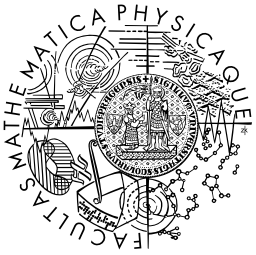


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

**BAKALÁŘSKÁ PRÁCE**

Iva Hamerníková

# Analýza rozptylu s náhodnými efekty

Katedra pravděpodobnosti a matematické statistiky

| | |
|---|---|
| Vedoucí bakalářské práce: | doc. RNDr. Arnošt Komárek, Ph.D. |
| Studijní program: | Matematika |
| Studijní obor: | Obecná matematika |

Praha 2018

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

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

V ..... dne ..... Podpis autora

Chtěla bych poděkovat vedoucímu práce doc. RNDr. Arnoštu Komárkovi, Ph.D. za veškeré rady, konzultace a trpělivost při vedení práce.

Název práce: Analýza rozptylu s náhodnými efekty

Autor: Iva Hamerníková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá popisem a odvozením metody analýzy rozptylu s náhodnými efekty. Nejprve uvedeme souhrn poznatků z teorie pravděpodobnosti, které budou důležité v dalším odvozování. Poté zavedeme model jednoduchého třídění s pevnými efekty a navrhneme testovou statistiku pro test shody středních hodnot skupin. V další části zavedeme model jednoduchého třídění s náhodnými efekty a odvodíme vlastnosti pozorování v tomto modelu. Za předpokladu vyváženého třídění definujeme součty čtverců a odvodíme jejich vlastnosti, díky kterým je pak můžeme použít k sestavení testové statistiky pro testování shody podmíněných středních hodnot skupin. Na závěr práce budeme pomocí simulací v programu R ověřovat, jak test analýzy rozptylu s náhodnými efekty dodržuje hladinu při porušení předpokladu normality.

Klíčová slova: analýza rozptylu, ANOVA, náhodné efekty

Title: Analysis of Variance with Random Effects

Author: Iva Hamerníková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of this thesis is to describe and derive the test of analysis of variance with random effects. At first we introduce a summary of results from the theory of probability which will be important in future derivations. Then we define the one-way classification model with fixed effects and propose the test statistics to test the equality of group means. In the following part we define the one-way classification model with random effects and derive properties of observations in this model. Under the assumption of balanced data we define sums of squares and derive their properties, which allow us to use them to create the test statistic. Finally we will use simulations in R to verify whether the ANOVA test with random effects observes the significance level when normality assumptions are violated.

Keywords: analysis of variance, ANOVA, random effects

# Obsah

**Seznam použitých zkratek** **2**

**Úvod** **3**

**1 Souhrn z teorie pravděpodobnosti** **4**
1.1 Podmíněná střední hodnota a rozptyl . . . . . 4
1.2 $\chi^2$ rozdělení a kvadratické formy . . . . . 5

**2 Analýza rozptylu s pevnými efekty** **7**
2.1 Jednoduché třídění . . . . . 7

**3 Analýza rozptylu s náhodnými efekty** **11**
3.1 Jednoduché třídění - zavedení modelu . . . . . 11
3.2 Odvození vlastností . . . . . 12
3.3 Součty čtverců . . . . . 14
3.4 Testování pro analýzu rozptylu s náhodnými efekty . . . . . 16

**4 Simulace** **17**

**Závěr** **21**

**Seznam použité literatury** **22**

# Seznam použitých zkratek

| | |
|---|---|
| $\mathbb{I}_k$ | jednotková matice o rozměrech $k \times k$ |
| $\mathbf{1}_k$ | sloupcový vektor tvořený $k$ prvky 1 |
| $\mathsf{diag}(a_1, \ldots, a_n)$ | diagonální matice s prvky $a_1, \ldots, a_n$ na diagonále |
| $\mathsf{rank}(\mathbb{A})$ | hodnost matice $\mathbb{A}$ |
| $\mathsf{tr}(\mathbb{A})$ | stopa matice $\mathbb{A}$ |
| $\mathrm{Var}(\boldsymbol{X})$ | varianční matice pro náhodný vektor $\boldsymbol{X}$ |

# Úvod

Analýza rozptylu, z anglického analysis of variance často zkracovaná jako ANOVA, je hojně používaná metoda pro porovnávání skupinových výběrů, respektive testování nulové hypotézy o rovnosti středních hodnot těchto skupin. Princip této metody spočívá v myšlence, že rozdělíme celkovou variabilitu pozorovaných dat na variabilitu mezi jednotlivými skupinami a na variabilitu uvnitř skupin. Tyto dvě veličiny pak využíváme k sestrojení testové statistiky, na základě které rozhodujeme o platnosti nulové hypotézy. Cílem této práce je podrobné odvození speciálního typu, analýzy rozptylu s náhodnými efekty.

V první kapitole shrneme užitečná tvrzení a vztahy z teorie pravděpodobnosti, jejichž výsledky pak budeme dále v práci používat při výpočtech a odvozeních. Nejprve zavedeme podmíněnou střední hodnotu a rozptyl, uvedeme vlastnosti střední hodnoty a větu o rozkladu nepodmíněného rozptylu. V další části pak připomeneme definci $\chi^2$ rozdělení a zaměříme se na tvrzení o rozdělení a nezávislosti kvadratických forem.

V druhé kapitole zavedeme a okomentujeme standardní model jednoduchého třídění s pevnými efekty. Definujeme základní značení a odvodíme vlastnosti součtů čtverců, na základě kterých navrhneme testovou statistiku a kritický obor pro test shody středních hodnot skupin.

V další části následuje zavedení modelu s náhodnými efekty na příkladu s populací myší. Podrobně odvodíme vztahy pro podmíněnou střední hodnotu a rozptyl v tomto rozdělení, porovnáme je s předchozím případem. Za předpokladu vyváženého třídění definujeme součty čtverců a odvodíme jejich vlastnosti, konkrétně střední hodnotu a za dodatečných předpokladů i jejich rozdělení. V závěru kapitoly využijeme předchozích výsledků k navržení testové statistiky pro test shody podmíněných středních hodnot skupin.

Na závěr práce budeme pomocí simulací v programu R zkoumat co se stane při porušení základního předpokladu normality pozorování. Provedeme simulace s různým počtem skupin a jiným než normálním rozdělením a budeme sledovat, zda test dodržuje hladinu.

# 1. Souhrn z teorie pravděpodobnosti

Při odvozování vlastností pozorování a testových statistik v dalších kapitolách budeme často používat různé závěry z teorie pravděpodobnosti, které pro přehlednost uvedeme v této kapitole.

## 1.1 Podmíněná střední hodnota a rozptyl

Nejprve shrneme základní poznatky o podmíněné střední hodnotě z teorie pravděpodobnosti. Věty a tvrzení v této části jsou převzaty z knih Lachout (2004) a Anděl (2007a).

K zavedení podmíněné střední hodnoty budeme potřebovat nejprve definovat podmíněnou hustotu.

**Definice 1** (Podmíněná hustota)**.** *Nechť X,Y jsou reálné náhodné veličiny, dále nechť existuje spojité rozdělení (X,Y) s hustotou f(x,y), absolutně spojitou vůči Lebesgueově míře, nechť existuje f(y), marginální hustota Y. Jako podmíněnou hustotu x při y označíme takovou funkci $f(x|y)$, která pro všechna $A, B \in \mathcal{B}$, kde $\mathcal{B}$ je borelovská $\sigma$-algebra, splňuje:*

$$P(X \in A, Y \in B) = \int_A \int_B f(x,y)dydx = \int_A \int_B f(x|y)f(y)dydx.$$

Podmíněnou hustotu také můžeme pro s.v. $x$ a $y$ taková, že $f(y) \neq 0$, zapisovat ve tvaru $f(x|y) = \frac{f(x,y)}{f(y)}$.

Nyní můžeme zavést podmíněnou střední hodnotu.

**Definice 2** (Podmíněná střední hodnota)**.** *Podmíněnou střední hodnotu veličiny X při daném Y definujeme předpisem*

$$\mathsf{E}[X|Y] = \int_{\mathbb{R}} xf(x|y)dx$$

.

V následující větě shrneme některé vlastnosti podmíněné střední hodnoty, které budeme později využívat při výpočtech.

**Věta 1** (Základní vlastnosti podmíněné střední hodnoty)**.** *Buďte náhodné veličiny X, Y, $Z \in L_1(\Omega, \mathcal{A}, P)$. Pak:*
*i)* $\forall a, b, c \in \mathbb{R} : \mathsf{E}[aX + bY + c|Z] = a\,\mathsf{E}[X|Z] + b\,\mathsf{E}[Y|Z] + c$ *s.j.*
*ii)* $X \leq Y$ *s.j.* $\implies \mathsf{E}[X|Z] \leq \mathsf{E}[Y|Z]$ *s.j.*
*iii)* $\mathsf{E}[\mathsf{E}[X|Y]] = \mathsf{E}[X]$
*iv) Je-li* $\sigma(X) \subset \sigma(Z)$ *pak* $\mathsf{E}[X|Z] = X$ *s.j.*
*v) Buď* $Y \in L(\Omega, \mathcal{F})$ *a* $XY \in L_1(\Omega, \mathcal{A}, P)$*, pak* $\mathsf{E}[XY|\mathcal{F}] = Y\,\mathsf{E}[X|\mathcal{F}]$
*vi) Jsou-li X a Y nezávislé veličiny, pak* $\mathsf{E}[X|Y] = \mathsf{E}[X]$ *s.j.*

Důkaz lze najít např. v (Lachout, 2004).

Dále nás bude zajímat podmíněný rozptyl. Podobně jako je nepodmíněný rozptyl definován pomocí střední hodnoty, využijeme k definici podmíněného rozptylu podmíněnou střední hodnotu.

**Definice 3** (Podmíněný rozptyl)**.** *Nechť S(Y,Z) je náhodná veličina a* $\mathsf{E}[S^2] < \infty$. *Pak podmíněný rozptyl definujeme předpisem* $\mathsf{var}(S|Y) = \mathsf{E}([S - \mathsf{E}(S|Y)]^2|Y)$.

