UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

# DIPLOMOVÁ PRÁCE

## Testy parametrických hypotéz pro kompoziční data

Vedoucí diplomové práce:
**RNDr. Karel Hron, Ph.D.**
Rok odevzdání: 2010

Vypracovala:
**Sandra Donevska**
AME, II. ročník

**Prohlášení**

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně za vedení RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 7. dubna 2010

**Poděkování**

Děkuji svému vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za vstřícný přístup, za čas strávený konzultacemi, cenné rady, připomínky a celkovou spolupráci. Dále bych chtěla poděkovat své rodině a přátelům, kteří mě po celou dobu studia podporovali a také Bc. Lucii Syslové za pomoc při korektuře českého jazyka.

# Obsah

# Úvod

Statistická analýza kompozičních dat (kompozice, angl. compositions), pozorování nesoucích pouze relativní informaci a se simplexem jako výběrovým prostorem, je mladá rozvíjející se oblast statistiky. Počet statistických metod přímo aplikovatelných na kompozice proto není moc velký ve srovnání se standardními mnohorozměrnými daty, které nesou pouze absolutní informaci.

Na konci 80. let minulého století objevil John Aitchison způsob, jak transformovat kompozice na standardní data pomocí tzv. logratio transformací a následně tak umožnit použití známých statistických technik.

V této práci tohoto faktu využijeme a budeme se zabývat testováním hypotéz o „střední hodnotě“ náhodné kompozice prostřednictvím standardního přístupu, aplikovaného na její logratio transformaci. Přitom budeme předpokládat normalitu kompozice na simplexu.

Za účelem vše dobře čtenáři nastínit jsem tuto práci rozdělila do tří kapitol.

V úvodní kapitole se hlouběji seznámíme se standardní teorií jednovýběrových testů o střední hodnotě náhodného výběru z mnohorozměrného normálního rozdělení. Zvláštní pozornost přitom budeme věnovat vlastnostem Wishartova rozdělení. Přitom předpokládáme, že čtenář má základní znalosti z oblasti mnohorozměrné statistické analýzy.

Ve druhé kapitole se již budeme věnovat samotným kompozicím, jejich specifickým vlastnostem a také výše zmíněným logratio transformacím (o kterých budeme hovořit v souvislosti s vyjádřením dat ve zvolené bázi na simplexu), normalitě kompozic a jejich číselným charakteristikám.

Tyto teoretické poznatky, uvedené v obou kapitolách, budeme nakonec demonstrovat na praktických příkladech.

# 1 Testování hypotéz o střední hodnotě $\boldsymbol{\mu}$ normálně rozděleného náhodného vektoru

V této úvodní kapitole budeme uvažovat především výběr, který pochází z vícerozměrného normálního rozdělení. Výběrový průměr je nejlepší lineární nestranný odhad, angl. best linear unbiased estimator (BLUE), střední hodnoty $\boldsymbol{\mu}$, o níž následně vyslovíme úsudek.

Při testovaní hypotéz o střední hodnotě uvažujeme dvě situace, buď je varianční matice známá nebo neznámá.

Nejprve popíšeme případ, kdy varianční matici známe, což je ovšem v praxi situace spíše méně častá. Tato teorie nám však poslouží jako základ, který postupně nadstavíme.

Testujeme-li hypotézu o střední hodnotě při neznámé varianční matici, tak ji musíme nejprve odhadhout. Pro odhad neznámé varianční matice použijeme výběrovou varianční matici, vyjádřenou jako násobek Wishartovy matice. Tato matice má Wishartovo rozdělení, kterému budeme také v této práci věnovat pozornost. Samotnou hypotézu o střední hodnotě při neznámé varianční matici pak budeme testovat pomocí Hotellingovy statistiky, kterou si též odvodíme.

## 1.1 Vektor výběrových průměrů jako BLUE $\boldsymbol{\mu}$

V této kapitole jsem převážně čerpala z literatury [15].

**Definice 1.1.** *Nechť $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ je náhodný výběr z nějakého p-rozměrného rozdělení. Výběrovou funkci*

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

*nazveme výběrovým průměrem.*

**Věta 1.1.** *(Reprodukční vlastnost mnohorozměrného normálního rozdělení) Nechť jsou dány náhodné vektory* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$*, které mají p-rozměrné normální rozdělení s libovolnými hodnotami parametrů, tedy* $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$*, kde* $\mathsf{E}(\mathbf{x}_i) = \boldsymbol{\mu}_i$*,* $\mathrm{var}(\mathbf{x}_i) = \boldsymbol{\Sigma}_i$*,* $i = 1, \ldots, n$*. Dále nechť jsou dány konstanty* $a_1, a_2, ..., a_n \in \mathbb{R}$*. Potom*

$$\sum_{i=1}^{n} a_i \mathbf{x}_i \sim N_p \left( \sum_{i=1}^{n} a_i \mathsf{E}(\mathbf{x}_i), \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j) \right),$$

*kde* $\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j)$ *označuje kovarianční matici vektorů* $\mathbf{x}_i$ *a* $\mathbf{x}_j$*.*

**Důsledek 1.1.** *Pokud* $\mathbf{x}_i$*,* $i = 1, ..., n$*, jsou navíc nezávislé, pak*

$$\sum_{i=1}^{n} a_i \mathbf{x}_i \sim N_p \left( \sum_{i=1}^{n} a_i \mathsf{E}(\mathbf{x}_i), \sum_{i=1}^{n} a_i^2 \mathrm{var}(\mathbf{x}_i) \right).$$

**Důkaz:** viz [15], str. 568, věta (VI). □

**Věta 1.2.** *(Rozdělení výběrového průměru)*
*Nechť jsou dány nezávislé náhodné vektory* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ *takové, že* $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*,* $i = 1, ..., n$*. Pak výběrový průměr* $\overline{\mathbf{x}}$ *má p-rozměrné normální rozdělení s parametry* $\boldsymbol{\mu}$ *a* $\frac{1}{n}\boldsymbol{\Sigma}$*, tj.*

$$\overline{\mathbf{x}} \sim N_p \left( \boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma} \right).$$

**Důkaz:** Nejprve dokážeme, že $\overline{\mathbf{x}}$ má $p$-rozměrné normální rozdělení.
Využijeme větu 1.1 a zvolíme za $a_i = \frac{1}{n}$, $i = 1, ..., n$, tj.

$$\sum_{i=1}^{n} a_i \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \overline{\mathbf{x}},$$

pak $\overline{\mathbf{x}}$ je normálně rozdělený a jeho parametry dle důsledku 1.1 jsou

$$\mathsf{E}(\overline{\mathbf{x}}) = \sum_{i=1}^{n} a_i \mathsf{E}(\mathbf{x}_i) = \frac{1}{n} n \boldsymbol{\mu} = \boldsymbol{\mu},$$

$$\mathrm{var}(\overline{\mathbf{x}}) = \sum_{i=1}^{n} a_i \mathrm{var}(\mathbf{x}_i) = \frac{1}{n^2} n\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{\Sigma}.$$

□

Ve zbytku této podkapitoly budeme předpokládat, že náhodný výběr $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ je z $p$-rozměrného normálního rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$. Poznamenejme ovšem, že uvedené vlastnosti by bylo možné dokázat i bez tohoto předpokladu. To platí zejména pro následující větu.

**Věta 1.3.** *Výběrový průměr $\overline{\mathbf{x}}$ je nestranným odhadem střední hodnoty $\boldsymbol{\mu}$.*

**Důkaz:** Nestrannost výběrové funkce $\overline{\mathbf{x}}$ ověříme následovně

$$\mathsf{E}(\overline{\mathbf{x}}) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(\mathbf{x}_i) = \frac{1}{n}n\boldsymbol{\mu} = \boldsymbol{\mu}.$$

□

**Definice 1.2.** *Fisherova informační matice pro $q$-rozměrný parametr $\boldsymbol{\theta}$ spojitého rozdělení s hustotou $f(\mathbf{x}, \boldsymbol{\theta})$ je definována vztahem*

$$\mathbf{J}(\boldsymbol{\theta})^{(q\times q)} = \mathsf{E}\left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{T}.$$

**Definice 1.3.** *Nestranný odhad $\mathbf{t}$ se nazývá eficientní, jestliže*

$$\mathrm{var}(\mathbf{t}) = \frac{1}{n}\mathbf{J}^{-1}(\boldsymbol{\theta}).$$

**Věta 1.4.** *Výběrový průměr $\overline{\mathbf{x}}$ je nejlepším lineárním nestranným odhadem (BLUE) parametru $\boldsymbol{\mu}$.*

**Důkaz:** Ve větě 1.3 jsme dokázali nestrannost odhadu $\boldsymbol{\mu}$. Přitom linearita $\overline{\mathbf{x}}$ je zřejmá, zbývá tedy ještě dokázat, že $\overline{\mathbf{x}}$ je nejlepším (eficientním) odhadem.

Tedy nejprve vypočítáme Fisherovu informační matici pro parametr $\boldsymbol{\mu}$ s hustotou

$$f(\mathbf{x}, \boldsymbol{\mu}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Za tímto účelem nejprve hustotu $f(\mathbf{x}, \boldsymbol{\mu})$ zlogaritmujeme, tj.

$$\ln f(\mathbf{x}, \boldsymbol{\mu}) = -\frac{p}{2} \ln (2\pi) + \frac{1}{2} \ln \left|\boldsymbol{\Sigma}^{-1}\right| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

V dalším kroku vypočítáme parciální derivaci hustoty $f(\mathbf{x}, \boldsymbol{\mu})$ dle vektoru $\boldsymbol{\mu}$ takto

$$\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -\frac{1}{2} 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Fisherova informační matice pro $\boldsymbol{\mu}$ s hustotou $f(\mathbf{x}, \boldsymbol{\mu})$ má tvar

$$\begin{aligned} \mathbf{J}(\boldsymbol{\mu}) &= \mathsf{E}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\right] = \boldsymbol{\Sigma}^{-1} \mathsf{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right] \boldsymbol{\Sigma}^{-1} = \\ &= \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Nyní ověříme, že $\overline{\mathbf{x}}$ je eficientním odhadem $\boldsymbol{\mu}$, tj.

$$\frac{1}{n} \mathbf{J}^{-1}(\boldsymbol{\mu}) = \frac{1}{n} \boldsymbol{\Sigma} = \mathsf{var}(\overline{\mathbf{x}}).$$

□

## 1.2 Test hypotézy o vektoru $\boldsymbol{\mu}$ při známé matici $\boldsymbol{\Sigma}$

V praxi se nejčastěji setkáváme s případem, že matice $\boldsymbol{\Sigma}$ je neznámá. Navzdory tomu budeme nejprve uvažovat situaci, kdy matici $\boldsymbol{\Sigma}$ známe, např. z nějakých hlubších úvah o podstatě experimentu. To nám následně poslouží jako základ pro určení vlastností testu hypotézy o střední hodnotě při neznámé varianční matici $\boldsymbol{\Sigma}$.

V následujícím jsem čerpala převážně z literatury [3], [9] a [16].

**Věta 1.5.** *Nechť $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je regulární (pozitivně definitní) matice. Pak $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \sim \chi^2(\delta)$, kde $\delta$ je parametr necentrality a $\delta = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$. Pro $\boldsymbol{\mu} = \mathbf{0}$ je rozdělení $\chi^2$ centrální.*

**Důkaz:** Budeme uvažovat skeletní rozklad matice $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$, kde matice $\mathbf{A}$ je čtvercová řádu $p$. Z regularity $\boldsymbol{\Sigma}$ vyplývá, že matice $\mathbf{A}$ je také regulární. Můžeme psát, že

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} = \mathbf{x}^T (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{x} = (\mathbf{A}^{-1}\mathbf{x})^T \mathbf{A}^{-1} \mathbf{x} = \mathbf{y}^T \mathbf{y},$$

kde $\mathbf{y} = \mathbf{A}^{-1}\mathbf{x}$. Vektor $\mathbf{y}$ má taktéž $p$-rozměrné normální rozdělení. Neznáme však jeho parametry, proto je vypočítáme,

$$\mathsf{E}(\mathbf{y}) = \mathsf{E}(\mathbf{A}^{-1}\mathbf{x}) = \mathbf{A}^{-1}\mathsf{E}(\mathbf{x}) = \mathbf{A}^{-1}\boldsymbol{\mu},$$

$$\begin{aligned}\mathrm{var}(\mathbf{y}) &= \mathrm{var}(\mathbf{A}^{-1}\mathbf{x}) = \mathbf{A}^{-1}\mathrm{var}(\mathbf{x})(\mathbf{A}^{-1})^T = (\mathbf{A}^{-1})(\mathbf{A}\mathbf{A}^T)(\mathbf{A}^{-1})^T = \\ &= (\mathbf{A}^{-1}\mathbf{A})(\mathbf{A}^{-1}\mathbf{A})^T = \mathbf{I}_p.\end{aligned}$$

Pokud je varianční matice náhodného vektoru $\mathbf{y} = (Y_1, Y_2, ..., Y_p)^T$ jednotková, pak náhodné veličiny $Y_j,\ j = 1, ..., p$, ze kterých se skládá náhodný vektor $\mathbf{y}$, mají jednorozměrné normální rozdělení s jednotkovým rozptylem a jsou navíc nezávislé. Podle definice $\chi^2$ rozdělení má $\mathbf{y}^T\mathbf{y} = \sum_{j=1}^{p} Y_j^2 \sim \chi^2(\delta)$, kde parametr necentrality $\delta$ vypadá následovně,

$$\delta = \sum_{j=1}^{n} [\mathsf{E}(Y_j)]^2 = [\mathsf{E}(\mathbf{y})]^T [\mathsf{E}(\mathbf{y})] = \boldsymbol{\mu}^T(\mathbf{A}^{-1})^T\mathbf{A}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

□

Nechť je dán náhodný výběr $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ z $p$-rozměrného normálního rozdělení se střední hodnotu $\boldsymbol{\mu}$ a pozitivně definitní maticí $\boldsymbol{\Sigma}$.

Dále, nechť je dán výběrový průměr $\overline{\mathbf{x}} = \dfrac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \sim N_p\left(\boldsymbol{\mu}, \dfrac{1}{n}\boldsymbol{\Sigma}\right)$.

Odtud $\sqrt{n}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Pokud ve větě 1.5 budeme předpokládat, že $\sqrt{n}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ namísto $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, tak dospějeme k tvrzení

$$n(\overline{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim \chi_p^2.$$

Testujeme nulovou hypotézu $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ naproti alternativě $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Složky vektoru $\boldsymbol{\mu}_0$ stanovíme na základě minulých zkušeností nebo představují naše cílové hodnoty. Pokud oba vektory rozepíšeme do složek dostaneme následující výraz,

$$H_0 : \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}, \qquad H_A : \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}.$$

Za předpokladu platnosti nulové hypotézy $H_0$ má tedy výběrová funkce

$$n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \chi^2_p.$$

Nulovou hypotézu $H_0$ zamítáme ve prospěch alternativy $H_A$, pokud hodnota testového kriteria $n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)$ je větší nebo rovna $(1-\alpha)$-kvantilu centrálního $\chi^2$ rozdělení o $p$ stupních volnosti, tj. zamítneme $H_0$, jestliže alespoň pro jedno $j$, $j = 1, ..., p$, neplatí rovnost $\mu_j = \mu_{0j}$.

Došlo-li k zamítnutí $H_0$, lze zjistit, které složky vektoru $\boldsymbol{\mu}$ toto zamítnutí způsobily. Provádí se to pomocí simultánních testů, kde testujeme $H_0^j : \mu_j = \mu_{0j}$, $j = 1, ..., p$. Protože ovšem toto testování nemá v případě kompozičních dat smysl, nebudeme se jím nyní dále zabývat.

## 1.3 Vznik a definice Wishartova rozdělení

Při tvorbě této kapitoly byly použity především zdroje [9], [10] a [15].

