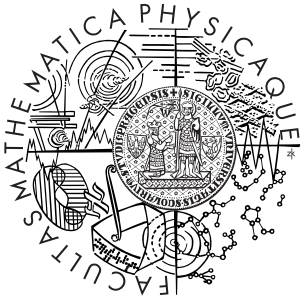
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Marie Turčičová

Jednoduché třídění s nestejnými rozptyly

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Karel Zvára, CSc.
Studijní program: Matematika, Obecná matematika

2009

Děkuji vřele doc. RNDr. Karlu Zvárovi, CSc. za pomoc při psaní této práce a poskytnutí potřebných písemných materiálů. Dále děkuji Pavlu Kuriščákovi za technickou pomoc a poradenství při práci s programy LaTeX a R.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 28. 5. 2009 Marie Turčičová

Obsah

Název práce: Jednoduché třídění s nestejnými rozptyly
Autor: Marie Turčičová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: doc. RNDr. Karel Zvára, CSc.
e-mail vedoucího: Karel.Zvara@mff.cuni.cz

Abstrakt: V předložené práci studujeme úlohu jednoduchého třídění. Nejprve se věnujeme situaci, kdy rozptyly jsou stejné, přičemž popíšeme obecný model této situace a pak odvodíme rozdělení F-statistiky, na základě které se testuje hypotéza o rovnosti středních hodnot. Pak již řešíme problém nestejných rozptylů. Opět popíšeme obecný model a ukážeme proč nestejné rozptyly vadí v klasickém řešení. Dále uvedeme tři konkrétní postupy testování hypotézy o rovnosti středních hodnot v případě nestejných rozptylů, a to metodou Welche, Boxe a Kruskala-Wallise. Pro ilustraci je práce doplněna simulacemi ukazujícími sílu těchto testů a jejich schopnost udržet hladinu.

Klíčová slova: jednoduché třídění, Welchův test, Boxův test, Kruskalův-Wallisův test

Title: One-way ANOVA under unequal variances
Author: Marie Turčičová
Department: Deptartment of Probability and Mathematical Statistics
Supervisor: doc. RNDr. Karel Zvára, CSc.
Supervisor's e-mail address: Karel.Zvara@mff.cuni.cz

Abstract: In this paper, we study the problem of one-way ANOVA. First, we focus on the simpler case of equal variances. We describe a general model and then obtain the F-statistic distribution, which is used for testing the hypothesis that the expected values of two random variables equal. Next we analyze the case of unequal variances. As before, we describe the general model and identify instances where unequal variances cause the classical solution to break down. Further, we present three specific methods for testing the hypothesis that the means equal in the case of unequal variances. Namely Welch test, Box test and Kruskal-Wallis test. For illustration, we append numerical simulations, which show the power of each of these tests and their ability to retain the significance level.

Keywords: one-way ANOVA, Welch test, Box test, Kruskal-Wallis test

Kapitola 1

Úvod

Cílem této práce je popsat úlohu jednoduchého třídění a některé její řešení v situaci, kdy rozptyly nejsou stejné, a pomocí simulací porovnat s klasickým řešením.

V mnoha přírodních i technických vědách je často zapotřebí porovnat několik sad naměřených hodnot a rozhodnout, zda se dané sady pohybují kolem stejné střední hodnoty. Jde vlastně o porovnání několika náhodných výběrů a test hypotézy o rovnosti středních hodnot. Předpokládejme, že výběry pocházejí z normálního rozdělení. Pokud jsou výběry jen dva, lze test provést snadno pomocí dvouvýběrového t-testu. Je-li výběrů více, použije se metoda zvaná jednoduché třídění. Její teoretické zázemí je popsáno v Kapitole 2, stejně jako odvození rozdělení F-statistiky, která se k testu hypotézy používá. Pro oba výše zmíněné postupy je ale nutné, aby všechny hodnoty, bez ohledu na to, z jaké sady pocházejí, byly naměřeny se stejnou přesností, tedy aby všechny náhodné výběry měly stejný rozptyl. Ne vždy je ovšem možné, například z technického hlediska, tento požadavek splnit. Jak tedy postupovat v situaci, kdy tomu tak není? Tomuto problému se věnujeme v Kapitole 3. Snažíme se zobecnit metodu jednoduchého třídění na novou situaci a ukazujeme, kde nestejné rozptyly vadí klasickému řešení. Dále uvidíme, že s rozdělením statistiky F už to pak není tak jednoduché a její rozdělení je možno pouze aproximovat. Dva různé způsoby aproximace jsou uvedeny v podkapitolách 3.2 a 3.3. Je to postup B. L. Welche a E. P. Boxe. Welchův postup je naprogramován coby protějšek k jednoduchému třídění v programu R a je velmi často používán. Řešení Boxe se už tak často nepoužívá a chceme-li ho použít, musíme si ho naprogramovat sami. Nakonec popíšeme ještě známý Kruskalův-Wallisův test, který je nepara-

metrickou obdobou jednoduchého třídění a není k němu tedy zapotřebí normalita výběrů. Místo F-statistiky se v něm používá veličina H, která má však s F-statistikou určitý vztah.