Nyní uvedeme důležitou větu o rozkladu nepodmíněného rozptylu.

**Věta 2.** *Pokud* $\mathsf{E}[S^2] < \infty$, *pak* $\mathsf{var}(S) = \mathsf{var}[E[S|Y]] + \mathsf{E}[\mathsf{var}(S|Y)]$.

*Důkaz.*

V důkazu využijeme vlastnosti podmíněné střední hodnoty z věty 1.
Nejprve rozepíšeme podmíněný rozptyl do vhodnějšího tvaru.

$$\mathsf{var}(S|Y) = \mathsf{E}([S - \mathsf{E}(S|Y)]^2|Y) = \mathsf{E}([S - \mathsf{E}[S] + \mathsf{E}[S] - \mathsf{E}(S|Y)]^2|Y) =$$

$$= \mathsf{E}\{[(S - \mathsf{E}[S])^2 + 2(S - \mathsf{E}[S])(\mathsf{E}[S] - \mathsf{E}(S|Y)]) + (\mathsf{E}[S] - \mathsf{E}(S|Y)])^2|Y\} =$$

$$= \mathsf{E}[(S - \mathsf{E}[S])^2|Y] + 2[\mathsf{E}[S] - \mathsf{E}(S|Y)]\,\mathsf{E}[(S - \mathsf{E}[S])|Y] + (\mathsf{E}[S] - \mathsf{E}(S|Y))^2 =$$

$$= \mathsf{E}[(S - \mathsf{E}[S])^2|Y] + 2(\mathsf{E}[S] - \mathsf{E}(S|Y))(\mathsf{E}(S|Y) - \mathsf{E}[S]) + (\mathsf{E}[S] - \mathsf{E}(S|Y))^2 =$$

$$= \mathsf{E}[(S - \mathsf{E}[S])^2|Y] - (\mathsf{E}[S] - \mathsf{E}(S|Y))^2$$

Nyní aplikujeme na výsledek operátor střední hodnoty.

$$\mathsf{E}[\mathsf{var}(S|Y)] = \mathsf{E}[\mathsf{E}[(S - \mathsf{E}[S])^2|Y] - (\mathsf{E}[S] - \mathsf{E}(S|Y))^2] =$$

$$= \mathsf{E}[\mathsf{E}[(S - \mathsf{E}[S])^2|Y]] - \mathsf{E}(\mathsf{E}[S] - \mathsf{E}(S|Y))^2 =$$

$$= \mathsf{E}(S - \mathsf{E}[S])^2 - \mathsf{E}(\mathsf{E}(S|Y) - \mathsf{E}[S])^2 =$$

$$= \mathsf{var}(S) - \mathsf{E}(\mathsf{E}(S|Y) - \mathsf{E}[\mathsf{E}(S|Y)])^2 = \mathsf{var}(S) - \mathsf{var}(S|Y)$$

Dohromady tedy získáváme vztah $\mathsf{E}[\mathsf{var}(S|Y)] = \mathsf{var}(S) - \mathsf{var}(S|Y)$,
ze kterého plyne dokazovaný vzorec.

□

## 1.2 $\chi^2$ rozdělení a kvadratické formy

Nejprve připomeňme definici $\chi^2$ rozdělení.

**Definice 4** ($\chi^2$ rozdělení)**.** *Nechť* $X_1, \dots X_n$ *jsou nezávislé náhodné veličiny s normovaným normálním rozdělením N(0,1). Pak náhodná veličina* $Y = \sum_{i=1}^{n} X_i^2$ *má* $\chi^2$ *rozdělení s n stupni volnosti, píšeme* $Y \sim \chi_n^2$.

V následujících kapitolách budeme často potřebovat zjistit rozdělení náhodných veličin ve tvaru kvadratických forem. Shrňme si proto několik užitečných vlastností, které pro ně platí, v následující větě.

**Věta 3.** *i) Nechť* $Y$, $Z$ *jsou nezávislé náhodné veličiny takové, že* $Y \sim \chi_{n_1}^2$, $Z \sim \chi_{n_2}^2$. *Pak náhodná veličina* $Y + Z \sim \chi_{n_1+n_2}^2$.
*ii) Nechť* $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, *kde* $\Sigma$ *je regulární varianční matice. Pak* $\boldsymbol{Y} = (\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_n^2$.
*iii) Nechť* $\boldsymbol{X} \sim N_n(\boldsymbol{0}, \Sigma)$ *a* $\mathbb{A}$ *buď čtvercová matice taková, že* $\mathbb{A}\Sigma$ *je nenulová idempotentní matice. Pak* $\boldsymbol{Y} = \boldsymbol{X}^T \mathbb{A} \boldsymbol{X} \sim \chi_{\mathsf{tr}\mathbb{A}\Sigma}^2$.

*Důkaz.*

i) Veličiny Y a Z lze přepsat jako $Y = \sum_{i=1}^{n_1} X_i^2$, $Z = \sum_{i=n_1+1}^{n_2} X_i^2$, kde $X_i$ jsou vzájemně nezávislé veličiny s N(0,1) rozdělením. Pak z definice vyplývá, že

$$Y + Z = \sum_{i=1}^{n_1+n_2} X_i^2 \sim \chi^2_{n_1+n_2}.$$

ii) Použijeme vlastnost, že $\boldsymbol{X}$ má stejné rozdělení, jako $\boldsymbol{\mu}+\mathbb{B}\boldsymbol{X}_0$, kde $\mathbb{B}\mathbb{B}^{\mathbb{T}} = \Sigma$ a $\boldsymbol{X}_0 \sim N_n(\boldsymbol{0}, \mathbb{I}_n)$. Z toho plyne, že $\boldsymbol{Y}$ má stejné rozdělení jako

$$(\mathbb{B}\boldsymbol{X}_0)^T\Sigma^{-1}(\mathbb{B}\boldsymbol{X}_0) = \boldsymbol{X}_0^T\mathbb{B}^T\mathbb{B}\mathbb{B}^{\mathbb{T}^{-1}}\mathbb{B}\boldsymbol{X}_0 = \boldsymbol{X}_0^T\boldsymbol{X}_0.$$

Pak $\boldsymbol{X}_0^T\boldsymbol{X}_0$ lze přepsat jako součet kvadrátů jednotlivých složek náhodného vektoru $\boldsymbol{X}_0$, což odpovídá definici $\chi^2_n$ rozdělení.

iii) Matice $\mathbb{A}\Sigma$ je nenulová, z toho plyne $rank(\mathbb{A}) \geq 1$, tedy existuje skeletní rozklad matice $\mathbb{A} = \mathbb{B}\mathbb{B}^{\mathbb{T}}$. Pak $\boldsymbol{X}^T\mathbb{A}\boldsymbol{X} = \boldsymbol{X}^T\mathbb{B}\mathbb{B}^{\mathbb{T}}\boldsymbol{X} = (\mathbb{B}^{\mathbb{T}}\boldsymbol{X})^T(\mathbb{B}^{\mathbb{T}}\boldsymbol{X}) = (\mathbb{B}^{\mathbb{T}}\boldsymbol{X})^T\mathbb{I}(\mathbb{B}^{\mathbb{T}}\boldsymbol{X})$, přičemž $\mathbb{B}^{\mathbb{T}}\boldsymbol{X} \sim N(\boldsymbol{0}, \mathbb{B}^T\Sigma\mathbb{B})$. Nyní chceme použít tvrzení ii), k tomu potřebujeme ukázat, že $\mathbb{B}^T\Sigma\mathbb{B}$ je idempotentní.

Vyjděme z předpokladu, že $\mathbb{A}\Sigma$ je idempotentní a přepišme ho pomocí skeletního rozkladu $\mathbb{A}$, tj. $(\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma)(\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma) = (\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma)$. Nyní vynásobíme rovnici zleva maticí $\mathbb{B}^{-1}$, z vlastností skeletního rozkladu plyne, že $\mathbb{B}^{-1}\mathbb{B} = \mathbb{I}$. Pak $(\mathbb{B}^{-1}\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma)(\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma) = (\mathbb{B}^{-1}\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma)$, neboli $(\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma)(\mathbb{B}^{\mathbb{T}}\Sigma\mathbb{B}) = (\mathbb{B}^{\mathbb{T}}\Sigma\mathbb{B})$. Dokázali jsme, že je matice $\mathbb{B}\Sigma\mathbb{B}$ idempotentní, tedy její pseudoinverzí je jednotková matice $\mathbb{I}$. Tvrzení pak vyplývá z části ii) a faktu, že $\mathsf{rank}(\mathbb{B}\Sigma\mathbb{B}) = \mathsf{tr}(\mathbb{B}\Sigma\mathbb{B}) = \mathsf{tr}(\mathbb{B}\mathbb{B}^{\mathbb{T}}\Sigma) = \mathsf{tr}(\mathbb{A}\Sigma)$.

□

Na závěr uvedeme užitečné tvrzení o nezávislosti.