Vznik Wishartova rozdělení je založený na myšlence maticového zobecnění $\chi^2$ rozdělení. Vyjdeme z toho, co již známe. Máme-li dáno $n$ nezávislých náhodných veličin $X_1, .., X_n$, které mají normované normální rozdělení $X_i \sim N(0, 1)$, $i = 1, 2, ..., n$, pak součet jejich druhých mocnin má (centrální) $\chi^2$ rozdělení o $n$ stupních volnosti,

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2_n.$$

Budeme-li uvažovat místo náhodných veličin náhodné vektory $\mathbf{x}_i,\ i = 1, \ldots, n$, které mají $p$-rozměrné normované normální rozdělení, tedy

$$\mathbf{x}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p), \mathbf{x}_2 \sim N_p(\mathbf{0}, \mathbf{I}_p), \ldots, \mathbf{x}_n \sim N_p(\mathbf{0}, \mathbf{I}_p),$$

můžeme výše uvedený součet čtverců zobecnit vytvořením pozitivně semidefinitní matice $\mathbf{S}^{(p\times p)} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Jak si dále uvedeme, takto vytvořená matice $\mathbf{S}$ má Wishartovo rozdělení a toto rozdělení je právě oním mnohorozměrným zobecněním $\chi^2$ rozdělení.

Ukážeme si také, že definice Wishartova rozdělení a jeho vlastnosti nejsou odvozeny z hustoty tohoto rozdělení.

Předpokládejme nyní, že je dána datová matice $\mathbf{X}^{(n\times p)} = (X_{ij})_{i=1}^{n}{}_{j=1}^{p}$, kde prvek $X_{ij}$ přestavuje hodnotu $j$-té proměnné pozorované u $i$-tého objektu,

$$\mathbf{X}^{(n\times p)} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}.$$

Nechť náhodné vektory $\mathbf{x}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), ..., \mathbf{x}_n \sim N_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma})$ a jsou nezávislé. Dále, nechť je dána matice $\boldsymbol{M}^{(n\times p)}$ tvořená středními hodnotami náhodných vektorů $\mathbf{x}_i,\ i = 1, \dots, n$,

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \boldsymbol{\mu}_n^T \end{pmatrix},$$

a $\boldsymbol{\Sigma}^{(p\times p)}$ je varianční matice.

**Definice 1.4.** *Sdružené rozdělení prvků matice* $\mathbf{S}^{(p\times p)} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T$ *nazveme p-rozměrným Wishartovým rozdělením o n stupních volnosti a s parametry* $\boldsymbol{\Sigma}, \boldsymbol{M}$. *Značíme* $\mathbf{S} \sim W_p(n, \boldsymbol{\Sigma}, \boldsymbol{M})$.

Výše uvedená matice $\mathbf{S}$ má následující tvar,

$$\mathbf{S} = (s_{ij}) = \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \dots & \mathbf{x}_1^T\mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{x}_n^T\mathbf{x}_1 & \dots & \mathbf{x}_n^T\mathbf{x}_n \end{pmatrix}.$$

**Poznámka 1.1.** *Wishartovo rozdělení je necentrální rozdělení pro* $\boldsymbol{M} \neq \mathbf{0}$ (*značíme* $W_p(n, \boldsymbol{\Sigma}, \boldsymbol{M})$) *a centrální pro* $\boldsymbol{M} = \mathbf{0}$ (*značíme* $W_p(n, \boldsymbol{\Sigma})$).

## 1.4 Vlastnosti Wishartova rozdělení

V následujícím se budeme podrobně věnovat vlastnostem Wishartova rozdělení, které následně využijeme při odvození Hotellingovy testovací statistiky. Při tvorbě této kapitoly jsem využila zejména [4], [9], [10] a [15].

**Věta 1.6.** *(O souvislosti $\chi^2$ a Wishartova rozdělení)*
*Nechť $\mathbf{S} \sim W_p(n, \mathbf{\Sigma}, \boldsymbol{M})$ a dále nechť je dán vektor konstant $\mathbf{k}^{(p\times 1)} \neq \mathbf{0}$. Pak*

$$\frac{\mathbf{k}^T\mathbf{S}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}} \sim \chi^2_n(\delta), \tag{1}$$

*kde parametr necentrality* $\delta = \dfrac{\mathbf{k}^T\boldsymbol{M}^T\boldsymbol{M}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}$.

**Důkaz:** Uvažujeme matici $\mathbf{S} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, kde vektory $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, $i = 1, \ldots, n$, jsou nezávislé. Tuto matici $\mathbf{S}$ dosadíme do (1), tj.

$$\frac{\mathbf{k}^T\mathbf{S}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}} = \frac{\mathbf{k}^T\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}} = \sum_{i=1}^n \left(\frac{\mathbf{k}^T\mathbf{x}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}\right)^2 = \sum_{i=1}^n U_i^2,$$

kde $U_i = \dfrac{\mathbf{k}^T\mathbf{x}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}$, $i = 1, \ldots, n$.
Víme, že náhodné veličiny $U_i$ jsou nezávislé a normálně rozdělené (vyplývá to z nezávislosti a rozdělení náhodných vektorů $\mathbf{x}_i,\ i = 1, ..., n$).

Odvodíme jejich střední hodnotu a rozptyl. Pro střední hodnotu náhodné veličiny $U_i$ dostaneme

$$\mathsf{E}(U_i) = \mathsf{E}\left(\frac{\mathbf{k}^T\mathbf{x}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}\right) = \frac{\mathbf{k}^T\mathsf{E}(\mathbf{x}_i)}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}} = \frac{\mathbf{k}^T\boldsymbol{\mu}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}$$

a pro rozptyl

$$\mathrm{var}(U_i) = \mathrm{var}\left(\frac{\mathbf{k}^T\mathbf{x}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}\right) = \frac{\mathbf{k}^T\mathrm{var}(\mathbf{x}_i)\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}} = \frac{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}} = 1.$$

Můžeme tedy psát $U_i \sim N_1\left(\dfrac{\mathbf{k}^T\boldsymbol{\mu}_i}{\sqrt{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}}, 1\right)$, $i = 1, \ldots, n$.

Jelikož náhodné veličiny $U_i$ jsou nezávislé a normálně rozdělené, můžeme využít znalostí o $\chi^2$ rozdělení. Pak totiž součet druhých mocnin náhodných veličin $U_i$ má právě $\chi^2$ rozdělení, tj.

$$\sum_{i=1}^n U_i^2 \sim \chi^2(\delta),$$

kde parametr necentrality $\delta$ se spočítá následovně,

$$\delta = \sum_{i=1}^{n} \left[\mathsf{E}(U_i)\right]^2 = \sum_{i=1}^{n} \frac{\left[\mathbf{k}^T \mathsf{E}(\mathbf{x}_i)\right]^2}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}} =$$

$$= \sum_{i=1}^{n} \frac{(\mathbf{k}^T \boldsymbol{\mu}_i)^2}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}} = \frac{\mathbf{k}^T \sum_{i=1}^{n} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i \mathbf{k}}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}} = \frac{\mathbf{k}^T \boldsymbol{M}^T \boldsymbol{M} \mathbf{k}}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}}.$$

□

**Lemma 1.1.** *Nechť je dán náhodný vektor* $\mathbf{x} = (X_1, ..., X_n)^T$, *kde* $X_i \sim N_1(\mu_i, 1)$, $i = 1, ..., n$, *jsou nezávislé náhodné veličiny. Pak kvadratická forma* $\mathbf{x}^T \mathbf{A} \mathbf{x}$ *má* $\chi^2$ *rozdělení právě tehdy, když matice* $\mathbf{A}$ *je symetrická a idempotentní* ($tj.$ $\mathbf{A} = \mathbf{A}^2$). *Počet stupňů volnosti pro rozdělení formy* $\mathbf{x}^T \mathbf{A} \mathbf{x}$ *je roven hodnosti matice* $\mathbf{A}$, *a tudíž stopě matice* $\mathbf{A}$.

**Důkaz:** viz [15], str. 219, věta (II). □

**Věta 1.7.** *Nechť jsou dány nezávislé náhodné vektory* $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$. *Dále nechť matice* $\mathbf{A}$ *je čtvercová řádu* $p$, *symetrická a idempotentní. Pak náhodná matice* $\mathbf{X}^T \mathbf{A} \mathbf{X} \sim W_p(r, \boldsymbol{\Sigma}, \boldsymbol{M})$ *právě tehdy, když pro* $p$*-rozměrný číselný vektor* $\mathbf{k}^{(p \times 1)} \neq \mathbf{0}$ *má náhodná matice*

$$\frac{\mathbf{k}^T \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{k}}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}} \sim \chi_r^2(\delta).$$

*Symbol* $r$ *označuje hodnost matice* $\mathbf{A}$ *a rozdělení* $\mathbf{X}^T \mathbf{A} \mathbf{X}$ *a* $\dfrac{\mathbf{k}^T \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{k}}{\mathbf{k}^T \boldsymbol{\Sigma} \mathbf{k}}$ *jsou buď centrální nebo necentrální.*

**Důkaz:** Případ, kdy z předpokladu věty vyplývá tvrzení, jsme dokázali již ve větě 1.6, volíme-li $\mathbf{S} = \mathbf{X}^T \mathbf{A} \mathbf{X}$. Nyní dokážeme obrácenou implikaci, přičemž využijeme výše uvedeného lemmatu 1.1.

Provedeme spektrální rozklad matice $\mathbf{A}^{(n \times n)} = \mathbf{V}^T \boldsymbol{\Lambda} \mathbf{V}$, kde $\mathbf{V}$ je ortogonální matice, jejíž sloupce jsou vlastní vektory matice $\mathbf{A}$, a $\boldsymbol{\Lambda}$ je diagonální matice, přičemž diagonální prvky $\lambda_j$, $j = 1, ..., n$, této matice jsou vlastní čísla matice $\mathbf{A}$.

Matice $\mathbf{A}$ je idempotentní, proto můžeme psát

$$\mathbf{A} = \mathbf{V}^T\mathbf{\Lambda}\mathbf{V} = \mathbf{A}^2 = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{2}$$

Matice $\mathbf{V}$ je ortogonální, platí tedy, že $\mathbf{V}\mathbf{V}^T = \mathbf{I}$ a vztah (2) vypadá následovně,

$$\mathbf{A} = \mathbf{V}^T\mathbf{\Lambda}\mathbf{V} = \mathbf{A}^2 = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T.$$

Z idempotence matice $\mathbf{A}$ dále vyplývá, že diagonální matice $\mathbf{\Lambda}$ je také idempotentní. Proto pro diagonální prvky $\mathbf{\Lambda}$ musí platit $\lambda_j = \lambda_j^2$ pro všechna $j = 1, ..., n$. Pro splnění této rovnosti je nutno volit $\lambda_j$ buď 0 nebo 1. Potom můžeme matici $\mathbf{A}$ přepsat do tvaru

$$\mathbf{A} = \sum_{j=1}^{r} \mathbf{v}_j\mathbf{v}_j^T\lambda_j = \sum_{j=1}^{r} \mathbf{v}_j\mathbf{v}_j^T = \mathbf{B}\mathbf{B}^T, \tag{3}$$

kde $\mathbf{B}^{(n\times r)} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ a $\mathbf{v}_j$ jsou ortonormální vlastní vektory. Hodnost matice $\mathbf{A}$ je $r$, což značí počet jedniček na diagonále matice $\mathbf{\Lambda}$.

Využitím vztahu (3) přepíšeme kvadratickou formu $\mathbf{X}^T\mathbf{A}\mathbf{X}$ do tvaru

$$\mathbf{X}^T\mathbf{A}\mathbf{X} = \sum_{j=1}^{r} \mathbf{X}^T\mathbf{v}_j\mathbf{v}_j^T\mathbf{X} = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{u}_j^T = \mathbf{U}\mathbf{U}^T,$$

kde $\mathbf{U}^{(p\times r)} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ a $\mathbf{u}_j = \mathbf{X}^T\mathbf{v}_j$, $j = 1, \ldots, r$.

Dále ukážeme, že když vektory $\mathbf{v}_j$ jsou ortonormální, pak vektory $\mathbf{u}_j$, $j = 1, \ldots, r$, mají $p$-rozměrné normální rozdělení a jsou nezávislé. Požadavek $p$-rozměrného normálního rozdělení vektorů $\mathbf{u}_j$ je splněn, jelikož při lineární kombinaci vektorů $\mathbf{x}_i,\ i = 1, \ldots, n$, se zachovává normalita (věta 1.1). Abychom dokázali, že vektory $\mathbf{u}_j,\ j = 1, \ldots, r$, jsou nezávislé, stačí ukázat, že $\mathrm{cov}(\mathbf{u}_j, \mathbf{u}_k) = \mathbf{0}$. Pro tento účel označíme

$$\mathbf{u}_j = \mathbf{X}^T\mathbf{v}_j = \sum_{i=1}^{n} \mathbf{x}_i v_{ji},$$

kde $v_{ji}$ je $i$-tý prvek vektoru $\mathbf{v}_j$. Kovarianční matice $\mathrm{cov}(\mathbf{u}_j, \mathbf{u}_k)$ tedy vypadá následovně,

$$\mathrm{cov}(\mathbf{u}_j, \mathbf{u}_k) = \mathrm{cov}\left(\sum_{i=1}^{n} \mathbf{x}_i v_{ji}, \sum_{i=1}^{n} \mathbf{x}_i v_{ki}\right) = \sum_{i=1}^{n} \mathrm{var}(\mathbf{x}_i) v_{ji} v_{ki} = \mathbf{\Sigma}\mathbf{v}_j\mathbf{v}_k^T = \mathbf{0}.$$

Dokázali jsme, že $\mathbf{u}_j$ jsou nezávislé náhodné vektory, které mají $p$-rozměrné normální rozdělení. Potom z definice 1.4 plyne, že kvadratická forma $\mathbf{X}^T\mathbf{A}\mathbf{X} = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{u}_j^T$ má Wishartovo rozdělení. Můžeme to psát ve tvaru $\mathbf{X}^T\mathbf{A}\mathbf{X} \sim W_p(n, \mathbf{\Sigma}, \boldsymbol{N})$, kde $\boldsymbol{N}$ je parametr necentrality vyjádřený vztahem

$$\boldsymbol{N} = \mathsf{E}(\mathbf{U}) = \mathsf{E}(\mathbf{X}^T\mathbf{B}) = \mathsf{E}(\mathbf{X}^T)\mathbf{B} = \boldsymbol{M}^T\mathbf{B}.$$

Aby bylo Wishartovo rozdělení centrální, tedy $\boldsymbol{N} = \mathbf{0}$, museli bychom ukázat, že matice $\dfrac{\mathbf{k}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}$ má centrální $\chi^2$ rozdělení s parametrem necentrality $\delta = 0$. Požadujeme, aby

$$\mathsf{E}\left(\frac{\mathbf{k}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{k}}{\mathbf{k}^T\mathbf{\Sigma}\mathbf{k}}\right) = \mathbf{k}^T\boldsymbol{M}^T\mathbf{A}\boldsymbol{M}\mathbf{k} = 0, \ \forall \mathbf{k} \neq \mathbf{0}.$$

Rovnost nastane, jestliže položíme $\boldsymbol{M}^T\mathbf{A}\boldsymbol{M} = \mathbf{0}$. Využitím vztahu (3) můžeme psát

$$\mathbf{A}\boldsymbol{M} = \mathbf{0} \Leftrightarrow \mathbf{B}\mathbf{B}^T\boldsymbol{M} = \mathbf{0} \Leftrightarrow \mathbf{B}^T\boldsymbol{M} = \boldsymbol{N}^T = \mathbf{0}.$$

□

**Důsledek 1.2.** *Wishartovo rozdělení je centrální právě tehdy, když* $\boldsymbol{N} = \mathbf{0}$*, tj.* $\boldsymbol{M}^T\mathbf{B} = \mathbf{0}$.

**Věta 1.8.** *Nechť je dáno* $n$ *nezávislých náhodných vektorů* $\mathbf{x}_i$, $i = 1, \ldots, n$, *takových, že* $\mathbf{x}_1 \sim N_p(\boldsymbol{\mu}_1, \mathbf{\Sigma}), \mathbf{x}_2 \sim N_p(\boldsymbol{\mu}_2, \mathbf{\Sigma}), \ldots, \mathbf{x}_n \sim N_p(\boldsymbol{\mu}_n, \mathbf{\Sigma})$.