Na konci je práce doplněna o simulace, které ilustrují sílu jednotlivých testů a jejich schopnost udržet danou hladinu. Budeme zde diskutovat, jak se jednotlivé testy hodí pro různé rozsahy výběrů a stejné či nestejné rozptyly. Vše bude ukázáno na normálním, logistickém a gamma rozdělení.

Kapitola 2

Jednoduché třídění pro případ stejných rozptylů

2.1 Model jednoduchého třídění

Máme k nezávislých výběrů $Y_{i1}, \ldots, Y_{in_i}$ z rozdělení $N(\mu_i, \sigma^2)$, kde $i = 1, \ldots, k$. Rozsah i-tého výběru je tedy n_i a parametr $\sigma^2 > 0$ není znám. Testujeme hypotézu $H_0 : \mu_1 = \ldots = \mu_k$. Jedná se o lineární model tvaru $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, maticově zapsáno

$$\begin{pmatrix} Y_{11} \\ \ldots \\ Y_{1n_1} \\ Y_{21} \\ \ldots \\ Y_{2n_2} \\ \ldots \\ \ldots \\ Y_{k1} \\ \ldots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix} . \begin{pmatrix} \mu_1 \\ \ldots \\ \ldots \\ \ldots \\ \mu_k \end{pmatrix} + \begin{pmatrix} e_{11} \\ \ldots \\ e_{1n_1} \\ e_{21} \\ \ldots \\ e_{2n_2} \\ \ldots \\ \ldots \\ e_{k1} \\ \ldots \\ e_{kn_k} \end{pmatrix}, \tag{2.1}$$

kde $\boldsymbol{\beta}$ je vektor neznámých parametrů a $\mathbf{e}$ je vektor chyb. Předpokládejme, že $n_i > 1$ pro všechna $i = 1, \ldots, k$ (pak $h(\mathbf{X}) = k$) a že $n > k$. Dále předpokládejme, že vektor chyb $\mathbf{e}$ má normální rozdělení $N(\mathbf{0}, \sigma^2\mathbf{I})$. Nulová střední hodnota odpovídá tomu, že pozorování vektoru $\mathbf{Y}$ nejsou zatížena systematickými chybami. Varianční matice $\sigma^2\mathbf{I}$ zase říká, že měření jed-

notlivých složek vektoru $\mathbf{Y}$ jsou prováděna se stejnou přesností a že chyby měření různých složek vektoru $\mathbf{Y}$ jsou nezávislé.

Často se používá i model

$$Y_{ip} = \mu + \eta_i + e_{ip}, \qquad p = 1, \ldots, n_i;\ i = 1, \ldots, k, \tag{2.2}$$

kde μ a η_i jsou neznámé parametry, přičemž $\mu_i = \mu + \eta_i$. Do modelu nám tímto přibyl další parametr, což bude mít za následek to, že dostaneme soustavu rovnic se singulární maticí a že parametry μ, η_1, $\ldots$, η_k nebudou odhadnutelné. Tento přístup je ale u složitějších modelů názornější.

Parametry $\mu_1, \ldots, \mu_k$ se odhadují metodou nejmenších čtverců, tj. z podmínky, že výraz

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

má být minimální [2, str.81].
Tento odhad označíme $\mathbf{b} = (b_1, \ldots, b_k)'$. Dle [1, věta 9.1] platí, že

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Vektor $\mathbf{b}$ se tedy může počítat ze soustavy normálních rovnic $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$.

Zavedeme-li si značení

$$\begin{aligned}
n &= \sum_{i=1}^{k} n_i, \\
y_{i.} &= \frac{1}{n_i} \sum_{p=1}^{n_i} Y_{ip}, \\
y_{..} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{p=1}^{n_i} Y_{ip}, \\
\mathsf{diag}\,(n_1, \ldots, n_k) &= \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n_k \end{pmatrix},
\end{aligned}$$

tak potom v případě našeho modelu je

$$\mathbf{X}'\mathbf{X} = \mathsf{diag}\,(n_1, \ldots, n_k), \qquad \mathbf{X}'\mathbf{Y} = (n_1 y_{1.}, \ldots, n_k y_{k.})',$$

a tudíž

$$\mathbf{b} = (y_{1.}, \ldots, y_{k.})'.$$

Výraz

$$R = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2$$

se nazývá reziduální součet čtverců a často se místo R označuje jako S_e.