**Věta 4.** *Nechť $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$ a $\mathbb{A}$ je pozitivně definitní matice typu $n \times n$.*

*i) Nechť $\mathbb{B}$ je libovolná obdelníková matice splňující $\mathbb{B}\Sigma\mathbb{A} = \boldsymbol{0}$. Pak náhodná veličina $\boldsymbol{X}^T\mathbb{A}\boldsymbol{X}$ a náhodný vektor $\mathbb{B}\boldsymbol{X}$ jsou vzájemně nezávislé.*

*ii) Nechť $\mathbb{B}$ je čtvercová pozitivně definitní matice splňující $\mathbb{B}\Sigma\mathbb{A} = \boldsymbol{0}$. Pak náhodné veličiny $\boldsymbol{X}^T\mathbb{A}\boldsymbol{X}$ a $\boldsymbol{X}^T\mathbb{B}\boldsymbol{X}$ jsou vzájemně nezávislé.*

*Důkaz.*

i) Využijeme skeletní rozklad $\mathbb{A} = \mathbb{L}\mathbb{L}^T$. Předpokládáme $\mathbb{B}\Sigma\mathbb{L}\mathbb{L}^T = \boldsymbol{0} \implies \mathbb{B}\Sigma\mathbb{L}\mathbb{L}^T(\mathbb{L}^T)^{-1} = \boldsymbol{0}$. Z toho plyne, že $\boldsymbol{0} = \mathbb{B}\Sigma\mathbb{L} = \mathsf{cov}(\mathbb{B}\boldsymbol{X}, \mathbb{L}^T\boldsymbol{X})$. Normalita vektoru $\boldsymbol{X}$ implikuje, že i vektory $\mathbb{B}\boldsymbol{X}$ a $\mathbb{L}^T\boldsymbol{X}$ mají normální rozdělení a protože jsou nekorelované, jsou i nezavislé. Z toho plyne i nezávislost vektorů $\mathbb{B}\boldsymbol{X}$ a $(\mathbb{L}^T\boldsymbol{X})^T\mathbb{L}^T\boldsymbol{X} = \boldsymbol{X}^T(\mathbb{L}^T)^T\mathbb{L}^T\boldsymbol{X} = \boldsymbol{X}^T\mathbb{A}\boldsymbol{X}$.

ii) Opět vyjdeme ze skeletního rozkladu matic $\mathbb{A} = \mathbb{L}\mathbb{L}^T$ a $\mathbb{B} = \mathbb{P}\mathbb{P}^T$. Analogicky dosadíme do předpokladu skeletní rozklady a získáme vztah $\mathbb{P}^T\Sigma\mathbb{L} = \boldsymbol{0} = \mathsf{cov}(\mathbb{P}^T\boldsymbol{X}, \mathbb{L}^T\boldsymbol{X})$. Z normality vektorů $\mathbb{P}^T\boldsymbol{X}$, $\mathbb{L}^T\boldsymbol{X}$ a jejich nekorelovanosti plyne, že jsou nezavislé. Pak jsou nezávislé i vektory $(\mathbb{P}^T\boldsymbol{X})^T\mathbb{P}^T\boldsymbol{X} = \boldsymbol{X}^T\mathbb{B}\boldsymbol{X}$ a $(\mathbb{L}^T\boldsymbol{X})^T\mathbb{L}^T\boldsymbol{X}$.

□

# 2. Analýza rozptylu s pevnými efekty

## 2.1 Jednoduché třídění

Základní situací v analýze rozptylu je takzvané jednoduché třídění. V tomto případě uvažujeme situaci, kdy máme $N$ vzájemně nezávislých náhodných výběrů $Y_{i,1}, \ldots, Y_{i,n_i}, i = 1, \ldots, N$, kde $Y_{i,j}, j = 1, \ldots, n_i$ pochází z rozdělení s distribuční funkcí $F_i$ s konečnou střední hodnotou $\mu_i$. Zajímá nás, zda je tato střední hodnota stejná pro všechny výběry. V případě zamítnutí nulové hypotézy také obvykle chceme vědět, které dvojice středních hodnot se od sebe statisticky signifikantně liší.

Příkladem může být situace, kdy máme $N$ odrůd ovocných stromů (např. jabloní), u kterých zjišťujeme celkovou hmotnost úrody. Chceme vědět, jestli všechny odrůdy mají shodnou střední hodnotu hmotnosti úrody (neboli zda celková hmotnost plodů nezávisí na skupině) a pokud ne, u kterých skupin nastávají významné odlišnosti.

**Model 1.** *Pro zbytek kapitoly předpokládejme, že jednotlivé náhodné výběry jsou vzájemně nezávislé a pochází z normálních rozdělení s neznámými středními hodnotami $\mu_i, i = 1, \ldots, N$ a shodným rozptylem $\sigma^2$.*
*Tedy: $Y_{i,j} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2), j = 1, \ldots, n_i$, přičemž $\boldsymbol{Y}_i := (Y_{i,1}, \ldots, Y_{i,n_i})^T$ jsou vzájemně nezávislé pro $i = 1, \ldots, N$.*

Budeme testovat hypotézu $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_N$ proti alternativní hypotéze $H_1$: Existuje dvojice středních hodnot, které se nerovnají.

Výše zavedený model 1 je ekvivalentní následujícímu modelu:

$$Y_{i,j} = \mu + \beta_i + \varepsilon_{i,j}, \text{ kde } \varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Toto vyjádření získáme, položíme-li $\mu_i = \mu + \beta_i$, přičemž $\beta_i$ zde reprezentuje odchylku mezi $\mu$ a střední hodnotou $\mu_i$ v $(i,j)$ - tém výběru. Veličina $\varepsilon_{i,j}$ pak představuje variabilitu dat danou přírodními podmínkami. Tento model se označuje jako přeparametrizovaný, protože máme zavedeno o parametr více než je zapotřebí.

Nulovou hypotézu pak zapíšeme v ekvivalentním tvaru:

$$H_0 : \beta_1 = \cdots = \beta_N$$

. Za platnosti nulové hypotézy bude náš model tvaru: $Y_{ij} = \mu + \beta + \varepsilon_{ij}$, kde $\beta = \beta_1 = \cdots = \beta_N$. Místo porovnávání střední hodnoty $\mu_i$ se tedy zaměříme na porovnávání parametrů $\beta_i$.

Abychom sestavili testovou statistiku, zavedeme napřed několik pojmů:

**Definice 5** (Součty čtverců)**.** *Označme $M = \sum_{i=1}^{N} n_i$.*

*Zavedeme značení pro součet a výběrový průměr pozorování ve skupině:*
$Y_{i,+} = \sum_{j=1}^{n_i} Y_{i,j}$ *a* $\overline{Y}_{i,+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$,
*a dále značení pro celkový součet a celkový průměr pozorování:*

$Y_{++} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Y_{i,j}$ *a* $\overline{Y}_{++} = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{n_i} Y_{i,j}$.
*Položme:* $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})^T$ *vektor pozorování ve skupinách,*
$\boldsymbol{Y} = (\boldsymbol{Y_1}^T, \ldots, \boldsymbol{Y_N}^T)^T$ *vektor všech pozorování a*
$\overline{\boldsymbol{Y}} = (\overline{Y}_{1,+}, \ldots, \overline{Y}_{N,+})^T$ *vektor skupinových průměrů.*

*Nyní položme:*
$SS_c = \sum_{i=1}^{N} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y}_{++})^2$ celkový součet čtverců,
$SS_a = \sum_{i=1}^{N} n_i (\overline{Y}_{i,+} - \overline{Y}_{++})^2$ součet čtverců skupin,
$SS_e = \sum_{i=1}^{N} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y}_{i,+})^2$ residuální součet čtverců.

Takto zavedený součet čtverců skupin reprezentuje rozptyl mezi jednotlivými skupinami, naopak residuální součet představuje rozptyl uvnitř skupin. Součty čtverců lze rozepsat jako kvadratické formy, které mají za platnosti uvažované nulové hypotézy a výše zavedeného modelu rozdělení chí-kvadrát a jsou nestrannými odhady rozptylu $\sigma^2$. Podrobněji jsou tyto vztahy popsány v následující větě.

**Věta 5.** *i) Za platnosti modelu 1 máme:* $\frac{SS_e}{\sigma^2} \sim \chi^2_{M-N}$ *a* $\mathsf{E}\, \frac{SS_e}{M-N} = \sigma^2$.
*ii) Za platnosti modelu 1 a nulové hypotézy splňuje celkový součet čtverců* $\frac{SS_c}{\sigma^2} \sim \chi^2_{M-1}$ *a* $\mathsf{E}\, \frac{SS_c}{M-1} = \sigma^2$*, pro součet čtverců skupin máme:* $\frac{SS_a}{\sigma^2} \sim \chi^2_{N-1}$ *a* $\mathsf{E}\, \frac{SS_a}{N-1} = \sigma^2$.
*iii)* $SS_a$ *a* $SS_e$ *jsou nezávislé.*

*Důkaz.*

V důkazu budeme využívat tvrzení 3 o rozdělení kvadratických forem.

i) Platí $\frac{SS_e}{\sigma^2} = \sum_{i=1}^{N} \frac{1}{\sigma^2} \boldsymbol{Y}_i^T \mathbb{A}_i \boldsymbol{Y}_i$, kde $\mathbb{A}_i$ je čtvercová matice, která má na diagonále prvky $1 - \frac{1}{n_i}$ a prvky $-\frac{1}{n_i}$ mimo diagonálu. Víme, že $\frac{1}{\sigma^2} \boldsymbol{Y}_i^T \mathbb{A}_i \boldsymbol{Y}_i \sim \chi^2_{n_i-1}$, s použitím nezávislosti $\boldsymbol{Y}_i$ dostáváme $\frac{SS_e}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{N}(n_i-1)} = \chi^2_{M-N}$.

Střední hodnota veličiny s $\chi^2$ rozdělení je rovna počtu stupňů volnosti, tzn. $\mathsf{E}\, \frac{SS_e}{\sigma^2} = M - N$, z čehož plyne, že po vydělení $M - N$ je residuální součet čtverců nestranným odhadem rozptylu $\sigma^2$.

ii) Za platnosti nulové hypotézy lze všechna pozorování $Y_{i,j}$ uspořádat do vektoru $\boldsymbol{Y} \sim N_M(\boldsymbol{\mu 1}_M, \sigma^2 \mathbb{I}_M)$, pak jde o výběr o rozsahu $M$. $\frac{SS_c}{M-1}$ je jeho výběrový rozptyl, z vlastností výběrového rozptylu pak plyne tvrzení pro $SS_c$.

