UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Mnohorozměrná analýza rozptylu a její aplikace

Vedoucí diplomové práce: **doc. RNDr. Eva Fišerová Ph.D.**
Vypracovala: **Bc. Eliška Calábková**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Eliška Calábková

Název práce: Mnohorozměrná analýza rozptylu a její aplikace

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová Ph.D.

Rok obhajoby práce: 2017

Abstrakt: V diplomové práci je popsána metoda mnohorozměrné analýzy rozptylu, post-hoc testy a možnosti ověřování předpokladů mnohorozměrné analýzy rozptylu. Praktické použití této metody je demonstrováno na reálných datech pomocí softwaru R.

Klíčová slova: mnohorozměrná analýza rozptylu, MANOVA, post-hoc testy, Boxův test

Počet stran: 76

Počet příloh: 8

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Eliška Calábková

Title: Multivariate analysis of variance and its aplication

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová Ph.D.

The year of presentation: 2017

Abstract: The thesis describes the method of multivariate analysis of variance, post-hoc tests and options of verifying the assumptions of multivariate analysis of variance. Practical application of this method is demonstrated on real data sets using software R.

Key words: multivariate analysis of variance, MANOVA, post-hoc tests, Box's test

Number of pages: 76

Number of appendices: 8

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové Ph.D. s použitím uvedené literatury.

V Olomouci dne 21. dubna 2017 .....

podpis

Obsah

Poděkování

Ráda bych poděkovala vedoucí diplomové práce paní doc. RNDr. Evě Fišerové Ph.D. za spolupráci i za čas, který mi věnovala při konzultacích.

Úvod

Tématem mé diplomové práce je popsat metodu mnohorozměrné analýzy rozptylu, která porovnává skupiny pozorování, kdy pro každé pozorování byly zaznamenány hodnoty několika proměnných. Jde o zobecnění jednorozměrné analýzy rozptylu, která skupiny srovnává pouze v jedné proměnné. Metoda mnohorozměrné analýzy rozptylu tedy slouží k ověření, zda jsou pro dané skupiny pozorování střední hodnoty několika proměnných stejné.

V první kapitole je popsán model mnohorozměrné analýzy rozptylu s jedním faktorem a dvěma faktory a následně testy hypotéz vztahující se k definovaným modelům. Druhá kapitola obsahuje testy, které dále provádíme pokud pomocí metody mnohorozměrné analýzy rozptylu zjistíme rozdíly mezi skupinami. Jde o testování rozdílů skupin v jednotlivých proměnných, testování dvojic skupin, testování jednotlivých proměnných pro dvojice skupin a testování mnohorozměrných kontrastů, které slouží k ověření složitějších vztahů mezi skupinami. V předposlední kapitole jsou uvedeny předpoklady modelu mnohorozměrné analýzy rozptylu. Všechny předpoklady jsou popsány a pro předpoklad normality a shody variančních matic jsou uvedeny i postupy testů, které slouží k jejich ověření. V poslední kapitole jsou uvedeny reálné příklady řešené pomocí softwaru R a simulace, které budou sloužit k prozkoumání chování mnohorozměrné analýzy rozptylu a post-hoc testů z druhé kapitoly.

1. Mnohorozměrná analýza rozptylu

Pomocí jednorozměrné analýzy rozptylu testujeme, zda jsou střední hodnoty sledované proměnné v několika skupinách stejné. Mnohorozměrná analýza rozptylu představuje zobecnění jednorozměrné analýzy rozptylu, které spočívá v tom, že závislá proměnná je vícerozměrná. Na každém zkoumaném objektu tedy měříme více znaků.

Mnohorozměrnou analýzu rozptylu lze využít v případě, kdy máme data, která se skládají z mnohorozměrných pozorování, ta jsou rozdělena do H skupin na základě nějaké nominální proměnné. Jednotlivé skupiny jsou nezávislé. Pozorování pocházejí z p-rozměrných normálních rozdělení se shodnými variančními maticemi, a našim cílem je ověřit hypotézu, že se střední hodnoty zkoumaných p proměnných v jednotlivých skupinách neliší. Jinými slovy otestovat, že hodnoty p-rozměrných pozorování nezávisí na tom, ze které skupiny pozorování pochází. Mějme například tři skupiny studentů. V první skupině jsou studenti, kteří spí denně 6 hodin, v druhé jsou studenti, kteří spávají 8 hodin a ve třetí 10 hodin, sledujeme jejich známky z šesti předmětů: z matematiky, češtiny, angličtiny, biologie, chemie a fyziky. Pomocí mnohorozměrné analýzy zkoumáme, zda má délka spánku vliv na výsledné známky z šesti předmětů.

Nominální proměnnou, podle které jsou pozorování rozdělována do skupin, označujeme jako faktor. V našem příkladě je faktorem délka spánku. Na základě počtu faktorů rozlišujeme mnohorozměrnou analýzu rozptylu s jedním nebo více faktory. V této kapitole bude popsána mnohorozměrná analýza rozptylu s jedním a dvěma faktory.

Místo dlouhého názvu mnohorozměrná analýza rozptylu se běžně používá označení MANOVA, které je zkratkou anglického názvu *Multivariate Analysis of Variance* pro tuto metodu . Při tvorbě této kapitoly jsem vycházela především ze zdrojů [1], [2], [3], [4] a [5].

1.1. MANOVA s jedním faktorem

Základní situace MANOVy s jedním faktorem je následující. Uvažujeme p-rozměrné vektory pozorování, kdy p odpovídá počtu sledovaných proměnných. Pozorování uspořádáme do datové matice $\mathbf{Y}$ o rozměrech $n \times p$, kde n značí celkový rozsah souboru. Tudíž řádky matice $\mathbf{Y}$ jsou tvořené p-rozměrnými vektory $\mathbf{y}_{hi}^T = (y_{hi1}, y_{hi2}, \ldots, y_{hip})$, kde $\mathbf{y}_{hi}$ představuje vektor pozorování pro i-tou jednotku v h-té skupině, $i = 1, 2, \ldots, n_h$, $h = 1, 2, \ldots, H$, n_h značí rozsah h-té skupiny a H označuje celkový počet skupin. Rozsahy skupin n_h jsou obecně různé. Sloupce matice $\mathbf{Y}$ tvoří hodnoty p proměnných. Matice $\mathbf{Y}$ je horizontálně členěna na H submatic odpovídající jednotlivým úrovním zkoumaného faktoru, tedy skupinám. Schéma dat, na která se dá použít metoda mnohorozměrné analýzy rozptylu, je ukázáno v tabulce 1, kde pravá část tabulky představuje matici $\mathbf{Y}$.

první skupina	$\mathbf{y}_{11}^T = (y_{111}, y_{112}, \ldots, y_{11p})$ $\mathbf{y}_{12}^T = (y_{121}, y_{122}, \ldots, y_{12p})$ $\vdots$ $\mathbf{y}_{1n_1}^T = (y_{1n_11}, y_{1n_12}, \ldots, y_{1n_1p})$
druhá skupina	$\mathbf{y}_{21}^T = (y_{211}, y_{212}, \ldots, y_{21p})$ $\mathbf{y}_{22}^T = (y_{221}, y_{222}, \ldots, y_{22p})$ $\vdots$ $\mathbf{y}_{2n_2}^T = (y_{2n_21}, y_{2n_22}, \ldots, y_{2n_2p})$
$\vdots$	$\vdots$
H-tá skupina	$\mathbf{y}_{H1}^T = (y_{H11}, y_{H12}, \ldots, y_{H1p})$ $\mathbf{y}_{H2}^T = (y_{H21}, y_{H22}, \ldots, y_{H2p})$ $\vdots$ $\mathbf{y}_{Hn_H}^T = (y_{Hn_H1}, y_{Hn_H2}, \ldots, y_{Hn_Hp})$

Tabulka 1: Schéma dat připravených na použití mnohorozměrné analýzy rozptylu.

Abychom mohli formulovat základní model MANOVy s jedním faktorem, mu-

síme nejprve definovat několik pojmů a stanovit předpoklady, které jsou na daný model kladeny. Nechť vektory $\boldsymbol{\mu}$, $\boldsymbol{\mu}_h$, $\boldsymbol{\alpha}_h$ a $\boldsymbol{\varepsilon}_{hi}$ jsou p-rozměrné vektory, kde $\boldsymbol{\mu}$ je vektor celkových průměrů, $\boldsymbol{\mu}_h$ je vektor průměrů v h-té skupině, $\boldsymbol{\alpha}_h$ je vektor efektů h-té úrovně zkoumaného faktoru a $\boldsymbol{\varepsilon}_{hi}$ je vektor náhodných složek pro i-té pozorování v h-té skupině. Předpokládáme nezávislost mezi vektory náhodných složek $\boldsymbol{\varepsilon}_{hi}$ a že náhodné složky pocházejí z p-rozměrných normálních rozdělení s nulovou střední hodnotou a stejnou varianční maticí pro všechny skupiny. To označíme $\boldsymbol{\varepsilon}_{hi} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Požadujeme tedy H nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_H$, z p-rozměrných normálních rozdělení se stejnou varianční maticí ve všech skupinách. Vektor $\mathbf{y}_{hi}$ má potom rozdělení $N(\boldsymbol{\mu}_h, \boldsymbol{\Sigma})$. Posledním důležitým předpokladem je pozitivní definitnost varianční matice $\boldsymbol{\Sigma}$. Pro ověření nulové hypotézy, že vektory středních hodnot se v H skupinách rovnají, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_H$, použijeme model

$$\mathbf{y}_{hi} = \boldsymbol{\mu}_h + \boldsymbol{\varepsilon}_{hi}, \tag{1}$$

$$i = 1, 2, \ldots, n_h,\ h = 1, 2, \ldots, H.$$

Model MANOVy (1) též můžeme zapsat jako

$$\mathbf{y}_{hi} = \boldsymbol{\mu} + \boldsymbol{\alpha}_h + \boldsymbol{\varepsilon}_{hi}, \tag{2}$$

$$i = 1, 2, \ldots, n_h,\ h = 1, 2, \ldots, H.$$

Nulová hypotéza pak říká, že efekty jednotlivých úrovní faktoru jsou ve všech skupinách nulové. V modelu (2) potom nulovou hypotézu zapíšeme ve tvaru $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_H = 0$.

Maticový zápis modelu (1) je následující

$$\begin{pmatrix} \mathbf{y}_{11}^T \\ \vdots \\ \mathbf{y}_{1n_1}^T \\ \mathbf{y}_{21}^T \\ \vdots \\ \mathbf{y}_{2n_2}^T \\ \vdots \\ \mathbf{y}_{H1}^T \\ \vdots \\ \mathbf{y}_{Hn_H}^T \end{pmatrix} = \begin{pmatrix} 1\ 0 \cdots 0 \\ \vdots\ \vdots\ \vdots\ \vdots \\ 1\ 0 \cdots 0 \\ 0\ 1 \cdots 0 \\ \vdots\ \vdots\ \vdots\ \vdots \\ 0\ 1 \cdots 0 \\ \vdots\ \vdots\ \vdots\ \vdots \\ 0\ 0 \cdots 1 \\ \vdots\ \vdots\ \vdots\ \vdots \\ 0\ 0 \cdots 1 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \boldsymbol{\mu}_H^T \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{11}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{1n_1}^T \\ \boldsymbol{\varepsilon}_{21}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{2n_2}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{H1}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{Hn_H}^T \end{pmatrix}.$$

Označme

$$\overline{\mathbf{y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{y}_{hi} \qquad \text{a} \qquad \mathbf{S}_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\mathbf{y}_{hi} - \overline{\mathbf{y}}_h\right) \left(\mathbf{y}_{hi} - \overline{\mathbf{y}}_h\right)^T,$$

vektor průměrů pro h-tou skupinu a výběrovou varianční matici v h-té skupině, dále

$$\overline{\mathbf{y}} = \frac{1}{n} \sum_{h=1}^{H} \overline{\mathbf{y}}_h n_h \qquad \text{a} \qquad \mathbf{S} = \frac{1}{n - H} \sum_{h=1}^{H} S_h \left(n_h - 1\right),$$

kde

$$n = \sum\nolimits_{h=1}^{H} n_h,$$

vektor celkových průměrů a celkovou výběrovou varianční matici, která je váženým průměrem výběrových variančních matic pro jednotlivé skupiny.

Mnohorozměrná analýza rozptylu, která slouží k ověření nulových hypotéz v modelu (1) nebo (2), je stejně jako ta jednorozměrná založena na rozkladu celkové variability na variabilitu vnitroskupinovou a meziskupinovou. V případě MANOVy však nejde o rozptyly, ale varianční matice, kdy celkovou varianční matici $\mathbf{T}$ rozdělíme na matici $\mathbf{E}$ vyjadřující vnitroskupinovou variabilitu a matici $\mathbf{B}$ vyjadřující meziskupinovou variabilitu. Platí vztah

$$\mathbf{T} = \mathbf{B} + \mathbf{E},$$

kde

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

popisuje celkovou variabilitu,

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

vnitroskupinovou variabilitu a meziskupinová variabilita je vyjádřena pomocí matice

$$\mathbf{B} = \sum_{h=1}^{H} n_h (\overline{\mathbf{y}}_h - \overline{\mathbf{y}}) (\overline{\mathbf{y}}_h - \overline{\mathbf{y}})^T .$$

1.1.1. Test věrohodnostním poměrem

K otestování nulové hypotézy vztahující se k modelu (1) nebo modelu (2), lze využít test věrohodnostním poměrem.

V případě jednorozměrné analýzy rozptylu se výpočty zapisují do tabulky. Podobně jako u ANOVy si vytvoříme výslednou tabulku 2. U ANOVy by obsahovala ještě dva sloupce navíc pro hodnotu testovací statistiky F a k ní příslušnou p-hodnotu.

Na rozdíl od ANOVy se však tabulka neskládá z číselných hodnot, ale z matic. Analogií statistiky F, kterou známe z ANOVy, je u MANOVy matice $\mathbf{BE}^{-1}$. Tato matice úzce souvisí se statistikou testu věrohodnostním poměrem. Lze dokázat, že maximálně věrohodný odhad varianční matice $\mathbf{\Sigma}$ je za platnosti nulové hypotézy roven $\mathbf{S}_0 = \frac{\mathbf{T}}{n}$, v případě platnosti alternativy je odhad tvaru $\mathbf{S}_1 = \frac{\mathbf{E}}{n}$ [14]. Statistika testu věrohodnostním poměrem je tvaru

variabilita	součet čtverců	stupně volnosti	podíl
meziskupinová	$\mathbf{B}$	$H-1$	$\mathbf{B}/(H-1)$
vnitroskupinová	$\mathbf{E}$	$n-H$	$\mathbf{E}/(n-K)$
celková	$\mathbf{T}$	$n-1$	$\mathbf{T}/(n-1)$

Tabulka 2: Výsledná tabulka pro MANOVu s jedním faktorem

$$V = n \cdot \ln \frac{|\mathbf{S}_0|}{|\mathbf{S}_1|} = n \cdot \ln \frac{|\mathbf{T}|}{|\mathbf{E}|}.$$

Testovací statistiku můžeme dále upravit jako

$$V = n \cdot \ln \frac{|\mathbf{E}+\mathbf{B}|}{|\mathbf{E}|} = n \cdot \ln \left|\mathbf{B}\mathbf{E}^{-1}+\mathbf{I}\right| = n \cdot \ln \left|\mathbf{T}\mathbf{E}^{-1}\right|.$$

Označme vlastní čísla matice $\mathbf{B}\mathbf{E}^{-1}$ jako λ_j, kde $j = 1, \ldots, p$. Pak $(1+\lambda_j)$ jsou vlastní čísla matice $\mathbf{T}\mathbf{E}^{-1}$, neboť platí $\mathbf{T}\mathbf{E}^{-1} = \mathbf{B}\mathbf{E}^{-1} + \mathbf{I}$. Dále označíme počet kladných vlastních čísel R, tedy $\mathrm{R} = h\left(\mathbf{B}\mathbf{E}^{-1}\right)$. Testovací statistiku pak můžeme rozepsat s využitím vlastních čísel matice $\mathbf{B}\mathbf{E}^{-1}$ takto

$$V = n \cdot \ln \prod_{r=1}^{R} (1+\lambda_r) = n \cdot \sum_{r=1}^{R} \ln(1+\lambda_r). \tag{3}$$

Výsledná statistika (3) má za platnosti nulové hypotézy přibližně chí-kvadrát rozdělení s $p \cdot (H-1)$ stupni volnosti. Nulovou hypotézu zamítáme na hladině významnosti α, jestliže hodnota testovací statistiky (3) překročí $(1-\alpha)$-kvantil chí-kvadrát rozdělení s $p \cdot (H-1)$ stupni volnosti. Test věrohodnostním se však používá poměrně zřídka kdy. Statistické softwary obvykle využívají testovací statistiky, které si popíšeme v další kapitole. Při psaní sekce jsem vycházela převážně ze zdroje [14].

1.1.2. Pillaiův, Wilksův, Lawleyho-Hottelingův a Royův test

V této kapitole jsou popsány nejčastěji používané testy, které slouží k ověření nulové hypotézy u modelu 1, respektive u modelu 2. Testování probíhá s pomocí čtyř testovacích statistik, jejichž výpočet vychází z vlastních čísel matice $\mathbf{BE}^{-1}$. Mezi tyto testovací statistiky se řadí:

$$\mathrm{P} = \mathrm{st}\left(\mathbf{B}\left(\mathbf{B}+\mathbf{E}^{-1}\right)\right) = \sum\nolimits_{r=1}^{R} \frac{\lambda_r}{1+\lambda_r}, \tag{4}$$

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{B}|} = \prod\nolimits_{r=1}^{R} \frac{1}{1+\lambda_r}, \tag{5}$$

$$\mathrm{LH} = \mathrm{st}\left(\mathbf{B}^{-1}\mathbf{E}\right) = \sum\nolimits_{r=1}^{R} \lambda_r \ , \tag{6}$$

$$\Theta = \frac{\lambda_{max}}{1+\lambda_{max}}, \tag{7}$$

kde st () značí stopu matice. Statistiky se nazývají Pillaiova statistika (4), Wilksova Λ (5), Lawley-Hottelingova statistika (6) a Royova statistika (7). Dále jsou uvedeny testy, které ze zmíněných čtyř testovacích statistik vycházejí.