Za platnosti naší hypotézy H_0 máme podmodel $\mathbf{Y} = \mathbf{U}\boldsymbol{\gamma} + \mathbf{e}$, kde $\mathbf{U}_{n \times 1} = (1, \ldots, 1)'$ a $\boldsymbol{\gamma}$ je typu 1×1. Nyní $\mathbf{U}'\mathbf{U} = n$ a $\mathbf{U}'\mathbf{Y} = \sum_{i=1}^{k} \sum_{p=1}^{n_i} Y_{ip}$, takže odhad $\mathbf{g}$ parametru $\boldsymbol{\gamma}$ metodou nejmenších čtverců činí

$$\mathbf{g} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y} = y_{..}.$$

Příslušný reziduální součet čtverců je roven

$$R_1 = (\mathbf{Y} - \mathbf{Ug})'(\mathbf{Y} - \mathbf{Ug}) = \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{..})^2.$$

Veličina R_1 se často značí S_T a nazývá se celkový součet čtverců. Rozdíl $R_1 - R$ se nazývá řádkový součet čtverců. Je to

$$\begin{aligned}
S_A &= R_1 - R \\
&= \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{..})^2 - \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 \\
&= \sum_{i=1}^{k} \sum_{p=1}^{n_i} [(Y_{ip} - y_{i.}) + (y_{i.} - y_{..})]^2 - \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 \\
&= \sum_{i=1}^{k} \sum_{p=1}^{n_i} (y_{i.} - y_{..})^2 + 2 \sum_{i=1}^{k} \sum_{p=1}^{n_i} (y_{i.} - y_{..})(Y_{ip} - y_{i.}) \\
&= \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2 + 2 \sum_{i=1}^{k} \left[(y_{i.} - y_{..}) \sum_{p=1}^{n_i} (Y_{ip} - y_{i.}) \right] \\
&= \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2.
\end{aligned}$$

Stačí si jen uvědomit, že

$$\sum_{p=1}^{n_i} (Y_{ip} - y_{i.}) = \sum_{p=1}^{n_i} Y_{ip} - \sum_{p=1}^{n_i} y_{i.} = n_i y_{i.} - n_i y_{i.} = 0,$$

a tudíž celý výraz $\sum_{i=1}^{k}[(y_{i.} - y_{..})\sum_{p=1}^{n_i}(Y_{ip} - y_{i.})]$ je roven nule. Z výpočtu S_A plyne, že pro veličiny S_A, S_T a S_e platí

$$S_T = S_A + S_e.$$

Celkový součet čtverců tedy vyjádříme jako součet řádkového součtu čtverců (který vyjadřuje variabilitu průměrů v jednotlivých výběrech) a reziduálního součtu čtverců (ten vyjadřuje variabilitu uvnitř výběrů).

K testování hypotézy H_0 se používá statistika

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}}. \tag{2.3}$$

2.2 Odvození rozdělení statistiky F

Věta 1 Nechť $\mathbf{X} \sim N_k(\mathbf{0}, \mathbf{A})$, kde $\mathbf{A}$ je symetrická a idempotentní matice, $h(\mathbf{A}) = r > 0$. Potom platí $\|\mathbf{X}\|^2 \sim \chi^2_r$. [1, věta 4.14]

Lemma 1 Nechť a a b jsou reálná čísla. Je-li $X \sim N(\mu, \sigma^2)$, pak $a + bX \sim N(a + b\mu, b^2\sigma^2)$. [2, věta 4.1]

Lemma 2 Nechť $\mathbf{X} = (X_1, \ldots, X_n)' \sim N(\boldsymbol{\mu}, \mathbf{V})$. Pak $\mathbf{Y} = \mathbf{a}_{m\times 1} + \mathbf{B}_{m\times n}\mathbf{X} \sim N(\mathbf{a} + \boldsymbol{B\mu}, \mathbf{BVB}')$. [2, věta 4.4]

Věta 2 Nechť $\mathbf{X} \sim N(\mu\mathbf{1}, \mathsf{diag}\,(1/n_1, \ldots, 1/n_k))$, kde $\mathbf{1}$ je vektor samých jedniček délky n. Dále nechť $n = \Sigma_{i=1}^{k} n_i, \mathbf{n} = (n_1, \ldots, n_k)', \bar{X} = \frac{1}{n}\Sigma_{i=1}^{k} n_i X_i$. Potom

$$\mathbf{Z} = \mathsf{diag}\,(\sqrt{n_1}, \ldots, \sqrt{n_k})(\mathbf{X} - \bar{X}\mathbf{1}) \sim N(\mathbf{0}, \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}\,'),$$

kde $\sqrt{\mathbf{n}} = (\sqrt{n_1}, \ldots \sqrt{n_k})'$. Přitom jsou $\mathbf{Z}$ a $\bar{X}$ nezávislé náhodné veličiny.