*(i) Pak kvadratické formy* $\mathbf{X}^T\mathbf{A}_1\mathbf{X}$ *a* $\mathbf{X}^T\mathbf{A}_2\mathbf{X}$ *jsou nezávislé a mají Wishartovo rozdělení tehdy a jen tehdy, když pro libovolný číselný vektor* $\mathbf{k}^{(p\times 1)}$ *jsou náhodné veličiny* $\mathbf{k}^T\mathbf{X}^T\mathbf{A}_1\mathbf{X}\mathbf{k}$ *a* $\mathbf{k}^T\mathbf{X}^T\mathbf{A}_2\mathbf{X}\mathbf{k}$ *nezávislé a mají* $\chi^2$ *rozdělení.*

*(ii) Náhodná matice* $\mathbf{X}^T\mathbf{A}\mathbf{X}$ *s Wishartovým rozdělením a náhodný vektor* $\mathbf{X}^T\mathbf{b}$ *s p-rozměrným normálním rozdělením jsou nezávislé tehdy a jen tehdy, když pro každý číselny vektor* $\mathbf{k}^{(p\times 1)}$ *má náhodná veličina* $\mathbf{k}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{k}$ $\chi^2$ *rozdělení a náhodná veličina* $\mathbf{k}^T\mathbf{X}^T\mathbf{b}$ *jednorozměrné normální rozdělení a jsou navíc nezávislé.*

**Důkaz:**

(i) Důkaz první implikace se provádí obdobně jako ve větě 1.7. Při transformaci kvadratických forem $\mathbf{X}^T\mathbf{A}_1\mathbf{X}$ a $\mathbf{X}^T\mathbf{A}_2\mathbf{X}$ přenásobením vektorem $\mathbf{k}$ se neporuší nezávislost náhodných veličin $\mathbf{k}^T\mathbf{X}^T\mathbf{A}_1\mathbf{X}\mathbf{k}$ a $\mathbf{k}^T\mathbf{X}^T\mathbf{A}_2\mathbf{X}\mathbf{k}$.

Nyní dokážeme obrácenou implikaci.

Podobně jako ve větě 1.7 vyjdeme z lemmatu 1.1. Nyní ovšem budeme předpokládat, že postačující podmínkou pro to, aby kvadratické formy $\mathbf{x}^T\mathbf{A}_1\mathbf{x}$ a $\mathbf{x}^T\mathbf{A}_2\mathbf{x}$, kde $\mathbf{x} = \mathbf{X}\mathbf{k}$, měly $\chi^2$ rozdělení, je idempotentnost matic $\mathbf{A}_1$ a $\mathbf{A}_2$.

Provedeme skeletní rozklad matic $\mathbf{A}_1$ a $\mathbf{A}_2$ následovně,

$$\mathbf{A}_1 = \mathbf{P}\mathbf{P}^T = \sum_{j=1}^{r} \mathbf{p}_j\mathbf{p}_j^T,$$

$$\mathbf{A}_2 = \mathbf{Q}\mathbf{Q}^T = \sum_{j=1}^{s} \mathbf{q}_j\mathbf{q}_j^T,$$

kde $\mathbf{p}_j,\ j = 1, \ldots, r$, a $\mathbf{q}_j,\ j = 1, \ldots, s$, jsou ortonormální vlastní vektory.

Matice $\mathbf{A}_1$ a $\mathbf{A}_2$ mají takto zřejmě hodnosti $h(\mathbf{A}_1) = r$ a $h(\mathbf{A}_2) = s$.

Dále využijeme tvrzení, podle kterého jsou kvadratické formy $\mathbf{x}^T\mathbf{A}_1\mathbf{x}$ a $\mathbf{x}^T\mathbf{A}_2\mathbf{x}$ nezávislé právě tehdy, když $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$ ([15], str. 221, věta (V)).

V našem případě $\mathbf{A}_1\mathbf{A}_2 = \mathbf{P}\mathbf{P}^T\mathbf{Q}\mathbf{Q}^T = \mathbf{0}$ právě tehdy, když $\mathbf{P}^T\mathbf{Q} = \mathbf{0}$ a znamená to, že vektory $\mathbf{p}_j$ a $\mathbf{q}_k$ jsou ortonormální pro $\forall j, k$. Pak pro $\mathbf{X}^T\mathbf{A}_1\mathbf{X}$ a $\mathbf{X}^T\mathbf{A}_2\mathbf{X}$ dostaneme

$$\mathbf{X}^T\mathbf{A}_1\mathbf{X} = \sum_{j=1}^{r} \mathbf{X}^T\mathbf{p}_j\mathbf{p}_j^T\mathbf{X} = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{u}_j^T = \mathbf{U}^T\mathbf{U},$$

$$\mathbf{X}^T\mathbf{A}_2\mathbf{X} = \sum_{j=1}^{s} \mathbf{X}^T\mathbf{q}_j\mathbf{q}_j^T\mathbf{X} = \sum_{j=1}^{s} \mathbf{v}_j\mathbf{v}_j^T = \mathbf{V}^T\mathbf{V},$$

kde $\mathbf{U}^{(r\times p)} = \mathbf{P}^T\mathbf{X} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)^T$ a $\mathbf{V}^{(s\times p)} = \mathbf{Q}^T\mathbf{X} = (\mathbf{v}_1, \ldots, \mathbf{v}_s)^T$.

Víme, že při lineární kombinaci se zachovává normalita a ve větě 1.7 jsme dokázali, že to platí i pro zachování nezávislosti. Vektory $\mathbf{u}_j$ a $\mathbf{v}_j$ pak mají $p$-rozměrné normální rozdělení a $\mathbf{u}_1, \ldots, \mathbf{u}_r$, resp. $\mathbf{v}_1, \ldots, \mathbf{v}_s$ a jsou nezávislé. Z definice 1.4 je zřejmé, že kvadratické formy $\mathbf{X}^T\mathbf{A}_1\mathbf{X} = \mathbf{U}^T\mathbf{U}$ a $\mathbf{X}^T\mathbf{A}_2\mathbf{X} = \mathbf{V}^T\mathbf{V}$ mají Wishartovo rozdělení, a to jsme chtěli dokázat.

Dále dokážeme, že kvadratické formy $\mathbf{X}^T\mathbf{A}_1\mathbf{X} = \mathbf{U}^T\mathbf{U}$ a $\mathbf{X}^T\mathbf{A}_2\mathbf{X} = \mathbf{V}^T\mathbf{V}$ jsou nezávislé. Proto vypočítáme

$$\operatorname{cov}(\mathbf{u}_j, \mathbf{v}_k) = \operatorname{cov}\left(\sum_{i=1}^{n} \mathbf{x}_i p_{ji}, \sum_{i=1}^{n} \mathbf{x}_i q_{ki}\right) = \sum_{i=1}^{n} \operatorname{var}(\mathbf{x}_i) p_{ji} q_{ki} = \boldsymbol{\Sigma}\mathbf{p}_j^T\mathbf{q}_k = \mathbf{0}.$$

Dokázali jsme nekorelovanost a normalitu vektorů $\mathbf{u}_j$ a $\mathbf{v}_k$, tj. nezávislost $\mathbf{u}_j$ a $\mathbf{v}_k$, a tím i nezávislost kvadratických forem $\mathbf{X}^T\mathbf{A}_1\mathbf{X}$ a $\mathbf{X}^T\mathbf{A}_2\mathbf{X}$.

(ii) Důkaz se provádí obdobně jako u (i).

□

**Poznámka 1.2.** *Větu 1.8 můžeme zformulovat i následovně.*
*Bez ohledu na to, zda kvadratické formy* $\mathbf{X}^T\mathbf{A}_1\mathbf{X}$ *a* $\mathbf{X}^T\mathbf{A}_2\mathbf{X}$ *mají nebo nemají Wishartovo rozdělení, jsou nezávislé právě tehdy, když* $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$. *Také funkce* $\mathbf{X}^T\mathbf{A}\mathbf{X}$ *a* $\mathbf{X}^T\mathbf{b}$ *bez ohledu na jejich rozdělení jsou nezávislé právě tehdy, když* $\mathbf{b}^T\mathbf{A} = \mathbf{0}$.

**Důsledek 1.3.** *Nechť je dán náhodný výběr* $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, ..., n$. *Pak výběrový průměr* $\overline{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$, *Wishartova matice* $\mathbf{W} \sim W_p(n-1, \boldsymbol{\Sigma})$ *a navíc* $\overline{\mathbf{x}}$ *a* $\mathbf{W}$ *jsou nezávislé.*

**Důkaz:** Ve větě 1.2 o rozdělení výběrového průměru jsme dokázali, že

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n}\mathbf{X}^T\mathbf{1}_n \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right).$$

Wishartova matice $\mathbf{W}$ je definovaná následovně,

$$\begin{aligned}\mathbf{W} &= \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \\ &= \mathbf{X}^T\mathbf{X} - n\overline{\mathbf{x}}\overline{\mathbf{x}}^T = n\mathbf{X}^T\mathbf{X} - n\frac{1}{n}\mathbf{X}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}\frac{1}{n} = \\ &= \mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{X}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{X} = \mathbf{X}^T\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{X} \\ &= \mathbf{X}^T\mathbf{H}\mathbf{X},\end{aligned}$$

kde $\mathbf{H}$ je symetrická idempotentní matice s hodností $h(\mathbf{H}) = n - 1$. Tím jsou splněny předpoklady Wishartova rozdělení, tj. $\mathbf{W} \sim W_p(n-1, \mathbf{\Sigma})$.
Musíme ještě dokázat, že se jedná o centrální Wishartovo rozdělení, kde $\boldsymbol{M} = \mathbf{0}$. K dokázání centrality využijeme důsledek 1.2 a ukážeme, že $\mathbf{H}\boldsymbol{M} = \mathbf{0}$, kde

$$\boldsymbol{M} = \begin{pmatrix}\boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{\mu}^T\end{pmatrix} = \mathbf{1}_n\boldsymbol{\mu}^T,$$

$$\mathbf{H}\boldsymbol{M} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{1}_n\boldsymbol{\mu}^T = \mathbf{1}_n\boldsymbol{\mu}^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n\boldsymbol{\mu}^T = \mathbf{1}_n\boldsymbol{\mu}^T - n\frac{1}{n}\mathbf{1}_n\boldsymbol{\mu}^T = \mathbf{0}.$$

Na závěr dokážeme, že $\overline{\mathbf{x}}$ a $\mathbf{W}$ jsou nezávislé. Pokud výběrový průměr formulujeme $\overline{\mathbf{x}} = \mathbf{X}\mathbf{k}^T$, kde $\mathbf{k}^{(n\times 1)} = \frac{1}{n}\mathbf{1}_n$, a Wishartovu matici $\mathbf{W} = \mathbf{X}^T\mathbf{H}\mathbf{X}$, pak podle poznámky 1.2 stačí ověřit, že $\mathbf{H}\mathbf{k} = \mathbf{0}$, tj.

$$\mathbf{H}\mathbf{k} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\frac{1}{n}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n - \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n - \frac{1}{n^2}n\mathbf{1}_n = \mathbf{0}.$$

□

**Věta 1.9.** *(Reprodukční vlastnost)*
*Jestliže náhodné matice* $\mathbf{S}_i$, $i = 1, \ldots, r$, *jsou nezávislé a mají p-rozměrné Wishartovo rozdělení, tj.* $\mathbf{S}_i \sim W_p(n_i, \mathbf{\Sigma})$, *pak také*

$$\sum_{i=1}^{r}\mathbf{S}_i \sim W_p\left(\sum_{i=1}^{r}n_i, \mathbf{\Sigma}\right).$$

**Důkaz:** Na základě Wishartova rozdělení matice $\mathbf{S}_i,\ i = 1, \ldots, r$, můžeme $\mathbf{S}_i$ zapsat ve tvaru

$$\mathbf{S}_i = \sum_{k=n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_i} \mathbf{x}_k \mathbf{x}_k^T.$$

Z nezávislosti náhodných matic $\mathbf{S}_i,\ i = 1, \ldots, r$, plyne nezávislost $\mathbf{x}_k$, $k = 1, \ldots, \sum_{i=1}^{r} n_i$. Navíc $\mathbf{x}_k \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, $k = 1, \ldots, \sum_{i=1}^{r} n_i$.
Protože $\sum_{i=1}^{r} \mathbf{S}_i = \sum_{k=1}^{n_1+\cdots+n_r} \mathbf{x}_k \mathbf{x}_k^T$, má $\sum_{i=1}^{r} \mathbf{S}_i \sim W_p\left(\sum_{i=1}^{r} n_i, \mathbf{\Sigma}\right)$.

□

**Věta 1.10.** *(Kvadratická transformace)*
*Nechť je dána náhodná matice* $\mathbf{S} \sim W_p(n, \mathbf{\Sigma})$ *a matice* $\mathbf{A}^{(k\times p)}$. *Pak má kvadratická forma*

$$\mathbf{ASA}^T \sim W_k\left(n, \mathbf{A\Sigma A}^T\right).$$

**Důkaz:** Náhodná matice $\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ sestává z nezávislých $p$-rozměrných normálně rozdělených náhodných vektorů $\mathbf{x}_i,\ i = 1, \ldots, n$, s nulovou střední hodnotou a varianční maticí $\mathbf{\Sigma}$, tj. $\mathbf{x}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma})$.

Kvadratickou formu $\mathbf{ASA}^T$ vyjádříme pomocí jejích prvků,

$$\mathbf{ASA}^T = \sum_{i=1}^{n} \mathbf{A}\mathbf{x}_i \mathbf{x}_i^T \mathbf{A}^T = \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^T,$$

kde jsme dosadili za $\mathbf{A}\mathbf{x}_i = \mathbf{u}_i,\ i = 1, \ldots, n$. Je zřejmé, že $\mathbf{u}_i$ jsou nezávislé a mají $k$-rozměrné normální rozdělení, tj.

$$\mathbf{u}_i \sim N_k\left(\mathbf{0}, \mathbf{A\Sigma A}^T\right),\ i = 1, \ldots, n.$$

Pak analogicky z definice 1.4 bude mít

$$\mathbf{ASA}^T \sim W_k\left(n, \mathbf{A\Sigma A}^T\right).$$

□

**Věta 1.11.** *(O reziduálním součtu čtverců)*
*Nechť* $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, *kde* $\mathbf{X}^{(n\times p)}$ *je matice konstant (regresorů) a* $\boldsymbol{\beta}^{(p\times 1)}$ *je vektor regresních parametrů. Potom náhodná veličina představující reziduální součet čtverců, kterou značíme*

$$R_0^2 = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \sim \sigma^2\chi^2_{n-r},$$

*kde* $r$ *je hodnost matice* $\mathbf{X}$.

**Důkaz:** viz [15], str. 223, věta (I). □

**Poznámka 1.3.** *V následujících větách budeme uvažovat náhodný vektor* $\mathbf{x}^{(p\times 1)}$ *složený ze dvou podvektorů* ${}_1\mathbf{x}^{(r\times 1)}$ *a* ${}_2\mathbf{x}^{(s\times 1)}$, *kde* $p = r + s$.
*Vektor středních hodnot* $\boldsymbol{\mu}^{(p\times 1)}$ *je pak tvořený dvěma podvektory* ${}_1\boldsymbol{\mu}^{(r\times 1)}$ *a* ${}_2\boldsymbol{\mu}^{(s\times 1)}$.
*Bloky, které tvoří rozklad varianční matice* $\boldsymbol{\Sigma}^{(p\times p)}$, *označíme* $\boldsymbol{\Sigma}_{11}^{(r\times r)}$,$\boldsymbol{\Sigma}_{12}^{(r\times s)}$,$\boldsymbol{\Sigma}_{21}^{(s\times r)}$, $\boldsymbol{\Sigma}_{22}^{(s\times s)}$.
*Náhodnou matici* $\mathbf{S}$ *zapíšeme ve tvaru* $\mathbf{S} = (\mathbf{S}_{rm})_{i,j=1}^2$, *kde* $\mathbf{S}_{rm}$, $r, m = 1, 2$, *jsou její bloky.*

**Lemma 1.2.** *(O podmíněném rozdělení)*
*Nechť je dán náhodný vektor* $\mathbf{x}$, *který má p-rozměrné regulární normální rozdělení, tj.* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Pak podmíněné rozdělení* ${}_1\mathbf{x}$ *při pevném* ${}_2\mathbf{x}$ *je*

$${}_1\mathbf{x}|{}_2\mathbf{x} \sim N_r({}_1\boldsymbol{\mu} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}({}_2\mathbf{x} - {}_2\boldsymbol{\mu}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

**Důkaz:** viz [4], str. 62, věta 4.12. □

**Věta 1.12.** *Nechť je dána náhodná matice* $\mathbf{S} \sim W_p(n, \boldsymbol{\Sigma})$ *a k ní inverzní matice* $\mathbf{S}^{-1} = (S^{rm})_{r,m=1}^p$. *Dále nechť k dané pozitivně definitní varianční matici* $\boldsymbol{\Sigma} = (\sigma_{rm})_{r,m=1}^p$ *existuje inverze* $\boldsymbol{\Sigma}^{-1} = (\sigma^{rm})_{r,m=1}^p$ *a* $n > p - 1$.