Pillaiův test

Transformací první zmíněné statistiky, vzorec (4), dostaneme Pillaiovu testovací statistiku v tomto tvaru

$$\frac{\nu_2 \cdot \mathrm{P}}{\nu_1 \cdot (\mathrm{R}-\mathrm{P})}, \tag{8}$$

kde ν_1 a ν_2 představují stupně volnosti. Tyto stupně volnosti vypočítáme pomocí vztahů:

$$\nu_1 = \mathrm{R} \cdot (2 \cdot \mathrm{m}_p + \mathrm{R} + 1) \qquad \text{a} \qquad \nu_2 = \mathrm{R} \cdot (2 \cdot \mathrm{n}_p + \mathrm{R} + 1)\,,$$

kde

$$\mathrm{R} = \min(p, \nu_H)\ , \qquad \mathrm{m}_p = \frac{|p - \nu_H - 1|}{2} \qquad \text{a} \qquad \mathrm{n}_p = \frac{\nu_E - p - 1}{2}.$$

Stupně volnosti ν_H a ν_E při analýze rozptylu s jedním faktorem spočítáme jako

$$\nu_H = H - 1 \qquad \text{a} \qquad \nu_E = n - H.$$

Statistika (8) má přibližně Fisherovo F-rozdělení s ν_1 a ν_2 stupni volnosti. Nulovou hypotézu o shodě vektorů středních hodnot zamítáme na hladině významnosti α, jestliže je hodnota testovací statistiky větší než $(1-\alpha)$-kvantil Fisherova F-rozdělění s ν_1 a ν_2 stupni volnosti.

Alternativně můžeme rozhodnout i na základě kritických hodnot pro danou statistiku. Jestliže je $\mathrm{P} \geq \mathrm{P}_\alpha$, kde P_α je kritická hodnota, tak zamítáme nulovou hypotézu o shodě vektorů středních hodnot na hladině významnosti α. Kritické hodnoty jsou tabelovány například v [3].

Pillaiova testovací statitika je rozšířením Royovi testovací statistiky $\Theta = \frac{\lambda_{max}}{1+\lambda_{max}}$, kterou pro vlastní čísla seřazená od největšího po nejmenší, tedy tak aby platilo $\lambda_1 > \lambda_2 > \ldots > \lambda_R$, můžeme přepsat jako $\Theta = \frac{\lambda_1}{1+\lambda_1}$. Pak ze vztahu (4) vidíme, že Pillaiova statistika využívá nejen informaci, kterou obsahuje Royova testovací statistika, ale i informaci v ostatních výrazech $\frac{\lambda_i}{1+\lambda_i}$, pro $i = 2, \ldots, R$.

Wilksův test

Wilksův test provádíme pomocí Wilksovi testovací statistiky, která vznikne transformací Wilksovy Λ, ta je daná vztahem (5), a má následující tvar:

$$\frac{\nu_2 \cdot \left(1 - \Lambda^{1/t}\right)}{\nu_1 \cdot \Lambda^{1/t}}, \tag{9}$$

kde stupně volnosti ν_1 a ν_2 jsou dány vztahy

$$\nu_1 = p \cdot \nu_H \qquad \text{a} \qquad \nu_2 = f \cdot t - \frac{p \cdot \nu_H}{2} + 1,$$

hodnoty f a t spočítáme jako

$$f = \nu_E - \frac{p + 1 - \nu_H}{2},$$

$$t = \begin{cases} \sqrt{\frac{p^2\nu_H^2 - 4}{p^2 + \nu_H^2 - 5}} & \text{pro } p^2 + \nu_H^2 - 5 > 0 \\ 1 & \text{jinak.} \end{cases}$$

Statistika (9) má přibližně Fisherovo F-rozdělení s ν_1 a ν_2 stupni volnosti. Nulovou hypotézu zamítáme na hladině významnosti α, jestliže je hodnota testovací statistiky větší než kvantil $F_{\nu_1,\nu_2}(1-\alpha)$, který značí $(1-\alpha)$-kvantil Fisherova F-rozdělení s ν_1 a ν_2 stupni volnosti. Pokud bychom měli v daných datech pouze dvě nebo tři skupiny pozorování nebo bychom sledovali jen jednu nebo dvě závislé proměnné, bude mít statistika (9) přesně Fisherovo F-rozdělení.

O nulové hypotéze můžeme rozhodnout i na základě kritických hodnot pro Wilksovu Λ, kdy nulovou hypotézu zamítáme na hladině významnosti α pro malé hodnoty Λ, tedy pokud je $\Lambda \leq \Lambda_{\alpha,p,\nu_H,\nu_E}$. Tabulku s kritickými hodnotami $\Lambda_{\alpha,p,\nu_H,\nu_E}$ můžeme nalézt v [3]. V této tabulce si můžeme všimnout, že kritické hodnoty klesají se zvyšujícím se počtem závislých proměnných. To značí, že přidání dalších proměnných sníží sílu testu, tj. zmenší se pravděpodobnost, že zamítneme nulovou hypotézu, pokud neplatí, i když proměnné přispívají k zamítnutí nulové hypotézy.

Aby bylo zaručeno, že determinanty ve vzorci (5) budou kladné, musí být splněno, že $\nu_E \geq p$. Wilksova Λ nabývá hodnot z intervalu $\langle 0, 1\rangle$. Pokud by výběrové vektory středních hodnoty byly stejné pro všechny skupiny, tak by matice $\mathbf{B}$, vyjadřující variabilitu mezi skupinami, byla nulová a Wilksova Λ by se zjednodušila na výraz $\Lambda = |\mathbf{E}| \,/\, |\mathbf{E}+\mathbf{0}|$. Pro shodné vektory středních hodnot je tedy Wilksova Λ rovna jedné. Naopak pokud by se vektory středních hodnot lišily, tak se zvětšujícím se rozdílem poroste hodnota meziskupinového rozptylu a hodnoty matice

B se tedy budou zvyšovat. Čím bude rozdíl mezi výběrovými vektory průměrů větší, tím se bude zvětšovat i rozdíl mezi maticí **B** a maticí **E**. Pro zvyšující se **B** se hodnota Λ blíží k nule.

Lawleyho-Hottelingův test

Tento test je založen na Lawley-Hottelingově statistice LH dané vzorcem (6), která se též nazývá Hottelingova zobecněná statistika T^2. Transformací této statistiky dostaneme testovací statistiku ve tvaru

$$\frac{\nu_2}{R \cdot \nu_1} \cdot \mathrm{LH}, \tag{10}$$

kde stupně volnosti ν_1 a ν_2 vypočítáme podle následujících vztahů

$$\nu_1 = p \cdot \nu_H \qquad \text{a} \qquad \nu_2 = \mathrm{R} \cdot (\nu_E - p - 1) + 2.$$

Statistika (10) má přibližně Fisherovo F-rozdělení s ν_1 a ν_2 stupni volnosti. Nulovou hypotézu zamítáme na hladině významnosti α, jestliže je hodnota testovací statistiky větší než kvantil $F_{\nu_1,\nu_2}(1-\alpha)$. Opět můžeme Lawleyho-Hottelingovu statistiku otestovat i na základě tabelovaných kritických hodnot, které jsou uvedeny v [3]. Nulovou hypotézu budeme zamítat pro velké hodnoty statistiky LH.

Royův test

Posledním z uvedených používaných testů je Royův test, jehož testovací statistika vznikne transformací statistiky Θ, dané výrazem (7), na tvar

$$\frac{\nu_2}{\nu_1} \cdot \Theta, \tag{11}$$

kde ν_1 a ν_2 jsou stupně volnosti dány vztahy

$$\nu_1 = 2 \cdot \mathrm{m}_p + \mathrm{R} + 1 \qquad \text{a} \qquad \nu_2 = 2 \cdot \mathrm{n}_p + \mathrm{R} + 1.$$

Statistika (11) má přibližně Fisherovo F-rozdělení s ν_1 a ν_2 stupni volnosti. Nulovou hypotézu zamítáme na hladině významnosti α, jestliže je hodnota testovací statistiky větší než kvantil $F_{\nu_1,\nu_2}(1-\alpha)$. Test nulové hypotézy můžeme provést i na základě kritických hodnot pro Royovu statistiku Θ. Jestliže je $\Theta \geq \Theta_{\alpha,R,m_p,n_p}$, tak zamítáme nulovovu hypotézu na hladině α. Kritické hodnoty $\Theta_{\alpha,R,m_p,n_p}$ jsou tabelovány a jsou dostupné v [3].

1.2. MANOVA se dvěma faktory

V předchozí kapitole byla pozorování rozdělena do skupin na základě jednoho faktoru. Nyní se budeme věnovat úloze, kdy jsou tyto faktory dva. Mnohorozměrnou analýzu rozptylu se dvěma faktory můžeme použít na příklad se studenty ze začátku 1. kapitoly s přidaným druhým faktorem. Tím bude faktor popisující fyzickou zdatnost studenta o třech úrovních, které popisují, jak často student cvičí: pravidelně, občas, nikdy. Budeme tedy zkoumat vliv délky spánku a četnosti cvičení na známky z několika předmětů ve škole.

Pro jednotnost se omezíme na vyvážené třídění, kdy obecně uvažujeme model

$$\boldsymbol{y}_{hki} = \boldsymbol{\mu} + \boldsymbol{\alpha}_h + \boldsymbol{\beta}_k + (\boldsymbol{\alpha} \cdot \boldsymbol{\beta})_{hk} + \boldsymbol{\varepsilon}_{hki} \ ,$$

$$h = 1, \ldots, H,\ k = 1, \ldots, K,\ i = 1, \ldots, I.$$

Písmenem H označujeme počet úrovní faktoru A, K počet úrovní faktoru B a I počet pozorování. Výraz $\boldsymbol{y}_{hki}$ představuje mnohorozměrné pozorování na i-té jednotce, patřící do h-té úrovně faktoru A a k-té úrovně faktoru B. Vektor $\boldsymbol{\mu}$ označuje vektor celkových průměrů. Vektory $\boldsymbol{\alpha}_h$, $\boldsymbol{\beta}_k$, $(\boldsymbol{\alpha} \cdot \boldsymbol{\beta})_{hk}$ jsou efekty faktorů A a B a interakce mezi faktory A a B, které popisují, jaký je dopad obou skupin najednou. Vektor $\boldsymbol{\varepsilon}_{hki}$ označuje vektor náhodných složek pro i-té pozorování, patřící do h-té úrovně faktoru A a k-té úrovně faktoru B. Pro různá i předpokládáme nezávislost vektorů $\boldsymbol{\varepsilon}_{hki}$, tedy, že $\boldsymbol{\varepsilon}_{hki} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Nejprve si definujeme celkový průměr

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{I}\mathbf{y}_{hki},\ n = HKI$$

a dílčí průměry počítané pro třídění podle různých úrovní obou faktorů

$$\overline{\boldsymbol{y}}_{h..} = \frac{1}{KI}\sum_{k=1}^{K}\sum_{i=1}^{I}\mathbf{y}_{hki}, \qquad \overline{\boldsymbol{y}}_{.k.} = \frac{1}{HI}\sum_{h=1}^{H}\sum_{i=1}^{I}\mathbf{y}_{hki} \qquad \text{a} \qquad \overline{\boldsymbol{y}}_{hk.} = \frac{1}{I}\sum_{i=1}^{I}\mathbf{y}_{hki}.$$

Pro dva faktory rozložíme celkovou varianční matici $\mathbf{T}$,

$$\mathbf{T} = \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(\boldsymbol{y}_{hki} - \overline{\boldsymbol{y}}\right)\left(\boldsymbol{y}_{hki} - \overline{\boldsymbol{y}}\right)^{T},$$

do tvaru

$$\mathbf{T} = \mathbf{B}_A + \mathbf{B}_B + \mathbf{B}_{AB} + \mathbf{E},$$

kde

$$\mathbf{B}_A = K \cdot I \cdot \sum_{h=1}^{H}\left(\overline{\boldsymbol{y}}_{h..} - \overline{\boldsymbol{y}}\right)\left(\overline{\boldsymbol{y}}_{h..} - \overline{\boldsymbol{y}}\right)^{T}$$

značí variabilitu příslušnou faktoru A,

$$\mathbf{B}_B = H \cdot I \cdot \sum_{k=1}^{K}\left(\overline{\boldsymbol{y}}_{.k.} - \overline{\boldsymbol{y}}\right)\left(\overline{\boldsymbol{y}}_{.k.} - \overline{\boldsymbol{y}}\right)^{T}$$

vyjadřuje variabilitu příslušnou faktoru B,

$$\mathbf{B}_{AB} = I \cdot \sum_{h=1}^{H}\sum_{k=1}^{K}\left(\overline{\boldsymbol{y}}_{hk.} - \overline{\boldsymbol{y}}_{h..} - \overline{\boldsymbol{y}}_{.k.} + \overline{\boldsymbol{y}}\right)\left(\overline{\boldsymbol{y}}_{hk.} - \overline{\boldsymbol{y}}_{h..} - \overline{\boldsymbol{y}}_{.k.} + \overline{\boldsymbol{y}}\right)^{T}$$

značí variabilitu příslušnou interakci AB a

$$\mathbf{E} = \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(\boldsymbol{y}_{hki} - \overline{\boldsymbol{y}}_{hk.}\right)\left(\boldsymbol{y}_{hki} - \overline{\boldsymbol{y}}_{hk.}\right)^{T}$$

představuje variabilitu způsobenou náhodnými vlivy. Testované nulové hypotézy mají tento tvar

$$\mathrm{H}_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_H = 0$$

$$\mathrm{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_K = 0$$

$$\mathrm{H}_0 : (\boldsymbol{\alpha\beta})_{11} = (\boldsymbol{\alpha\beta})_{12} = \ldots = (\boldsymbol{\alpha\beta})_{HK} = 0.$$

První hypotéza testuje, zda má faktor **A** vliv na závislé proměnné, druhá, zda má vliv faktor **B**, a poslední testuje vliv interakce faktorů **A** a **B**. Postup pro testování je podobný jako pro mnohorozměrnou analýzu rozptylu s jedním faktorem, pouze místo matice **B**, při výpočtu vlastních čísel, použijeme jednu z matic $\mathbf{B}_A$, $\mathbf{B}_B$ a $\mathbf{B}_{AB}$ a změní se stupně volnosti přibližného Fisherova F-rozdělení. Tedy vlastní čísla ve vzorcích v kapitole 1.1 odpovídají při testování efektu faktoru **A** vlastním číslům matice $\mathbf{B}_A\mathbf{E}^{-1}$, stupně volnosti se pro analýzu rozptylu s dvěma faktory změní na $\nu_A = H - 1$ a $\nu_E = HK\,(I - 1)$. Pro testování efektu faktoru **B** musíme vypočítat vlastní čísla matice $\mathbf{B}_B\mathbf{E}^{-1}$ a použít stupeň volnosti $\nu_B = K - 1$. Pro testování interakce obou faktorů vypočítáme vlastní čísla matice $\mathbf{B}_{AB}\mathbf{E}^{-1}$ a použijeme stupeň volnosti $\nu_{AB} = (H - 1)\,(K - 1)$.

2. Další testování při zamítnutí nulové hypotézy o shodě vektorů středních hodnot

Pokud provádíme testování pomocí mnohorozměrné analýzy rozptylu a dojdeme k závěru, že nulovou hypotézu o shodě vektorů středních hodnot zamítáme, tak nás bude nejspíše zajímat, čím je to způsobeno. K tomu nám slouží několik post-hoc testů, které jsou v této sekci popsány. Testy z kapitol 2.1, 2.2 a 2.4 jsou simultánní s hladinou významnosti α. Při psaní kapitoly jsem vycházela hlavně ze zdrojů [3], [4] a [5].

2.1. Simultánní testy o složkách vektorů středních hodnot

Při zamítnutí hypotézy o shodě vektorů středních hodnot můžeme provést testy složek vektoru, kdy otestujeme jednotlivé závislé proměnné odděleně. Tímto postupem získáme představu o tom, na základě kterých proměnných jsme hypotézu zamítali, tedy zda se skupiny v některých proměnných liší. Testujeme simultánně p hypotéz

$$H_{01} : \mu_{11} = \ldots = \mu_{1H}, \ldots, H_{0p} : \mu_{p1} = \ldots = \mu_{pH}.$$

Alternativní hypotézy zní, že se pro danou proměnnou alespoň jedna dvojice středních hodnot liší. K testování použijeme následující statistiky

$$K_j = -\left(n - \frac{p+H}{2} - 1\right) \ln \frac{e_{jj}}{t_{jj}}, \qquad j = 1, \ldots, p, \tag{12}$$

kde e_{jj} je j-tý diagonální prvek matice $\mathbf{E}$ a t_{jj} je j-tý diagonální prvek matice $\mathbf{T}$. Statistika K_j má za platnosti nulové hypotézy přibližně χ^2 rozdělení s $(p\,(H-1))$ stupni volnosti. Nulovou hypotézu H_{0j} zamítáme na hladině významnosti α, pokud hodnota testové statistiky K_j překročí $(1-\alpha)$-kvantil chí-kvadrát rozdělení s $(p\,(H-1))$ stupni volnosti.

Občas i přes zamítnutí hypotézy o shodě vektorů středních hodnot vyjde pomocí simultánních testů, že by se jednotlivé složky vektoru měly shodovat.

V tomto případě rozdíl mezi skupinami způsobuje nějaký složitější vztah mezi proměnnými.

2.2. Vícerozměrná obdoba mnohonásobného porovnávání

Vícerozměrná obdoba mnohonásobného porovnávání slouží k testování, zda se liší dvojice skupin. Porovnáváme dva vektory středních hodnot pro všechny kombinace dvojic skupin. Jelikož všech možných dvojic je $\binom{H}{2}$, tak budeme testovat současně $\frac{H\cdot(H-1)}{2}$ hypotéz. Nulová hypotéza je tvaru

$$H_0 : \boldsymbol{\mu_i} = \boldsymbol{\mu_j}$$

pro všechna $i, j = 1, \ldots, H$, kdy $i \neq j$, proti alternativě, že se dané vektory nerovnají. Využijeme k tomu testovou statistiku

$$F = \frac{n - H - p + 1}{(H-1)\,p} \frac{n_i n_j}{n_i + n_j} \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}_j\right)^T \boldsymbol{E^{-1}} \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}_j\right), \tag{13}$$

která má za platnosti nulové hypotézy F-rozdělení se stupni volnosti

$$\nu_1 = \frac{(H-1)\,p\,(n-H-p)}{n-2-(H-1)\,p} \qquad \text{a} \qquad \nu_2 = n - H - p + 1.$$

Nulovou hypotézu o shodě $\boldsymbol{\mu_i}$ a $\boldsymbol{\mu_j}$ zamítáme na hladině významnosti α, jestliže je hodnota f testovací statistiky F větší než $(1-\alpha)$-kvantil F-rozdělení s ν_1 a ν_2 stupni volnosti.