Důkaz: Vektor $\mathbf{Z}$ dostaneme, když vektor $(\mathbf{I} - \frac{1}{n}\mathbf{1n}')\mathbf{X}$ vynásobíme maticí

$$\mathbf{D} := \mathsf{diag}\,(\sqrt{n_1}, \ldots, \sqrt{n_k}),$$

neboť $(\mathbf{I} - \frac{1}{n}\mathbf{1n}')\mathbf{X} = \mathbf{X} - \frac{1}{n}\mathbf{1n}'\mathbf{X} = \mathbf{X} - \frac{1}{n}\mathbf{1}n\bar{X} = \mathbf{X} - \bar{X}\mathbf{1}$.
Potom tedy $\mathbf{Z} = \mathbf{D}(\mathbf{X} - \bar{X}\mathbf{1}) = \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1n}')\mathbf{X}$ a dle lemmatu 2 platí, že

$$\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}\,').$$

Střední hodnotu náhodného vektoru $\mathbf{Z}$ snadno vypočteme pomocí linearity střední hodnoty:

$$\begin{aligned}
\mathsf{E}\,\mathbf{Z} &= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\,\mathsf{E}\,\mathbf{X} \\
&= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mu\mathbf{1} \\
&= \mathbf{D}(\mu\mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{n}'\mu\mathbf{1}) \\
&= \mathbf{D}(\mu\mathbf{1} - \frac{1}{n}\mathbf{1}\Sigma_{i=1}^{k}\mu n_i) \\
&= \mathbf{D}(\mu\mathbf{1} - \frac{1}{n}\mathbf{1}\mu n) \\
&= \mathbf{D}(\mu\mathbf{1} - \mu\mathbf{1}) = 0.
\end{aligned}$$

Nyní vypočteme varianční matici vektoru $\mathbf{Z}$:

$$\begin{aligned}
\mathsf{var}\,\mathbf{Z} &= \mathsf{var}[\mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{X}] \\
&= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\ \mathsf{var}\,\mathbf{X}\ (\mathbf{I} - \frac{1}{n}\mathbf{n}\mathbf{1}')\mathbf{D} \\
&= \mathbf{I} - \frac{1}{n}\mathbf{D}\mathbf{D}^{-2}\mathbf{n}\mathbf{1}'\mathbf{D} - \frac{1}{n}\mathbf{D}\mathbf{1}\mathbf{n}'\mathbf{D}^{-2}\mathbf{D} + \frac{1}{n^2}\mathbf{D}\mathbf{1}\mathbf{n}'\mathbf{D}^{-2}\mathbf{n}\mathbf{1}'\mathbf{D} \\
&= \mathbf{I} - \frac{1}{n}\mathbf{D}^{-1}\mathbf{n}\mathbf{1}'\mathbf{D} - \frac{1}{n}\mathbf{D}\mathbf{1}\mathbf{n}'\mathbf{D}^{-1} + \frac{1}{n^2}n\mathbf{D}\mathbf{1}\mathbf{1}'\mathbf{D} \\
&= \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}' - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}' + \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}' \\
&= \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}'.
\end{aligned}$$

Předposlední rovnost plyne z toho, že: $\mathbf{D}^{-1}\mathbf{n} = \sqrt{\mathbf{n}}$ a $\mathbf{D}\mathbf{1} = \sqrt{\mathbf{n}}$.

Zbývá už jen dokázat nezávislost. Víme, že

$$\begin{aligned}
\mathbf{Z} &= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{X} \qquad \text{a} \\
\bar{X} &= \frac{1}{n}\mathbf{n}'\mathbf{X}.
\end{aligned}$$

Tudíž

$$\begin{aligned}\mathsf{cov}(\mathbf{Z}, \bar{X}) &= \mathsf{cov}[\mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{X}, \frac{1}{n}\mathbf{n}'\mathbf{X}] \\ &= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\,\mathsf{var}\,\mathbf{X}\frac{1}{n}\mathbf{n} \\ &= \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{D}^{-2}\frac{1}{n}\mathbf{n} \\ &= \frac{1}{n}\mathbf{D}^{-1}\mathbf{n} - \frac{1}{n^2}\mathbf{D}\mathbf{1}\mathbf{n}'\mathbf{D}^{-2}\mathbf{n} \\ &= \frac{1}{n}\mathbf{D}^{-1}\mathbf{n} - \frac{1}{n}\mathbf{D}\mathbf{1} = \mathbf{0}.\end{aligned}$$

Navíc $\mathbf{Z}$ a $\bar{X}$ jsou lineární funkce téhož vektoru $\mathbf{X}$, který má normální rozdělení, a tedy i sdružené rozdělení vektoru $(\mathbf{Z}, \bar{X})'$ je normální. V takovém případě je pak nekorelovanost ekvivalentní nezávislosti, a tudíž $\mathbf{Z}$ a $\bar{X}$ jsou nezávislé.
□

Lemma 3 Je-li $\mathbf{A}_{n\times n}$ idempotentní matice hodnosti r, pak $\mathbf{I} - \mathbf{A}$ je rovněž idempotentní a $h(\mathbf{I} - \mathbf{A}) = n - r$. [2, věta A.13]

Lemma 4 Nechť Y a Z jsou nezávislé náhodné veličiny takové, že $Y \sim \chi^2_r$ a $Z \sim \chi^2_s$. Pak $Y + Z \sim \chi^2_{r+s}$. [2, věta 4.14]