*(i) Pak platí, že* $\dfrac{\sigma^{pp}}{S^{pp}} \sim \chi^2_{n-p+1}$ *a nezávisí na* $(S_{rm})_{r,m=1}^{p-1}$.

*(ii) Pro libovolný pevný vektor* $\mathbf{a}$ *platí*

$$\frac{\mathbf{a}^T\boldsymbol{\Sigma}^{-1}\mathbf{a}}{\mathbf{a}^T\mathbf{S}^{-1}\mathbf{a}} \sim \chi^2_{n-p+1}. \tag{4}$$

**Důkaz:**

(i) Budeme předpokládat, že je dána náhodná matice $\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X}^T\mathbf{X}$, kde $\mathbf{x}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma}),\ i = 1, ..., n$, jsou nezávislé.

Vezmeme $i$-tý řádek datové matice $\mathbf{X}$,

$$\mathbf{x}_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$$

a označme

$$\mathbf{x}_i^* = (X_{i1}, X_{i2}, ..., X_{i,p-1})^T,\ i = 1, ..., n.$$

Podmíněné rozdělení náhodné veličiny $X_{ip}$ při pevných hodnotách náhodných veličin $X_{i1}, X_{i2}, ..., X_{ip-1}$ je podle lemmatu 1.2

$$X_{ip}|X_{i1}, X_{i2}, ..., X_{ip-1} \sim N_1\left(\sum_{j=1}^{p-1} \beta_j X_{ij}, \frac{1}{\sigma^{pp}}\right).$$

Z věty 1.11 o reziduálním součtu čtverců víme, že náhodná veličina

$$R_0^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (X_{ip} - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{ip-1})^2$$

má $\frac{1}{\sigma^{pp}}\chi^2_{n-p+1}$ rozdělení, které je podmíněno vektory $\mathbf{x}_1^*, ..., \mathbf{x}_n^*$.

Víme ovšem, že toto podmíněné rozdělení nezávisí na podmínce, tj. jedná se o nepodmíněné rozdělení. Navíc $R_0^2$ je nezávislé na $(X_{ij})_{i=1,j=1}^{n,p-1}$.

Z podmínky $\mathbf{x}_1^*, ..., \mathbf{x}_n^*$ vytvoříme matici $(S_{rm})_{r,m=1}^{p-1}$ a dokážeme, že $R_0^2 = \frac{1}{S^{pp}}$. Datovou matici $\mathbf{X}^{(n\times p)}$ rozdělíme na bloky $\mathbf{A}^{(n\times(p-1))} = (X_{ij})_{i=1,j=1}^{n,p-1}$ a $\mathbf{b}^{(n\times 1)} = (X_{1p}, ..., X_{np})^T$, tj.

$$\mathbf{X} = (\mathbf{A}|\mathbf{b}).$$

Reziduální součet čtverců při regresi $\mathbf{b}$ na $\mathbf{A}$, kde $|\mathbf{A}^T\mathbf{A}| \neq 0$, je možné zapsat jako podíl

$$R_0^2 = \begin{vmatrix} \mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A} & \mathbf{b}^T\mathbf{b} \end{vmatrix} / |\mathbf{A}^T\mathbf{A}|.$$

Přitom

$$\mathbf{S} = \mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{b} \\ \mathbf{b}^T\mathbf{A} & \mathbf{b}^T\mathbf{b} \end{pmatrix},$$

$$|\mathbf{S}| = |\mathbf{A}^T\mathbf{A}||\mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}|.$$

Pak dle [4], str. 319, věta A.19 obdržíme $R_0^2 = \dfrac{|\mathbf{S}|}{|\mathbf{A}^T\mathbf{A}|} = \dfrac{1}{S^{pp}}$, což jsme chtěli dokázat. Poznamenejme přitom, že analogicky je možno obdržet obecněji $\dfrac{\sigma^{jj}}{S^{jj}}$, $j = 1, ..., p$.

(ii) Budeme uvažovat kvadratickou formu $\mathbf{ASA}^T$.

Námi zvolený vektor $\mathbf{a} \neq \mathbf{0}$ je prvním řádkem ortogonální matice $\mathbf{A}^{(p\times p)}$. Využitím věty 1.10 o kvadratické transformaci můžeme psát

$$\mathbf{ASA}^T \sim W_p(n, \mathbf{A\Sigma A}^T).$$

První diagonální prvek kvadratických forem

$$\left(\mathbf{ASA}^T\right)^{-1} = \mathbf{AS}^{-1}\mathbf{A}^T \text{ je } \mathbf{a}^T\mathbf{S}^{-1}\mathbf{a},$$
$$\left(\mathbf{A\Sigma A}^T\right)^{-1} = \mathbf{A\Sigma}^{-1}\mathbf{A}^T \text{ je } \mathbf{a}^T\mathbf{\Sigma}^{-1}\mathbf{a}.$$

Vycházíme z tvrzení (i) a dostáváme $\dfrac{\mathbf{a}^T\mathbf{\Sigma}^{-1}\mathbf{a}}{\mathbf{a}^T\mathbf{S}^{-1}\mathbf{a}} \sim \chi^2_{n-p+1}$.

□

**Věta 1.13.** *(Marginální rozdělení)*
*Nechť je dána náhodná matice* $\mathbf{S} \sim W_p(n, \mathbf{\Sigma})$. *Pak*

$$\mathbf{S}_{11} \sim W_r(n, \mathbf{\Sigma}_{11}).$$

**Důkaz:** Nejprve vezmeme náhodné vektory, z nichž se skládá náhodná matice $\mathbf{S} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, kde $\mathbf{x}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, ..., n$, jsou nezávislé. Rozložíme vektory $\mathbf{x}_i$ na složky,

$$\mathbf{x}_i = \begin{pmatrix} {}_1\mathbf{x}_i \\ {}_2\mathbf{x}_i \end{pmatrix}.$$

Pak ${}_1\mathbf{x}_i \sim N_r(\mathbf{0}, \mathbf{\Sigma}_{11})$, $i = 1, ..., n$, jsou také nezávislé a

$$\mathbf{S}_{11} = \sum_{i=1}^{n} {}_1\mathbf{x}_i \; {}_1\mathbf{x}_i^T \sim W_r(n, \mathbf{\Sigma}_{11}).$$

□

**Poznámka 1.4.** *Nakonec se též zmíníme o hustotě a charakteristické funkci Wishartova rozdělení.*

*(i) Hustota Wishartova rozdělení* $\mathbf{S} \sim W_p(n, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ *pozitivně definitní, je dána vztahem*

$$f(\mathbf{S}) = \frac{|\mathbf{S}|^{\frac{n-p-1}{2}} \exp\left\{-\frac{1}{2} tr(\mathbf{S}\mathbf{\Sigma}^{-1})\right\}}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} |\mathbf{\Sigma}|^{\frac{n}{2}} \prod_{j=1}^{p} \Gamma(\frac{n+1-j}{2})},$$

*kde* $tr(\mathbf{S}\mathbf{\Sigma}^{-1})$ *značí stopu matice* $\mathbf{S}\mathbf{\Sigma}^{-1}$.

*(ii) Charakteristická funkce Wishartova rozdělení (i pro singulární* $\mathbf{\Sigma}$*) je dána vztahem*

$$\varphi_{\mathbf{S}}(\mathbf{T}) = \mathsf{E} \exp\{\mathrm{i}, tr(\mathbf{S}\mathbf{T})\} = |\mathbf{I}_p - 2\mathrm{i}\mathbf{T}\mathbf{\Sigma}|^{-\frac{n}{2}},$$

*kde* $\mathbf{T}^{(p\times p)}$ *je symetrická matice.*

## 1.5 Definice Hotellingovy statistiky

Hotellingova statistika představuje klíčový pojem při testování hypotézy o střední hodnotě při výběrech z $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Odvozením jejího rozdělení též vrcholí úvahy, provedené v předchozí kapitole.

Při tvorbě této kapitoly jsem použila [9], [10] a [15].

**Definice 1.5.** *Nechť je dána náhodná matice* $\mathbf{S} \sim W_p(n, \mathbf{\Sigma})$ *a na* $\mathbf{S}$ *nezávislý p-složkový vektor* $\mathbf{z}$ *s rozdělením* $N_p\left(\boldsymbol{\nu}, \frac{1}{c}\mathbf{\Sigma}\right)$. *Potom*

$$T^2 = nc\mathbf{z}^T\mathbf{S}^{-1}\mathbf{z} = \frac{n\mathbf{z}^T\mathbf{S}^{-1}\mathbf{z}}{\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{z}} c\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{z}$$

*se nazývá Hottelingova* $T^2$ *statistika.*

**Věta 1.14.** *Nechť $n > p - 1$. Potom*

$$F = \frac{n-p+1}{p}\frac{T^2}{n} \sim F_{p,n-p+1}(\delta), \tag{5}$$

*kde $\delta = c\boldsymbol{\nu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$ je parametr necentrality.*

**Důkaz:** Víme, že Fisherovo rozdělení je definováno jako podíl dvou nezávislých náhodných veličin, které mají $\chi^2$ rozdělení. Vztah (5) přepíšeme do tvaru

$$F = \frac{n-p+1}{p}\frac{\mathbf{z}^T\mathbf{S}^{-1}\mathbf{z}}{\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}}c\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}.$$

Ve větě 1.12 jsme dokázali, že podíl daný vztahem (4) má $\chi^2$ rozdělení o $n - p + 1$ stupních volnosti, které je nezávislé na $\mathbf{z}$. Z toho vyplývá, že i tento podíl daný vztahem (4) je nezávislý na $\mathbf{z}$.

Také $c\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} \sim \chi^2_p(\delta)$, $\delta = c\boldsymbol{\nu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$; z toho je tedy zřejmé, že $F \sim F_{p,n-p+1}$ necentrální s parametrem necentrality $\delta = c\boldsymbol{\nu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$.

□

## 1.6 Test hypotézy o vektoru $\boldsymbol{\mu}$ při neznámé matici $\boldsymbol{\Sigma}$

V této kapitole jsem čerpala zejména z [3], [9], [10] a [16].

Nechť $\mathbf{x}_i = (X_{i1}, ..., X_{ip})^T$, $i = 1, ..., n$, je náhodný výběr z $p$-rozměrného normálního rozdělení, kde neznáme ani střední hodnotu $\boldsymbol{\mu}$ ani pozitivně definitní varianční matici $\boldsymbol{\Sigma}$.

Dále nechť je dán výběrový průměr $\overline{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$ a Wishartova matice $\mathbf{W} \sim W_p(n-1, \boldsymbol{\Sigma})$, které, jak jsme dokázali v důsledku 1.3, jsou nezávislé.

Ověřujeme nulovou hypotézu $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ proti alternativě $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ pomocí Hotellingovy statistiky

$$T^{2*} = (n-1)n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T\mathbf{W}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0), \tag{6}$$

kde $\frac{1}{n-1}\mathbf{W}$ představuje výběrovou varianční matici.

Podle věty 1.12 je nulová hypotéza $H_0$ správná, když testovací statistika

$$\frac{n-p}{p}\frac{T^{2*}}{n-1} \sim F_{p,n-p}. \tag{7}$$

V neprospěch $H_0$ svědčí hodnoty statistiky $\frac{n-p}{p}\frac{T^{2*}}{n-1}$ větší než $(1-\alpha)$-kvantil $F_{p,n-p}$ rozdělení při testu na hladině $\alpha$.

Nulovou hypotézu můžeme také testovat ve tvaru $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\mu}_0$ proti $H_A : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{C}\boldsymbol{\mu}_0$, kde $\mathbf{C}^{(q\times p)}$ je matice konstant s hodností $h(\mathbf{C}) = q,\ (q \leq p)$. Hotellingova statistika po transformaci dostane tvar

$$T_C^2 = n(n-1)(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}_0)^T(\mathbf{C}\mathbf{W}\mathbf{C}^T)^{-1}(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}_0).$$

Platí-li nulová hypotéza, pak testovací statistika $\frac{n-q}{n}\frac{T_C^2}{n-1} \sim F_{q,n-q}$. Nulovou hypotézu zamítáme, překročí-li hodnota testovací statistiky kvantil $F_{q,n-q;1-\alpha}$.

Pokud bychom chtěli testovat jen prvních $q$ složek vektoru $\boldsymbol{\mu}$, volíme matici $\mathbf{C}$ ve tvaru

$$\mathbf{C} = \begin{pmatrix} 1 & 0\cdots0\cdots0 \\ 0 & 1\cdots0\cdots0 \\ \cdots\cdots\cdots\cdots\cdots \\ 0 & 0\cdots1\cdots0 \end{pmatrix}.$$

# 2 Kompoziční data

V této kapitole se budeme věnovat kompozičním datům jako specifickému druhu mnohorozměrných pozorování, s kvantifikovaně vyjádřenými relativními příspěvky částí na celku.

Popíšeme výběrový prostor, na kterém jsou tato data definována, tj. simplex, společně se speciální geometrií, která je kompozičním datům přirozená.

Abychom mohli kompozice statisticky analyzovat jako standardní mnohorozměrná data je zapotřebí tyto nejprve transformovat prostřednictvím tzv. logratio transformací (log-ratio = logaritmus podílu). Budeme uvažovat především izometrickou logratio (ilr) transformaci, která umožní kompozice vyjádřit v ortonormálním souřadnicovém systému. Důležité přitom je, že uvedené logratio transformace jsou bijekce, je tedy vždy možné aplikovat zpětnou (inverzní) transformaci, jejímž výsledkem je původní kompozice.

Následně se již budeme věnovat samotným aspektům testování hypotéz o centru distribuce náhodné kompozice. V tomto ohledu se zmíníme o BLUE centra distribuce, zavedení normality na simplexu i invariantnosti testovacích statistik na volbu ortonormálních souřadnic na simplexu.

## 2.1 Definice a výběrový prostor

Při psaní této kapitoly jsem čerpala z [1] a [13].

První ucelený pohled na statistickou analýzu kompozičních dat zveřejnil v 80. letech minulého století John Aitchison v knize [1]. Do té doby byly pokusy o jejich statistické zpracování charakterizovány především paradoxními výsledky s diskutabilní možností relativních závěrů.

**Definice 2.1.** *Sloupcový vektor* $\mathbf{x} = (x_1, ..., x_D)^T$ *se nazývá D-složková kompozice, jestliže všechny jeho složky jsou kladná čísla nesoucí pouze relativní informaci.*

Takzvaná uzavřená forma kompozičních dat, kde součet složek kompozice je roven předepsané konstantě $k$, je charakteristickou reprezentací kompozičních dat.

Konstantu $k$ nejčastěji volíme 1 (složky kompozice vyjadřují proporcionální části na celku) nebo 100 (procentuální podíly na celku). Z definice 2.1 je přitom zřejmé, že konstantu je možné zvolit bez ztráty informace, obsažené pouze v podílech mezi složkami kompozice.

