrné statistické metody [2]. Praha, Informatorium 2004
4. RAO, C.: Lineární metody statistické indukce a jejich aplikace. Praha, Academia 1978
5. RIMARČÍK, M.: Štatistika pre prax. Marián Rimarčík 2007.
6. <http://atrey.karlin.mff.cuni.cz>
7. <http://ibgwww.colorado.edu>
8. <http://www2.imm.dtu.dk>
9. <http://www.kirp.chtf.stuba.sk>
10. <http://www.wikipedia.org>