Dále definujme vektor $\overline{\boldsymbol{Y}} = \mathsf{diag}^{-1}(n_1, \ldots, n_N)(Y_{1,+}, \ldots, Y_{N,+})^T$. Pak

$$\frac{SS_a}{\sigma^2} = \frac{1}{\sigma^2} (\overline{\boldsymbol{Y}} - \boldsymbol{\mu 1_N})^T \mathbb{C} (\overline{\boldsymbol{Y}} - \boldsymbol{\mu 1_N}),$$

přičemž $\mathbb{C} = \mathsf{diag}(\boldsymbol{n}) - \frac{1}{M} \boldsymbol{n}\boldsymbol{n}^T$. Vektor $(\overline{\boldsymbol{Y}} - \boldsymbol{\mu 1_N}) \sim N_N(\boldsymbol{0}, \sigma^2 \mathsf{diag}^{-1}(\boldsymbol{n}))$, kde $\boldsymbol{n} = (n_1, \ldots, n_N)$, takže pokud dokážeme, že je matice $\mathbb{C}\mathsf{diag}^{-1}(\boldsymbol{n})$ idempotentní, bude platit $\frac{SS_a}{\sigma^2} \sim \chi^2_{\mathsf{tr}\mathbb{C}\mathsf{diag}^{-1}(\boldsymbol{n})}$. Idempotenci snadno ověříme s použitím vztahu

$$\mathbb{C}\mathsf{diag}^{-1}(\boldsymbol{n}) = \mathbb{I}_{\mathbb{N}} - \frac{1}{M} \boldsymbol{n}\boldsymbol{n}^T \mathsf{diag}^{-1}(\boldsymbol{n}) = \mathbb{I}_{\mathbb{N}} - \frac{1}{M} \boldsymbol{n}\boldsymbol{1}_N.$$

Tento přepis je také vhodný pro výpočet stopy matice,

$$\mathsf{tr}(\mathbb{I}_{\mathbb{N}} - \frac{1}{M} \boldsymbol{n}\boldsymbol{n}^T \mathsf{diag}^{-1}(\boldsymbol{n})) = \mathsf{tr}\mathbb{I}_{\mathbb{N}} - \frac{1}{M} \mathsf{tr}(\boldsymbol{n}\boldsymbol{1}_N) = N - 1.$$

Dohromady dostáváme $\frac{SS_a}{\sigma^2} \sim \chi^2_{N-1}$. Pro střední hodnotu opět platí, že $\mathsf{E}\, \frac{SS_a}{\sigma^2} = N - 1$, z čehož plyne $\mathsf{E}\, \frac{SS_a}{N-1} = \sigma^2$.

iii) V předchozích částech důkazu jsme ukázali, že $SS_a$ i $SS_e$ lze přepsat jako kvadratické formy, konkrétně

$$SS_e = \boldsymbol{Y}^T \mathbb{A} \boldsymbol{Y},$$

$$SS_a = \boldsymbol{Y}^T \mathbb{B} \boldsymbol{Y},$$

kde $B = \mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}, \ldots, \frac{1}{n_N}\mathbf{1}_{n_N})\mathbb{C}\mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}^T, \ldots, \frac{1}{n_N}\mathbf{1}_{n_N}^T)$. K ověření nezávislosti nám tedy podle věty 4 stačí ukázat, že $\mathbb{B}\mathrm{Var}\boldsymbol{Y}\mathbb{A} = \sigma^2\mathbb{B}\mathbb{A} = 0$. Nemusíme provádět celý výpočet, stačí si všimnout, že $\mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}^T, \ldots, \frac{1}{n_N}\mathbf{1}_{n_N}^T)\mathbb{A} = \mathbf{0}$, pak je i celý součin nulový a nezávislost veličin $SS_a$ a $SS_e$ je dokázána.

□

Na základě vztahů z předchozí věty navrhneme testovou statistiku jako podíl nestranných odhadů rozptylu:

$$F_A = \frac{\frac{SS_a}{N-1}}{\frac{SS_e}{M-N}}$$

Z věty 5 dále plyne, že za platnosti nulové hypotézy má $F_A$ Fisherovo rozdělení s $N-1$, $M-N$ stupni volnosti.

Navrhli jsme testovou statistiku a nyní nás zajímá, jak volit kritický obor. Pokud platí nulová hypotéza o rovnosti středních hodnot skupin, pak skupinový průměr $\overline{Y}_{i,+}$ představuje odhad středních hodnot skupin a celkový průměr $\overline{Y}_{++}$ představuje odhad celkové střední hodnoty, které by měly být stejné. Očekáváme tedy, že rozdíl $(\overline{Y}_{i,+} - \overline{Y}_{++})$ bude nabývat velmi malých hodnot, a proto i skupinový součet čtverců $SS_a$ bude nabývat menší hodnotu, než residuální součet čtverců $SS_e$. Platnost nulové hypotézy by se měla projevit v malých hodnotách testové statistiky $F_A$, velké hodnoty naznačují platnost alternativní hypotézy. To nás vede k myšlence zvolit kritický obor tak, aby test zamítal nulovou hypotézu pro vysoké hodnoty $F_A$. Za účelem získání testu s hladinou $\alpha$ pak $H_0$ zamítáme $\iff F_A \geq F_{N-1,M-N}(1-\alpha)$, kde $F_{N-1,M-N}(1-\alpha)$ je $(1-\alpha)$ kvantil Fisherova rozdělení s $N-1$, $M-N$ stupni volnosti.

Pro zápis výsledků jednoduchého třídění se často používá následující tabulka:

| **Zdroj měnlivosti** | **Součet čtverců** | **Počet stupňů volnosti** | **Podíl** | **Testová statistika** |
|---|---|---|---|---|
| Skupiny | $SS_a$ | $N-1$ | $\frac{SS_a}{N-1}$ | $\frac{SS_a}{N-1} / \frac{SS_e}{M-N}$ |
| Residuální | $SS_e$ | $M-N$ | $\frac{SS_e}{M-N}$ | |
| Celkový | $SS_c$ | $M-1$ | | |

Tabulka 2.1: Analýza rozptylu jednoduchého třídění

Alternativně se model jednoduchého třídění zavádí s využitím teorie lineárních modelů. Tento přístup je popsán např. v knize (Anděl, 2007b, kapitola 9.2).

# 3. Analýza rozptylu s náhodnými efekty

## 3.1 Jednoduché třídění - zavedení modelu

V minulé kapitole jsme zkoumali případ, kdy máme $N$ vzájemně nezávislých náhodných výběrů, o kterých za platnosti nulové hypotézy předpokládáme, že pochází z rozdělení se stejnou distribuční funkcí. Zavedli jsme model, ve kterém jsme tyto výběry popsali pomocí normálního rozdělení s neznámou střední hodnotou a pro všechny výběry shodným neznámým rozptylem $\sigma^2$. Jednotlivá pozorování jsme popsali pomocí působení několika veličin, které byly všechny až na člen $\varepsilon_{i,j}$ pevnými neznámými konstanty. Tento případ se nazýval model s pevnými efekty.

Nyní zavedeme mírně odlišnou situaci. Uvažujme, že skupin, ze kterých bereme naše pozorování (různé odrůdy jabloní v příkladu z předchozí kapitoly), není jen pevně dané $N$, ale může jich být až spočetně mnoho. Nepůjde o pevně daný počet skupin, ale o náhodný výběr z větší populace. Tato úvaha nás vede k zavedení jiného typu modelu (bude se jednat o speciální případ lineárního smíšeného modelu). Nový model vysvětlíme na příkladu.

Mějme $k$ laboratorních myší, každá z nich má $n_i, i = 1, \ldots, k$ mláďat. Budeme měřit rozdíly ve hmotnosti mláďat po dvou týdnech věku a zajímá nás, jestli mláďata mají stejnou střední hodnotu hmotnosti nezávisle na matce.

V tomto případě předpokládáme, že našich $k$ myší bylo náhodně vybráno z populace myší. Hmotnost mláděte pro danou $i$-tou myš má střední hodnotu $\mu + B_i$. Předpokládáme, že, náhodný výběr myši se projeví jako jistá variabilita $B_i$. Dále předpokládejme, že $\mathsf{E}\ B_i = 0, \mathsf{var}\, B_i = d^2$.

Zároveň hmotnosti všech mláďat dané matky mají také určitou variabilitu. Stejně jako v předchozím modelu jako původce této variability označíme veličinu $\varepsilon_{i,j}$, která reprezentuje odchylku hmotnosti konkrétního mláděte od střední hodnoty hmotnosti pro mláďata dané myši a budeme předpokládat, že $\mathsf{E}\ \varepsilon_{i,j} = 0$, $\mathsf{var}\, \varepsilon_{i,j} = \sigma^2$. Hmotnost $j$-tého mláděte $i$-té myši pak popíšeme vztahem
$Y_{i,j} = \mu + B_i + \varepsilon_{i,j}$.

Narozdíl od modelu s pevnými efekty, kdy $B_i$, resp. $\beta_i$, byla neznámou konstantou, jde nyní o náhodnou proměnnou. Tato situace se proto označuje jako model s pevnými efekty. Obecně se podle knihy (Khuri, 2010, kapitola 1.5) jako model s náhodnými efekty označuje model, kde jsou všechny členy kromě prvního náhodné.

Výše zmíněné požadavky společně s předpoklady o rozdělení zformulujeme jako následující model.

**Model 2.** *Mějme náhodný vektor* $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_k^T)^T$, *kde pro všechna* $i = 1, \ldots, k$ *jsou* $\boldsymbol{Y}_i = (Y_{i,1}, \ldots Y_{i,n_i})^T$ *vzájemně nezávislé skupiny pozorování. Řekneme, že* $\boldsymbol{Y}$ *se řídí modelem s náhodnými efekty, pokud lze psát* $Y_{i,j} = \mu + B_i + \varepsilon_{i,j}$, *kde* $\varepsilon_{i,j} \overset{i.i.d.}{\sim} N(0, \sigma^2), 0 < \sigma^2 < \infty$ *a* $B_i \overset{i.i.d.}{\sim} N(0, d^2)$, $0 < d^2 < \infty$, *přičemž* $\mu, \sigma^2$ *a* $d^2$ *jsou neznámé parametry.*

*Pro zbytek textu předpokládejme, že* $B_i$ *a* $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i})^T$ *jsou nezávislé.*

## 3.2 Odvození vlastností

Nejprve odvodíme vlastnosti pozorování při použití našeho modelu, budou nás zajímat hlavně střední hodnota a rozptyl $(i,j)$ - tého pozorování. Kromě toho potřebujeme znát tyto hodnoty i pro $(i,j)$ - té pozorování podmíněné veličinou $B_i$.

V průběhu odvozování vycházíme z modelu 2, budeme využívat předpokládané vlastnosti složek $Y_{i,j}$ (známé rozdělení a nezávislost $B_i$ a $\varepsilon_{i,j}$). Dále budeme používat vlastnosti podmíněné střední hodnoty a podmíněného rozptylu, které byly uvedeny v kapitole 1.