Kompoziční data jsou tudíž obvykle představována tzv. uzávěrem:

**Definice 2.2.** *Uzávěr kompozice* $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T \in \mathbb{R}_+^D$ *je vektor*

$$\mathcal{C}(\mathbf{x}) = \left( \frac{kx_1}{\sum_{i=1}^D x_i}, \frac{kx_2}{\sum_{i=1}^D x_i}, ..., \frac{kx_D}{\sum_{i=1}^D x_i} \right)^T .$$

Následně již můžeme přikročit k definici výběrového prostoru (reprezentací) kompozičních dat:

**Definice 2.3.** *Výběrový prostor kompozičních dat se nazývá simplex a je definován vztahem*

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)^T | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=}^{D} x_i = k \right\} .$$

Simplex tak vlastně představuje $(D-1)$-dimenzionální podprostor $D$-rozměrného reálného prostoru.

Z $D$-složkové kompozice $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T$ můžeme vytvořit podkompozici $\mathbf{x}_s = (x_{i_1}, x_{i_2}, \ldots, x_{i_s})^T$, kde indexy $i_1, i_2, ..., i_s$ vyjadřují, které složky z původní kompozice jsou vybrány.

## 2.2 Aitchisonova geometrie na simplexu

V této kapitole bylo použito zdrojů [1], [2], [6], [11], [13] a [14].

Nejprve budeme definovat operace vhodné pro práci s kompozičními daty přímo na simplexu. Euklidovská geometrie, která je základem většiny mnohorozměrných statistických metod, se totiž pro kompoziční data ukázala jako zcela nevhodná [1], [2], [7] a [8].

Operace perturbace kompozic na simplexu je analogická operaci sčítání vektorů v reálném prostoru. Mocninná transformace kompozice reálným číslem je pak obdobou operace násobení vektoru skalárem.

**Definice 2.4.** *Perturbace kompozice* $\mathbf{x} \in \mathcal{S}^D$ *kompozicí* $\mathbf{y} \in \mathcal{S}^D$ *je kompozice*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, ..., x_D y_D\right)^T.$$

**Definice 2.5.** *Mocninná transformace kompozice* $\mathbf{x} \in \mathcal{S}^D$ *reálným číslem* $\alpha \in \mathbb{R}$ *je kompozice*

$$\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha\right)^T.$$

Je patrné, že tyto základní operace na simplexu jsou definovány pomocí uzávěru kompozice. Ten je chápan jako projekce vektoru s kladnými složkami na simplex.

Simplex spolu s operacemi na něm zavedenými, perturbací a mocninnou transformací, tvoří lineární prostor, který označíme $(\mathcal{S}^D, \oplus, \odot)$. Znamená to, že libovolné kompozice $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^D$ vzhledem k perturbaci splňují tyto podmínky:

(i) komutativita, tj. $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;

(ii) asociativita, tj. $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;

(iii) neutrálním prvkem perturbace je kompozice $\mathbf{n} = \mathcal{C}(1, 1, ..., 1)^T = \left(\frac{1}{D}, \frac{1}{D}, ..., \frac{1}{D}\right)^T$;

(iv) inverzním prvkem perturbace je neutrální prvek, tj. $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$ , kde $\mathbf{x}^{-1} = \mathcal{C}(x_1^{-1}, x_2^{-1}, ..., x_D^{-1})^T$.

Pak zapisujeme $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y}) = \mathbf{x} \oplus \mathbf{y}^{-1}$.

Dále uvedeme vlastnosti, kterým vyhovuje operace mocninná transformace pro libovolné kompozice $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ a konstanty $\alpha, \beta \in \mathbb{R}$:

(i) asociativita, tj. $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha\beta) \odot \mathbf{x}$;

(ii) distributivita, tj.

$$\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y}),$$
$$(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \oplus \mathbf{x});$$

(iii) neutrální prvek: $1 \odot \mathbf{x} = \mathbf{x}$; neutrální prvek je jediný.

Nyní se zmíníme o důvodech, které vedou k nezahrnutí nulových složek do definice kompozice.

Pokud se přitom jedná o úplnou absenci složek v pozorování, hovoříme o tzv. strukturních nulách. Tyto se nejčastěji vyskytují při výběrových šetřeních; je zřejmé, že např. vybrané osoby - nekuřáci budou mít při zjišťování struktury výdajů domácností na jednotlivé komodity nulové výdaje za cigarety.

Na druhé straně, jestliže nulová hodnota ukazuje na přítomnost komponenty, která je ovšem zastoupená jen v malém, nedetekovatelném množství (např. určitý chemický prvek ve sloučenině), hovoří se o tzv. zaokruhlené nule.

Poznamenejme též, že kdybychom zahrnuli nulovou složku do kompozice, pak nebude splněna vlastnost existence inverzního prvku pro každé $\mathbf{x} \in \mathcal{S}^D$, a tím porušíme vektorovou strukturu na simplexu.

V následujícím ukážeme, že lze na simplexu dokonce zavést reálný normovaný lineární prostor se skalárním součinem. Tento prostor je navíc úplný v metrice indukované skalárním součinem, je tedy Hilbertovým prostorem, který se souhrnně nazývá Aitchisonova geometrie.

Začněme definicí skalárního součinu kompozic.

**Definice 2.6.** *(Aitchisonův skalární součin)*
*Skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *je definovaný vztahem*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})},$$

*kde* $g(\mathbf{x}) = \prod_{i=1}^{D} (x_i)^{\frac{1}{D}}$, *resp.* $g(\mathbf{y}) = \prod_{i=1}^{D} (y_i)^{\frac{1}{D}}$ *je geometrický průměr složek kompozice* $\mathbf{x}$, *resp* $\mathbf{y}$.

Definováním skalárního součinu na lineárním prostoru získáme lineární prostor se skalárním součinem.

**Definice 2.7.** *(Aichisonova norma)*
*Normu kompozice* $\mathbf{x} \in \mathcal{S}^D$ *zavádíme jako*

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j}\right)^2} = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_a}.$$

Definicí normy kompozice obdržíme normovaný lineární prostor se skalárním součinem; povšimněme si přitom její souvislosti se skalárním součinem.

**Definice 2.8.** *Řekneme, že lineární prostor se skalárním součinem je Hilbertův prostor, je-li úplný normovaný prostor v normě* $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$.

**Definice 2.9.** *(Aitchisonova vzdálenost, Aitchisonova metrika)*
*Vzdálenost mezi kompozicemi* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *definujeme jako*

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j} - \frac{y_i}{y_j}\right)^2}.$$

Aitchisonova metrika má tyto vlastnosti:

(i) je invariantní na změnu škály: $d_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_a(\mathbf{x}, \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$;

(ii) je invariantní funkcí vzhledem k permutaci složek kompozice;

(iii) je invariantní vzhledem k permutaci, tj. $d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{z} \oplus \mathbf{x}, \mathbf{z} \oplus \mathbf{y})$.

Důkaz výše uvedených vlastností nalezeme v literatuře [14].

Aitchisonova geometrie na simplexu tedy zahrnuje Aitchisonův skalární součin, normu a metriku, které jsou definované na $(\mathcal{S}^D, \oplus, \odot)$, má tak stejné vlastnosti jako euklidovská geometrie v reálnému prostoru.

Následující definice souvisí s dimenzí zavedeného Hilbertova prostoru:

**Definice 2.10.** *Řekneme, že kompozice* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{D-1} \in \mathcal{S}^D$ *jsou lineárně nezávislé, jestliže*

$$(\alpha_1 \odot \mathbf{x}_1) \oplus (\alpha_2 \odot \mathbf{x}_2) \oplus \cdots \oplus (\alpha_{D-1} \odot \mathbf{x}_{D-1}) = \mathbf{n}$$

*tehdy a jen tehdy, když* $\alpha_1 = \alpha_2 = \cdots = \alpha_{D-1} = 0$.

Na simplexu tedy máme $D-1$ lineárně nezávislých kompozic, které tvoří bázi, dimenze Hilbertova prostoru je proto rovna $D-1$.

Dalším argumentem proti zahrnutí nulových složek do kompozice je tedy skutečnost, že logaritmus nuly neexistuje, nemohli bychom takto definovat Aitchisonovu metriku, normu a skalární součin.

Obecné vlastnosti invariantnosti na permutaci složek a na změnu škály kompozic, které jsme si uvedli u Aitchisonovy metriky, tvoří zároveň dvě ze tří vyžadovaných vlastností, které musí splňovat každá statistická metoda, aby byla aplikovatelná na kompoziční data.

Třetí vlastností je pak podkompoziční soudržnost, což znamená, že informace získaná z kompozice o $D$ složkách by neměla být ve sporu s informací získanou z podkompozice o $d$ složkách, $d \leq D$.

Bohužel, ani jeden z uvedených požadavků není dodržen, použijeme-li pro kompoziční data standardní statistické metody. Tato je totiž nejprve potřeba transformovat do reálného prostoru, jak si ukážeme v další kapitole.

## 2.3 Ortonormální souřadnice pro kompoziční data

V této kapitole jsem převážně čerpala z [5], [8] a [13].

Hlavní motivací Johna Aitchisona bylo najít vhodný způsob, jak pracovat s kompozičními daty, aby bylo možné použít známé statistické metody, a aby přitom kompozice zachovávaly relativní informaci, kterou v sobě nesou.

Výsledkem jeho práce bylo zavedení tzv. logratio transformací, aditivní logratio (alr) a centrované logratio (clr) transformace. Nicméně tyto transformace

mají velkou nevýhodou, nemají totiž jednotnou možnost použití. Pro statistické modelování využíváme alr transformaci, která ale není izometrická, tedy např. nezachovává vzdálenost kompozic. Pracujeme-li se statistickými metodami, které jsou založeny na metrice, používáme clr transformaci. Ta oproti alr transformaci sice zachovává metrické vlastnosti, ale nevýhodou je, že ve výsledku dostaneme singulární varianční matici.

Hlavním důvodem, který ale vedl k definici nové transformace, nazvané izometrická logratio (ilr) transformace, byl nemožnost pracovat s clr a alr transformacemi v ortogonálním souřadnicovém systému, na něhož jsme v praxi zvyklí.

Pro definování žádoucí ortonormální báze na simplexu je zapotřebí na počátku nalézt vhodnou generující množinu.
Jako množinu generátorů zvolíme $W = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_D\}$, kde

$$\mathbf{w}_i = \mathcal{C}(\exp(\vec{\mathbf{e}_i}))^T = \mathcal{C}(1, 1, ..., e, ..., 1)^T,$$

a $\vec{\mathbf{e}_i}$, $i = 1, ..., D$ je kanonická báze v $\mathbb{R}^D$. Můžeme si povšimnout, že na $i-$té pozici kompozice $\mathbf{w}_i$ se nachází Eulerovo číslo $e$.

Díky základním operacím na simplexu a známým vlastnostem Hilbertova prostoru můžeme každou kompozici $\mathbf{x} \in \mathcal{S}^D$ vyjádřit ve tvaru lineární kombinace kompozic z množiny generátorů,

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln x_i \odot \mathbf{w}_i = \ln x_1 \odot \mathbf{w}_1 \oplus \ln x_2 \odot \mathbf{w}_2 \oplus \cdots \oplus \ln x_D \odot \mathbf{w}_D \tag{8}$$

nebo také

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \odot \mathbf{w}_i = \ln \frac{x_1}{g(\mathbf{x})} \odot \mathbf{w}_1 \oplus \ln \frac{x_2}{g(\mathbf{x})} \odot \mathbf{w}_2 \oplus \cdots \oplus \ln \frac{x_D}{g(\mathbf{x})} \odot \mathbf{w}_D. \tag{9}$$

Všimněme si, že geometrický průměr složek kompozic $g(\mathbf{x})$ ze vztahu (9) je nahrazen konstantou 1 ve vztahu (8). Ukazuje to na nejednoznačnost koeficientů vzhledem ke generující množině.

Koeficienty $\ln \frac{x_i}{g(\mathbf{x})}$, $i = 1, ..., D$, ze vztahu (9) představují složky reálného vektoru $\mathbf{y} \in \mathbb{R}^D$, tj.

$$\mathbf{y} = \left( \ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, ..., \ln \frac{x_D}{g(\mathbf{x})} \right)^T = (y_1, y_2, ..., y_D)^T.$$

Takto je zároveň definována clr transformace kompozice $\mathbf{x}$ z $\mathcal{S}^D$ do $\mathbb{R}^D$, jinak řečeno $\mathbf{y} = clr(\mathbf{x})$.

Chceme-li poté opět získat kompozici $\mathbf{x} \in \mathcal{S}^D$, musíme provést zpětnou transformaci,

$$clr^{-1}(\mathbf{y}) = \mathcal{C} \left( \exp(y_1), \exp(y_2), ..., \exp(y_D) \right)^T = \mathbf{x}.$$

Vektor $\mathbf{y} \in \mathbb{R}^D$ ovšem nevyjadřuje souřadnice v bázi, nýbrž pouze v generujícím systému. To plyne i z dimenze $D - 1$ Aitchisonovy geometrie.

Jedna z důležitých vlastností clr transformace je symetrie ve složkách, která usnadňuje interpretaci nových souřadnic. Problémem je však singularita dat, tj. $\sum_{i=1}^{D} y_i = 0$.

Naopak, předností clr transformace jsou vztahy mezi operacemi definovanými na simplexu a na vektorovém prostoru, obecně pro $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ a $\alpha, \beta \in \mathbb{R}$ platí

$$\text{(i)} \quad clr\left(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}\right) = \alpha \cdot clr(\mathbf{x}) + \beta \cdot clr(\mathbf{y}); \tag{10}$$

$$\text{(ii)} \quad \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_a ; \tag{11}$$

$$\text{(iii)} \quad \|clr(\mathbf{x})\| = \|\mathbf{x}\|_a ; \tag{12}$$

$$\text{(iv)} \quad d(clr(\mathbf{x}), clr(\mathbf{y})) = d_a(\mathbf{x}, \mathbf{y}), \tag{13}$$

kde se ve výrazech na levé straně rovností postupně objeví standardní euklidovský skalární součin, norma a vzdálenost.

Vyjmutím $\mathbf{w}_D$ z množiny generátorů $W$ dostaneme bázi na $\mathcal{S}^D$. Důvodem je fakt, že na simplexu máme maximálně $D - 1$ lineárně nezávislých kompozic, které tvoří bázi.

Potom můžeme $\mathbf{x} \in \mathcal{S}^D$ zapsat ve tvaru

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} \ln \frac{x_i}{x_D} \odot \mathbf{w}_i =$$
$$= \ln \frac{x_1}{x_D} \odot (e, 1, ..., 1)^T \oplus \cdots \oplus \ln \frac{x_{D-1}}{x_D} \odot (1, 1, ..., e, 1)^T.$$

Alr transformace je tak definována právě koeficienty $\ln \frac{x_i}{x_D},\ i = 1, ..., D-1$, tj.

$$alr(\mathbf{x}) = \left( \ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, ..., \ln \frac{x_{D-1}}{x_D} \right)^T.$$

Euklidovská vzdálenost vypočtená z alr transformovaných kompozic a Aitchisonova vzdálenost určená z původních dat nejsou stejné. Příčinou je, že souřadnice vektoru alr transformace jsou vyjádřeny vzhledem k bázi na simplexu, která není ortonormální. Toto můžeme ověřit, spočítáme-li skalární součin jakýchkoliv dvou kompozic báze $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{D-1}\}$.

Označme $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{D-1}\}$ ortonormální bázi na simplexu, kde

$$\mathbf{v}_i = \mathcal{C}\left[ \exp\left( \frac{1}{\sqrt{(D-i+1)(D-i)}} \right), ..., \exp\left( \frac{1}{\sqrt{(D-i+1)(D-i)}} \right), \exp\left( \sqrt{\frac{D-i+1}{D-i}} \right), \underbrace{1, ..., 1}_{D-i} \right]^T$$

pro $i = 1, ..., D-1$. Tuto bázi dostaneme z $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{D-1}\}$ využitím Gram-Schmidtovy ortonormalizační metody, modifikované pro kompoziční data.