2.3. Mnohorozměrné kontrasty

Mnohorozměrné kontrasty využijeme, pokud porovnání dvojic jednotlivých vektorů středních hodnot nestačí, a máme za cíl testovat kontrasty těchto vektorů. Kontrast vektorů průměrů pro jednotlivé skupiny definujeme obecně jako lineární kombinaci vektorů středních hodnot

$$\boldsymbol{\delta} = c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \cdots + c_H\boldsymbol{\mu}_H,$$

kde $\sum_{i=1}^{H} c_i = 0$. Odhad tohoto kontrastu dostaneme pomocí lineární kombinace výběrových vektorů středních hodnot skupin

$$\hat{\boldsymbol{\delta}} = c_1\overline{\mathbf{y}}_1 + c_2\overline{\mathbf{y}}_2 + \cdots + c_H\overline{\mathbf{y}}_H.$$

Jelikož předpokládáme, že pozorování pocházejí z nezávislých p-rozměrných normálních rozdělení se stejnou varianční maticí $\boldsymbol{\Sigma}$ pro všechny skupiny, jsou nezávislé i výběrové vektory středních hodnot $\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2, \ldots, \overline{\mathbf{y}}_H$ a varianční matice pro i-tý vektor průměrů je rovna $\frac{\boldsymbol{\Sigma}}{n_i}$, pro $i = 1, \ldots, H$. Varianční matice pro odhad kontrastu $\hat{\boldsymbol{\delta}}$ je daná výrazem

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\delta}}} = c_1^2\frac{\boldsymbol{\Sigma}}{n_1} + c_2^2\frac{\boldsymbol{\Sigma}}{n_2} + \cdots + c_H^2\frac{\boldsymbol{\Sigma}}{n_H},$$

který můžeme přepsat takto

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\delta}}} = \boldsymbol{\Sigma} \cdot \sum_{i=1}^{H} \frac{c_i^2}{n_i}.$$

Varianční matici $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\delta}}}$ odhadneme pomocí výběrové varianční matice $\boldsymbol{S}$. Varianční matici odhadu kontrastu odhadneme pomocí vztahu

$$\boldsymbol{S}_{\hat{\boldsymbol{\delta}}} = \boldsymbol{S} \cdot \sum_{i=1}^{H} \frac{c_i^2}{n_i} = \frac{\boldsymbol{E}}{\nu_E} \cdot \sum_{i=1}^{H} \frac{c_i^2}{n_i}.$$

Nulová hypotéza pro testování kontrastů má obecně tvar

$$H_0 : \boldsymbol{\delta} = \mathbf{0}$$

nebo ekvivalentně

$$H_0 : c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \cdots + c_H\boldsymbol{\mu}_H = \mathbf{0}.$$

Například pokud bychom testovali, zda platí $\boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 = \mathbf{0}$, což je ekvivalentní výrazu $\boldsymbol{\mu}_2 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3)$, tak bychom porovnávali vektor $\boldsymbol{\mu}_2$ s průměrem vektorů $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_3$. Z čehož plyne, že i-tý prvek vektoru $\boldsymbol{\mu}_2$ musí být roven i-tému prvku vektoru $\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3)$, musí tedy platit

$$\begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left(\mu_{11} + \mu_{31}\right) \\ \frac{1}{2}\left(\mu_{12} + \mu_{32}\right) \\ \vdots \\ \frac{1}{2}\left(\mu_{1p} + \mu_{3p}\right) \end{pmatrix}.$$

Za splnění předpokladu mnohorozměrného normálního rozdělení lze nulovou hypotézu otestovat pomocí statistiky T^2, která má tvar

$$T^2 = \hat{\boldsymbol{\delta}}^T \left(\boldsymbol{S} \cdot \sum_{i=1}^{H} \frac{c_i^2}{n_i} \right)^{-1} \hat{\boldsymbol{\delta}},$$

přičemž tento tvar můžeme zjednodušit na

$$T^2 = \left(\sum_{i=1}^{H} c_i \overline{\boldsymbol{y}}_i \right)^T \left(\frac{\boldsymbol{E}}{\nu_E} \cdot \sum_{i=1}^{H} \frac{c_i^2}{n_i} \right)^{-1} \left(\sum_{i=1}^{H} c_i \overline{\boldsymbol{y}}_i \right).$$

Následující transformací statistiky T^2

$$F = \frac{(\nu_E - p + 1)}{\nu_E \cdot p} \cdot T^2$$

dostaneme statistiku, která má Fisherovo F-rozdělení s p a $(\nu_E - p + 1)$ stupni volnosti, kdy $\nu_E = n - H$. Nulovou hypotézu zamítáme na hladině významnosti α, pokud hodnota f testové statistiky F překročí $(1 - \alpha)$-kvantil Fisherova F-rozdělení s p a $(\nu_E - p + 1)$ stupni volnosti.

Druhou možností, jak provést test nulové hypotézy, je pomocí Wilksovy Λ. Matice odpovídající nulové hypotéze kontrastů má tvar

$$\mathbf{H}_1 = \sum_{i=1}^{H} \frac{n_i}{c_i^2} \left(\sum_{i=1}^{H} c_i \overline{\boldsymbol{y}}_i \right) \left(\sum_{i=1}^{H} c_i \overline{\boldsymbol{y}}_i \right)^T.$$

Tato matice má hodnost jedna a Wilksova Λ je daná vztahem

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_1|},$$

s rozdělením $\Lambda_{p,1,\nu_E}$. Nulovou hypotézu zamítáme na hladině významnosti α, pokud je hodnota testové statistiky Λ menší nebo rovna kritické hodnotě $\Lambda_{p,1,\nu_E}$.

2.4. Simultánní testy v mnohonásobném porovnávání

Simultánní testy v mnohonásobném porovnávání slouží k zjišťování rozdílů mezi proměnnými pro dvojice skupin. Pokud se prokázal statisticky významný rozdíl mezi nějakou dvojicí skupin, tj. $\boldsymbol{\mu_i} \neq \boldsymbol{\mu_j}$, pro $i, j = 1, \ldots, H, i \neq j$, pomocí testů z kapitoly 2.2, tak chceme zjistit, které proměnné to způsobují. Počet všech možných dvojic je $\binom{H}{2}$ a pro každou dvojici dále provedeme p testů, pro jednotlivé složky vektorů. Výsledný počet testů je tudíž $p \cdot \binom{H}{2}$, což můžeme přepsat jako $p \cdot \frac{H \cdot (H-1)}{2}$. Jde o simultánní testy následujících hypotéz

$$H_0 : \mu_{ir} = \mu_{jr} \qquad \text{proti alternativě} \qquad H_A : \mu_{ir} \neq \mu_{jr},$$

pro všechna $i, j = 1, \ldots, H$, kdy $i \neq j$, $r = 1, \ldots, p$. Testová statistika má tvar

$$F = \frac{n - H - p + 1}{(H-1)\, p\, (n-H)} \frac{n_i n_j}{n_i + n_j} \frac{\left(\overline{y}_{ir} - \overline{y}_{jr}\right)^2}{S_r}, \tag{14}$$

kde $\overline{y}_{ir}$ značí výběrový průměr r-té proměnné v i-té skupině a S_r je r-tý diagonální prvek celkové výběrové varianční matice $\mathbf{S}$. Tato testovací statistika má za platnosti nulové hypotézy přibližně Fisherovo F-rozdělení s ν_1 a ν_2 stupni volnosti. Nulovou hypotézu zamítáme na hladině významnosti α, pokud hodnota f testovací statistiky F překročí $(1-\alpha)$-kvantil Fisherova F-rozdělení s ν_1 a ν_2 stupni volnosti,

$$\nu_1 = \frac{(H-1)\, p\, (n-H-p)}{n - 2 - (H-1)\, p} \qquad \text{a} \qquad \nu_2 = n - H - p + 1.$$

3. Ověřování předpokladů

Stejně jako v případě jednorozměrné analýzy rozptylu, při výpočtu MANOVy požadujeme splnění některých předpokladů, které jsou kladeny na vytvořený matematický model, pomocí nějž se snažíme popsat realitu. V praxi ovšem obvykle tyto předpoklady splněny nejsou. Jejich drobné porušení však obvykle nevede k chybným záměrům. Samozřejmě záleží na tom, jaký předpoklad je porušen a za jakých podmínek. Předpoklady pro správné testování pomocí mnohorozměrné analýzy rozptylu a některé postupy pro jejich ověření jsou shrnuty do několika podkapitol. Při tvorbě této kapitoly jsem vycházela především ze zdrojů [1], [2], [4], [5] a [6].

3.1. Nezávislost pozorování

Předpoklad nezávislosti jednotlivých pozorování je velmi důležitý. I velmi malá závislost mezi pozorováními ve skupině způsobí, že zadaná pravděpodobnost chyby prvního druhu α bude několikanásobně větší.

Závislá pozorování se vyskytují velmi často například ve výzkumech z oblasti sociologie nebo psychologie. Například pokud budeme uvažovat studenty, kteří jsou rozděleni do několika skupin, ve kterých se učí a pracují na úkolech, a následně jsou všichni individuálně ohodnoceni. Tehdy nemůžeme jednotlivá pozorování známek studentů považovat za nezávislá, jelikož v jednotlivých skupinách probíhají interakce mezi studenty. Závislosti mezi pozorováními bychom se měli vyhnout už při sbírání dat. Případně bychom se měli ujistit, zda byla data získána tak, že jednotlivá pozorování můžeme označit za nezávislá.

3.2. Normalita

Předpokladem normality se rozumí to, že pozorování v jednotlivých skupinách pocházejí z mnohorozměrného normálního rozdělení. I když při ověřování tohoto předpokladu chceme posoudit normalitu vícerozměrnou, doporučuje se nejprve podívat, zda jednotlivé závislé proměnné splňují normalitu jednorozměrnou. Je

však třeba mít na paměti, že jednorozměrná normalita nám negarantuje normalitu mnohorozměrnou. Protože posouzení mnohorozměrné normality je často obtížné, tak se v praxi většinou spokojíme s ověřením jednorozměrné normality pro jednotlivé proměnné. Podle [4] má odchylka od mnohorozměrného normálního rozdělení za následek pouze malý dopad na chybu prvního druhu.

Posouzení jednorozměrné i vícerozměrné normality můžeme provést pomocí testů normality nebo vizuálně s využitím grafických zobrazení. Nejprve se stručně podíváme na několik testů pro testování jednorozměrné normality.

Jelikož tvar normálního rozdělení lze charakterizovat pomocí koeficientu šikmosti a špičatosti, kdy normální rozdělení má šikmost rovnu nule a špičatost rovnu třem, tak jednou z možností ověření jednorozměrné normality je provést **testy normality založené na koeficientu šikmosti a špičatosti**. Výběrové koeficienty šikmosti a špičatosti jsou ve tvaru:

$$\sqrt{b_1} = \frac{\sqrt{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^3}{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^{\frac{3}{2}}},$$

$$b_2 = \frac{n \cdot \sum_{i=1}^{n} (x_i - \overline{x})^4}{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^2}.$$

Pak k ověření nulové hypotézy, že data nejsou sešikmená, $H_0 : \sqrt{b_1} = 0$, a hypotézy, že špičatost dat odpovídá špičatosti normálního rozdělení, $H_0 : b_2 = 3$, lze použít testovací statistiky

$$U_1 = \sqrt{\frac{b_1 (n+1)(n+3)}{6 \cdot (n-2)}} \text{ a}$$

$$U_2 = \sqrt{\frac{(n+1)^2 (n+3)(n+5)}{24 \cdot n \cdot (n-2)(n-3)}} \left(b_2 - 3 + \frac{6}{n+1} \right),$$

které mají za platnosti nulové hypotézy a dostatečného rozsahu výběru normované normální rozdělení. Hypotézy tedy zamítáme na hladině významnosti

α, pokud je $\left(1-\frac{\alpha}{2}\right)$-kvantil normovaného normálního rozdělení menší než hodnota statistiky v absolutní hodnotě, tedy v případě, že $|u_1| > u_{1-\frac{\alpha}{2}}$ nebo $|u_2| > u_{1-\frac{\alpha}{2}}$. Z dalších testů založených na momentech bychom mohli provést například **Jarqueho-Berův test** popsaný v [6]. Výpočet tohoto testu je založen jak na koeficientu špičatosti, tak i šikmosti. Testová statistika má tvar

$$JB = \frac{n}{6}\left[\frac{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^3\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)^3} + \frac{\left(\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^4\right)^2-3\right)^2}{4}\right].$$

Nulovou hypotézu, že náhodná veličina X pochází z normálního rozdělení, zamítáme na hladině významnosti α, pokud hodnota testové statistiky neleží v intervalu vymezeném $\left(1-\frac{\alpha}{2}\right)$-kvantilem a $\left(\frac{\alpha}{2}\right)$-kvantilem chí-kvadrát rozdělení o dvou stupních volnosti. Dále můžeme použít **Andersonův-Darlingův test** [8], jehož testování je založeno na distribuční funkci. Nulová hypotéza zní, že náhodná veličina X s distribuční funkcí $F\left(x\right)$ pochází z normálního rozdělení. Postupujeme tak, že hodnoty náhodné veličiny X normujeme, nově získané hodnoty označme u_i. Následně si určíme hodnoty distribuční funkce normovaného normálního rozdělení v bodech u_i, které označíme jako F_i. Pak už můžeme provést test, který vychází ze statistiky

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}\left(2i-1\right)\left[\ln F_i + \ln\left(1-F_{n-i+1}\right)\right].$$

Nulovou hypotézu zamítáme na hladině významnosti α, jestliže je hodnota testovací statistiky větší než $(1-\alpha)$-kvantil rozdělení statistiky AD za platnosti nulové hypotézy, tedy za předpokladu, že data jsou normálně rozdělená. Pro 95% kvantil statistiky AD platí při velkých rozsazích souboru následující přibližný vztah

$$ad_{0,95} = 1,0348 \cdot \left(1 - \frac{1,013}{n} - \frac{0,93}{n^2}\right).$$

Z dalších testů založených na empirických distribučních funkcích můžeme použít **Kolmogorovův-Smirnovův test** [6] nebo **Lillieforsův test** [6], který představuje vylepšení Kolmogorovova-Smirnovova testu. Další skupinou testů jsou

testy založené na regresi, tedy vzdálenosti bodů od regresní přímky, čehož využívají i některé vizualizační nástroje. Mezi tyto testy patří **Fillibenův test** nebo **Shapirův-Wilkův test** dostupné v [2] a [6]. U testu Shapira a Wilka porovnáváme hodnoty testované proměnné s kvantily normovaného normálního rozdělení výběrové distribuční funkce $F_n(x)$. Označme $X_{(i)}$ uspořádané veličiny X_i, $q_{(i)}$ značí uspořádané kvantily normovaného normálního rozdělení a platí $q_{(i)} = F^{-1}\left(\frac{i}{n}\right)$. Jestliže náhodná veličina X pochází z normálního rozdělení, budou body $\left(x_{(i)}, q_{(i)}\right)$ ležet na přímce, která prochází počátkem a svírá úhel 45° s x-ovou i y-ovou osou. Shapiro a Wilk odvodili následující testovací statistiku, která zohledňuje i to, že jednotlivé veličiny $X_{(i)}$ nejsou nezávislé,

$$SW = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2},$$

kde a_i představují hodnoty koeficientů, které se dají spolu s kritickými hodnotami pro statistiku SW najít v článku Shapira a Wilka z roku 1965 [7]. Nulovou hypotézu zamítáme na hladině významnosti α, pokud je hodnota testovací statistiky menší než kritická hodnota. Z dalších testů, které se využívají, stojí za zmínku **chí-kvadrát test dobré shody** nebo **D'Agostinův test**.

Pro ověřování vícerozměrné normality můžeme taktéž použít testy založené na šikmosti a špičatosti. Zobecněné vícerozměrné koeficienty šikmosti a špičatosti mají tvar

$$B_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T \mathbf{S}^{-1} \left(\mathbf{x}_j - \overline{\mathbf{x}}\right) \right]^3,$$

$$B_2 = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T \mathbf{S}^{-1} \left(\mathbf{x}_i - \overline{\mathbf{x}}\right) \right]^2.$$

Test šikmosti provádíme tak, že zjišťujeme, zda hodnota testovací statistiky

$$V = \frac{n \cdot B_1}{6}$$

překročí $\left(1-\frac{\alpha}{2}\right)$-kvantil chí-kvadrát rozdělení s $\frac{p\cdot(p+1)\cdot(p+2)}{6}$ stupni volnosti. V tomto případě označíme odchylku od očekávané střední hodnoty, $E(B_1)=0$, za signifikantní. U testu špičatosti porovnáváme testovací statistiku

$$U_3 = \sqrt{\frac{n}{8p\cdot(p+2)}}\,(B_2 - p\cdot(p+2))$$

s $\left(1-\frac{\alpha}{2}\right)$-kvantilem normovaného normálního rozdělení. Odchylku od střední hodnoty, $E(B_2)=p\cdot(p+2)$, označíme za významnou, jestliže absolutní hodnota $|u_3|$ testovací statistiky u_3 překročí $\left(1-\frac{\alpha}{2}\right)$-kvantil normovaného normálního rozdělení.

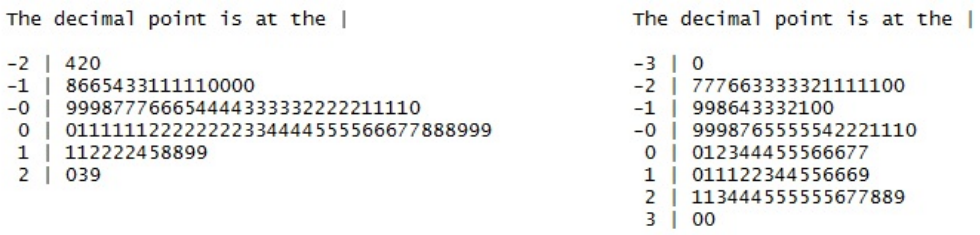
Dalším používaným testem pro ověření mnohorozměrné normality je **zobecnění Shapirova-Wilkova testu** na mnohorozměrný případ od Alvy a Estrada. Test je popsaný v [9]. Nechť $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ jsou nezávislé a stejně rozdělené p-rozměrné náhodné vektory, $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, a $\boldsymbol{\Sigma}$ je pozitivně definitní. Potom náhodné vektory $\boldsymbol{Z}_i = \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)$ mají přibližně rozdělení $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ a jednotlivé prvky vektoru $\boldsymbol{Z}_i$, které označíme jako $Z_{1i}, Z_{2i}, \ldots, Z_{pi}$, pocházejí z přibližně jednorozměrného normálního rozdělení. Výběrový vektor hodnot $\boldsymbol{Z}_i = \boldsymbol{S}^{-\frac{1}{2}}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)$ a má přibližně rozdělení $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Nulovou hypotézu, že data jsou výběrem z p-rozměrného normálního rozdělení, testujeme pomocí statistiky

$$SW_{AVE} = \frac{1}{p}\sum_{j=1}^{p} SW_{Z_{ji}},$$

kde $SW_{Z_{ji}}$ značí Shapiro-Wilkovu statistiku pro j-té složky vektorů $\boldsymbol{Z}_i$. Nulovou hypotézu zamítáme na hladině významnosti α, pokud je hodnota testové statistiky SW_{AVE} menší než kritická hodnota $C_{\alpha,n,p}$. Lze si všimnout, že pro jednu proměnnou dostaneme Shapiro-Wilkovu statistiku pro ověřování jednorozměrné normality.