Nejprve odvodíme střední hodnotu pro $(i,j)$ - té pozorování. Využijeme linearity střední hodnoty a toho, že už známe střední hodnotu $B_i$ a $\varepsilon_{i,j}$.

$$\mathsf{E}\left(Y_{i,j}\right) = \mathsf{E}\left(\mu + B_i + \varepsilon_{i,j}\right) = \mu + \mathsf{E}\left(B_i\right) + \mathsf{E}\left(\varepsilon_{i,j}\right) = \mu$$

Dále odvodíme střední hodnotu pro $(i,j)$ - té pozorování podmíněné veličinou $B_i$. Využíváme znalostí o podmíněných veličinách, zejména vztahů o linearitě podmíněné střední hodnoty. Důležitý je také předpoklad, že $\varepsilon_{i,j}$ a $B_i$ jsou na sobě nezávislé veličiny.

$$\mathsf{E}\left(Y_{i,j}|B_i\right) = \mathsf{E}\left(\mu + B_i + \varepsilon_{i,j}|B_i\right) = \mathsf{E}\left(\mu|B_i\right) + \mathsf{E}\left(B_i|B_i\right) + \mathsf{E}\left(\varepsilon_{i,j}|B_i\right) =$$

$$= \mu + B_i + \mathsf{E}\left(\varepsilon_{i,j}\right) = \mu + B_i$$

Zajímá nás také podmíněný rozptyl $(i,j)$ - tého pozorování. Při odvození použijeme invariantnost rozptylu vůči konstantě (při podmínění veličinou $B_i$ vnímáme $B_i$ jako zafixovanou na určité hodnotě) a dále předpoklad o nezávislosti $\varepsilon_{i,j}$ a $B_i$.

$$\mathsf{var}(Y_{i,j}|B_i) = \mathsf{var}(\mu+B_i+\varepsilon_{i,j}|B_i) = \mathsf{var}(B_i+\varepsilon_{i,j}|B_i) = \mathsf{var}(\varepsilon_{i,j}|B_i) = \mathsf{var}(\varepsilon_{i,j}) = \sigma^2$$

Když jsme odvodili podmíněný rozptyl, můžeme použít vzorec 2 o rozkladu rozptylu a zjistit hodnotu nepodmíněného rozptylu.

$$\mathsf{var}(Y_{i,j}) = \mathsf{E}\left(\mathsf{var}(Y_{i,j}|B_i)\right) + \mathsf{var}(\mathsf{E}\left(Y_{i,j}|B_i\right)) = \mathsf{E}\left(\sigma^2\right) + \mathsf{var}(\mu + B_i) =$$

$$= \sigma^2 + \mathsf{var}(B_i) = \sigma^2 + d^2$$

Nyní jsme odvodili vlastnosti pro jednotlivá pozorování, pro srovnání výsledků s předchozím modelem uvádíme následující tabulku:

| **Model 1** | **Model 2** |
|---|---|
| $Y_{i,j} = \mu + \beta_i + \varepsilon_{i,j}, \beta_i \in \mathbb{R}$ | $Y_{i,j} = \mu + B_i + \varepsilon_{i,j}, B_i \sim N(0, d^2)$ |
| $\mathsf{E}\left(Y_{i,j}\right) = \mu + \beta_i$ | $\mathsf{E}\left(Y_{i,j}\right) = \mu$ |
| $\mathsf{var}(Y_{i,j}) = \sigma^2$ | $\mathsf{var}(Y_{i,j}) = \sigma^2 + d^2$ |

Tabulka 3.1: Srovnání vlastností $Y_{i,j}$

Obdobně jako v přechozí kapitole položíme $\boldsymbol{Y} = (\boldsymbol{Y_1}^T, \ldots, \boldsymbol{Y_k}^T)^T$ vektor všech pozorování. Chtěli bychom s použitím předchozích výsledků sestrojit rozptylovou matici $\mathrm{Var}(\boldsymbol{Y})$.

Nejprve zjistíme rozptylovou matici pro vektor $\boldsymbol{Y}_i$, který reprezentuje $i$-tý náhodný výběr o rozsahu $n_i$. Struktura rozptylové matice je známá.

$$\mathrm{Var}(\boldsymbol{Y}_i) = \begin{pmatrix} \mathsf{var}(Y_{i,1}), & \mathsf{cov}(Y_{i,1}, Y_{i,2}), & \dots, & \mathsf{cov}(Y_{i,1}, Y_{i,n_i}) \\ \mathsf{cov}(Y_{i,2}, Y_{i,1}), & \mathsf{var}(Y_{i,2}), & \dots, & \mathsf{cov}(Y_{i,2}, Y_{i,n_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{cov}(Y_{i,n_i}, Y_{i,1}), & \mathsf{cov}(Y_{i,n_i}, Y_{i,2}), & \dots, & \mathsf{var}(Y_{i,n_i}) \end{pmatrix}$$

Z předchozích výpočtů už víme, že na diagonále budou hodnoty $\sigma^2 + d^2$ (tuto hodnotu jsme odvodili pro libovolné $(i,j)$ - té pozorování). Zbývá už zjistit jen příslušné hodnoty kovariance. Pro následující výpočet nás zajímají pouze hodnoty mimo diagonálu, tedy zkoumáme jen situace kdy $j \neq l$.

$$\mathsf{cov}(Y_{i,j}, Y_{i,l}) = \mathsf{E}\,(Y_{i,j}, Y_{i,l}) - \mathsf{E}\,(Y_{i,j})\,\mathsf{E}\,(Y_{i,l}) = \mathsf{E}\,((\mu + B_i + \varepsilon_{i,j})(\mu + B_i + \varepsilon_{i,l})) - \mu^2 =$$
$$= \mathsf{E}\,(\mu^2) + \mathsf{E}\,(2\mu B_i) + \mathsf{E}\,(\mu\varepsilon_{i,j}) + \mathsf{E}\,(\beta_i^2) + \mathsf{E}\,(B_i\varepsilon_{i,l}) + \mathsf{E}\,(\mu\varepsilon_{i,l}) + \mathsf{E}\,(B_i\varepsilon_{i,j}) + E(\varepsilon_{i,j}\varepsilon_{i,l}) - \mu^2$$
$$= \mu^2 + \mathsf{var}(B_i) - \mu^2 = d^2$$

V průběhu odvození jsme opakovaně využívali nezávislost veličin $B_i$ a $\varepsilon_{i,j}$ a z toho plynoucího vztahu $\mathsf{E}\,(B_i\varepsilon_{i,j}) = 0$. Díky zavedenému modelu jsme také mohli využít znalost momentů (nulová střední hodnota a známý rozptyl) u obou veličin.

Dohromady dostáváme, že

$$\mathrm{Var}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma^2 + d^2, & d^2, & \dots, & d^2 \\ d^2, & \sigma^2 + d^2, & \dots, & d^2 \\ \vdots & \vdots & \ddots & \vdots \\ d^2, & d^2, & \dots, & \sigma^2 + d^2 \end{pmatrix}.$$

Rozptylovou matici můžeme přepsat i pomocí vztahu $\mathrm{Var}(\boldsymbol{Y}_i) = \sigma^2\mathbb{I}_{n_i} + d^2\mathbb{J}_{n_i}$, kde $\mathbb{I}_{n_i}$ je jednotková matice o délce hrany $n_i$ a $\mathbb{J}_{n_i}$ je čtvercová matice o délce hrany $n_i$, tvořená samými jedničkami. Tento úspornější zápis využijeme pro zápis matice celkového rozptylu.

Nyní můžeme přistoupit k odvození matice rozptylu všech pozorování $\mathrm{Var}(\boldsymbol{Y})$. Známe už rozptyl a kovariance pro jeden náhodný výběr, zbývá nám zjistit mezivýběrové kovariance. V modelu 2 jsme definovali, že všechny náhodné výběry jsou vzájemně nezávislé, z toho vyplývá důležitý vztah: $\mathsf{cov}(Y_{i,j}, Y_{k,l}) = 0$, kdykoliv $i \neq k$.

Vidíme, že matice $\mathrm{Var}(\boldsymbol{Y})$ bude blokově diagonální, diagonála obsahuje rozptylové matice jednotlivých výběrů $\mathrm{Var}(\boldsymbol{Y}_i), i = 1,2,\dots,k$, což jsou čtvercové matice s délkou hrany $n_i, i = 1,2,\dots,k$ a strukturou popsanou výše. Zbytek matice je tvořen nulami. Můžeme si rovněž povšimnout, že hlavní diagonála je tvořena hodnotami $\sigma^2 + d^2$.

Matice vypadá následovně (indexy $n_1,\dots,n_k$ značí rozměry bloků na diagonále, symboly 0 reprezentují nulové bloky vhodných rozměrů):

$$Var(\boldsymbol{Y}) = \begin{pmatrix} \sigma^2\mathbb{I}_{n_1} + d^2\mathbb{J}_{n_1} & 0 & \dots & 0 \\ 0 & \sigma^2\mathbb{I}_{n_2} + d^2\mathbb{J}_{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2\mathbb{I}_{n_k} + d^2\mathbb{J}_{n_k} \end{pmatrix}.$$

## 3.3 Součty čtverců

Obdobně jako v kapitole 2 nejprve zavedeme součty čtverců, nyní ovšem budeme předpokládat tzv. vyvážené třídění, tedy že všechny skupiny obsahují stejný počet ($n$) vzorků.

**Definice 6** (Součty čtverců)**.** *Položme:*
$SS_c = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j} - \overline{Y}_{++})^2$ celkový součet čtverců,
$SS_a = \sum_{i=1}^{k} n(\overline{Y}_{i,+} - \overline{Y}_{++})^2$ součet čtverců skupin,
$SS_e = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j} - \overline{Y}_{i,+})^2$ residuální součet čtverců,
*kde* $\overline{Y}_{i,+}$ *a* $\overline{Y}_{++}$ *značí skupinový a celkový průměr jako v definici 5.*

Opět platí, že skupinový součet čtverců reprezentuje variabilitu mezi jednotlivými skupinami a residuální součet čtvreců celkovou variabilitu uvnitř skupin. Chceme zjistit, jaká je v novém modelu střední hodnota součtů čtverců.