Námi zvolená ortonormální báze je jednou z nekonečně mnoha ortonormálních bází, které můžeme definovat na $\mathcal{S}^D$. Pro ortonormální báze s využitím (11) platí

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_a = \langle clr(\mathbf{v}_i), clr(\mathbf{v}_j) \rangle = \delta_{ij} = \begin{cases} 0, & \text{pro } i \neq j \\ 1, & \text{pro } i = j. \end{cases}$$

Pro vyjádření kompozice $\mathbf{x} \in \mathcal{S}^D$ v souřadnicích vzhledem k vybrané bázi $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{D-1}\}$ budeme potřebovat funkci, která zobrazí $\mathcal{S}^D$ do $\mathbb{R}^{D-1}$. Oba prostory jsou stejné dimenze, proto můžeme hovořit o izometrii těchto prostorů.

Výše uvedená funkce pak představuje ilr transformaci, definovanou jako

$$\begin{aligned} ilr(\mathbf{x}) &= \left(\langle \mathbf{x}, \mathbf{v}_1 \rangle_a, \langle \mathbf{x}, \mathbf{v}_2 \rangle_a, ..., \langle \mathbf{x}, \mathbf{v}_{D-1} \rangle_a\right)^T = \\ &= \left( \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}, \sqrt{\frac{2}{3}} \ln \frac{\sqrt{x_1 x_2}}{x_3}, \ldots, \sqrt{\frac{D-1}{D}} \ln \frac{\sqrt[D-1]{\prod_{i=1}^{D-1} x_i}}{x_D}, \right)^T = \\ &= (x_1^*, x_2^*, ..., x_{D-1}^*)^T = \mathbf{x}^*, \ \mathbf{x}^* \in \mathbb{R}^{D-1}. \end{aligned}$$

Ilr transformovanou kompozici $\mathbf{x} \in \mathcal{S}^D$ zpětně vyjádříme pomocí inverzní ilr transformace,

$$\mathbf{i}lr^{-1}(\mathbf{x}^*) = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{v}_i = x.$$

Nyní uvedeme vlastnosti ilr transformace. Uvažujme $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ a $\alpha, \beta \in \mathbb{R}$:

(i) $ilr(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y}) = \alpha \cdot \mathbf{x}^* + \beta \cdot \mathbf{y}^*$;

(ii) $\langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle = \langle \mathbf{x}^*, \mathbf{y}^* \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_a$;

(iii) $\|ilr(\mathbf{x})\| = \|\mathbf{x}^*\| = \|\mathbf{x}\|_a$;

(iv) $d\left(ilr(\mathbf{x}), ilr(\mathbf{y})\right) = d(\mathbf{x}^*, \mathbf{y}^*) = d_a(\mathbf{x}, \mathbf{y})$.

Prostřednictvím zvolené ortonormální báze pak můžeme vyjádřit každou kompozici ze simplexu v ortonormálním souřadnicovém systému. Následně pak lze aplikovat běžné mnohorozměrné metody pro jejich statistickou analýzu. Jedinou možnou „nevýhodou“ ilr transformace je nutnost komplexního pohledu na interpretaci nově vzniklých souřadnic. V případě testování hypotéz, jak jsme uvedli v první kapitole, ovšem tohoto není potřeba. V dalším textu tedy bude ilr transformace tvořit osu našich úvah.

## 2.4 Elementy statistické analýzy na $\mathcal{S}^D$

V této kapitole jsem převážně vycházela z [12], [13] a [14].

Rozdělení pravděpodobnosti náhodné kompozice $\mathbf{x} = (x_1, \ldots, x_D)^T$ je charakterizováno jejím centrem a metrickým rozptylem kolem centra kompozice. Tyto

charakteristiky jsou definovány pomocí Aitchisonovy metriky jako analogie definice střední hodnoty a varianční matice náhodného vektoru.

**Definice 2.11.** *Metrický rozptyl náhodné kompozice* $\mathbf{x} \in \mathcal{S}^D$ *kolem (nenáhodné)* $\boldsymbol{\xi} \in \mathcal{S}^D$ *je definován vztahem*

$$\mathrm{Mvar}(\mathbf{x}, \boldsymbol{\xi}) = \mathsf{E}[d_a^2(\mathbf{x}, \boldsymbol{\xi})]$$

*za předpokladu existence uvedené střední hodnoty.*

Pokud metrický rozptyl náhodné kompozice $\mathbf{x} \in \mathcal{S}^D$ kolem $\boldsymbol{\xi} \in \mathcal{S}^D$ existuje, můžeme definovat centrum rozdělení následovně:

**Definice 2.12.** *Centrum rozdělení náhodné kompozice* $\mathbf{x} \in \mathcal{S}^D$ *je kompozice* $\boldsymbol{\xi} \in \mathcal{S}^D$*, která minimalizuje* $\mathrm{Mvar}(\mathbf{x}, \boldsymbol{\xi})$*. Zkráceně ji nazýváme centrum* $\mathbf{x}$ *a značíme* $\mathrm{cen}(\mathbf{x})$ *nebo* $\boldsymbol{\gamma}$*.*

**Definice 2.13.** *Metrický rozptyl kolem centra* $\mathrm{cen}(\mathbf{x}) = \boldsymbol{\gamma}$ *je definován jako*

$$\mathrm{Mvar}(\mathbf{x}, \boldsymbol{\gamma}) = \mathsf{E}[d_a^2(\mathbf{x}, \boldsymbol{\gamma})].$$

*Zkráceně jej nazýváme metrický rozptyl a značíme* $\mathrm{Mvar}(\mathbf{x})$*.*

Dále uvedeme některé důležité vlastnosti centra a metrického rozptylu náhodné kompozice.

**Věta 2.1.** *Centrum* $\mathrm{cen}(\mathbf{x})$ *představuje inverzní ilr transformaci střední hodnoty ilr transformované kompozice, tedy*

$$\mathrm{cen}(\mathbf{x}) = ilr^{-1}(\mathsf{E}[ilr(\mathbf{x})]).$$

**Důkaz:** viz [14], str. 271, věta 5. □

**Věta 2.2.** *Metrický rozptyl je roven součtu rozptylů logaritmů podílů (= log-ratios) složek náhodné kompozice* $\mathbf{x}$*, tj.*

$$\mathrm{Mvar}(\mathbf{x}) = \frac{1}{D} \sum_{i<j} \mathrm{var}\left(\ln \frac{x_i}{x_j}\right).$$

**Důkaz:** viz [14], str. 272, věta 6. □

**Věta 2.3.** *Nechť jsou dány náhodné kompozice* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathcal{S}^D$. *Pak*

$$\mathrm{cen}(\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_n) = \mathrm{cen}(\mathbf{x}_1) \oplus \mathrm{cen}(\mathbf{x}_2) \oplus \cdots \oplus \mathrm{cen}(\mathbf{x}_n).$$

**Důkaz:** Důkaz provedeme matematickou indukcí.
Nejprve ukážeme, že rovnost platí pro libovolné náhodné kompozice $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}^D$.
Využijeme přitom větu 2.1 a vlastnosti ilr transformace.

$$\begin{aligned}
\mathrm{cen}(\mathbf{x}_1 \oplus \mathbf{x}_2) &= ilr^{-1}\left(\mathsf{E}[ilr(\mathbf{x}_1 \oplus \mathbf{x}_2)]\right) = \\
&= ilr^{-1}\left(\mathsf{E}[ilr(\mathbf{x}_1) + ilr(\mathbf{x}_2)]\right) = \\
&= ilr^{-1}\left(\mathsf{E}[ilr(\mathbf{x}_1)]\right) \oplus ilr^{-1}\left(\mathsf{E}[ilr(\mathbf{x}_2)]\right) = \\
&= \mathrm{cen}(\mathbf{x}_1) \oplus \mathrm{cen}(\mathbf{x}_2).
\end{aligned}$$

Dále předpokládejme, že rovnost platí i pro $k$-náhodných kompozic, tj.

$$\mathrm{cen}(\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_k) = \mathrm{cen}(\mathbf{x}_1) \oplus \mathrm{cen}(\mathbf{x}_2) \oplus \cdots \oplus \mathrm{cen}(\mathbf{x}_k)$$

a ukážeme, že platí i pro $k+1$ náhodných kompozic.

$$\begin{aligned}
\mathrm{cen}(\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_k \oplus \mathbf{x}_{k+1}) &= ilr^{-1}\left(\mathsf{E}[ilr(\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_k \oplus \mathbf{x}_{k+1}]\right) = \\
&= ilr^{-1}(\mathsf{E}[ilr(\mathbf{x}_1) + ilr(\mathbf{x}_2) + \cdots + ilr(\mathbf{x}_k) + \\
&+ ilr(\mathbf{x}_{k+1})]) = \\
&= \mathrm{cen}(\mathbf{x}_1) \oplus \mathrm{cen}(\mathbf{x}_2) \oplus \cdots \oplus \mathrm{cen}(\mathbf{x}_k) \oplus \mathrm{cen}(\mathbf{x}_{k+1}).
\end{aligned}$$

□

**Věta 2.4.** *Nechť je dána náhodná kompozice* $\mathbf{x} \in \mathcal{S}^D$, *nenáhodná kompozice* $\mathbf{b} \in \mathcal{S}^D$ *a číslo* $a \in \mathbb{R}$.
*Potom*

$$\mathrm{cen}[(a \odot \mathbf{x}) \oplus \mathbf{b}] = [a \odot \mathrm{cen}(\mathbf{x})] \oplus \mathbf{b}.$$

**Důkaz:** Na základě věty 2.3 můžeme psát

$$\mathrm{cen}[(a \odot \mathbf{x}) \oplus \mathbf{b}] = \mathrm{cen}(a \odot \mathbf{x}) \oplus \mathrm{cen}(\mathbf{b}).$$

Pokud podle nahradíme náhodnou kompozici $\mathbf{x}$ vektorem $\mathbf{b}$, dostaneme $\mathrm{cen}(\mathbf{b}) = \mathbf{b}$. Tudíž

$$\mathrm{cen}[(a \odot \mathbf{x}) \oplus \mathbf{b}] = \mathrm{cen}(a \odot \mathbf{x}) \oplus \mathbf{b}.$$

Dále pomocí věty 2.3 a vlastností ilr transformace obdržíme

$$\mathrm{cen}(a \odot \mathbf{x}) = a \odot \mathrm{cen}(\mathbf{x}).$$

□

**Věta 2.5.** *Nechť jsou dány nezávislé náhodné kompozice* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$. *Potom*

$$\mathrm{Mvar}(\mathbf{x} \oplus \mathbf{y}) = \mathrm{Mvar}(\mathbf{x}) + \mathrm{Mvar}(\mathbf{y}).$$

**Důkaz:** viz [14], str. 273, věta 9. □

**Věta 2.6.** *Nechť jsou dány stejné předpoklady jako ve větě 2.6. Pak*

$$\mathrm{Mvar}[(a \odot \mathbf{x}) \oplus \mathbf{b}] = a^2 \mathrm{Mvar}(\mathbf{x}).$$

**Důkaz:** viz [14], str. 273, věta 10. □

Vzhledem k tomu, že metrický rozptyl je definován jako střední hodnota ze vzdálenosti centra a náhodné kompozice, můžeme základní číselné charakteristiky náhodné kompozice na simplexu geometricky interpretovat. Metrický rozptyl je invariantní vzhledem k perturbaci, což je ekvivalentní s posunutím náhodného vektoru, vzniklého ilr transformací kompozice.

Existenci analogie mezi vlastnostmi střední hodnoty náhodného vektoru a centra náhodné kompozice na simplexu nelze přehlédnout. Stejně tak jako vztah mezi vlastnostmi rozptylu náhodné veličiny a metrického rozptylu náhodné kompozice.

Pro provádění testů, uvedených v kapitolách 1.2 a 1.7, na souborech kompozičních dat je nutné, aby byl příslušný výběr kompozic z normálního rozdělení na simplexu. Jeho definice je intuitivní:

**Definice 2.14.** *(Normální rozdělení na* $\mathcal{S}^D$*)*
*Řekneme, že náhodná kompozice* $\mathbf{x} \in \mathcal{S}^D$ *má normální rozdělení na simplexu, jestliže ilr transformace kompozice* $\mathbf{x}$ *(tj.* $ilr(\mathbf{x}) = \mathbf{x}^* \in \mathbb{R}^{D-1}$*) má mnohorozměrné normální rozdělení na* $\mathbb{R}^{D-1}$.

Pro další vlastností normality na simplexu lze odkázat na [12]. Je tedy zřejmé, že normalita na simplexu je ekvivalentní standardnímu normálnímu rozdělení ilr transformovaných souřadnic.

Dále uvažujme soubor $n$ nezávislých náhodných kompozic ze stejného rozdělení jako náhodná kompozice $\mathbf{x}$. Tento soubor tvoří náhodný výběr $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$. Ukážeme si, co je v tomto případě analogií výběrového průměru jako BLUE střední hodnoty náhodného vektoru.

Předešleme již nyní, že tímto je kompozice, složená z geometrických průměrů složek kompozic ve výběru, která je nejlepším lineárním nestranným odhadem centra $\mathbf{x}$ ve smyslu Aitchisonovy geometrie. Proto na začátek uvedeme definice těchto vlastností odhadu, které jsou do určité míry analogické vlastnostem odhadů parametrů rozdělení v reálném prostoru.

Abychom tyto vlastnosti odhadů parametrů na simplexu specifikovali, budeme před nimi používat označení „c“ (compositional = kompoziční).

Dále uvažujeme kompoziční odhad $\widehat{\boldsymbol{\theta}} \in \mathcal{S}^D$ neznámého kompozičního parametru $\boldsymbol{\theta} \in \mathcal{S}^D$ z rozdělení náhodné kompozice $\mathbf{x}$ [14].

**Definice 2.15.** *$\widehat{\boldsymbol{\theta}}$ je c-nestranný kompoziční odhad parametru $\boldsymbol{\theta} \in \mathcal{S}^D$ právě tehdy, když* $\operatorname{cen}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$*, nebo ekvivalentně, když* $\operatorname{cen}(\widehat{\boldsymbol{\theta}} \ominus \boldsymbol{\theta}) = \mathbf{n}$*, kde $\mathbf{n}$ je neutrální prvek na simplexu.*

V následující definici označuje $\boldsymbol{\Theta}$ třídu všech c-nestranných odhadů parametru $\boldsymbol{\theta} \in \mathcal{S}^D$.

**Definice 2.16.** *Řekneme, že c-nestranný odhad $\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ je c-nejlepší odhad vzhledem k Aitchisonově metrice, jestliže navíc*

$$\operatorname{Mvar}(\widehat{\boldsymbol{\theta}}) < \operatorname{Mvar}(\widetilde{\boldsymbol{\theta}}),\ \widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta},$$

*tj. $\widehat{\boldsymbol{\theta}}$ je c-nestranný odhad a má nejmenší metrický rozptyl v $\boldsymbol{\Theta}$.*

Je zřejmé, že i další vlastnosti kompozičního odhadu mohou být zformulované analogicky nahrazením euklidovské metriky Aitchisonovou a střední hodnoty centrem.

Ostatní vlastnosti kompozičního odhadu nebudeme v této práci uvádět. Ukážeme nyní, že kompozice složená z geometrických průměrů složek kompozic ve výběru (tzv. výběrové centrum) je c-nejlepší c-nestranný c-lineární odhad centra náhodné kompozice.

Nejprve uvedeme definici linearity kompozičního odhadu na simplexu.

**Definice 2.17.** *Řekneme, že kompoziční odhad* $\widehat{\boldsymbol{\theta}}$ *parametru* $\boldsymbol{\theta}$ *je c-lineárním odhadem, jestliže*

$$\widehat{\boldsymbol{\theta}} = (\alpha_1 \odot \mathbf{x}_1) \oplus (\alpha_2 \odot \mathbf{x}_2) \oplus \cdots \oplus (\alpha_n \odot \mathbf{x}_n) = \bigoplus_{i=1}^{n} (\alpha_i \odot \mathbf{x}_i), \quad \alpha_1, \alpha_2, ..., \alpha_n \in \mathbb{R}. \tag{14}$$

Jestliže ve vztahu (14) zvolíme za $\alpha_i = \dfrac{1}{n}$, $i = 1, ..., n$, pak obdržíme právě výběrové centrum. S využitím první vlastnosti distributivity operace mocninné transformace můžeme psát

$$\bigoplus_{i=1}^{n} \left( \frac{1}{n} \odot \mathbf{x}_i \right) = \frac{1}{n} \odot \left( \bigoplus_{i=1}^{n} \mathbf{x}_i \right).$$

**Věta 2.7.** *Výběrové centrum dané vztahem*

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{n} \odot \left( \bigoplus_{i=1}^{n} \mathbf{x}_i \right)$$

*je c-lineární a c-nestranný odhad centra náhodné kompozice* $\mathbf{x}$*, tj.* $\mathrm{cen}(\mathbf{x})$.