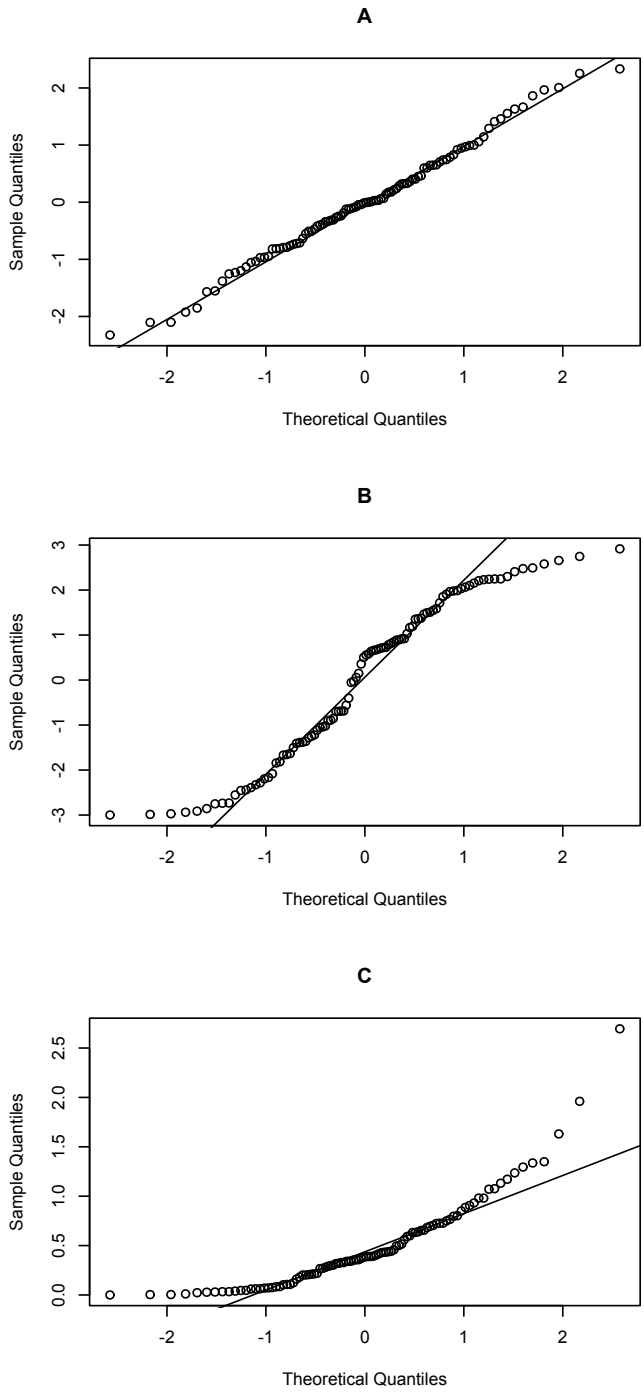
Z uvedených testů normality dojdeme k závěru, že na dané hladině α zamítáme nulovou hypotézu o původu dat z normálního rozdělení nebo ji nemůžeme zamítnout. Neposkytují však žádnou další informaci. Proto se velmi často využívají k posuzování normality i grafická zobrazení dat. Ta nám pomohou udělat

si představu nejen o tom, zda je rozdělení normální, ale i o tvaru rozdělení nebo o výskytu odlehlých hodnot v datech. Z vizuálních nástrojů, které se využívají k posuzování jednorozměrné normality, se často používá **histogram**, z nějž se snažíme vyčíst tvar hustoty posuzované proměnné. Z dalších grafických zobrazení bychom mohli použít **stem-and-leaf** diagram. Stem-and-leaf diagram stejně jako histogram porovnává tvar hustoty předpokládaného rozdělení s tvarem hustoty četností. Ukázky tohoto grafu vidíme v obrázku 1. V porovnání s histogramem se z něj dají vyčíst i konkrétní hodnoty dané proměnné.



Obrázek 1: Ukázka dvou steam-and-leaf diagramů. Vlevo vidíme graf 100 hodnot vygenerovaných z rozdělení $\mathcal{N}(0,1)$ a vpravo graf 100 hodnot vygenerovaných z rozdělení $\mathrm{Ro}(-3,3)$.

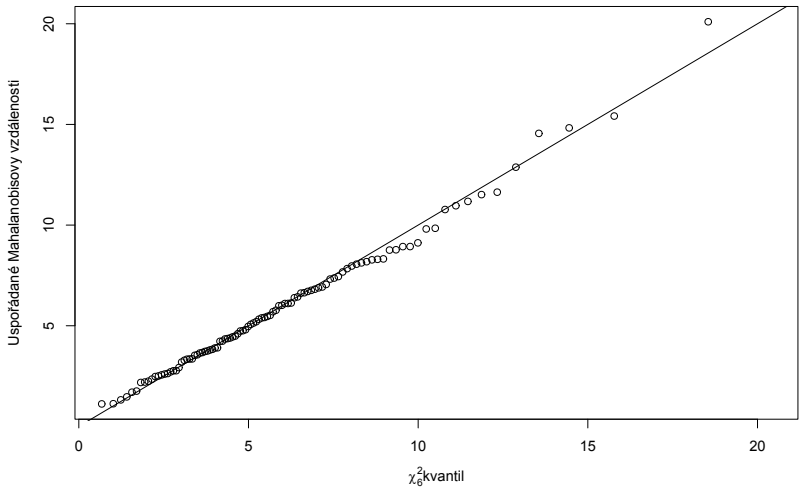
Zřejmě nejpoužívanějším nástrojem je však **q-q graf**. Tento diagram srovnává teoretické kvantily normovaného normálního rozdělení na x-vé ose s výběrovými kvantily seřazeného výběru na y-vé ose. Pokud data pocházejí z normálního rozdělení, tak by body jednotlivých pozorování měly ležet přibližně na přímce. Ukázku q-q grafů zobrazuje obrázek 2. Z q-q grafů si můžeme udělat i představu o šikmosti nebo špičatosti dat. Jestliže data nekopírují přímku, ale utvoří konvexní křivku, pak je považujeme za kladně sešikmená. Naopak pro záporně sešikmená data vytvoří konkávní křivku. Konvexně konkávní křivka značí menší špičatost a konkávně konvexní křivka špičatost větší než tři. Grafické znázornění mnohorozměrné normality je náročnější. Nejprve si musíme vypočítat Mahalanobisovy vzdálenosti d_i^2 podle vztahu



Obrázek 2: Ukázka q-q grafů pro 100 hodnot generovaných z různých rozdělení. V případě **A** jsou hodnoty náhodným výběrem z rozdělení $\mathcal{N}(0,1)$, **B** z rozdělení $\mathrm{Ro}(-3,3)$ a **C** z rozdělení $\mathrm{Exp}(2)$.

$$d_i^2 = (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}),$$

což znamená, že p-rozměrná pozorování nahradíme číslem. Tyto vzdálenosti následně uspořádáme a porovnáme s odpovídajícími kvantily chí-kvadrát rozdělení o p stupních volnosti, které by tyto vzdálenosti měly splňovat v případě, že data pocházejí z vícerozměrného normálního rozdělení. Ukázku výsledného grafu vidíme v obrázku 3. V případě mnohorozměrné normality dat by body měly ležet přibližně na přímce.



Obrázek 3: Ukázka grafu, který slouží k ověření vícerozměrné normality. Graf byl vytvořen ze 100 pozorování, která byla generována z $\mathcal{N}_6(\mathbf{0}, \boldsymbol{I})$.

3.3. Shodnost variančních matic

Jedním z dalších předpokladů, který musíme u mnohorozměrné analýzy rozptylu ověřit, je shoda variančních matic p-rozměrných normálních rozdělení v jednotlivých srovnávaných skupinách. Hypotéza, kterou musíme otestovat, je v následujícím tvaru

$$\mathrm{H}_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_H,$$

a vypovídá o tom, že se všech H variančních matic rovná. Alternativní hypotéza říká, že alespoň jedna ze dvojic variančních matic se neshoduje. Na tento test je kladen předpoklad, že výběry o rozsazích $n_1, n_2, \ldots, n_H$ jsou vzájemně nezávislé a pocházejí z mnohorozměrného normálního rozdělení. Testování shodnosti variančních matic vychází z následující statistiky:

$$\mathrm{M} = \frac{|\mathbf{S}_1|^{\frac{n_1-1}{2}} |\mathbf{S}_2|^{\frac{n_2-1}{2}} \ldots |\mathbf{S}_H|^{\frac{n_H-1}{2}}}{|\mathbf{S}|^{\frac{n-H}{2}}},$$

tedy

$$\mathrm{M} = \frac{\prod_{h=1}^{H} |\mathbf{S}_h|^{\frac{n_h-1}{2}}}{|\mathbf{S}|^{\frac{n-H}{2}}},$$

kde $\mathbf{S}_h$, $h = 1, \ldots, H$, je výběrová varianční matice pro h-tou skupinu a $\mathbf{S}$ je celková výběrová varianční matice

$$\mathbf{S} = \frac{\sum_{h=1}^{H} \mathbf{S}_h (n_h - 1)}{n - H} = \frac{\sum_{h=1}^{H} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\mathbf{y}_{hi} - \overline{\mathbf{y}}_h) (\mathbf{y}_{hi} - \overline{\mathbf{y}}_h)^T (n_h - 1)}{n - H}$$

$$= \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} (\mathbf{y}_{hi} - \overline{\mathbf{y}}_h) (\mathbf{y}_{hi} - \overline{\mathbf{y}}_h)^T}{n - H} = \frac{\mathbf{E}}{n - H}.$$

Výraz $n - H$ lze přepsat jako $\sum_{h=1}^{H} (n_h - 1)$, z čehož je zřejmé, že $(n_i - 1)$ musí být větší než p. Jinak by pro nějaké h vyšlo $|\mathbf{S}_h|$ rovno nule, tím pádem i M a nemohli bychom použít dále definované statistiky.

Nulovou hypotézu testujeme pomocí dvou různých aproximativních rozdělení statistiky M. První veličina

$$\mathrm{V} = -2 (1 - c_1) \ln (\mathrm{M}),$$

kde

$$c_1 = \frac{2p^2 + 3p - 1}{6 (p + 1) (H - 1)} \left(\sum_{h=1}^{H} \frac{1}{n_h - 1} - \frac{1}{n - H} \right),$$

má přibližně rozdělení chí-kvadrát s $\frac{p(p+1)(H-1)}{2}$ stupni volnosti.

Aproximace může být použita pouze, pokud počet proměnných p a počet skupin H není větší než 5 a rozsahy výběrů ve skupinách jsou alespoň 20. Tento postup je označován jako Boxův test a představuje zobecnění Bartlettova testu shody rozptylů. Bartlettův test je tedy speciálním případem Boxova testu pro $p = 1$.

Druhou testovací statistikou je veličina

$$F = (-2b_1) \ln (\mathrm{M}),$$

která má pro $c_2 > c_1^2$ přibližně F-rozdělení s d_1 a d_2 stupni volnosti, kde

$$d_1 = \frac{p\,(p+1)\,(H-1)}{2}\ ,\ d_2 = \frac{d_1+2}{|c_2 - c_1^2|}\ ,\ b_1 = \frac{1 - c_1 - \frac{d_1}{d_2}}{d_1}$$

$$\text{a} \qquad c_2 = \frac{(p-1)\,(p+2)}{6\,(H-1)} \left(\sum_{h=1}^{H} \frac{1}{(n_h - 1)^2} - \frac{1}{(n-H)^2} \right).$$

V případě, že $c_2 < c_1^2$, tak má statistika F také přibližně F-rozdělení s d_1 a d_2 stupni volnosti, ale je tvaru

$$F = -\frac{b_2 d_2 \ln \mathrm{M}}{d_1 + 2 b_2 d_1 \ln \mathrm{M}},$$

kde

$$b_2 = \left(1 - c_1 - \frac{2}{d_2}\right) d_2.$$

I tento postup je Boxovým vylepšením postupu Bartleta. Boxův test je velmi citlivý na nesplnění předpokladu normality. To se projeví například pokud špičatost dat neodpovídá normálnímu rozdělení, i když MANOVA je k danému porušení normality vcelku robustní. Dalším nedostatkem je, že Boxův test často dojde k závěru zamítnutí nulové hypotézy i při malých rozdílech ve variančních maticích, které nebrání v provedení MANOVy, proto bychom v praxi měli zvážit i použití jiných testů shody variančních matic.

Pokud je ve všech skupinách stejný počet pozorování, tak je mnohorozměrná analýza rozptylu odolná vůči porušení předpokladu o shodnosti variančních matic. Jinak, jak je popsáno v [4], mají odlišné varianční matice významný dopad na chybu prvního druhu.

3.4. Nezávislost skupin

Splnění tohoto předpokladu vychází z designu experimentu. Například při vyšetřování dvou skupin mužů a žen, kdy se u osob zjišťují proměnné výška a váha, by byl jeden z postupů, jak se vyhnout závislosti mezi souborem mužů a žen, že bychom nezahrnuli muže a ženy, kteří jsou v příbuzenském vztahu.

3.5. Velikost výběru

Rozsah výběru ve všech skupinách by měl být větší než počet závislých proměnných. Větší rozsahy výběru zajistí, že je výpočet odolnější vůči porušení předpokladů. Měli bychom mít též na paměti, že pokud je třídění vyvážené, tak je MANOVA odolná vůči porušení předpokladu shody variančních matic.

3.6. Odlehlá pozorování

Mnohorozměrná analýza rozptylu je citlivá na vliv odlehlých pozorování. Pokud je odlehlých pozorování mnoho nebo jsou jejich hodnoty příliš extrémní, měli bychom zvážit, jak s těmito hodnotami naložit. K identifikování odlehlých pozorování nám mohou posloužit například vizualizační nástroje z kapitoly 3.2 nebo testy, které slouží k odhalení odlehlých pozorování. Například test založený na Mahalanobisových vzdálenostech v [3].

3.7. Linearita

Podstatou mnohorozměrné analýzy rozptylu je vytvoření lineárního modelu, proto musí být splněn předpoklad, že v každé skupině existuje mezi závislými proměnnými přibližně lineární vztah. Tento předpoklad lze ověřit tím, že si pro danou skupinu vytvoříme matici bodových grafů všech dvojic proměnných. Z

té můžeme vizuálně posoudit, zda jsou vztahy mezi proměnnými lineární. Při rozhodování je vhodné si pomoci hodnotou korelačního koeficientu pro danou dvojici. Jestliže se ve skupinách nelineární vztahy vyskytují, snižuje se síla testů mnohonásobné analýzy rozptylu.

4. Praktická část

V této části nejprve uvedeme dvě reálné aplikace na použití metody mnohorozměrné analýzy rozptylu. Dále pak budeme pomocí simulací zkoumat chování MANOVy a post-hoc testů. První simulace slouží k zjištění skutečné hladiny významnosti testů z kapitoly 1.1.2. Následně pomocí simulací vytvoříme řezy silofunkcemi pro testy z kapitoly 1.1.2 a budeme sledovat, zda má porušení předpokladu shodných variančních matic nějaký dopad. Uvažovali jsme dva případy. V prvním mají skupiny rozdílné rozsahy a v druhém jsou rozsahy skupin stejné. Nakonec jsme zjišťovali skutečnou hladinu významnosti pro simultánní testy z kapitoly 2. Všechny výpočty budou prováděny pomocí softwaru R.

4.1. Řešené příklady

Tato kapitola obsahuje dva řešené příklady na mnohorozměrnou analýzu rozptylu. U každého příkladu je uveden postup řešení daného problému včetně ověřování předpokladů, které jsou na model MANOVy kladeny.

Shoda variančních matic je testována pomocí Boxova testu, který je obsažen v balíčku `biotools` [10] pod příkazem s názvem `boxM()`. Pro ověření mnohorozměrné normality je využit balíček `mvShapiroTest` [9], ve kterém se nachází zobecněný test Shapira a Wilka na mnohorozměrný případ pod názvem `mvShapiro.Test()`.

Za účelem grafického ověřování mnohorozměrné normality byla využita funkce, jejímž výstupem bude graf založený na Mahalanobisových vzdálenostech, který je popsaný v kapitole 3.2. Vstupem funkce s názvem `graf_mah_vzdalenosti` je datová matice $\boldsymbol{X}$, která obsahuje data, u kterých chceme mnohorozměrnou normalitu prozkoumat, a popis grafu. Kód pro vytvoření grafu je uveden v příloze A.

Pro další testování v případě zamítnutí nulové hypotézy o shodě vektorů středních hodnot byly vytvořeny tři funkce. Jde o postupy z kapitoly 2. První funkce slouží k testování jednotlivých závislých proměnných zvlášť. Do jejího argumentu

se zadává datová matice $\boldsymbol{X}$. Tato matice má rozměry $n \times p$ a musí být uspořádaná tak, aby řádky matice tvořila jednotlivá pozorování, která jsou v matici seřazena tak, aby pozorování ze stejné skupiny byla v matici za sebou. V matici jsou uloženy pouze naměřené hodnoty jednotlivých proměnných, ne kategorická proměnná, která rozděluje pozorování do skupin. Dále se zadává počet skupin v datech H a vektor rozsahů souborů skupin N, kdy rozsahy skupin jsou uspořádány tak, aby odpovídaly pořadí skupin v matici $\boldsymbol{X}$. Výstupem je vektor p hodnot testovací statistiky (12) pro jednotlivé proměnné a odpovídající kvantil chí-kvadrát rozdělení. Popsaná funkce nese název `simultanni_testy_slozek_vektoru` a její kód je obsažen v příloze B.

Další funkce se nazývá `vicerozmerna_obdoba_mnohonasobneho_porovnavani` a slouží k výpočtu vícerozměrné obdoby mnohonásobného porovnávání, kdy porovnáváme všechny možné kombinace dvojic vektorů středních hodnot skupin. Vstupy funkce jsou stejné, jako u funkce předchozí. Výstupem je matice, ve které jsou pro všechny kombinace dvojic skupin uvedeny hodnoty testovací statistiky (13) a je uveden i příslušný kvantil Fisherova F-rozdělení. V příloze C je uveden kód funkce.

Poslední z funkcí sloužících k dalšímu testování při zamítnutí nulové hypotézy o shodě vektorů středních hodnot je funkce pro výpočet simultánních testů v mnohonásobném porovnávání. Funkce má název `simultanni_testy_v_mnohonasobnem_porovnavani`. První tři vstupní argumenty jsou stejné jako u předchozích dvou funkcí. Přidány byly argumenty i a j, které značí, pro které skupiny budeme testování provádět. Výstupem tedy bude p hodnot testovací statistiky (14) pro dvojice výběrových průměrů proměnných zadaných dvou skupin i a j. Dále je ve výstupu uveden i kvantil Fisherova F-rozdělení, podle kterého rozhodneme o nulové hypotéze. Kód funkce obsahuje příloha D.