Věta 3 Nechť jsou splněny předpoklady modelu jednoduchého třídění, tj. $Y_{ip} \sim N(\mu_i, \sigma^2)$ (kde $\sigma^2 > 0$) jsou pro $p = 1, \ldots, n_i, i = 1, \ldots, k$, nezávislé náhodné veličiny a $n_i > 1$, $\forall i = 1, \ldots, k$. Platí-li $\mu_1 = \ldots = \mu_k$, pak

$$\frac{\dfrac{S_A}{k-1}}{\dfrac{S_e}{n-k}} \sim F_{k-1,n-k}. \tag{2.4}$$

Důkaz: Stačí ukázat, že $\frac{S_A}{\sigma^2} \sim \chi^2_{k-1}$, $\frac{S_e}{\sigma^2} \sim \chi^2_{n-k}$ a že S_A, S_e jsou nezávislé. Pak (2.4) plyne triviálně z definice F-rozdělení.

Nejprve ukážeme, že $\frac{S_A}{\sigma^2} \sim \chi^2_{k-1}$. Zvolíme $X_i = \frac{y_{i.}}{\sigma}$ a použijeme větu 2. Máme tedy

$$\mathbf{X} = (X_1, \ldots, X_k)' = \frac{1}{\sigma}(y_{1.}, \ldots, y_{k.})'$$

a

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i X_i = \frac{1}{n}\sum_{i=1}^{k} n_i \frac{1}{\sigma n_i}\sum_{p=1}^{n_i} Y_{ip} = \frac{y_{..}}{\sigma}.$$

Náhodný vektor $\mathbf{Z}$ má potom tvar:

$$\begin{aligned}
\mathbf{Z} &= \mathbf{D}(\mathbf{X} - \bar{X}\mathbf{1}) \\
&= \frac{1}{\sigma}\,\mathsf{diag}\,(\sqrt{n_1},\ldots,\sqrt{n_k})(y_{1.} - y_{..},\ldots,y_{k.} - y_{..})' \\
&= \frac{1}{\sigma}[\sqrt{n_1}(y_{1.} - y_{..}),\ldots,\sqrt{n_k}(y_{k.} - y_{..})]'.
\end{aligned}$$

Nyní si spočteme čtverec normy

$$\begin{aligned}
\|\mathbf{Z}\|^2 &= \left(\sqrt{z_1^2 + \ldots + z_k^2}\right)^2 \\
&= z_1^2 + \ldots + z_k^2 \\
&= \frac{1}{\sigma^2}[n_1(y_{1.} - y_{..})^2 + \ldots + n_k(y_{k.} - y_{..})^2] \\
&= \frac{S_A}{\sigma^2},
\end{aligned}$$

což je přesně výraz, který potřebujeme. Dle věty 1 potom $\frac{S_A}{\sigma^2} \sim \chi^2_r$, kde $r = h(\mathsf{var}\,\mathbf{Z})$. Varianční matice $\mathsf{var}\,\mathbf{Z}$ má dle věty 2 tvar $\mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}'$ a je idempotentní, tedy její hodnost se rovná její stopě.

Tudíž

$$\begin{aligned}
r &= h(\mathsf{var}\,\mathbf{Z}) \\
&= \mathsf{tr}(\mathsf{var}\,\mathbf{Z}) \\
&= \mathsf{tr}(\mathbf{I}_k - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}') \\
&= \mathsf{tr}(\mathbf{I}_k) - \frac{1}{n}\,\mathsf{tr}(\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}') \\
&= k - \frac{1}{n}\,\mathsf{tr}(\sqrt{\mathbf{n}}'\sqrt{\mathbf{n}}) \\
&= k - \frac{1}{n}n \\
&= k - 1.
\end{aligned}$$

Zde jsme ještě využili toho, že $\mathsf{tr}(AB) = \mathsf{tr}(BA)$ pro každé dvě matice A, B.

Ještě bychom měli ověřit, zda použití obou vět bylo legální, tj. zda byly splněny jejich předpoklady.

Ověření předpokladů věty 2: Víme, že $Y_{ip} \sim N(\mu_i, \sigma^2)$. Dle lemmatu 1 a z nezávislosti Y_{ip} má

$$X_i = \frac{y_{i.}}{\sigma} = \frac{1}{n_i \sigma} \sum_{p=1}^{n_i} Y_{ip}$$

rozdělení $N(\frac{1}{n_i\sigma} n_i \mu_i, \frac{1}{n_i^2 \sigma^2} n_i \sigma^2)$, což po úpravě dává $N(\frac{\mu_i}{\sigma}, \frac{1}{n_i})$. Platí-li $\mu_1 = \ldots = \mu_k$, pak $X_i \sim N(\frac{\mu}{\sigma}, \frac{1}{n_i})$ a $\mathbf{X} \sim N(\frac{\mu}{\sigma}\mathbf{1}, \mathsf{diag}\,(\frac{1}{n_1}, \ldots, \frac{1}{n_k}))$, neboť složky $\mathbf{X}$ jsou nezávislé náhodné veličiny.