**Důkaz:** Z definice 2.17 je patrné, že $\widehat{\boldsymbol{\gamma}}$ je c-lineární odhad.
Vektorová struktura na simplexu $(\mathcal{S}^D, \oplus, \odot)$ nám umožní, abychom mohli s využitím věty 2.3 a věty 2.4 dokázat, že $\widehat{\boldsymbol{\gamma}}$ je opravdu c-nestranným odhadem centra náhodné kompozice $\mathbf{x}$,

$$\mathrm{cen}(\widehat{\boldsymbol{\gamma}}) = \frac{1}{n} \odot \left( \bigoplus_{i=1}^{n} \mathrm{cen}(\mathbf{x}_i) \right) = \frac{1}{n} \odot [n \odot \mathrm{cen}(\mathbf{x})] = \mathrm{cen}(\mathbf{x}).$$

□

**Věta 2.8.** *Výběrové centrum $\widehat{\boldsymbol{\gamma}}$, které je c-lineárním a c-nestranným odhadem, je také c-nejlepším odhadem $\boldsymbol{\gamma} = \mathrm{cen}(\mathbf{x})$ a navíc* $\mathrm{Mvar}(\widehat{\boldsymbol{\gamma}}) = \frac{1}{n}\mathrm{Mvar}(\mathbf{x})$.

**Důkaz:** Uvažujme jiný libovolný odhad $\widetilde{\boldsymbol{\gamma}} = \bigoplus_{i=1}^{n} (\alpha_i \odot \mathbf{x}_i)$ centra $\boldsymbol{\gamma}$, který je také c-lineární a c-nestranný.

Pomocí vět 2.3 a 2.4 ověříme c-nestrannost odhadu $\widetilde{\boldsymbol{\gamma}}$; využitím druhé vlastnosti distributivity operace mocninná transformace na simplexu získáme

$$\mathrm{cen}(\widetilde{\boldsymbol{\gamma}}) = \bigoplus_{i=1}^{n} (\alpha_i \odot \mathrm{cen}(\mathbf{x}_i)) = \bigoplus_{i=1}^{n} (\alpha_i \odot \boldsymbol{\gamma}) = \left( \sum_{i=1}^{n} \alpha_i \right) \odot \boldsymbol{\gamma}.$$

Pro nestrannost odhadu $\widetilde{\boldsymbol{\gamma}}$ je nutné, aby $\sum_{i=1}^{n} \alpha_i = 1$.
Nyní vypočítáme metrický rozptyl odhadu $\widetilde{\boldsymbol{\gamma}}$ a určíme, kdy má tento odhad nejmenší metrický rozptyl. Na základě tohoto rozhodneme, zda je c-nejlepším odhadem $\boldsymbol{\gamma}$.

Užitím věty 2.5 a věty 2.6 spočítáme metrický rozptyl $\mathrm{Mvar}(\widetilde{\boldsymbol{\gamma}})$. Nesmíme totiž zapomenout, že pracujeme s náhodným výběrem. Obdržíme

$$\mathrm{Mvar}(\widetilde{\boldsymbol{\gamma}}) = \mathrm{Mvar}\left[ \bigoplus_{i=1}^{n} (\alpha_i \odot \mathbf{x}_i) \right] = \mathrm{Mvar}(\mathbf{x}) \sum_{i=1}^{n} \alpha_i^2.$$

$\mathrm{Mvar}(\widetilde{\boldsymbol{\gamma}})$ nabývá minima pro $\alpha_i = \frac{1}{n}$, $i = 1, ..., n$. Pro dosažení minimálního metrického rozptylu je tedy zapotřebí, aby $\widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}$ a $\mathrm{Mvar}(\widehat{\boldsymbol{\gamma}}) = \frac{1}{n}\mathrm{Mvar}(\mathbf{x})$.

□

Ve větách 2.7 a 2.8 jsme dokázali, že výběrové centrum je c-nejlepším c-lineárním a c-nestranným odhadem centra vzhledem k Aitchisonově geometrii na simplexu, tj. je c-BLUE.

Již dříve jsme se zmínili, že testy uvedené v kapitole 1.7 budeme aplikovat na ilr transformované kompozice. Z nich následně spočítáme výběrový průměr, který je nejlepším lineárním nestranným odhadem parametru $\boldsymbol{\mu}$.

Jak víme, testy o střední hodnotě $\boldsymbol{\mu}$ byly postaveny na předpokladu normality.

Abychom tedy mohli uvedené testy aplikovat, musíme po provedení ilr transformace kompozic ve výběru ověřit, zda mají normální rozdělení na $\mathbb{R}^{D-1}$. Toto již přesahuje cíle a zaměření naší práce.

Parametry normálního rozdělení v $\mathbb{R}^{D-1}$ jsou v praxi obvykle neznámé. Používáme proto jejich odhady, tj. pro odhad střední hodnoty $\boldsymbol{\mu}$ výběrový průměr $\overline{\mathbf{x}}$ a pro odhad varianční matice $\boldsymbol{\Sigma}$ výběrovou varianční matici $\mathbf{S} = \frac{1}{n-1}\mathbf{W}$, kde $\mathbf{W}$ je Wishartova matice.

## 2.5 Invariantnost testovacích statistik

Když testujeme (za předpokladu normality) hypotézu $H_0 : \text{cen}(\mathbf{x}) = \boldsymbol{\gamma}_0$ proti $H_A : \text{cen}(\mathbf{x}) \neq \boldsymbol{\gamma}_0$, prostřednictvím volby ortonormální báze v ilr transformaci zobrazíme náhodné kompozice ve výběru z $\mathcal{S}^D$ na náhodné vektory z $\mathbb{R}^{D-1}$ a testujeme tak ekvivalentní hypotézu $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ proti $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

Uvažujme ovšem případ, kdy souřadnice ilr transformované kompozice $\mathbf{x}$ vzniknou volbou různých ortonormálních bází na $\mathcal{S}^D$. Nechť tyto volby reprezentují vektory $\mathbf{x}^* = ilr_1(\mathbf{x})$ a $\mathbf{y}^* = ilr_2(\mathbf{x})$. Vztah mezi $\mathbf{x}^*$ a $\mathbf{y}^*$ lze potom vyjádřit jako lineární transformaci

$$\mathbf{y}^* = \mathbf{P}\mathbf{x}^*.$$

Matice přechodu $\mathbf{P}^{(D-1)\times(D-1)}$ je ortogonální a platí $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ neboli $\mathbf{P}^T = \mathbf{P}^{-1}$ [5].

Za předpokladu normality na simplexu má náhodný vektor $\mathbf{x}^*$ $(D-1)$-rozměrné normální rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$. Pro parametry $\boldsymbol{\nu}$ a $\boldsymbol{\Psi}$ $(D-1)$-rozměrného normálního rozdělení náhodného vektoru $\mathbf{y}^*$ tedy dostaneme $\boldsymbol{\nu} = \mathbf{P}\boldsymbol{\mu}$ a
$\boldsymbol{\Psi} = \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^T$.

Chceme dokázat, že nulová hypotéza $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ je ekvivalentní s nulovou hypotézu $H_0 : \boldsymbol{\nu} = \boldsymbol{\nu}_0$, kde $\boldsymbol{\nu}_0 = \mathbf{P}\boldsymbol{\mu}_0$.

**Věta 2.9.** *Nechť pro náhodný výběr* $\mathbf{x}_i^* \sim N_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, ..., n$, *se známou maticí* $\boldsymbol{\Sigma}$ *testujeme nulovou hypotézu* $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *pomocí statistiky*

$$V_1 = n(\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0),$$

*a dále nechť pro náhodný výběr* $\mathbf{y}_i^* \sim N_{D-1}(\boldsymbol{\nu}, \boldsymbol{\Psi})$, $i = 1, ..., n$, *se známou maticí* $\boldsymbol{\Psi}$ *testujeme nulovou hypotézu* $H_0 : \boldsymbol{\nu} = \boldsymbol{\nu}_0$ *pomocí statistiky*

$$V_2 = n(\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0)^T \boldsymbol{\Psi}^{-1} (\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0).$$

*Pak* $V_1 = V_2$.

**Důkaz:** Pro vztah mezi $\overline{\mathbf{x}}^*$ a $\overline{\mathbf{y}}^*$ zřejmě platí $\overline{\mathbf{y}}^* = \mathbf{P}\overline{\mathbf{x}}^*$.

Samotný důkaz invariantnosti uvedené statistiky je jednoduchý a vypadá následovně,

$$\begin{aligned} n(\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0)^T \boldsymbol{\Psi}^{-1} (\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0) &= n(\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0)^T \left(\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^T\right)^{-1} (\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0) = \\ &= n(\overline{\mathbf{x}}^{*T}\mathbf{P}^T - \boldsymbol{\mu}_0^T\mathbf{P}^T)(\mathbf{P}^T)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{P}^{-1}(\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0) = \\ &= n(\overline{\mathbf{x}}^{*T} - \boldsymbol{\mu}_0^T)\mathbf{P}^T\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^T\mathbf{P}(\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0) = \\ &= n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T (\boldsymbol{\Sigma})^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_0). \end{aligned}$$

□

**Věta 2.10.** *Nechť pro náhodný výběr* $\mathbf{x}_i^* \sim N_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, ..., n$, *s neznámou maticí* $\boldsymbol{\Sigma}$ *testujeme nulovou hypotézu* $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *pomocí Hotellingovy statistiky*

$$T_1^{2*} = n(n-1)(\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0)^T \mathbf{W}_1^{-1} (\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0),$$

*a dále nechť pro náhodný výběr* $\mathbf{y}_i^* \sim N_{D-1}(\boldsymbol{\nu}, \boldsymbol{\Psi})$, $i = 1, ..., n$, *s neznámou maticí* $\boldsymbol{\Psi}$ *testujeme nulovou hypotézu* $H_0 : \boldsymbol{\nu} = \boldsymbol{\nu}_0$ *pomocí Hotellingovy statistiky*

$$T_2^{2*} = n(n-1)(\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0)^T \mathbf{W}_2^{-1} (\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0).$$

*Pak* $T_1^{2*} = T_2^{2*}$.

**Důkaz:** Při důkazu uplatníme vztah mezi Wishartovými maticemi obou výběrů, tedy $\mathbf{W}_2 = \mathbf{P}\mathbf{W}_1\mathbf{P}^T$. Dále budeme postupovat obdobně jako v předcházející větě,

$$
\begin{aligned}
n(n-1)(\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0)^T \mathbf{W}_2^{-1} (\overline{\mathbf{y}}^* - \boldsymbol{\nu}_0) = \\
&= n(n-1)(\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0)^T \left(\mathbf{P}\mathbf{W}_1\mathbf{P}^T\right)^{-1} (\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0) = \\
&= n(n-1)(\overline{\mathbf{x}}^{*T}\mathbf{P}^T - \boldsymbol{\mu}_0^T\mathbf{P}^T)(\mathbf{P}^T)^{-1}\mathbf{W}_1^{-1}\mathbf{P}^{-1}(\mathbf{P}\overline{\mathbf{x}}^* - \mathbf{P}\boldsymbol{\mu}_0) = \\
&= n(n-1)(\overline{\mathbf{x}}^{*T} - \boldsymbol{\mu}_0^T)\mathbf{P}^T\mathbf{P}\mathbf{W}_1^{-1}\mathbf{P}^T\mathbf{P}(\overline{\mathbf{x}}^* - \boldsymbol{\mu}_0) = \\
&= n(n-1)(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T\mathbf{W}_1^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0).
\end{aligned}
$$

□

# 3 Praktické příklady

Teoretické poznatky týkající se testu střední hodnoty transformovaných kompozic nyní budeme demonstrovat na praktických příkladech. Pro jejich vyřešení použijeme statistický software R (`www.r-project.org`). Práce s kompozicemi v R-ku je umožněna pomocí knihovny `compositions`. Popíšeme také ve zkratce užívané příkazy.

**Příklad 3.1.** *Podle údajů na obalu má 350g balíček mražené zeleniny obsahovat* 50% *mrkve,* 30% *kukuřice a* 20% *hrášku. Tyto údaje tak vlastně představují hypotetické centrum náhodné kompozice* $\mathbf{x}$, $\mathrm{cen}(\mathbf{x}) = (50, 30, 20)^T$*; skutečné procentuální podíly ve výběru ovšem kolísají a předpokládáme, že jde o výběr z normálně rozdělené třísložkové kompozice. Chceme ověřit údaje na obalu na základě výsledků přesného zjišťování hmotnosti u 11 náhodně vybraných balíčků mražené zeleniny. Hmotnosti jsou uvedeny v tabulce 1.*

***Tabulka 1***

| | *mrkev* | *kukuřice* | *hrášek* |
|---|---|---|---|
| *1* | *164.62* | *98.26* | *77.46* |
| *2* | *173.43* | *103.45* | *69.45* |
| *3* | *172.48* | *103.64* | *70.34* |
| *4* | *165.76* | *104.01* | *73.65* |
| *5* | *177.32* | *96.49* | *68.89* |
| *6* | *171.98* | *100.79* | *70.98* |
| *7* | *165.54* | *98.78* | *76.93* |
| *8* | *167.20* | *97.91* | *77.96* |
| *9* | *170.67* | *102.11* | *68.35* |
| *10* | *164.08* | *98.45* | *77.59* |
| *11* | *163.87* | *98.84* | *78.12* |

**Řešení:** V R-ku si nejprve nastavíme pracovní adresář pomocí funkce `setwd` a následně „zavoláme“ knihovnu, ve které pracujeme s kompozicemi, příkazem

```
> library(compositions)
```

V prvním kroku načteme kompozice z datového souboru `zelenina1.txt` do matice `A`,

```
> A=matrix(scan("zelenina1.txt"),ncol=3,byrow=T)
Read 33 items
```

Sloupce matice `A` ještě pojmenujeme pomocí `colnames`.

```
> colnames(A)=c("mrkev","kukuřice","hrášek").
> A
      mrkev kukuřice hrášek
 [1,] 164.62   98.26   77.46
 [2,] 173.43  103.45   69.45
 [3,] 172.48  103.64   70.34
 [4,] 165.76  104.01   73.65
 [5,] 177.32   96.49   68.89
 [6,] 171.98  100.79   70.98
 [7,] 165.54   98.78   76.93
 [8,] 167.20   97.91   77.96
 [9,] 170.67  102.11   68.35
[10,] 164.08   98.45   77.59
[11,] 163.87   98.84   78.12
```

Dále kompozice uzavřeme, tj. 3-složkové kompozice vyjádříme jako procentuální podíly na celku (celý obsah sáčku mražené zeleniny). To nám umožní příkaz `acomp`.

```
> P=acomp(A,total=100)
```

```
> P
      mrkev  kukuřice  hrášek
 [1,] 48.36928 28.87113 22.75959
 [2,] 50.07652 29.87035 20.05313
 [3,] 49.78352 29.91399 20.30249
 [4,] 48.26743 30.28653 21.44604
 [5,] 51.74205 28.15582 20.10213
 [6,] 50.03055 29.32073 20.64873
 [7,] 48.50989 28.94652 22.54359
 [8,] 48.73641 28.53937 22.72423
 [9,] 50.03078 29.93287 20.03635
[10,] 48.24180 28.94567 22.81254
[11,] 48.07969 28.99979 22.92052
```

Také vytvoříme funkci, kterou následně provedeme ilr transformací kompozic dle kapitoly 2.3,

```
> ilr1=function(w1,w2,w3)
+ dat=cbind((1/sqrt(2))*log(w1/w2),(2/sqrt(6))*log(sqrt(w1*w2)/w3))
+ return(dat).
```

Nyní provedeme ilr transformaci kompozic z matice `A` prostřednictvím funkce `ilr1` následovně,

```
> X=ilr1(P[,1],P[,2],P[,3])
```

```
> X
          [,1]      [,2]
 [1,] 0.3648832 0.4048735
 [2,] 0.3653519 0.5362947
 [3,] 0.3601704 0.5244045
 [4,] 0.3295499 0.4720900
 [5,] 0.4302862 0.5235267
 [6,] 0.3778348 0.5044402
 [7,] 0.3650917 0.4149093
 [8,] 0.3784025 0.4045119
 [9,] 0.3632275 0.5374586
[10,] 0.3611939 0.4029516
[11,] 0.3574927 0.3984845
```

Poznamenejme, že alternativně by bylo možné využít též přednastavené funkce `ilr` z knihovny `compositions`.

```
> X=-ilr(P)
```

Testujeme nulovou hypotézu $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$ naproti alternativě $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$ pomocí testovací statistiky dané vztahem (7) na hladině testu $\alpha = 0.05$, kde $n = 11$ je počet pozorovaní, $p = 2$ je dimenze ilr transformovaných kompozic a $T^{2*}$ je Hotellingova statistika, daná vztahem (6).