4.1.1. Analýza srdečního tepu

Data pro následující příklad jsou získána z [12]. Pocházejí ze studie, která měla za cíl vyšetřit, zda mají dva léky vliv na hodnoty srdečního tepu. Datový

soubor se skládá z třiceti vybraných žen, které byly náhodně rozděleny do třech skupin po deseti lidech. Ženy byly ošetřeny v závislosti na skupině placebem, lékem A nebo lékem B a ve čtyřech časech jim byl změřen srdeční tep. Časy jsou označeny jedna až čtyři tak, jak šly chronologicky po sobě. Cílem je pomocí MANOVy zjistit, zda se hodnoty srdečního tepu na hladině významnosti 0,05 liší u třech uvedených typů ošetření.

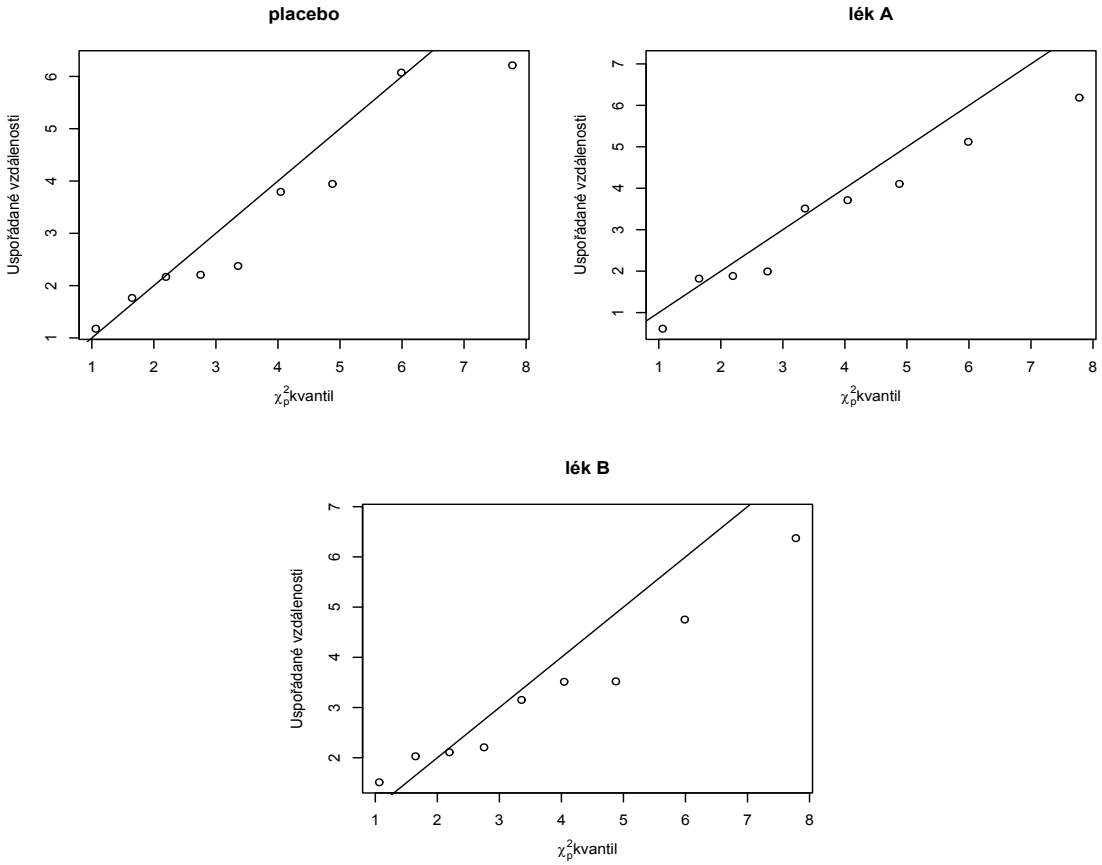
Ze začátku musíme ověřit, zda jsou splněny předpoklady, kterými je mnohorozměrná analýza rozptylu zatížena. Nejprve se podíváme na předpoklad, že pozorování v jednotlivých skupinách pocházejí z mnohorozměrného normálního rozdělení. Pomocí příkazu `shapiro.test`, který provede Shapiro-Wilkův test normality, jsme otestovali jednorozměrnou normalitu ve skupinách. Výsledné p-hodnoty jsou zaznamenány v tabulce 3. Na základě získaných p-hodnot nemůžeme zamítnout ani jednu hypotézu o normalitě proměnných.

skupina	čas	p-hodnota
placebo	1	0.3591
placebo	2	0.6011
placebo	3	0.1595
placebo	4	0.9636
lék A	1	0.7989
lék A	2	0.842
lék A	3	0.06236
lék A	4	0.1677
lék B	1	0.2223
lék B	2	0.2084
lék B	3	0.3574
lék B	4	0.3915

Tabulka 3: P-hodnoty Shapiro-Wilkova testu normality pro data týkající se srdečního tepu.

Mnohorozměrnou normalitu budeme posuzovat na základě grafů Mahalanobiso-

vých vzdáleností pro dané typy ošetření. Výsledné grafy jsou zobrazeny v obrázku 4 a na jejich základě to nevypadá, že by byla mnohorozměrná normalita výrazně porušena. Budeme tedy dále předpokládat, že data ve skupinách pocházejí ze čtyřrozměrných normálních rozdělení.



Obrázek 4: Mahalanobisovy vzdálenosti vzhledem k odpovídajícím kvantilům chí-kvadrát rozdělení pro všechny tři druhy ošetření.

Ověření shody tří variačních matic bylo provedeno následovně:

```
> boxM(x[,2:5],x[,1])

 Box´s M-test for Homogeneity of Covariance Matrices

data:  x[, 2:5]
Chi-Sq (approx.) = 24.408, df = 20, p-value = 0.225
```

Výsledná p-hodnota je rovna 0,225. Jelikož je tato hodnota větší než 0,05, tak nelze zamítnout nulovou hypotézu o shodě variančních matic, a budeme je tak dále považovat za shodné. MANOVu provedeme následovně:

```
> fit=manova(cbind(time1,time2,time3,time4) ~ group, data=x)
> summary(fit, test="Wilks")
          Df    Wilks approx F num Df den Df    Pr(>F)
group      2 0.062801   17.942      8     48 4.824e-12 ***
Residuals 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Pillai")
          Df Pillai approx F num Df den Df    Pr(>F)
group      2 1.4371   15.958      8     50 2.181e-11 ***
Residuals 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Hotelling-Lawley")
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
group      2           6.9625   20.017      8     46 1.317e-12 **
   *
Residuals 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Roy")
          Df    Roy approx F num Df den Df    Pr(>F)
group      2 5.5204   34.502      4     25 7.681e-10 ***
Residuals 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Na základě p-hodnot zamítáme nulovou hypotézu o shodě vektorů středních hodnot pro všechny čtyři testy. To znamená, že je alespoň mezi dvěma ošetřeními rozdíl. Nejprve provedeme simultánní testy o složkách vektorů středních hodnot a vícerozměrnou obdobu mnohonásobného porovnávání. Pomocí simultánních testů se budeme snažit zjistit, ve kterých časech se ošetření liší. Mnohonásobné porov-

nání bude sloužit k porovnání vektorů středních hodnot dvojic ošetření.

```
> simultanni_testy_slozek_vektoru(X,3,c(10,10,10))
[1] "Hodnoty testovací statistiky pro proměnné 1 až p"
[1] 20.97993 16.10987 20.63397 31.60601
[1] "Odpovídající kvantil chí-kvadrát rozdělení"
[1] 15.50731
> vicerozmerna_obdoba_mnohonasobneho_porovnavani(X,3,c(10,10,10))
[1] "Hodnoty testovací statistiky pro jednotlivé dvojice vektorů
    středních hodnot"
     [,1]     [,2]      [,3]
[1,]   NA 6.631984  8.205825
[2,]   NA       NA 16.493240
[3,]   NA       NA        NA
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 2.290481
```

Na základě výsledků simultánních testů zamítáme nulovou hypotézu o shodě středních hodnot třech typů ošetření v jednotlivých časech. Stejně tak zamítáme podle výsledků mnohonásobného porovnávání, že se některé dvojice ošetření shodují. Pomocí simultánních testů v mnohonásobném porovnávání zjistíme, ve kterých časech se dvojice ošetření liší. Výsledky jsou zobrazeny v tomto pořadí: skupina žen užívající placebo a lék A, placebo a lék B a nakonec lék A a lék B.

```
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(10,10,10)
    ,1,2)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1] 2.591051 2.606634 3.604660 6.250394
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 2.290481
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(10,10,10)
    ,1,3)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1] 0.02534418 0.94447174 1.60207101 0.11377953
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 2.290481
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(10,10,10)
    ,2,3)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1] 3.1289111 0.4130221 0.4005178 4.6775591
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 2.290481
```

Podle výsledku simultánního testu pro skupiny žen ošetřené pomocí placeba a

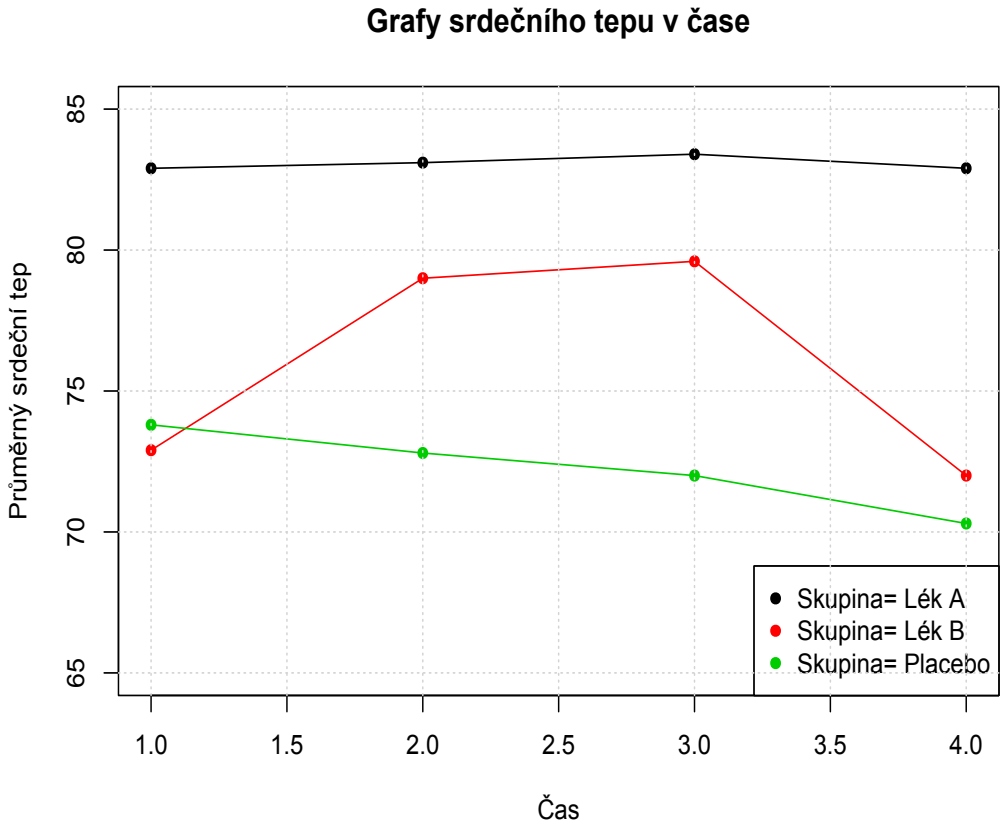
léku A zamítáme nulovou hypotézu o shodě středních hodnot srdečního tepu pro všechny časy měření. Hodnoty srdečního tepu se tedy pro ženy, které dostaly placebo a ty, které dostaly lék A, liší ve všech časech. Naopak u skupin pacientek s placebem a lékem B není signifikantní rozdíl mezi středními hodnotami srdečního tepu ani v jednom čase měření. Pro skupiny s lékem A a lékem B zamítáme nulovou hypotézu o shodě středních hodnot srdečního tepu pro první a poslední měřený čas. Situace se dá přehledně shrnout do grafu, který se vytvoří pomocí následujícího příkazu:

```
prumery=t(aggregate(x[, 2:5], by = list(x[,1]),FUN = mean))
prumery
for (i in 1:3){
  if (i == 1){
    plot(prumery[-c(1), i], type = "l", col = i, ylim = c(65, 85)
        ,
         main = "Grafy srdečního tepu v čase",
         ylab = "Průměrný srdeční tep",xlab = "Čas")
    points(prumery[-c(1), i], type = "p", pch = 16, col = i)
  } else {
    points(prumery[-c(1), i], type = "l", col = i)
    points(prumery[-c(1), i], type = "p", pch = 16, col = i)
  }
}

legend("bottomright", pch = 16,legend = mylegend, col = (1:3))

labels1=c("Lék A", "Lék B", "Placebo")
mylegend <- paste(paste("Skupina=", labels1))
legend("bottomright", pch = 16, legend = mylegend, col = (1:3))
grid()
```

Výsledný graf je zobrazen v obrázku 5. V grafu jsou zaznamenány průměrné hodnoty pro všechny tři skupiny ošetření ve čtyřech časech, kdy bylo prováděno měření. Vidíme z něj například, že skupina s lékem A a skupina s placebem se liší ve všech proměnných, jelikož hodnoty srdečního tepu jsou u skupiny s lékem A ve všech časech výrazně vyšší. Ke grafu je přiložena tabulka 4, ve které je znázorněno, zda byl pomocí simultánních testů v mnohonásobném porovnávání zjištěn rozdíl mezi dvojicemi středních hodnot srdečních tepů v jednotlivých časech měření.



Obrázek 5: Průměrné hodnoty srdečního tepu v časech jedna až čtyři pro tři skupiny ošetření.

	skupina	**lék A**				**lék B**			
skupina	**čas**	**1.**	**2.**	**3.**	**4.**	**1.**	**2.**	**3.**	**4.**
placebo	**1.**	✓				✗			
	2.		✓				✗		
	3.			✓				✗	
	4.				✓				✗
lék B	**1.**	✓							
	2.		✗						
	3.			✗					
	4.				✓				

Tabulka 4: Shrnutí výsledků simultánních testů v mnohonásobném porovnávání pro data týkající se srdečního tepu.

4.1.2. Analýza velikosti kališních a okvětních lístků kosatců

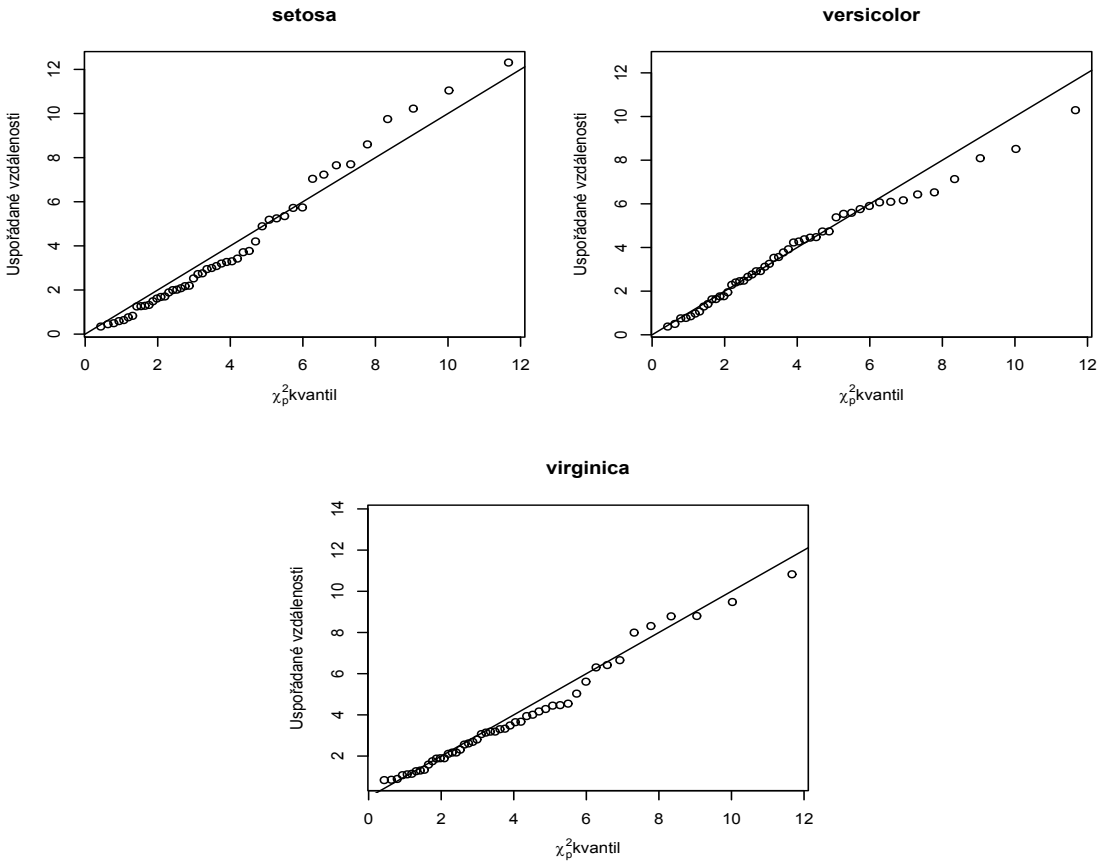
Datový soubor Iris, byl představen Ronaldem Fisherem v roce 1936, jako příklad lineární diskriminační analýzy [13], a je implementovaný v softwaru R pod názvem `iris`. Skládá se ze 150 pozorování rozdělených do tří skupin. Pozorování představují kosatce a třídění do skupin je vyvážené. Skupiny jsou označeny pojmy setosa, versicolor a virginica, které udávají druh kosatce. U každého kosatce se měřily čtyři proměnné: délka kališního lístku, šířka kališního lístku, délka okvětního lístku a šířka okvětního lístku. Cílem je pomocí MANOVy zjistit na hladině významnosti 0,05, zda se rozměry okvětních a kališních lístků kosatců liší pro tři dané druhy kosatců.

Při ověřování předpokladů začneme s mnohorozměrnou normalitou. Pomocí příkazu `shapiro.test` jsme testovali proměnné v rámci tří skupin na normalitu jednorozměrnou. Výsledné p-hodnoty Shapiro-Wilkova testu jsou uvedeny v tabulce 5.

skupina (druh kosatce)	proměnná	p-hodnota
setosa	délka kališního lístku	0.4595
setosa	šířka kališního lístku	0.2715
setosa	délka okvětního lístku	0.05481
setosa	šířka okvětního lístku	8.659e-07
versicolor	délka kališního lístku	0.4647
versicolor	šířka kališního lístku	0.338
versicolor	délka okvětního lístku	0.1585
versicolor	šířka okvětního lístku	0.02728
virginica	délka kališního lístku	0.2583
virginica	šířka kališního lístku	0.1809
virginica	délka okvětního lístku	0.1098
virginica	šířka okvětního lístku	0.08695

Tabulka 5: P-hodnoty Shapiro-Wilkova testu normality pro datový soubor Iris.

Vidíme, že test zamítá hypotézu o normalitě proměnné šířka okvětního lístku u druhů setosa a versicolor. Z obrázku 6, kde jsou zobrazeny grafy Mahalanobisových vzdáleností vzhledem k odpovídajícím kvantilům pro tři druhy kosatců, to však nevypadá, že by byla mnohorozměrná normalita ve skupinách výrazně porušena.



Obrázek 6: Mahalanobisovy vzdálenosti vzhledem k odpovídajícím kvantilům chí-kvadrát rozdělení pro všechny tři druhy kosatců.

Pomocí Shapiro-Wilkova testu mnohorozměrné normality, kdy $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ a $\boldsymbol{X}_3$ označují datové matice pro skupiny setosa, versicolor a virginica, dostaneme tento výsledek:

```
> X1=as.matrix(x[1:50,1:4])
> X2=as.matrix(x[51:100,1:4])
> X3=as.matrix(x[101:150,1:4])
>
> mvShapiro.Test(X1)

  Generalized Shapiro-Wilk test for Multivariate Normality by
     Villasenor-Alva and Gonzalez-Estrada

data:  X1
MVW = 0.96003, p-value = 0.01203

> mvShapiro.Test(X2)

  Generalized Shapiro-Wilk test for Multivariate Normality by
     Villasenor-Alva and Gonzalez-Estrada

data:  X2
MVW = 0.97415, p-value = 0.3183

> mvShapiro.Test(X3)

  Generalized Shapiro-Wilk test for Multivariate Normality by
     Villasenor-Alva and Gonzalez-Estrada

data:  X3
MVW = 0.98521, p-value = 0.9652
```

Vidíme, že na základě tohoto testu zamítáme nulovou hypotézu o mnohorozměrném rozdělení dat pro kosatec druhu setosa. Jelikož víme z kapitoly 3.2, že odchylka od mnohorozměrného normálního rozdělení má za následek pouze malý dopad na chybu prvního druhu, tak budeme předpokládat, že data pro druh kosatce setosa jsou normálně rozdělená. Dále provedeme Boxův test shody variačních matic:

```
> boxM(x[,1:4],x[,5])

  Box´s M-test for Homogeneity of Covariance Matrices

data:  x[, 1:4]
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

Na základě p-hodnoty $2,2 \cdot 10^{-16}$ zamítáme nulovou hypotézu o shodě variačních matic. Z kapitoly 3.3 však víme, že MANOVA je při vyváženém třídění odolná vůči porušení předpokladu shody variančních matic, tak budeme pokračovat v

analýze dál a provedeme MANOVu.

```
> fit=manova(cbind(Sepal.Length,Sepal.Width,Petal.Length,Petal.
    Width) ~ Species, data=x)
> summary(fit, test="Wilks")
            Df    Wilks approx F num Df den Df    Pr(>F)
Species      2 0.023439   199.15      8    288 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Pillai")
            Df Pillai approx F num Df den Df    Pr(>F)
Species      2 1.1919   53.466      8    290 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Hotelling-Lawley")
            Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
Species      2           32.477   580.53      8    286 < 2.2e-16 *
    **
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fit, test="Roy")
            Df    Roy approx F num Df den Df    Pr(>F)
Species      2 32.192     1167      4    145 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Z výsledných p-hodnot pro všechny čtyři testy zamítáme nulovou hypotézu o shodě velikostí kališních a okvětních lístků třech druhů kosatců. Provedeme další vyšetření pomocí simultánních testů jednotlivých složek vektoru.

```
> simultanni_testy_slozek_vektoru(X,3,c(50,50,50))
[1] "Hodnoty testovací statistiky pro proměnné 1 až p"
[1] 140.28875  74.51509 412.71629 384.61876
[1] "Odpovídající kvantil chí-kvadrát rozdělení"
[1] 15.50731
```

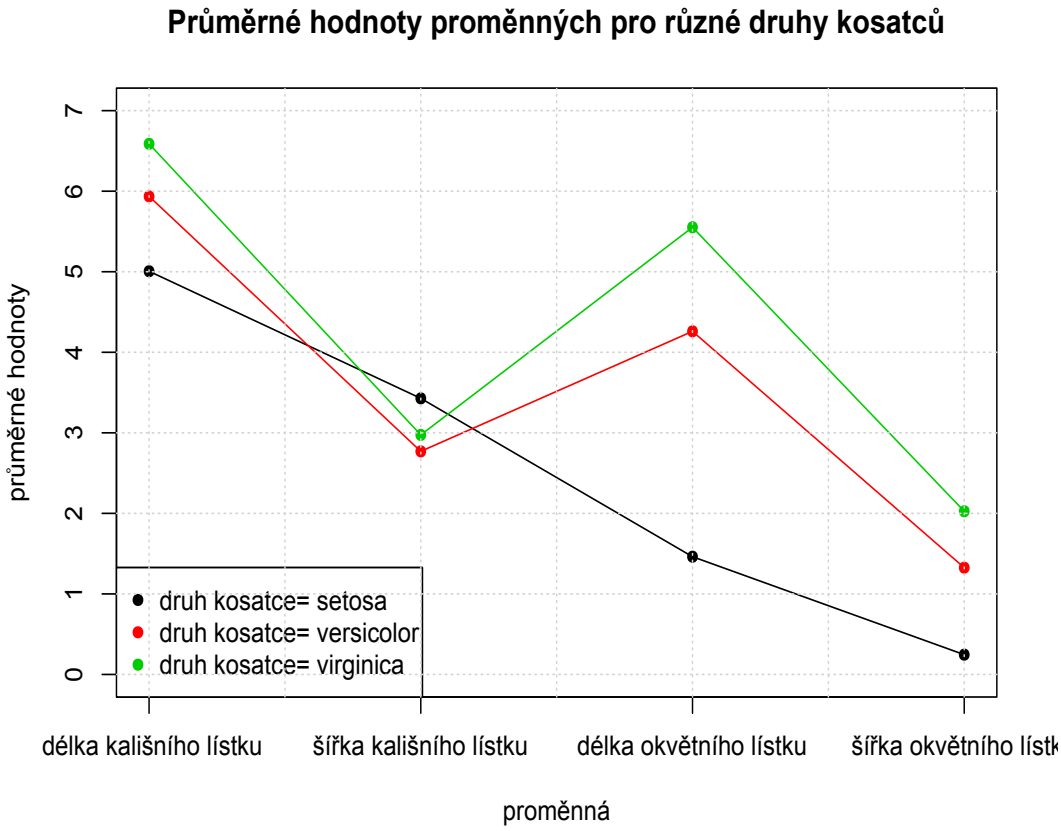
Nulovou hypotézu o shodě středních hodnot zamítáme pro všechny proměnné. Druhy kosatců se tedy neshodují ani v jedné z měřených proměnných, tedy ani v délce a ani v šířce okvětních i kališních lístků. Při porovnání dvojic vektorů pomocí vícerozměrné obdoby mnohonásobného porovnávání dojdeme ke stejnému závěru. Hodnoty se liší i pro všechny dvojice vektorů středních hodnot druhů kosatců.

```
> vicerozmerna_obdoba_mnohonasobneho_porovnavani(X,3,c(50,50,50))
[1] "Hodnoty testovací statistiky pro jednotlivé dvojice vektorů
    středních hodnot"
      [,1]     [,2]      [,3]
[1,]   NA 275.0944 549.13688
[2,]   NA       NA  52.65633
[3,]   NA       NA        NA
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 1.992552
```

Nakonec byly provedeny i simultánní testy pro dvojice vektorů středních hodnot, a to postupně pro dvojice: setosa a versicolor, setosa a virginica a nakonec versicolor a virginica.