Ověření předpokladů věty 1: Víme, že

$$\mathbf{Z} = \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{X} \sim N(\mathbf{0}, \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}\,').$$

Tedy $\mathbf{A} = \mathsf{var}\,\mathbf{Z} = \mathbf{I} - \frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}\,'$. Matice $\mathbf{A}$ je zřejmě symetrická a dle lemmatu 3 je také idempotentní, neboť $\frac{1}{n}\sqrt{\mathbf{n}}\sqrt{\mathbf{n}}\,'$ je idempotentní.

Nyní se podíváme na druhou část důkazu, a to na tvrzení $\frac{S_e}{\sigma^2} \sim \chi^2_{n-k}$. Zde stačí ukázat, že

$$\frac{1}{\sigma^2} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 \sim \chi^2_{n_i - 1}.$$

Pak z lemmatu 4 dostáváme, že

$$\begin{aligned} \frac{S_e}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 &\sim \chi^2_{n_1 - 1 + n_2 - 1 + \ldots + n_k - 1} \\ &\sim \chi^2_{n_1 + n_2 + \ldots + n_k - k} \\ &\sim \chi^2_{n-k}. \end{aligned}$$

Přitom je zde důležité, že veličiny Y_{ip}, a tedy i veličiny $\frac{1}{\sigma^2} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2$, jsou nezávislé, tudíž použití lemmatu 4 je oprávněné.

To, že

$$\frac{1}{\sigma^2} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 \sim \chi^2_{n_i - 1}$$

plyne z věty o vlastnostech náhodného výběru z $N(\mu_i, \sigma^2)$ [2, věta 4.21 b)]. Ta nám pro naši konkrétní situaci dává, že

$$\frac{(n_i - 1) S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}, \tag{2.5}$$

kde

$$S_i^2 = \frac{1}{n_i - 1} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2. \tag{2.6}$$

Potom

$$\frac{(n_i - 1)S_i^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 \sim \chi^2_{n_i - 1},$$

což je přesně výraz, který jsme potřebovali.

Zbývá už jen dokázat nezávislost S_A a S_e. Veličiny $Y_{i1}, \ldots, Y_{in_i}$ mají rozdělení $N(\mu_i, \sigma^2)$ a jsou mezi sebou nezávislé pro všechna $i = 1, \ldots, k$. Z vlastnosti náhodného výběru z $N(\mu_i, \sigma^2)$ [2, věta 4.21 c)] máme, že $y_{i.}$ a S_i^2 jsou nezávislé. Navíc dle [2, str. 33] i jejich měřitelné funkce jsou nezávislé. Tedy S_A a S_e jsou nezávislé, neboť $S_A = \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2$ je funkcí průměrů $y_{1.}, \ldots, y_{k.}$ a $S_e = \sum_{p=1}^{n_i} (Y_{ip} - y_{i.})^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$ je funkcí S_i.
□

Dokázali jsme, že statistika

$$F = \frac{\dfrac{S_A}{k-1}}{\dfrac{S_e}{n-k}},$$

pomocí které se analýza rozptylu jednoduchého třídění provádí, má za platnosti hypotézy H_0 rozdělení $F_{k-1,n-k}$. Jednoduché třídění tedy patří do kategorie tzv. F-testů a jeho kritický obor má tvar:

$$\{F \geq F_{k-1,n-k}(\alpha)\}.$$

Překročí-li tudíž veličina F kritickou hodnotu $F_{k-1,n-k}(\alpha)$, pak zamítáme na hladině α hypotézu o rovnosti středních hodnot.

Kapitola 3

Jednoduché třídění pro případ nestejných rozptylů

3.1 Momenty S_A a S_e

Uvolněme si nyní požadavky modelu jednoduchého třídění a uvažujme v jednotlivých výběrech různé rozptyly. Předpokládejme tedy, že

$$\mathbf{Y} \sim N\left(\begin{pmatrix}\mu_1\mathbf{1}_{n_1}\\ \dots\\ \dots\\ \mu_k\mathbf{1}_{n_k}\end{pmatrix}, \begin{pmatrix}\sigma_1^2\mathbf{I}_{n_1} & 0 & \dots & 0\\ \dots & \sigma_2^2\mathbf{I}_{n_2} & \dots & \dots\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \sigma_k^2\mathbf{I}_{n_k}\end{pmatrix}\right), \tag{3.1}$$

kde $\mathbf{1}_{n_i}$ je vektor samých jedniček o délce n_i, $\mathbf{I}_{n_i}$ je jednotková matice typu $n_i \times n_i$ a $\mathbf{Y}$ je jako v (2.1). V dalším budeme varianční matici $\mathbf{Y}$ značit $\boldsymbol{V}$.