V softwaru R, jako v mnoha jiných, se nemohou zadávat určité znaky jako např. $\overline{\mathbf{x}}$, $\boldsymbol{\mu}$, $\boldsymbol{\mu_0}$, $T^{2*}$, atd. Proto budeme používat pro výběrový průměr $\overline{\mathbf{x}}$ označení `v,` pro Hotellingovu statistiku $T^{2*}$ pouze `T` a hypotetickou (cílovou) střední hodnotu $\boldsymbol{\mu}_0$ označíme `mi0.`

Ilr transformací hypotetického centra $\boldsymbol{\gamma}_0$ náhodné kompozice (označeno jako `y0`) dostaneme cílovou střední hodnotu $\boldsymbol{\mu}_0$.

```
> y0=c(50, 30, 20)
```

```
> y0
[1] 50 30 20
> mi0=c(ilr(mi0[1],mi0[2],mi0[3])
> mi0
[1] 0.3612083 0.5396046
```

Pro získaní hodnoty Hotellingovy statistiky potřebujeme vypočítat výběrový průměr a inverzi Wishartovy matice. Nejprve určíme výběrový průměr `v` ilr transformovaných kompozic z matice `X` využitím příkazu `colMeans`.

```
> v=c(colMeans(X, na.rm = FALSE, dims = 1))
> v
[1] 0.3684986 0.4658132
```

Wishartovu matici `W` obdržíme takto:

```
> W=t(X)%*%X-11*v%*%t(v).
> W
          [,1]       [,2]
[1,] 182.51655 −17.69447
[2,] −17.69447  28.78388
```

Potřebnou inverzi Wishartovy matice (označenou jako `Winv`) dostaneme využitím příkazu `solve`.

```
> Winv=solve(W)
> Winv
          [,1]       [,2]
[1,] 182.51655 −17.69447
[2,] −17.69447  28.78388
```

Poté, co jsme provedli všechny potřebné mezivýpočty, obdržíme hodnotu Hotellingovy statistiky dosazením do vztahu (6).

```
> b=v-mi0
> T=10*11*t(b)%*%Winv%*%b
```

```
> T
        [,1]
[1,] 20.40187
```

Hodnota testovací statistiky, potřebné pro úsudek o střední hodnotě, je spočítána dosazením do vztahu (7).

```
> (9/2)*(T/10)
        [,1]
[1,] 8.82272
```

Funkce `qf` poskytne hodnotu $(1-\alpha)$-kvantilu Fisherova rozdělení o $(2, 9)$ stupních volnosti.

```
> kvantilF=qf(0.95, 2, 9, lower.tail = TRUE, log.p = FALSE)
> kvantilF
[1] 4.256495
```

Na základě použitého jednovýběrového testu na hladině testu 0,05 zamítáme nulovou hypotézu ve prospěch alternativy, jelikož se testovací statistika realizuje v kritickém oboru, který je tvořen hodnotami přesahujícími právě $(1-\alpha)$ kvantil Fisherova rozdělení o $(2, 9)$ stupních volnosti, tj.

$$8.82272 > 4.256495.$$

Znamená to, že námi zvolený výrobek mražené zeleniny má nevyhovující deklarovaný obsah, tj. složení uvedené na obalu není ze strany výrobce dodržováno.

Zjistíme ještě příslušnou P-hodnotu, tj. nejnižší hladinu na které bychom zamítly nulovou hypotézu. Pomocí funkce `pf` obdržíme

```
> 1-pf(8.82272, 2, 9, lower.tail = TRUE, log.p = FALSE)
[1] 0.00756464.
```

**Příklad 3.2.** *(dle [2], str. 382) Továrna dostala* 27 *nových přístrojů stejné značky. Z předchozích zkušeností známe údaje o chování stavů přístroje při užívání v osmihodinové směně, tj. každý přístroj by v ní měl ze* 60% *produkovat vysokou kvalitu výrobku (I.), z* 20% *nízkou kvalitu (II.),* 10% *je potřeba na seřizování (III.) a stejně tolik času je potřeba počítat s opravami (IV.). Tyto údaje se týkají centra, předpokládáme, že jde o normálně rozdělenou čtyřsložkovou kompozici. Skutečné proporce chování stavů přístroje ve výběru tak vykazují určité odchylky.*

*Chceme zjistit, zda tyto odchylky jsou pro továrnu akceptovatelné, nebo jinak řečeno, zda přístroje jsou produktivní tak, jak by měly být. Údaje o proporcích chování stavů* 27 *přístrojů jsou uvedeny v tabulce 2.*

**Řešení:** Pro vyřešení tohoto příkladu budeme používat stejné označení a příkazy ze statistického softwaru R jako v příkladu 3.1.

Nejprve načteme údaje do matice `A` ze souboru `pristroje.txt`,

```
> A=matrix(scan("pristroje.txt"),ncol=4,byrow=T)
Read 108 items
```

Kompozice jsou již v uzavřené formě, proto je jen zapotřebí provést jejich ilr transformaci. Výsledek následně zobrazíme v matici `X`.

**Tabulka 2**

| Přístroj | I. | II. | III. | IV. |
|---|---|---|---|---|
| O1 | 0.667 | 0.180 | 0.053 | 0.100 |
| O2 | 0.578 | 0.180 | 0.112 | 0.123 |
| O3 | 0.560 | 0.271 | 0.086 | 0.076 |
| O4 | 0.490 | 0.316 | 0.091 | 0.103 |
| O5 | 0.598 | 0.119 | 0.117 | 0.116 |
| O6 | 0.617 | 0.180 | 0.079 | 0.124 |
| O7 | 0.700 | 0.197 | 0.046 | 0.057 |
| O8 | 0.577 | 0.266 | 0.800 | 0.077 |
| O9 | 0.591 | 0.179 | 0.084 | 0.145 |
| O10 | 0.532 | 0.107 | 0.119 | 0.242 |
| O11 | 0.511 | 0.303 | 0.082 | 0.104 |
| O12 | 0.625 | 0.219 | 0.082 | 0.074 |
| O13 | 0.667 | 0.115 | 0.107 | 0.111 |
| O14 | 0.573 | 0.208 | 0.112 | 0.107 |
| O15 | 0.585 | 0.235 | 0.080 | 0.092 |
| O16 | 0.558 | 0.245 | 0.100 | 0.096 |
| O17 | 0.652 | 0.214 | 0.066 | 0.068 |
| O18 | 0.619 | 0.214 | 0.066 | 0.068 |
| O19 | 0.628 | 0.245 | 0.067 | 0.061 |
| O20 | 0.596 | 0.228 | 0.090 | 0.086 |
| O21 | 0.546 | 0.185 | 0.114 | 0.155 |
| O22 | 0.606 | 0.146 | 0.079 | 0.169 |
| O23 | 0.613 | 0.257 | 0.058 | 0.072 |
| O24 | 0.680 | 0.173 | 0.060 | 0.088 |
| O25 | 0.584 | 0.246 | 0.072 | 0.098 |
| 026 | 0.542 | 0.122 | 0.159 | 0.176 |
| O27 | 0.545 | 0.121 | 0.156 | 0.178 |

> X=-ilr(A)

> X

| | [, 1] | [, 2] | [, 3] |
|---|---|---|---|
| [1, ] | 0.9261919 | 1.5330389 | 0.5342015 |
| [2, ] | 0.8249228 | 0.8636627 | 0.5295678 |
| [3, ] | 0.5132308 | 1.2334655 | 0.9792446 |
| [4, ] | 0.3101817 | 1.1955260 | 0.7380904 |
| [5, ] | 1.1416007 | 0.6729428 | 0.4832761 |
| [6, ] | 0.8710935 | 1.1753183 | 0.4406421 |
| [7, ] | 0.8965241 | 1.7052536 | 1.0201120 |
| [8, ] | 0.5475453 | 0.5829316 | 1.6150029 |
| [9, ] | 0.8445897 | 1.1053601 | 0.3088297 |
| [10, ] | 1.1340682 | 0.5679654 | 0.2131050 |
| [11, ] | 0.3695600 | 1.2805377 | 0.6996472 |
| [12, ] | 0.7415287 | 1.2302092 | 0.9587904 |
| [13, ] | 1.2429933 | 0.7765146 | 0.5172944 |
| [14, ] | 0.7165450 | 0.9191408 | 0.6894821 |
| [15, ] | 0.6449000 | 1.2521563 | 0.7643708 |
| [16, ] | 0.5820201 | 1.0676823 | 0.7903183 |
| [17, ] | 0.7877654 | 1.4152789 | 0.9748999 |
| [18, ] | 0.7510388 | 1.3940748 | 0.9599063 |
| [19, ] | 0.6655869 | 1.4429181 | 1.1015466 |
| [20, ] | 0.6794554 | 1.1512467 | 0.8534259 |
| [21, ] | 0.7652756 | 0.8371449 | 0.3258848 |
| [22, ] | 1.0064062 | 1.0825075 | 0.1068786 |
| [23, ] | 0.6146800 | 1.5703495 | 0.9231501 |
| [24, ] | 0.9678886 | 1.4234374 | 0.6748412 |
| [25, ] | 0.6113429 | 1.3561601 | 0.6919532 |
| [26, ] | 1.0544694 | 0.3925220 | 0.1895843 |
| [27, ] | 1.0641923 | 0.4069682 | 0.1735173 |

Tak jako v předchozím příkladu budeme testovat nulovou hypotézu proti alternativě pomocí testovací statistiky dané vztahem (7); nyní $n = 27$ představuje počet přístrojů (rozsah souboru), $p = 3$ je dimenze ilr transformovaných kompozic.

Hypotetické centrum kompozice je představeno v uzavřené formě, kde $k = 1$.

```
> y0
[1] 0.6 0.2 0.1 0.1
```

Provedeme ilr transformaci centra kompozice `y0` a dostaneme cílovou střední hodnotu `mi0`.

```
> mi0
[1] 0.7768362 1.0144589 0.7173308
```

Dopočítáme výběrový průměr `v`,

```
> v
[1] 0.7879851 1.0543871 0.6604205,
```

který potřebujeme pro výpočet Wishartovy matice `W`,

```
> W
          [,1]       [,2]       [,3]
[1,]  1.4182963 −0.6252564 −1.3875357
[2,] −0.6252564  5.8427449  0.5108601
[3,] −1.3875357  0.5108601  3.6550039
```

Zajímá nás také její inverze `Winv`,

```
> Winv
          [,1]        [,2]        [,3]
[1,] 1.16371547 0.086970115 0.429621198
[2,] 0.08697012 0.179769626 0.007889734
[3,] 0.42962120 0.007889734 0.435590282
```

Máme potřebné mezivýpočty, můžeme je tedy dosadit do vztahu (6) a získat tím Hotellingovu statistiku `T`.

```
> T
          [,1]
[1,] 0.9395733.
```

Pomocí této statistiky vyjádříme testovací statistiku danou vztahem (7), tj.

```
> (24/3)*(T/26)
          [,1]
[1,] 0.2890995.
```

Dále zjistíme $(1-\alpha)$-kvantil Fisherova rozdělení o $(3, 24)$ stupních volnosti, pomocí kterého určíme kritický obor,

```
> kvantilF=qf(0.95, 3, 24, lower.tail = TRUE, log.p = FALSE)
> kvantilF
[1] 3.008787
```

Hodnota testovací statistiky nespadá do kritického oboru, jelikož

$$0.2890995 < 3.008787,$$

tzn. že nulovou hypotézu na dané hladině nelze zamítnout (příslušná P-hodnota je rovna `0.832807`). Produktivita strojů neodpozuje předpokladům.

# Závěr

V této práci jsem se nejprve zabývala jednovýběrovými testy o střední hodnotě náhodného vektoru z hlediska standardní mnohorozměrné statistické analýzy. Přitom jsem uvažovala dvě možné situace. První, kdy je varianční matice známá a druhou, kdy je neznámá, což je v praxi zřejmě běžnější. Na této kapitole jsem chtěla především vysvětlit některé aspekty práce s vícerozměrnými daty. Ve druhé kapitole jsem pak ukázala, že kompoziční data lze logratio transformacemi převést na standardní pozorování a na ně následně aplikovat výše uvedené testy. Vše jsem, doufám, názorně využila v poslední kapitole. Ta se zabývala konkrétními příklady z různých odvětví, ukazující na možné aplikace uvedené teorie.

Při psaní této práce jsem získala nové znalosti o kompozičních datech jako specifickém druhu mnohorozměrných pozorování a také jsem si prohloubila své znalosti z mnohorozměrné statistické analýzy. Kupodivu největším problémem a překážkou byl pro mě český jazyk.

Doufám, že tato práce bude přínosem nejen pro mě, ale i pro další zájemce o statistickou analýzu kompozičních dat a mnohorozměrnou statistickou analýzu vůbec.

# Literatura


[1] Aitchison, J., *The statistical analysis of compositional data*, London: Chapman and Hall, 1986.

[2] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., *Logratio analysis and compositional distance*, Mathematical Geology 32 (3), 271-275, (2000).

[3] Anderson, T.W., *angl. An introduction to multivariate statistical analysis (ruský překlad: Vvedenije v mnogomernyj statističeskij analiz)*. Moskva, 1963.

[4] Anděl, J. *Základy matematické statistiky*, 1.vydání. Praha: Matfyzpress, 2005.

[5] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., *Isometric logratio transformations for compositional data analysis*, Mathematical Geology 35, 279-300, (2003).

[6] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., *Reply to "On the Harker Variation Diagrams; . . ." by J.A. Cortés*, [online], dostupné z: http://www.springerlink.com/content/x08r2624vg95710g/fulltext.pdf, [citováno 13. 02. 2010].

[7] Filzmoser, P., Hron, K., *Outlier detection for compositional data using robust methods*, Mathematical Geosciences 40, 233-248, (2008).

[8] Filzmoser, P., Hron, K., Reimann, C., *Principal component analysis for compositional data with outliers*, Environmetrics 20 (6), 621-632, (2009).

[9] Giri, N.C., *Multivariate statistical analysis*, 2. vydání. New York, Basel: Marcel Dekker, 2004.

[10] Hebák, P., a kol. *Vícerozměrné statistické metody*, 1. vydání. Praha: Informatorium, 2004.

[11] *Hilbertův prostor*, [online], dostupné z: http://www.karlin.mff.cuni.cz/jvesely/fourier/kapi05.pdf [citováno 13. 02. 2010].

[12] Mateu-Figueras, G., Pawlowsky-Glahn, V., *A critical approach to probability laws in geochemistry*, Mathematical Geosciences 40 (5), 489-502, (2008).

[13] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, J., *Lecture notes on compositional data analysis*, [online], dostupné z: http://hdl.handle.net/10256/297, [citováno 08. 01. 2010].

[14] Pawlowsky-Glahn, V., Egozcue, J.J., *BLU estimators and compositional data*, Mathematical Geology, 34, 259-274, (2002).

[15] Rao, C.R., *Lineární metody statistické indukce a jejich aplikace*, 1. vydání. Praha: Academia, 1978.

[16] Rencher, A.C., *Methods of multivariate analysis*, 2. vydání. New York, Chichester: Wiley, 2000.