Nejprve uvedeme několik pomocných mezivýsledků. K jejich výpočtu nám pomohly předpoklady o nezávislosti veličin $B$ a $\varepsilon$ a z toho plynoucích vztahů pro střední hodnoty: $\mathsf{E}\,(B_iB_j) = 0$, $\mathsf{E}\,(B_iB_i) = d^2$, $\mathsf{E}\,(B_i\varepsilon_{k,j}) = 0$, $\mathsf{E}\,(\varepsilon_{i,j}\varepsilon_{i,l}) = 0$. Dále využíváme znalostí o rozptylu a kovarianci pozorování $Y_{i,j}$, odvozené v předchozích částech.

$$\mathsf{E}\,(Y_{i,j})^2 = \mu^2 + d^2 + \sigma^2$$
$$\mathsf{E}\,(Y_{i,j}Y_{i,k}) = \mu^2 + d^2$$
$$\mathsf{E}\,(Y_{i,j}Y_{u,v}) = \mu^2$$
$$\mathsf{E}\,(\overline{Y}^2_{i,+}) = \mu^2 + d^2 + \frac{1}{n}\sigma^2$$
$$\mathsf{E}\,(\overline{Y}^2_{++}) = \mu^2 + \frac{1}{k}d^2 + \frac{1}{kn}\sigma^2$$

Nyní přikročíme k výpočtu středních hodnot jednotlivých součtů čtverců.

$$\mathsf{E}\,(SS_c) = \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j} - \overline{Y}_{++})^2) = \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y^2_{i,j} - 2Y_{i,j}\overline{Y}_{++} + \overline{Y}^2_{++})) =$$

$$= \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y^2_{i,j})) - 2\,\mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}Y_{i,j}\overline{Y}_{++}) + \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}\overline{Y}^2_{++}) =$$

$$= kn\,\mathsf{E}\,(Y_{i,j})^2 - 2\,\mathsf{E}\,(kn\overline{Y}_{++}\overline{Y}_{++}) + kn\,\mathsf{E}\,(\overline{Y}^2_{++}) =$$

$$= kn(\mu^2 + d^2 + \sigma^2) - kn(\mu^2 + \frac{1}{k}d^2 + \frac{1}{kn}\sigma^2) = d^2(kn - n) + \sigma^2(kn - 1)$$

$$\mathsf{E}\,(SS_a) = \mathsf{E}\,(\sum_{i=1}^{k} n(\overline{Y}_{i,+} - \overline{Y}_{++})^2) = n\,\mathsf{E}\,(\sum_{i=1}^{k}\overline{Y}^2_{i,+} - 2\overline{Y}_{i,+}\overline{Y}_{++} + \overline{Y}^2_{++}) =$$

$$= nk\,\mathsf{E}\,(\overline{Y}^2_{i,+}) - 2n\,\mathsf{E}\,(k\overline{Y}_{++}\overline{Y}_{++}) + kn\,\mathsf{E}\,(\overline{Y}^2_{++}) = nk\,\mathsf{E}\,(\overline{Y}^2_{i,+}) - kn\,\mathsf{E}\,(\overline{Y}^2_{++}) =$$

$$= kn(\mu^2 + d^2 + \frac{1}{n}\sigma^2) - kn(\mu^2 + \frac{1}{k}d^2 + \frac{1}{kn}) = d^2(kn - n) + \sigma^2(k - 1)$$

$$\mathsf{E}\,(SS_e) = \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j} - \overline{Y}_{i,+})^2) = \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j}^2 - 2Y_{i,j}\overline{Y}_{i,+} + \overline{Y}_{i,+}^2)) =$$

$$= \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{i,j}^2)) - 2\,\mathsf{E}\,(\overline{Y}_{i,+}\sum_{i=1}^{k} n\overline{Y}_{i,+}) + \mathsf{E}\,(\sum_{i=1}^{k}\sum_{j=1}^{n}\overline{Y}_{i,+}^2) =$$

$$= kn\,\mathsf{E}\,(Y_{i,j})^2 - 2\sum_{i=1}^{k} n\,\mathsf{E}\,(\overline{Y}_{i,+}^2) + kn\,\mathsf{E}\,(\overline{Y}_{i,+}^2) = kn(\mu^2+d^2+\sigma^2) - kn(\mu^2+d^2+\frac{1}{n}\sigma^2)$$

$$= \sigma^2(kn-k)$$

Výsledky shrneme v tvrzení.

**Tvrzení 6.** *i) Za platnosti zavedeného modelu je* $(\frac{SS_a}{n(k-1)} - \frac{SS_e}{kn(n-1)})$ *nestranným odhadem* $d^2$, $\frac{SS_e}{k(n-1)}$ *je nestranným odhadem* $\sigma^2$ *a navíc* $\frac{SS_e}{\sigma^2} \sim \chi^2_{k(n-1)}$.

*ii) Pokud navíc platí* $d^2 = 0$, *pak* $\mathsf{E}\,\frac{SS_a}{k-1} = \sigma^2$ *a* $\frac{SS_a}{\sigma^2} \sim \chi^2_{k-1}$.

*iii) Platí-li* $d^2 = 0$, *jsou* $SS_a$ *a* $SS_e$ *nezávislé.*

*Důkaz.*

i) Důkaz nestrannosti odhadů plyne z předcházejících výpočtů, rozdělení $\frac{SS_e}{\sigma^2}$ pak bylo odvozeno v kapitole 2.

ii) Pokud je člen $d^2$ nulový, pak $\mathsf{E}\,(SS_a) = \sigma^2(k-1)$ plyne z předchozích výpočtů.

Pokusme se najít vyjádření $SS_a$ jako kvadratické formy a potom využít větu 3, která dává do souvislosti kvadratické formy a $\chi^2$ rozdělení. Položme $M = kn$. Za předpokladu $d^2 = 0$ můžeme všechna pozorování $Y_{i,j}$ upořádat do vektoru $\boldsymbol{Y}$, pro který platí $\boldsymbol{Y} \sim N_M(\boldsymbol{\mu}\mathbf{1}_M, \sigma^2\mathbb{I}_M)$. Struktura rozptylové matice se velmi zjednodušila, neboť vymizely členy $d^2$ a také máme vyvážené třídění. Budeme tedy postupovat analogicky jako u pevných efektů. Nejprve zadefinujeme vektor skupinových průměrů $\overline{\boldsymbol{Y}} = (\overline{Y}_{1,+}, \ldots, \overline{Y}_{k,+})$, pro který platí, že $\overline{\boldsymbol{Y}} - \boldsymbol{\mu}\mathbf{1}_k$ má rozdělení $N_k(\mathbf{0}, \sigma^2 diag(\frac{1}{n}, \ldots, \frac{1}{n}))$. Pro následující výpočet označme $\boldsymbol{n} = n\mathbf{1}_k$.

$$SS_a = \sum_{i=1}^{k} n(\overline{Y}_{i,+} - \overline{Y}_{++})^2 = \sum_{i=1}^{k} n(\overline{Y}_{i,+} - (\frac{1}{M}\boldsymbol{n}^T\overline{\boldsymbol{Y}}))^2 =$$

$$= (\overline{\boldsymbol{Y}} - \mathbf{1}_k\frac{1}{M}\boldsymbol{n}^T\overline{\boldsymbol{Y}})^T diag(\boldsymbol{n})(\overline{\boldsymbol{Y}} - \mathbf{1}_k\frac{1}{M}\boldsymbol{n}^T\overline{\boldsymbol{Y}}) =$$

$$= \overline{\boldsymbol{Y}}^T diag(\boldsymbol{n})\overline{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}^T\frac{1}{M}\boldsymbol{n}\boldsymbol{n}^T\overline{\boldsymbol{Y}} = \overline{\boldsymbol{Y}}^T\mathbb{C}\overline{\boldsymbol{Y}},$$

$$\text{kde } \mathbb{C} = diag(\boldsymbol{n}) - \frac{1}{M}\boldsymbol{n}\boldsymbol{n}^T.$$

Našli jsme vyjádření $SS_a$ coby kvadratické formy, nyní stačí ukázat, že je matice $diag^{-1}(\boldsymbol{n})\mathbb{C}$ idempotentní. Ověříme definici idempotence.

$$(diag^{-1}(\boldsymbol{n})\mathbb{C})(diag^{-1}(\boldsymbol{n})\mathbb{C}) = (\mathbb{I}_k - \frac{1}{M}\boldsymbol{n}\mathbf{1}_k^T)(\mathbb{I}_k - \frac{1}{M}\boldsymbol{n}\mathbf{1}_k^T) =$$

$$= \mathbb{I}_k\mathbb{I}_k - \frac{2}{M}\boldsymbol{n}\mathbf{1}_k^T + \frac{1}{M^2}\boldsymbol{n}\mathbf{1}_k^T\boldsymbol{n}\mathbf{1}_k^T = \mathbb{I}_k - \frac{1}{M}\boldsymbol{n}\mathbf{1}_k^T$$

S využitím věty o $\chi^2$ rozdělení a stopě matice pak dostáváme následující vztah pro rozdělení $SS_a$.

$$\frac{SS_a}{\sigma^2} = \frac{1}{\sigma^2}\sigma^2 \sim \chi^2_{tr(diag^{-1}(\boldsymbol{n})\mathbb{C})} = \chi^2_{k-1}.$$

iii) Součty čtverců $SS_a$ i $SS_e$ lze přepsat jako kvadratické formy, ověříme tedy předpoklad věty 4.

$$SS_e = \boldsymbol{Y}^T\mathbb{A}\boldsymbol{Y} = \boldsymbol{Y}^T diag(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\boldsymbol{Y},$$
$$SS_a = \boldsymbol{Y}^T\mathbb{H}^T diag^{-1}(\boldsymbol{n})\mathbb{C}diag^{-1}(\boldsymbol{n})\mathbb{H}\boldsymbol{Y},$$

kde $\mathbb{H} = diag(\mathbf{1}_n^T, \ldots, \mathbf{1}_n^T)$ je matice typu $k \times M$.
Potom $\mathbb{H}^T diag^{-1}(\boldsymbol{n})\mathbb{C}diag^{-1}(\boldsymbol{n})\mathbb{H}\sigma^2\mathbb{I}_M\mathbb{A} = \sigma^2\mathbb{H}^T diag^{-1}(\boldsymbol{n})\mathbb{C}diag^{-1}(\boldsymbol{n})\mathbb{H}\mathbb{A} = 0$, neboť $\mathbf{1}_n^T(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) = 0$. Tím je podmínka pro nezávislost ověřena.