```
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(50,50,50)
    ,1,2)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1]   9.990836  11.486487 129.413127  85.254848
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 1.992552
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(50,50,50)
    ,1,3)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1]  28.910053   5.468235 276.521897 231.585615
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 1.992552
> simultanni_testy_v_mnohonasobnem_porovnavani(X,3,c(50,50,50)
    ,2,3)
[1] "Hodnoty testovací statistiky pro dvojice průměrů
    jednotlivých proměnných 1 až p pro dvě zadané skupiny"
[1]  4.910561  1.104068 27.593573 35.815223
[1] "Odpovídající kvantil Fisherova F-rozdělení"
[1] 1.992552
```

Bylo zjištěno, že druhy kosatců setosa a versicolor se liší ve všech proměnných, stejně tak druhy setosa a virginica. Druhy versicolor a virginica se liší v proměnných délka kališního lístku, délka okvětního lístku a šířka okvětního lístku. Tabulka 6 je shrnutím zjištěných signifikantních rozdílů pomocí simultánních testů v mnohonásobném porovnávání. Průměrné hodnoty čtyř proměnných pro všechny tři druhy kosatců jsou zobrazeny v grafu na obrázku 7.



Obrázek 7: Průměrné hodnoty čtyř sledovaných proměnných pro tři druhy kosatců.

	skupina	versicolor				virginica			
skupina	**proměnná**	**DK**	**ŠK**	**DO**	**ŠO**	**DK**	**ŠK**	**DO**	**ŠO**
setosa	**DK**	✓				✓			
	ŠK		✓				✓		
	DO			✓				✓	
	ŠO				✓				✓
virginica	**DK**	✓							
	ŠK		✗						
	DO			✓					
	ŠO				✓				

Tabulka 6: Shrnutí výsledků simultánních testů v mnohonásobném porovnávání pro datový soubor Iris.

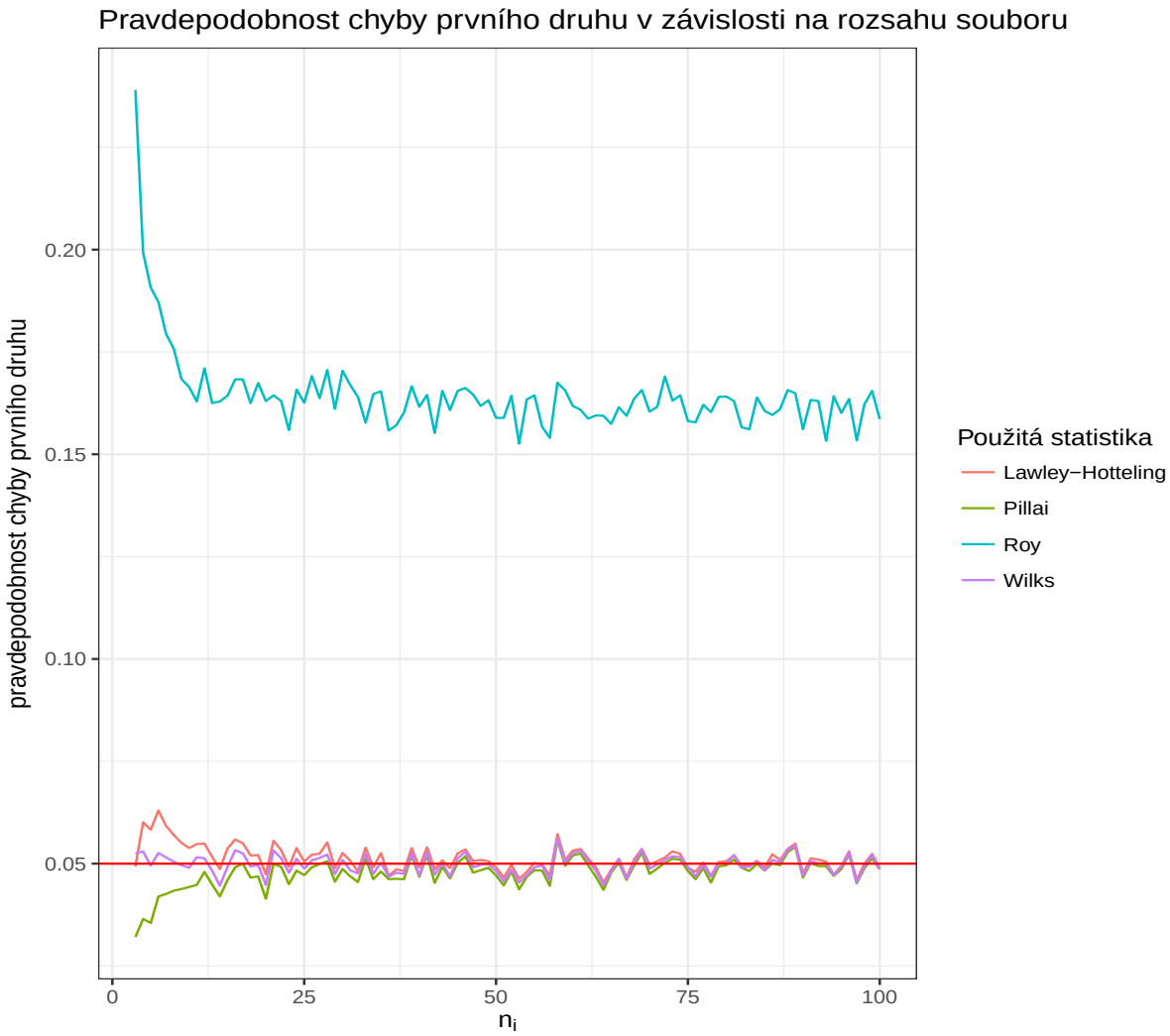
4.2. Simulace

4.2.1. Pravděpodobnost chyby prvního druhu

V této kapitole bude popsána simulace, která měla za cíl zjistit, jaká je skutečná hladina významnosti pro Pillaiův, Wilksův, Lawleyův-Hottelingův a Royův test za platnosti nulové hypotézy pro různě velké rozsahy souboru. Pomocí kódu uvedeného v příloze E byly získány skutečné hladiny významnosti pro různé rozsahy souboru a zaznamenány pro každý test zvlášť.

Uvažovali jsme jednofaktorovou MANOVU s třemi úrovněmi faktoru a u pozorování byly zjišťovány čtyři proměnné. Data ve skupinách byla generována ze čtyřrozměrných normálních rozdělení s nulovým vektorem středních hodnot a jednotkovou varianční maticí. Rozsahy výběrů pro všechny skupiny byly stejné a jejich hodnoty se pohybovali od 3 do 100. Bylo provedeno 10 000 opakování pro každou hodnotu rozsahu výběru. Skutečná hladina významnosti, na které byl test realizován, byla získána jako relativní četnost p-hodnot menších než stanovená hodnota $\alpha = 0,05$ při 10 000 opakováních testu.

Výsledné hodnoty hladin významnosti vzhledem k rozsahu souboru jsou vykresleny v grafu na obrázku 9. Vidíme, že hladina významnosti Royova testu se pohybuje kolem hodnoty 0,16, ostatní tři testy už dosahují pro dostatečné rozsahy výběru přibližně požadované hodnoty pravděpodobnosti chyby prvního druhu 0,05. Z toho vyvozujeme závěr, že Royův test, je ve srovnání s dalšími třemi testy, liberální, až do té míry, že velmi výrazně nedodržuje požadovanou hladinu významnosti. Royův test má tedy tendenci najít statisticky významný rozdíl mezi porovnávanými skupinami, ačkoliv to ve skutečnosti není pravda. Zbývající tři testy, Pillaiův, Wilksův a Lawleyho-Hottelingův test, dodržují požadovanou hladinu významnosti.



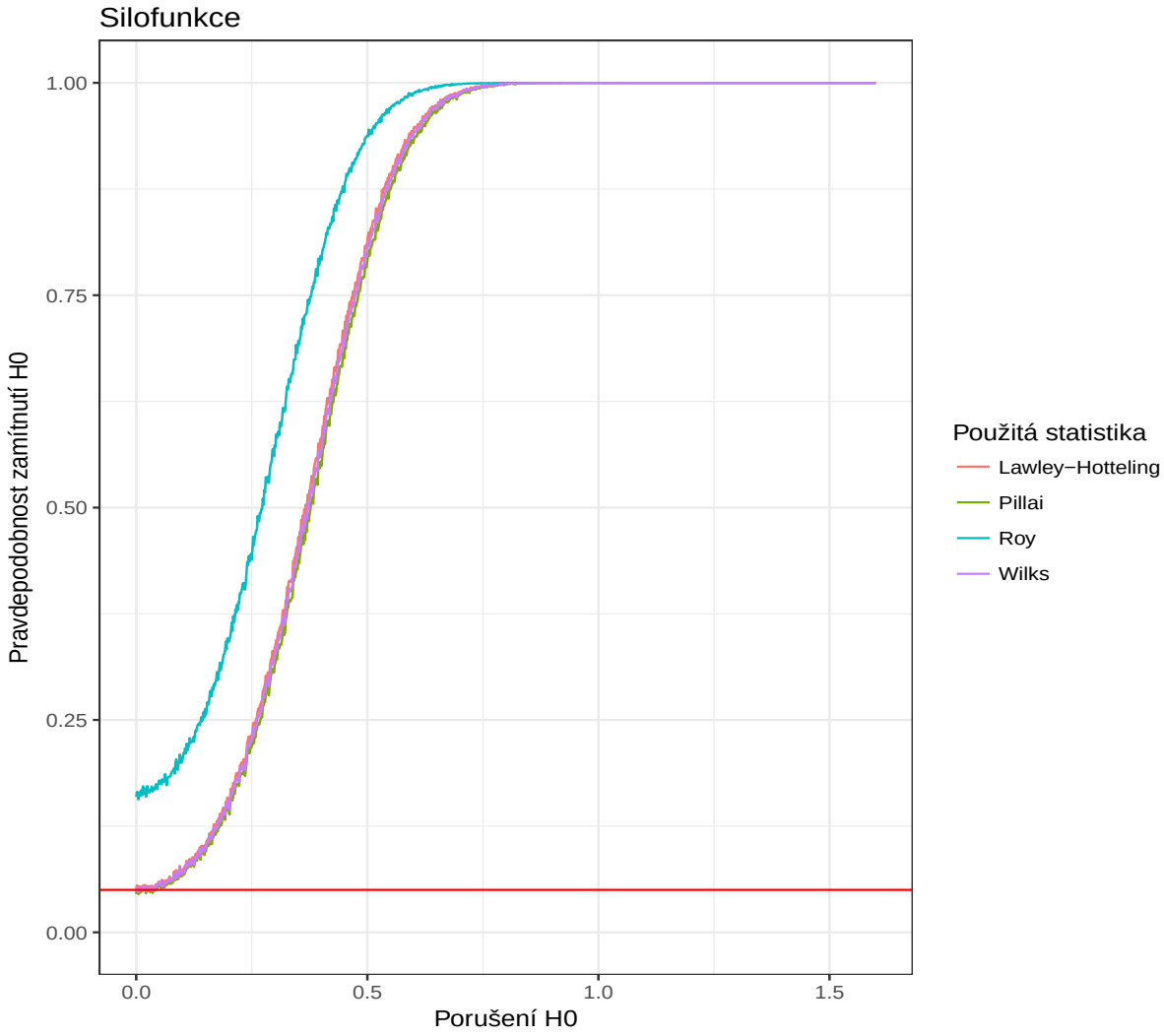
Obrázek 8: Pravděpodobnost chyby prvního druhu u čtyř testů pro jednofaktorovou MANOVu s třemi úrovněmi faktoru pro různé rozsahy skupin.

4.2.2. Síla testu v případě porušení předpokladu shody variančních matic pro různé rozsahy výběru

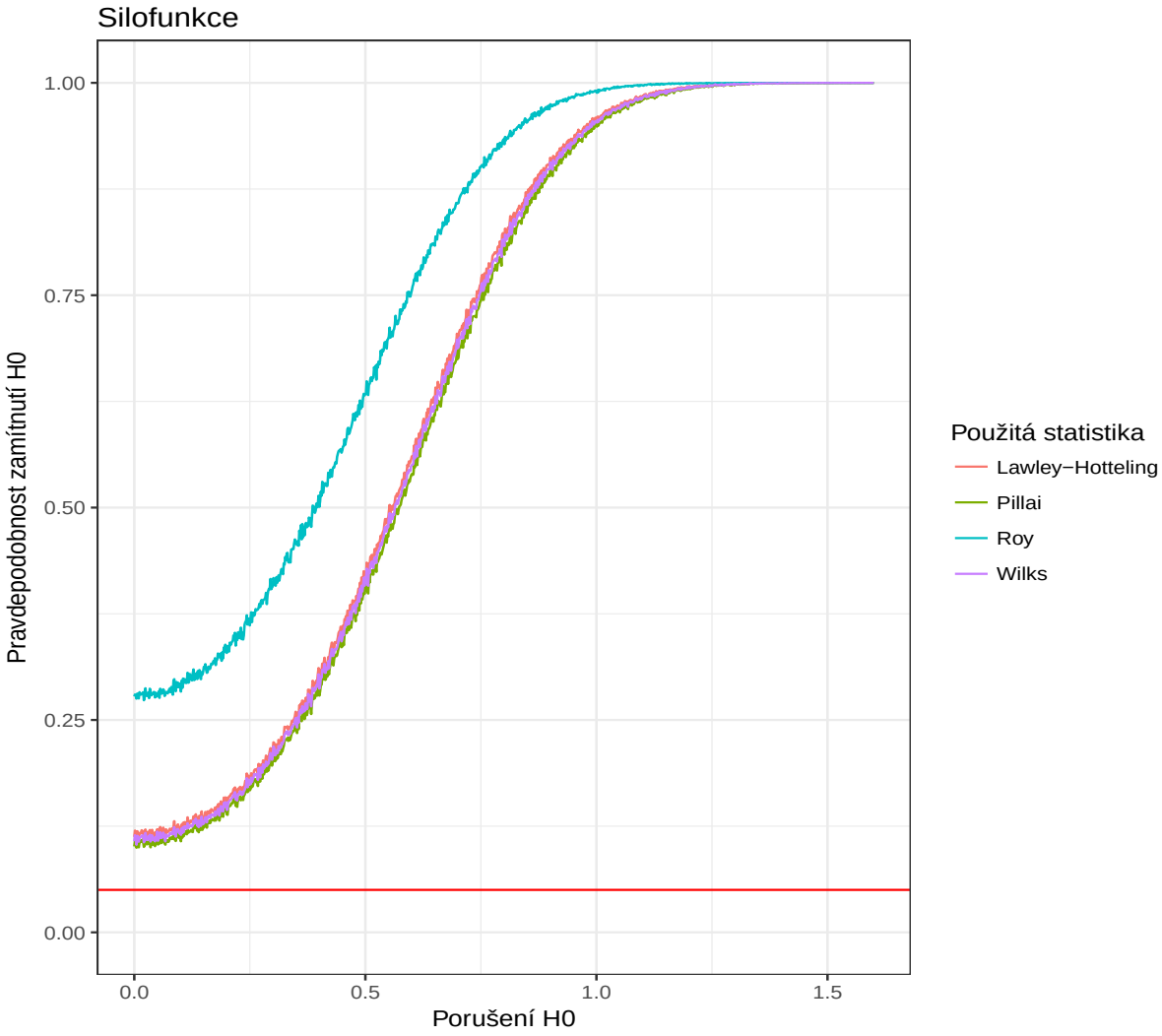
Cílem následující studie je porovnat silofunkce Pillaiova, Wilksova, Lawleyho-Hottelingova a Royova testu pro jednofaktorovou MANOVU pro shodné a rozdílné varianční matice porovnávaných náhodných výběrů.

Situace je taková, že jsme si vygenerovali data z čtyřrozměrných normálních rozdělení pro tři skupiny s rozsahy souboru 25, 20 a 30. U druhé a třetí skupiny jsme uvažovali nulové vektory středních hodnot. Vektor středních hodnot první skupiny se obecně skládal ze čtyř hodnot j. Hodnota j bude představovat hodnoty z intervalu 0 až 1.6 a jedná se o míru porušení nulové hypotézy. Čím vyšší hodnota, tím dochází k většímu porušení nulové hypotézy o shodě vektorů středních hodnot. Následně na hladině významnosti 0,05 provedeme pro všechny hodnoty j 10 000 opakování. Vypočítáme hodnoty všech čtyř testových statistik a zjistíme k nim odpovídající hodnoty pravděpodobností zamítnutí nulové hypotézy. Kód pro shodné varianční matice, kdy byly uvažovány varianční matice rovny jednotkovým maticím, je uveden v příloze F. Kód pro odlišné varianční matice by byl stejný, akorát se místo jednotkových matic zadají matice diagonální. Pro první skupinu uvedeme diagonální matici s hodnotou 3 na diagonále a nulovými ostatními prvky, stejně pro druhou, ale s hodnotu 5, a ve třetí skupině zůstane matice jednotková. V příloze G se nachází kód, kterým se data vynesou do grafů.

Výsledné silofunkce pro shodné varianční matice jsou zobrazeny v obrázku 9, silofunkce pro rozdílné varianční matice vidíme na obrázku 10. Z výsledků vyplývá, že v našem případě při porušení předpokladu shody variančních matic, výrazně klesá síla testu a pravděpodobnost chyby prvního druhu vzroste přibližně dvojnásobně.



Obrázek 9: Řezy silofunkcí testů pro jednofaktorovou MANOVu pro shodné varianční matice při $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \mathbf{0}$.



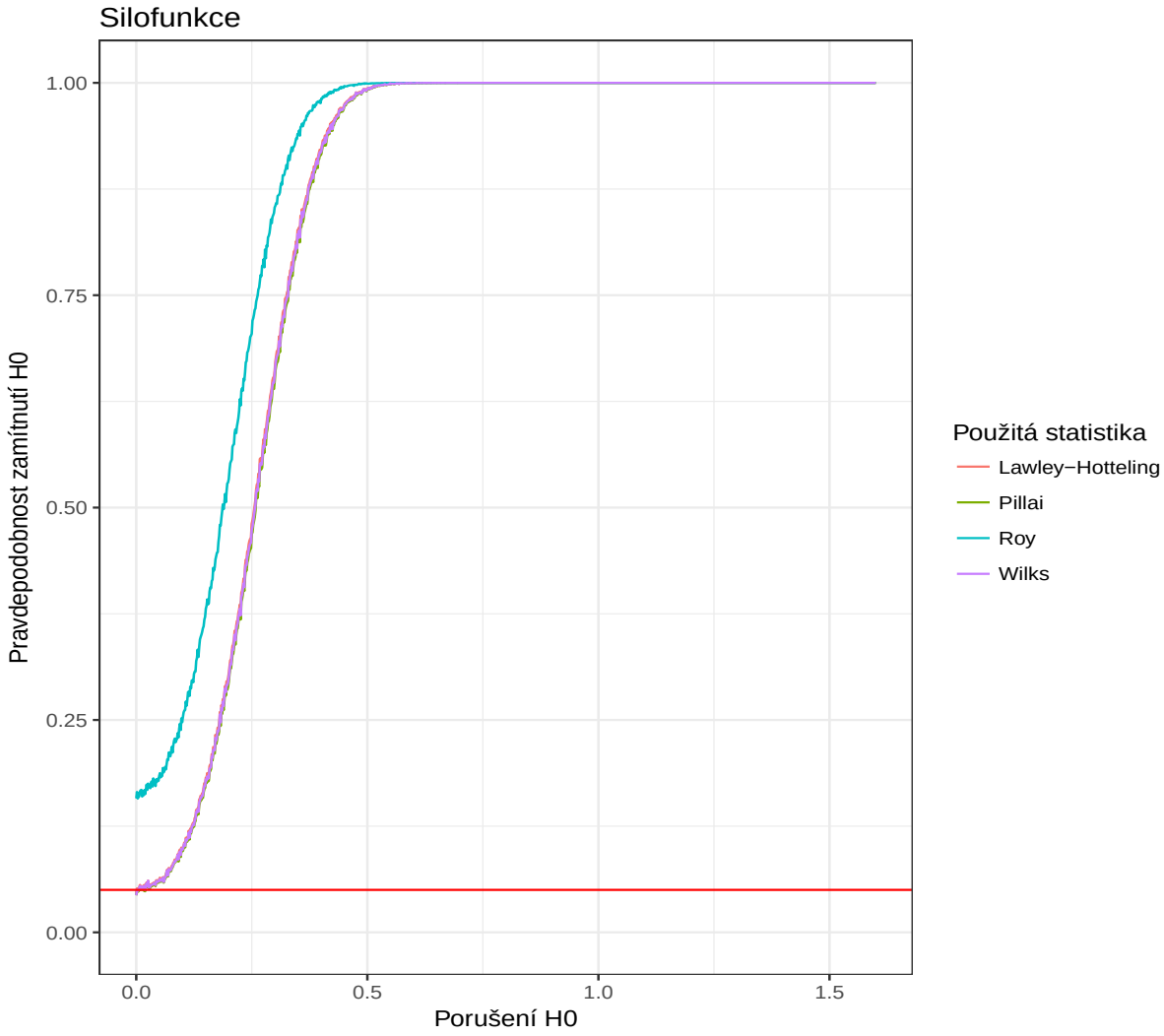
Obrázek 10: Řezy silofunkcí testů pro jednofaktorovou MANOVu pro rozdílné varianční matice při $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \mathbf{0}$.

4.2.3. Síla testu v případě porušení předpokladu shody variančních matic pro stejné rozsahy výběru

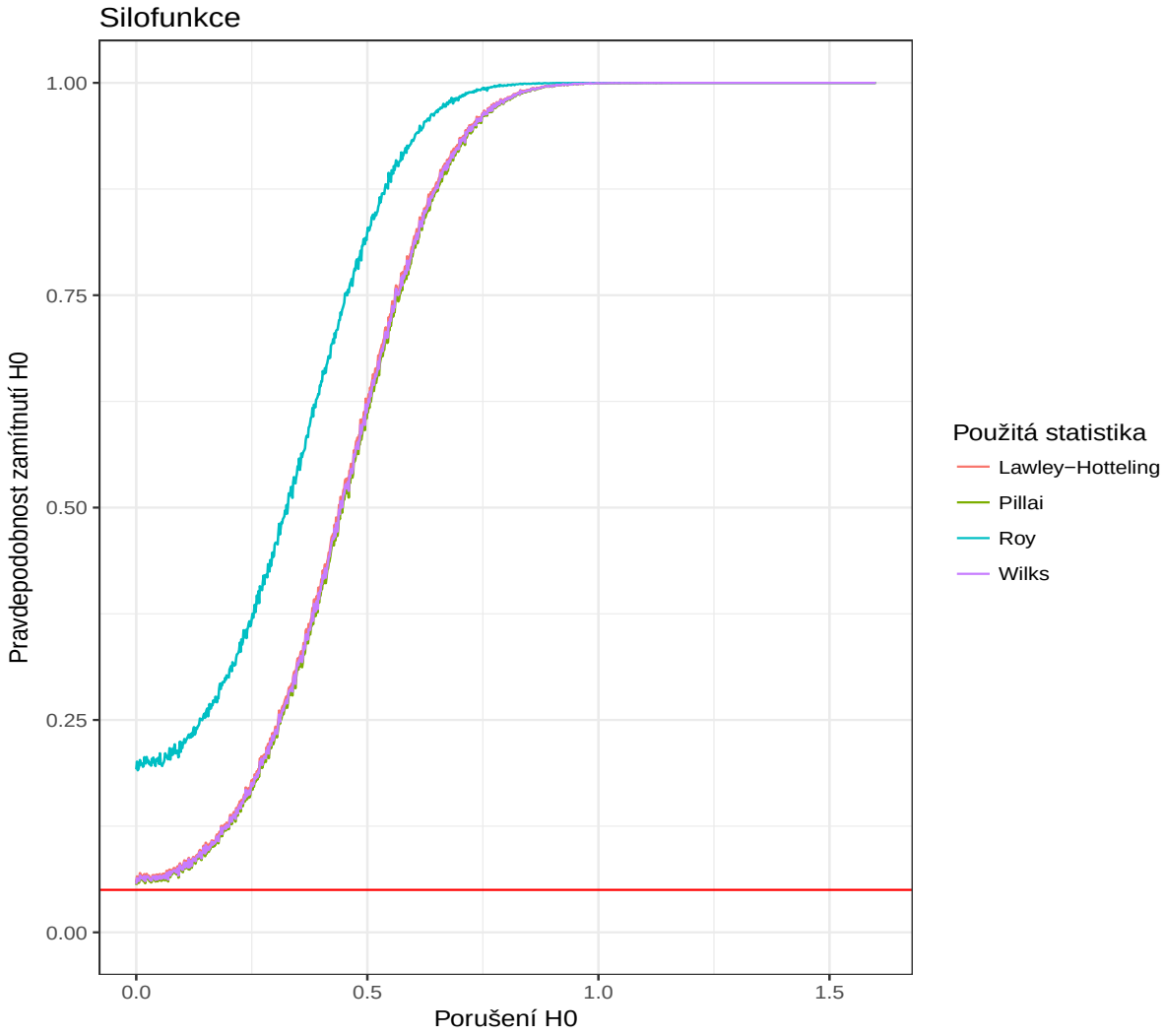
Cílem této simulace je ověřit, zda je MANOVA skutečně odolná na porušení předpokladu o shodě variančních matic pro stejné rozsahy výběru ve skupinách. Vytvoříme grafy řezů silofunkcí pro Pillaiův, Wilksův, Lawleyho-Hottelingův a Royův test, pro shodné a rozdílné varianční matice a výsledné grafy porovnáme. Situace je stejná jako v případě kapitoly 4.2.2, pouze budeme uvažovat vyvážené třídění s 50 pozorováními ve skupinách.

Řezy silofunkcí, pro případ bez porušení předpokladu shody variančních matic, jsou zobrazeny v obrázku 11. Při rozdílných variančních maticích mají řezy silofunkcemi tvar z obrázku 12.

Z výsledných silofunkcí vidíme, že MANOVA je při vyváženém třídění skutečně odolná na porušení předpokladu shody variančních matic, co se týče pravděpodobnosti chyby prvního druhu, která se pohybuje kolem zadané hladiny významnosti $\alpha = 0,05$. Pro rozdílné varianční matice se však snižuje síla testů, tzn. snižuje se pravděpodobnost zamítnutí nulové hypotézy o shodě vektorů středních hodnot v případě porušení nulové hypotézy. Rozdíly mezi skupinami tedy nebudou zachycovány tak dobře, jako v případě shodných variančních matic.



Obrázek 11: Řezy silofunkcí testů pro jednofaktorovou MANOVu pro shodné varianční matice při $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \mathbf{0}$.



Obrázek 12: Řezy silofunkcí testů pro jednofaktorovou MANOVu pro rozdílné varianční matice při $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \mathbf{0}$.

4.2.4. Hladina významnosti post-hoc testů

Další otázkou je skutečná hladina významnosti, na které jsou prováděny post-hoc testy z kapitol 2.1, 2.2 a 2.4. Obecně jsme uvažovali MANOVu s jedním faktorem a pozorování z p-rozměrných normálních rozdělení s nulovou střední hodnotou a korelační koeficienty mezi všemi dvojicemi proměnných rovny 0,3. Počet proměnných p jsme volili roven 2, 3, 5 a 10. Stejných hodnot nabývaly i počty skupin. Rozsah ve skupinách byl pevně zvolen a byl roven hodnotě 100. Pro každou kombinaci počtu proměnných a skupin bylo provedeno 10 000 opakování všech třech skupin post-hoc testů. Pro jednotlivé post-hoc testy byly získány relativní četnosti toho, že alespoň jedna p-hodnota testů z dané skupiny post-hoc testů je menší než stanovená hladina významnosti $\alpha = 0,05$ při 10 000 opakováních testů. Kód uvedený v příloze H slouží k získání výsledků pro 2 skupiny a 3 proměnné.

Výsledné hodnoty pro všechny kombinace a testy jsou zachyceny v tabulkách 7, 8 a 9. Tabulky tedy obsahují odhady pravděpodobnosti, že alespoň jeden ze skupiny post-hoc testů daného typu nesprávně zamítne nulovou hypotézu. Tento odhad se nazývá family-wise error rate a udává hladinu významnosti jednotlivých post-hoc testů. Hladina významnosti by se u simultánního testu měla pohybovat kolem zvolené hladiny významnosti, tedy 0,05.

Z hodnot pravděpodobností chyb prvního druhu v tabulkách 7, 8 a 9 vidíme, že všechny tři testy jsou ve většině případů velmi konzervativní, zadané hladiny významnosti 0,05 nedosahují. Jen test z kapitoly 2.2 se kolem zadané hladiny významnosti α pohybuje, ale pouze v případě, kdy pozorování pocházejí ze dvou skupin.

počet skupin \počet proměnných	2	3	5	10
2	0,0277	0,0149	0,0025	0,001
3	0,0164	0,0035	0,0002	0,0000
5	0,0085	0,0014	0,0000	0,0000
10	0,0017	0,0001	0,0000	0,0000

Tabulka 7: Pravděpodobnosti chyby prvního druhu testu z kapitoly 2.1 pro vyvážené třídění, kdy rozsah skupin je roven 100.

počet skupin \počet proměnných	2	3	5	10
2	0,0469	0,0520	0,0489	0,0469
3	0,0255	0,0147	0,0068	0,0022
5	0,0038	0,0013	0,0002	0,0000
10	0,0000	0,0000	0,0000	0,0000

Tabulka 8: Pravděpodobnosti chyby prvního druhu testu z kapitoly 2.2 pro vyvážené třídění, kdy rozsah skupin je roven 100.

počet skupin \počet proměnných	2	3	5	10
2	0,0277	0,0149	0,0025	0,001
3	0,0127	0,0020	0,0001	0,0000
5	0,0017	0,0002	0,0000	0,0000
10	0,0000	0,0000	0,0000	0,0000

Tabulka 9: Pravděpodobnosti chyby prvního druhu testu z kapitoly 2.4 pro vyvážené třídění, kdy rozsah skupin je roven 100.

Závěr

Cílem práce bylo uvést metodu mnohorozměrné analýzy rozptylu sloužící k porovnávání skupin několika mnohorozměrných pozorování. Metoda byla popsána v první kapitole pro třídění do skupin na základě jednoho a dvou faktorů. U obou postupů jsou uvedeny explicitní vzorce potřebné k provedení metody. Dále bylo představeno testování pro případ, kdy je nalezen rozdíl mezi skupinami pomocí metody mnohorozměrné analýzy rozptylu. Post-hoc testy jsou uvedeny v druhé kapitole spolu s potřebnými vzorci a postupy testování. Třetí kapitola obsahuje několik možných způsobů ověřování předpokladů, které jsou kladeny na model mnohorozměrné analýzy rozptylu.

Čtvrtá kapitola tvoří praktickou část, která byla vytvořena s využitím softwaru R. Nejprve jsou uvedeny dva příklady, na které byla aplikována MANOVA, a následně i post-hoc testy z druhé kapitoly. Na závěr je uvedeno několik simulací. První simulace byla provedena za účelem prozkoumat, jaká je skutečná hladina významnosti pro jednotlivé testovací statistiky popsané v první kapitole. Bylo zjištěno, že Royův test je liberálnější, a v nasimulovaném případě se skutečná hladina významnosti tohoto testu pohybovala kolem hodnoty 0,16, zatímco u ostatních tří testů kolem požadované hodnoty 0,05. Dále bylo potvrzeno, že porušení předpokladu shody variančních matic při různých rozsazích skupin, má za následek snížení síly testu. V případě stejných rozsahů výběru se hladina významnosti pohybuje i při nedodržení předpokladu shodných variančních matic kolem zadané hladiny významnosti testů. Na závěr bylo zjištěno, že post-hoc testy z druhé kapitoly jsou v některých situacích velmi konzervativní.

Literatura

[1] Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

[2] Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I. (2007). Vícerozměrné statistické metody 1. Praha: Informatorium.

[3] Rencher, A.C. (2002). Methods of Multivariate Analysis, 2nd Edition. New York: Wiley.

[4] Stevens, J. (2002). Applied Multivariate Statistics for the Social Sciences. Mahwah, N.J.: Lawrence Erlbaum Associates.

[5] Budíková, M. (2014). Využití vícerozměrné analýzy rozptylu v psychometrii. Kvaternion 3 (1), 3-15.

[6] Kotlorz, L. (2012). Testy normality (bakalářská práce). Praha.

[7] Shapiro, S. S., Wilk, M. B. (December 1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, Vol. 52, No. 3/4 , pp. 591-611. Biometrika Trust.

[8] Anderson, T.W., Darling, D.A. (December 1954). A Test of Goodness of Fit. Journal of the American Statistical Association, Vol. 49, No. 268, pp. 765-769.

[9] Sakthivel, R., William, M.L. (December 2015). An Extension of Shapiro-Wilk's Test for Multivariate Normality and Power Investigation for Comtaminated Alternatives. International Journal of Scientific and Research Publications, Vol. 5, No. 12.

[10] Package biotools [online]. [cit. 2017-03-25]. Dostupné z: https://cran.r-project.org/web/packages/biotools/biotools.pdf

[11] Package mvShapiroTest [online]. [cit. 2017-03-25]. Dostupné z: https://cran.r-project.org/web/packages/mvShapiroTest/mvShapiroTest.pdf

[12] Qeadan, F. (2015). On MANOVA using STATA, SAS & R. A short course in biostatistics for the Mountain West Clinical Translational Research Infrastructure Network (grant 1U54GM104944) and UNM Clinical & Translational Science Center (CTSC) (grant UL1TR001449). University of New Mexico Health Sciences Center. Albuquerque, New Mexico.

[13] Iris flower data set [online]. [cit. 2017-03-25]. Dostupné z: https://en.wikipedia.org/wiki/Iris_flower_data_set

[14] Lecture 9. MANOVA [online]. [cit. 2017-03-25]. Dostupné z: http://www.stat.pitt.edu/sungkyu/course/2221Fall13/lec9_manova.pdf

Přílohy

A. R-kód grafu založeného na Mahalanobisových vzdálenostech