Nejprve se podíváme na rozdělení veličiny S_A za hypotézy H_0: $\mathsf{E}\,\mathbf{Y} = \mu\mathbf{1}$ (kde $\mathbf{1}$ je stále jednotkový vektor délky n). Roznásobením lze ověřit, že

$$S_A = \mathbf{Y}'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{Y} = \left\|(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{Y}\right\|^2, \tag{3.2}$$

kde $\mathbf{H}_1 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathsf{diag}\,(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}'_{n_1}, \dots, \frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}'_{n_k})$ a $\mathbf{H}_0 = \frac{1}{n}\mathbf{1}\mathbf{1}'$. Bez újmy na obecnosti můžeme předpokládat, že $\mu = 0$. (To lze, neboť $(\mathbf{H}_1 - \mathbf{H}_0)\mu\mathbf{1} = \mathbf{0}$ pro všechna μ.) Potom $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{V})$ a $\mathbf{Z} = \boldsymbol{V}^{-1/2}\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$. Pak

$$\begin{aligned} S_A &= \mathbf{Y}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\mathbf{Y}\\ &= \mathbf{Z}'\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}\mathbf{Z} = \mathbf{Z}'\mathbf{A}\mathbf{Z}, \end{aligned} \tag{3.3}$$

kde $\mathbf{A} = \boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}$.

Věta 4 Nechť $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ má konečné druhé momenty a nechť $\mathsf{E}\,\mathbf{Z} = \boldsymbol{\mu}$, $\mathsf{var}\,\mathbf{Z} = \mathbf{V}$. Pak pro libovolnou matici $\mathbf{C}_{n\times n}$ platí $\mathsf{E}\,\mathbf{Z}'\mathbf{C}\mathbf{Z} = \mathsf{tr}(\mathbf{C}\mathbf{V}) + \boldsymbol{\mu}'\boldsymbol{C}\boldsymbol{\mu}$. [2, věta 4.18]

Lemma 5 Vlastní čísla $\lambda_1, \ldots, \lambda_n$ symetrické matice $\mathbf{C}_{n\times n}$ jsou vždy reálná. Bez újmy na obecnosti předpokládejme, že $\lambda_1 \geq \ldots \geq \lambda_n$. Položme $\boldsymbol{\Lambda} = \mathsf{diag}\,(\lambda_1, \ldots, \lambda_n)$. Pak existuje taková matice $\mathbf{Q}_{n\times n}$, že platí

$$\mathbf{C} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}', \qquad \mathbf{I} = \mathbf{Q}\mathbf{Q}'.$$

viz [2, věta A.6]

Věta 5 Nechť $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ a nechť $\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{I})$. Pak pro libovolnou symetrickou matici $\mathbf{C}_{n\times n}$ platí $\mathsf{var}\,\mathbf{Z}'\mathbf{C}\mathbf{Z} = 2\,\mathsf{tr}(\mathbf{C}^2)$.

Důkaz: K matici $\mathbf{C}$ najdeme z lemmatu 5 příslušné matice $\mathbf{Q}$ a $\boldsymbol{\Lambda}$ tak, že $\mathbf{C} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$. Pak

$$\mathbf{Z}'\mathbf{C}\mathbf{Z} = (\mathbf{Z}'\mathbf{Q})\boldsymbol{\Lambda}(\mathbf{Q}'\mathbf{Z}) = \mathbf{U}'\boldsymbol{\Lambda}\boldsymbol{U} = \sum_{i=1}^{n} \lambda_i U_i^2,$$

kde $\mathbf{U} = \mathbf{Q}'\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$.

Pro výpočet rozptylu budeme potřebovat $\mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z})$ a $\mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^2$.

$$\begin{aligned}
\mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z}) &= \mathsf{E}\left(\sum_{i=1}^{n} \lambda_i U_i^2\right) = \sum_{i=1}^{n} \lambda_i\,\mathsf{E}\,U_i^2 = \sum_{i=1}^{n} \lambda_i \\
\mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^2 &= \mathsf{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j U_i^2 U_j^2\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\,\mathsf{E}\,U_i^2U_j^2 \\
&= \sum_{i=1}^{n} \lambda_i^2\,\mathsf{E}\,U_i^4 + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \lambda_i\lambda_j\,\mathsf{E}\,U_i^2U_j^2 \\
&= 3\sum_{i=1}^{n} \lambda_i^2 + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \lambda_i\lambda_j \\
&= 2\sum_{i=1}^{n} \lambda_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j = 2\sum_{i=1}^{n} \lambda_i^2 + \left(\sum_{i=1}^{n} \lambda_i\right)^2.
\end{aligned}$$

Zde jsme využili znalosti momentů normovaného normálního rozdělení - $\mathsf{E}\, U_i^4 = 3$, $\mathsf{E}\, U_i^2 = 1$.