□

## 3.4 Testování pro analýzu rozptylu s náhodnými efekty

Chceme testovat, zda mezi (podmíněnými) středními hodnotami jednotlivých skupin panuje shoda, nebo se statisticky významně liší. Odvodili jsme, že pro danou $i$-tou skupinu je střední hodnota podmíněná veličinou $B_i$ rovna $\mu + B_i$, kde náhodná veličina $B_i$ nyní modeluje odlišnost mezi skupinami, navíc předpokládáme, že $\mathsf{E}\ B_i = 0$ a $\mathsf{var}\, B_i = d^2$. Tedy v případě, že všechny skupiny mají skoro jistě shodnou střední hodnotu, musí platit, že $B_i = 0\ \forall i$, což je nyní vyjádřeno stavem, kdy $d^2 = 0$. Vyslovíme tedy nulovou hypotézu pomocí rozptylu $d^2$. Budeme testovat:

$$H_0 : d^2 = 0 \text{ proti } H_1 : d^2 \neq 0.$$

Z tvrzení 6 plyne, že za platnosti $H_0$ mají (po vydělení $\sigma^2$) skupinový a residuální součet čtverců $\chi^2$ rozdělení s přislušnými stupni volnosti, navíc jsou nezávislé. Můžeme je tedy využít pro konstrukci testové statistiky

$$F = \frac{\frac{SS_a}{k-1}}{\frac{SS_e}{k(n-1)}} \overset{H_0}{\sim} F_{k-1,k(n-1)}.$$

Stejně jako v případu s pevnými efekty budou velké hodnoty $F$ svědčit pro neplatnost $H_0$. Nulovou hypotézu zamítneme právě tehdy, když $F$ je větší, než $(1-\alpha)$ kvantil Fisherova rozdělení s $k-1, k(n-1)$ stupni volnosti.

# 4. Simulace

Předchozí výsledky pro model s náhodnými efekty byly odvozeny za předpokladu normality pozorování $Y_{i,j}$, kterou implikovalo normální rozdělení $\varepsilon_{i,j}$ a $B_i$. Díky normalitě jsme mohli odvodit $\chi^2$ rozdělení součtů čtverců a sestrojit testovou statistiku. Co když byl ale předpoklad normality chybný a $\varepsilon_{i,j}$ má jiné rozdělení? Zajímá nás, zda bude testová statistika (alespoň asymptoticky) dodržovat hladinu.

Provedeme simulace v programu R. Předpokládáme, že pro naše data platí: $Y_{i,j} = \mu + B_i + \varepsilon_{i,j} \overset{H_0}{=} \mu + \varepsilon_{i,j}$, bez újmy na obecnosti položíme $\mu = 0$.

Porovnáme případ, kdy $\varepsilon_{i,j}$ pochází z normálního rozdělení s případy, kdy pochází z rozdělení s těžším chvostem a šikmého rozdělení. Konkrétně budeme simulovat nejprve případ $\varepsilon_{i,j} \sim N(0,3)$, poté $\varepsilon_{i,j} \sim t_3$ a $ln(\varepsilon_{i,j}) \sim N(0,3)$. Pro potřeby simulace přeškálujeme lognormální rozdělení (odečteme střední hodnotu a vydělíme směrodatnou odchylkou), aby $\mathsf{E}\ \varepsilon_{i,j} = 0$ a $\mathsf{var}\,\varepsilon_{i,j} = 3$.

Postupně budeme generovat data pro $k = 10, 50, 200, 500$ skupin a 10 pozorování v každé skupině, pro tato data pak spočítáme příslušnou F-statistiku. Pro každý případ vygenerujeme F-statistiku 10 000 krát, abychom měli reprezentativní počet výsledků. Dále budeme počítat odhad pravděpodobnosti, že za platnosti nulové hypotézy hodnota testové statistiky $F$ převýší $(1 - \alpha)$ kvantil Fisherova rozdělení s danými stupni volnosti, neboli pravděpodobnost, že i za platnosti $H_0$ dojde k jejímu zamítnutí (chyba 1. druhu).

Použijeme hladinu $\alpha = 0.05$, výsledné empirické odhady této pravděpodobnosti jsou shrnuty v následující tabulce. Vidíme, že i pro $t_3$ a lognormální rozdělení jsme obdrželi výsledky blízké hodnotě 0.05, z toho usuzujeme, že test dodržuje hladinu i při námi uvažovaném porušení normality.

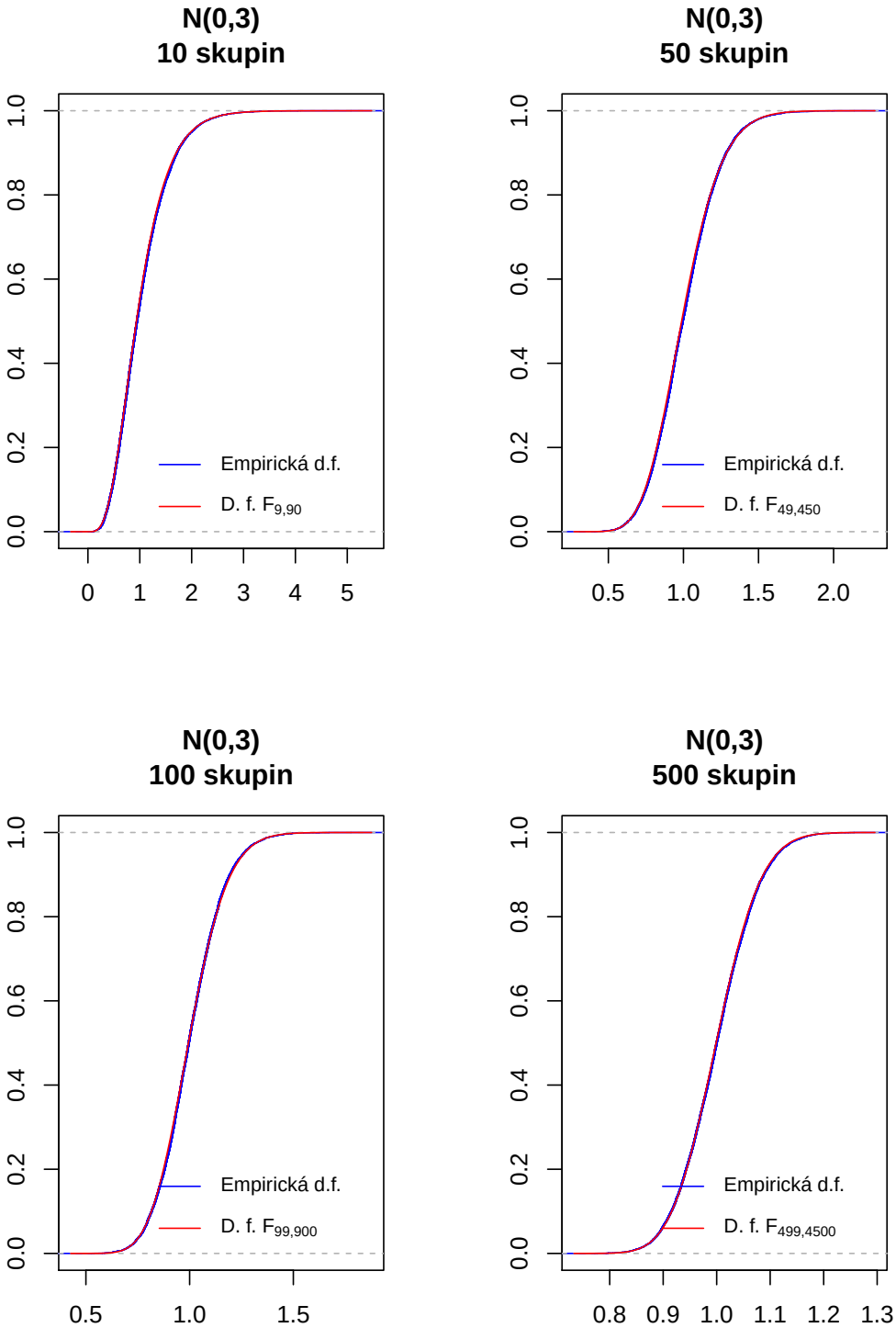
| | **Rozdělení** | | |
|---|---|---|---|
| **Počet skupin** | **N(0,3)** | $\boldsymbol{t_3}$ | **LN(0,3)** |
| 10 | 0.0512 | 0.0445 | 0.0311 |
| 50 | 0.0478 | 0.0474 | 0.0485 |
| 100 | 0.0475 | 0.0441 | 0.0469 |
| 500 | 0.0512 | 0.0483 | 0.054 |