```
graf_mah_vzdalenosti=function(X,popis){
p=ncol(X)
X_prumer=colMeans(X)
S=cov(X)
d=apply(X, 1, function(X) t(X-X_prumer)%*%solve(S)%*%(X-X_prumer)
   )
qc=qchisq((1:nrow(X))/nrow(X), df=p)
sd=sort(d)
plot(qchisq((1:nrow(X))/nrow(X), df=p),
     sort(d),
     xlab=expression(paste(chi[p]^2, "kvantil")),
     ylab="Uspořádané vzdálenosti",
     main=popis)
abline(a=0, b=1)}
```

B. R-kód pro výpočet simultánních testů složek vektorů středních hodnot

```
simultanni_testy_slozek_vektoru=function(X, H, N){

  n=sum(N)
  p=ncol(X)

  # matice vektorů průměrů jednotlivých skupin
  s=matrix(rep(0,H*p), nrow=H, ncol=p)
  nn=numeric(H)
  for (q in 2:H) {
    for (k in 1:(q-1)) {
      nn[q]=nn[q]+N[k]
    }}
  for (j in 1:H) {
     s[j,]=colSums(X[(nn[j]+1):(nn[j]+N[j]),])
    }

  y_skup_prum=matrix(rep(0,H*p), nrow=H, ncol=p)
  for (l in 1:H) {
      y_skup_prum[l,]=s[l,]/N[l]
    }
  # celkový vektor průměrů
  y_celk_prum=colSums(s[1:H,])/n

  # matice E a T
  T=matrix(rep(0,p*p), ncol=p, nrow=p)
  for (m in 1:n) {
    T=T+((X[m,]-y_celk_prum)%*%(t(X[m,]-y_celk_prum)))
  }

  B=matrix(rep(0,p*p), ncol=p)
  for (q in 1:H) {
  B=B+(N[q]*((y_skup_prum[q,]-y_celk_prum)%*%(t(y_skup_prum[q,]-y
      _celk_prum))))
  }
  E=T-B

  # hodnoty testovací statistiky
  test=numeric(p)
  for (i in 1:p) {
    test[i]=-(n-((p+H)/2)-1)*(log((E[i,i])/(T[i,i])))
  }
    print('Hodnoty testovací statistiky pro proměnné 1 až p')
    print(test)
    print('Odpovídající kvantil chí-kvadrát rozdělení')
    df=p*(H-1)
    test_kvantil=qchisq(0.95, df=df)
    print(test_kvantil) }
```

C. R-kód pro výpočet vícerozměrné obdoby mnohonásobného porovnávání

```
vicerozmerna_obdoba_mnohonasobneho_porovnavani=function(X, H, N){

  n=sum(N)
  p=ncol(X)

  # vektory průměrů jendotlivých skupin
  s=matrix(rep(0,H*p), nrow=H, ncol=p)
  nn=numeric(H)
  for (q in 2:H) {
    for (k in 1:(q-1)) {
      nn[q]=nn[q]+N[k]
    }}
  for (j in 1:H) {
    s[j,]=colSums(X[(nn[j]+1):(nn[j]+N[j]),])
  }

  y_skup_prum=matrix(rep(0,H*p), nrow=H, ncol=p)
  for (l in 1:H) {
    y_skup_prum[l,]=s[l,]/N[l]
  }

  # celkový vektor průměrů
  y_celk_prum=colSums(s[1:H,])/n

  # matice E
  T=matrix(rep(0,p*p), ncol=p, nrow=p)
  for (m in 1:n) {
    T=T+((X[m,]-y_celk_prum)%*%(t(X[m,]-y_celk_prum)))
  }

  B=matrix(rep(0,p*p), ncol=p)
  for (q in 1:H) {
    B=B+(N[q]*((y_skup_prum[q,]-y_celk_prum)%*%(t(y_skup_prum[q
       ,]-y_celk_prum))))
  }

  E=T-B

  # hodnoty testovací statistiky
  test=matrix(, nrow=H, ncol=H)
   for (i in 1:H){
    for (k in i:H){
      if (i!=k) {
        test[i,k]=((n-H-p+1)/((H-1)*p))*((N[i]*N[k])/(N[i]+N[k]))
           *
          ((t(y_skup_prum[i,]-y_skup_prum[k,]))%*%(solve(E))%*%(y
             _skup_prum[i,]-y_skup_prum[k,]))}
```

```
    }}
 print('Hodnoty testovací statistiky pro jednotlivé dvojice
    vektorů středních hodnot')
 print(test)
 ni1=((H-1)*p*(n-H-p))/(n-2-(H-1)*p)
 ni2=n-H-p+1
 print('Odpovídající kvantil Fisherova F-rozdělení')
 test_kriterium=qf(0.95, df1=ni1, df2=ni2)
 print(test_kriterium)
}
```

D. R-kód pro výpočet simultánních testů v mnohonásobném porovnávání

```
simultanni_testy_v_mnohonasobnem_porovnavani=function(X,H,N,i,j){

  n=sum(N)
  p=ncol(X)

  # vektory průměrů jednotlivých skupin
  s=matrix(rep(0,H*p), nrow=H, ncol=p)
  nn=numeric(H)
  for (q in 2:H) {
    for (k in 1:(q-1)) {
      nn[q]=nn[q]+N[k]
    }}
  for (l in 1:H) {
    s[l,]=colSums(X[(nn[l]+1):(nn[l]+N[l]),])
  }

  y_skup_prum=matrix(rep(0,H*p), nrow=H, ncol=p)
  for (m in 1:H) {
    y_skup_prum[m,]=s[m,]/N[m]
  }

  # celkový vektor průměrů
  y_celk_prum=colSums(s[1:H,])/n

  # matice S

  T=matrix(rep(0,p*p), ncol=p, nrow=p)
  for (o in 1:n) {
    T=T+((X[o,]-y_celk_prum)%*%(t(X[o,]-y_celk_prum)))
  }

  B=matrix(rep(0,p*p), ncol=p)
  for (r in 1:H) {
    B=B+(N[r]*((y_skup_prum[r,]-y_celk_prum)%*%(t(y_skup_prum[r
        ,]-y_celk_prum))))
  }

  E=T-B
  S=E/(n-H)

  # hodnoty testovací statistiky
  test=numeric(p)
  for(t in 1:p){
    test[t]=((n-H-p+1)/((H-1)*p*(n-H)))*((N[i]*N[j])/(N[i]+N[j]))
        *(((y_skup_prum[i,t]-y_skup_prum[j,t])^2)/S[t,t])
  }
```

```
45    print('Hodnoty testovací statistiky pro dvojice průměrů
          jednotlivých proměnných 1 až p pro dvě zadané skupiny')
46    print(test)
47    print("Odpovídající kvantil Fisherova F-rozdělení")
48    ni1=((H-1)*p*(n-H-p))/(n-2-(H-1)*p)
49    ni2=n-H-p+1
50    test_kriterium=qf(0.95, df1=ni1, df2=ni2)
51    print(test_kriterium)
52  }
```

E. R-kód k simulaci z kapitoly 4.2.1

```r
library(MASS)

alpha=0.05
pocet=10000 # počet opakování testu

v_p=c(rep(0,pocet))
v_w=c(rep(0,pocet))
v_lh=c(rep(0,pocet))
v_r=c(rep(0,pocet))

k=3
p=4
mu = c(0,0,0)

N = 3:100
P1 = numeric(length(N))
P2 = numeric(length(N))
P3 = numeric(length(N))
P4 = numeric(length(N))

for(l in 1:length(N)){
  n = N[l]
  X=data.frame(skupina = c(rep('A',n),rep('B',n),rep('C',n)),
      matrix(numeric(p*n*k),ncol=p, nrow=n*k))
  for(i in 1:pocet){
    for(j in 1:k){
      X[((j-1)*n+1):(j*n),2:(p+1)] = mvrnorm(n, mu=c(rep(mu[j],p)
          ), Sigma=diag(1,nrow=p))
    }
    m=manova(cbind(X1,X2,X3,X4)~skupina,data=X)

    test1=summary(m,test="Pillai")
    test2=summary(m,test="Wilks")
    test3=summary(m,test="Hotelling-Lawley")
    test4=summary(m,test="Roy")

    v_p[i]=test1$stats[1,6]
    v_w[i]=test2$stats[1,6]
    v_lh[i]=test3$stats[1,6]
    v_r[i]=test4$stats[1,6]
  }
  P1[l] = sum(v_p<=alpha)/pocet
  P2[l] = sum(v_w<=alpha)/pocet
  P3[l] = sum(v_lh<=alpha)/pocet
  P4[l] = sum(v_r<=alpha)/pocet
}
```

F. R-kód k simulaci z kapitoly 4.2.2 a 4.2.3

```
alpha=0.05 # hladina významnosti
pocet=10000 # počet opakování

dol=0 # dolní mez intervalu
hor=1.6 # horní mez intervalu
int=c(dol,hor) # interval, ze kterého budou brány hodnoty vektoru
    středních hodnot pro  první skupinu

mu=seq(int[1],int[2],0.006) # dělení intervalu

v_p=matrix(rep(0,it*length(mu)), ncol=pocet)
v_w=matrix(rep(0,it*length(mu)), ncol=pocet)
v_lh=matrix(rep(0,it*length(mu)), ncol=pocet)
v_r=matrix(rep(0,it*length(mu)), ncol=pocet)

k=3 # počet skupin
p=4 # počet proměnných
n1=25
n2=20
n3=30
n=n1+n2+n3 # celkový rozsah souboru

X=matrix(rep(0,k*n*p),nrow=k*n,ncol=p)

for(i in 1:pocet){
  for (j in 1:length(mu)){
  x1=mvrnorm(n1, mu=c(rep(mu[j],p)), Sigma=diag(1,nrow=p))
  x2=mvrnorm(n2, mu=c(rep(0,p)), Sigma=diag(1,nrow=p))
  x3=mvrnorm(n3, mu=c(rep(0,p)), Sigma=diag(1,nrow=p))

  x=rbind(x1,x2,x3)
  skupina=c(rep('A',n1),rep('B',n2),rep('C',n3))

  X=data.frame(skupina,x)

  m=manova(cbind(X1,X2,X3,X4)~skupina,data=X)

  test1=summary(m,test="Pillai")
  test2=summary(m,test="Wilks")
  test3=summary(m,test="Hotelling-Lawley")
  test4=summary(m,test="Roy")

  v_p[j,i]=round(test1$stats[1,6],6)
  v_w[j,i]=round(test2$stats[1,6],6)
  v_lh[j,i]=round(test3$stats[1,6],6)
  v_r[j,i]=round(test4$stats[1,6],6)
}}

s_p=c(rep(0,length(mu)))
```

```
s_w=c(rep(0,length(mu)))
s_lh=c(rep(0,length(mu)))
s_r=c(rep(0,length(mu)))

for (k in 1:length(mu)){
  s_p[k]=1-(sum(v_p[k,]>0.05))/pocet
  s_w[k]=1-(sum(v_w[k,]>0.05))/pocet
  s_lh[k]=1-(sum(v_lh[k,]>0.05))/pocet
  s_r[k]=1-(sum(v_r[k,]>0.05))/pocet
}

statistika=c(rep("Pillai",length(mu)),rep("Wilks",length(mu)),rep
    ("Lawley-Hotteling",length(mu)),rep("Roy",length(mu)))
S=c(s_p,s_w,s_lh,s_r)
P=c(rep(mu,4))
data=data.frame(statistika,P,S)
```

G. R-kód pro vytvoření grafů z kapitol 4.2.2 a 4.2.3

```
ggplot(data, aes(x=P, y=S, color=factor(statistika))) +
  geom_line() +
  scale_color_discrete(name="Použitá statistika") +
  ggtitle('Silofunkce')  +
  xlab('Porušení H0') +
  ylab('Pravděpodobnost zamítnutí H0') +
  geom_abline(intercept=0.05, slope=0, col='red')
  theme_bw()
```

H. R-kód pro výpočet hladiny významnosti post-hoc testů

```
r = 2 # počet skupin
p = 3 # počet proměnných
n = 100 # rozsah každé skupiny

rho = 0.3 # korelační koeficient každé dvojice závisle proměnných
pocet_cyklu = 10000 # počet opakování
alfa = 0.05

X = factor(rep(LETTERS[1:r], each=n))
N = r*n
dvojice_r = combn(1:r, 2)

pocty_zamitnuti_h0 = numeric(3)
names(pocty_zamitnuti_h0) = c("testy_z_kap_2.1","testy_z_kap_2.2"
    ,"testy_z_kap_2.4")

for(cyklus in 1:pocet_cyklu)
{

  z = rnorm(n*r,0,sqrt(rho))
  Y = matrix(rnorm(n*r*p,0,sqrt(1-rho)),ncol = p) + z

  S = apply(array(unlist(by(Y,X,var)),dim = c(p,p,r)),c(1,2),mean
      )
  M = matrix(unlist(lapply(1:p, function(i) by(Y[,i],X,mean))),
      ncol=p)

  res = matrix(unlist(lapply(1:p,function(i) Y[,i] - rep(M[,i],
      each=n))), ncol = p)

  T = cov(Y)*(N-1)
  E = cov(res)*(N-1)
  E_inv = solve(E)

  K21 = sapply(1:p, function(i) -(N-(p+r)/2-1)*log(E[i,i]/T[i,i])
       )
  p21 = 1-pchisq(K21,p*(r-1))

  K22 = sapply(1:(r*(r-1)/2), function(i) (N-r-p+1)/((r-1)*p) * n
      /2 * t(M[dvojice_r[1,i],]-M[dvojice_r[2,i],]) %*% E_inv %*%
      (M[dvojice_r[1,i],]-M[dvojice_r[2,i],]) )
  p22 = 1-pf(K22, (r-1)*p*(N-r-p)/(N-2-(r-1)*p) , N-r-p-1 )

  K24 = sapply(1:(r*(r-1)/2), function(i) (N-r-p+1)/((r-1)*p*(N-r
      )) * n/2 * (M[dvojice_r[1,i],]-M[dvojice_r[2,i],])^2/diag(S)
      )
  p24 = 1-pf(K24, (r-1)*p*(N-r-p)/(N-2-(r-1)*p) , N-r-p-1 )
```

```
39
40    pocty_zamitnuti_h0 = pocty_zamitnuti_h0 + c(min(p21)<alfa, min(
          p22)<alfa, min(p24)<alfa)
41    }
42
43  pocty_zamitnuti_h0/pocet_cyklu
```