Tabulka 4.1: Výsledné odhady na základě simulací

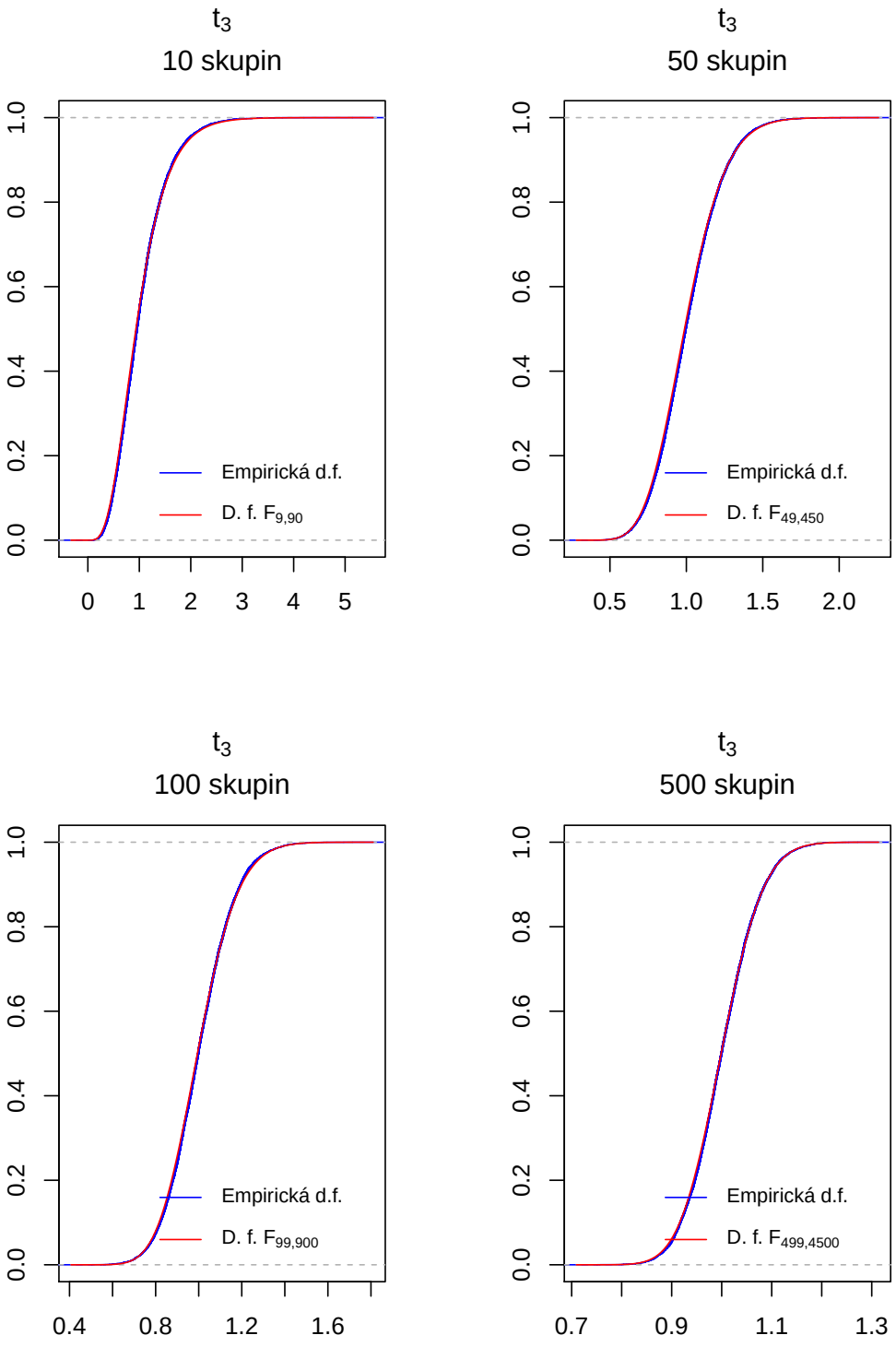
Dále pro každý případ porovnáme empirickou distribuční funkci testové statistiky F s distribuční funkcí Fisherova rozdělení s příslušnými stupni volnosti. Grafy s tímto porovnáním jsou na následujících stránkách.

Vidíme, že v případě $N(0,3)$ i $t_3$ rozdělení empirická distribuční funkce testové statistiky F téměř splývá se skutečnou distribuční funkcí Fisherova rozdělení, a to už pro 10 skupin. V případě lognormálního rozdělení mezi nimi pozorujeme rozdíl, ale pro rostoucí počet skupin $k$ se empirická funkce svým tvarem více podobá skutečné distribuční funkci F rozdělení s příslušnými stupni volnosti.

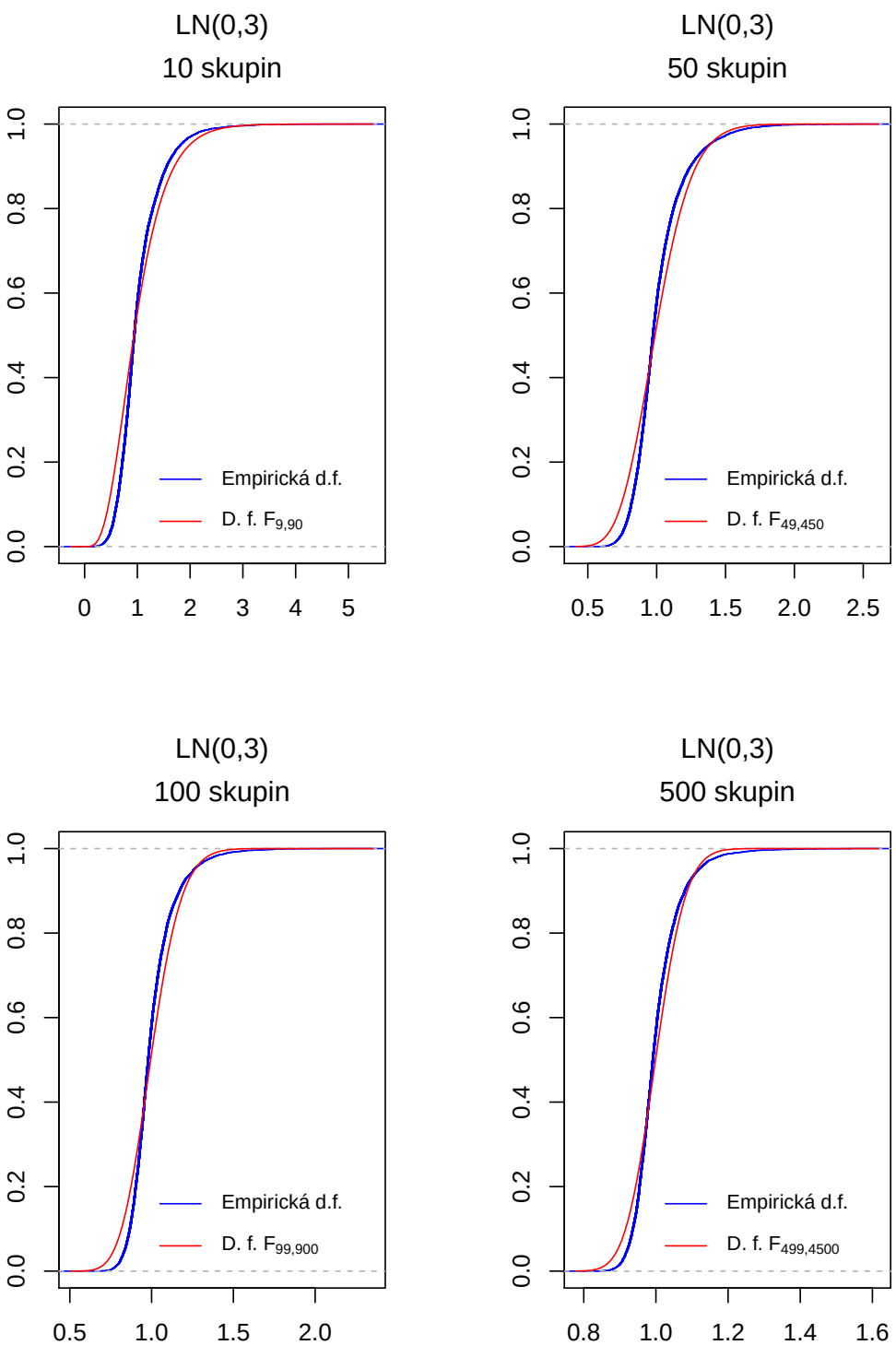
Můžeme tedy učinit závěr, že pro dostatečně velký počet pozorování dodržuje test analýzy rozptylu s náhodnými efekty hladinu i v případě, kdy pozorování nepochází z normálního rozdělení.




Obrázek 4.1: Porovnání distribučních funkcí pro normální rozdělení




Obrázek 4.2: Porovnání distribučních funkcí pro $t_3$ rozdělení




Obrázek 4.3: Porovnání distribučních funkcí pro lognormální rozdělení

# Závěr

V této práci jsme se zabývali analýzou rozptylu s náhodnými efekty. Vyšli jsme ze základních tvrzení z teorie pravděpodobnosti, podrobněji jsem se zaměřili na vztahy platící pro rozdělení kvadratických forem.

Dále jsme zavedli případ jednoduchého třídění s pevnými efekty. Podrobně jsme definovali součty čtverců a ukázali jejich vlastnosti, díky kterým jsme je mohli použít k sestavení testové statistiky.

V další části jsme nejprve zavedli model jednoduchého třídění s náhodnými efekty a podrobně jsme odvodili vlastnosti, která plynou pro pozorování a skupiny pozorování na základě tohoto modelu. S předpokladem vyváženého třídění, tj. všechny skupiny mají stejný počet pozorování, jsme zavedli součty čtverců, pro které jsme ukázali, že (za platnosti dodatečného předpokladu na nulovost rozptylu $d^2$ a po vydělení příslušnými stupni volnosti) jsou skupinový a residuální součet čtverců nezávislými odhady rozptylu $\sigma^2$. V závěru kapitoly jsme se zabývali testem na shodu podmíněných středních hodnot skupin, na základě vlastností součtů čtverců jsme setrojili testovou statistiku a kritický obor.

Na závěr jsme se zabývali otázkou, jak test dodržuje hladinu při porušení předpokladu normality. Provedli jsme simulace v programu R pro pevný počet pozorování ve skupině a různé počty skupin, data jsme nejprve generovali z normálního rozdělení a poté porovnávali jejich chování s daty generovanými z $t_3$ a lognormálního rozdělení, coby reprezentanty rozdělení s těžším chvostem a šikmého rozdělení. Zjistili jsme, že pro případ vyváženého třídění test dodržuje hladinu velmi dobře i v případě uvažovaného porušení normality.

# Seznam použité literatury


Anděl, J. (2007a). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Anděl, J. (2007b). *Statistické metody*. Čtvrté upravené vydání. Matfyzpress, Praha. ISBN 80-7378-003-8.

Khuri, A. I. (2010). *Linear model methodology*. Chapman and Hall/CRC. ISBN 978-1-58488-481-1.

Lachout, P. (2004). *Teorie pravděpodobnosti*. Druhé vydání. Karolinum, Praha. ISBN 80-246-0872-3.