Konečně

$$\begin{aligned}
\mathsf{var}(\mathbf{Z}'\mathbf{C}\mathbf{Z}) &= \mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^2 - [\mathsf{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z})]^2 \\
&= 2\sum_{i=1}^{n} \lambda_i^2 + \left(\sum_{i=1}^{n} \lambda_i\right)^2 - \left(\sum_{i=1}^{n} \lambda_i\right)^2 \\
&= 2\sum_{i=1}^{n} \lambda_i^2 = 2\ \mathsf{tr}\,\mathbf{C}^2.
\end{aligned}$$

□

Střední hodnotu S_A vypočteme z věty 4, přičemž si uvědomíme, že $\mathsf{E}\,\mathbf{Z} = \boldsymbol{\mu} = \mathbf{0}$ a $\mathsf{var}\,\mathbf{Z} = \mathbf{V} = \boldsymbol{I}$. Pak tedy

$$\begin{aligned}
\mathsf{E}\, S_A &= \mathsf{E}\,\mathbf{Z}'\mathbf{A}\mathbf{Z} = \mathsf{tr}\,\mathbf{A} = \mathsf{tr}[\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}] \\
&= \mathsf{tr}[\boldsymbol{V}(\mathbf{H}_1 - \mathbf{H}_0)] = \mathsf{tr}(\mathbf{V}\mathbf{H}_1) - \mathsf{tr}(\mathbf{V}\mathbf{H}_0) \\
&= \sum_{i=1}^{k} \sigma_i^2 - \frac{1}{n}\sum_{i=1}^{k} n_i\sigma_i^2 \\
&= k\overline{\sigma^2} - \overline{\sigma_n^2},
\end{aligned}$$

kde $\overline{\sigma^2} = \frac{1}{k}\sum_{i=1}^{k} \sigma_i^2$ a $\overline{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{k} n_i\sigma_i^2$.

Podle věty 5 vypočteme rozptyl S_A. Pro zjednodušení zápisu budeme počítat jeho polovinu. Je to

$$\begin{aligned}
\frac{1}{2}\,\mathsf{var}\, S_A &= \frac{1}{2}\,\mathsf{var}\,\mathbf{Z}'\mathbf{A}\mathbf{Z} = \mathsf{tr}\,\mathbf{A}^2 \\
&= \mathsf{tr}[\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}] \\
&= \mathsf{tr}[\mathbf{V}(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{V}(\mathbf{H}_1 - \mathbf{H}_0)] \\
&= \mathsf{tr}[(\mathbf{V}(\mathbf{H}_1 - \mathbf{H}_0))^2] \\
&= \mathsf{tr}[(\mathbf{V}\mathbf{H}_1 - \mathbf{V}\mathbf{H}_0)^2] \\
&= \mathsf{tr}[(\mathbf{V}\mathbf{H}_1)^2] - 2\,\mathsf{tr}[\mathbf{V}\mathbf{H}_1\mathbf{V}\mathbf{H}_0] + \mathsf{tr}[(\mathbf{V}\mathbf{H}_0)^2] \\
&= \sum_{i=1}^{k} \sigma_i^4 - \frac{2}{n}\sum_{i=1}^{k} n_i\sigma_i^4 + \left(\frac{1}{n}\sum_{i=1}^{k} n_i\sigma_i^2\right)^2 \\
&= k\overline{\sigma^4} - 2\overline{\sigma_n^4} + \overline{\sigma_n^2}^2,
\end{aligned}$$

kde $\overline{\sigma^4} = \frac{1}{k}\sum_{i=1}^k \sigma_i^4$ a $\overline{\sigma_n^4} = \frac{1}{n}\sum_{i=1}^k n_i\sigma_i^4$.

Z našich výsledků vidíme, že pokud $\sigma_i^2 = \sigma^2$ pro všechna $i = 1, \ldots, k$, pak

$$\begin{aligned} \mathsf{E}\, S_A &= k\sigma^2 - \frac{1}{n}\sigma^2 \sum_{i=1}^k n_i = k\sigma^2 - \frac{1}{n}\sigma^2 n = \sigma^2(k-1) \qquad \text{a} \\ \frac{1}{2}\,\mathsf{var}\, S_A &= \sum_{i=1}^k \sigma^4 - \frac{2}{n}\sum_{i=1}^k n_i\sigma^4 + \left(\frac{1}{n}\sum_{i=1}^k n_i\sigma^2\right)^2 \\ &= k\sigma^4 - 2\sigma^4 + \sigma^4 = \sigma^4(k-1), \end{aligned}$$

což přesně odpovídá vztahům pro střední hodnotu a rozptyl rozdělení $\sigma^2\chi^2_{k-1}$.

V obecném případě nic takového říct nelze. Muselo by totiž platit

$$(k\overline{\sigma^2} - \overline{\sigma_n^2})^2 = (k-1)(k\overline{\sigma^4} - 2\overline{\sigma_n^4} + \overline{\sigma_n^2}^2),$$

a to obecně neplatí.

Pokud mají všechny výběry stejný rozsah, tj. $n_i = m$ pro všechna $i = 1, \ldots, k$, potom

$$\begin{aligned} \mathsf{E}\, S_A &= \sum_{i=1}^k \sigma_i^2 - \frac{1}{n}\sum_{i=1}^k m\sigma_i^2 = \sum_{i=1}^k \sigma_i^2 - \frac{1}{k}\sum_{i=1}^k \sigma_i^2 \\ &= (k-1)\frac{1}{k}\sum_{i=1}^k \sigma_i^2 = (k-1)\overline{\sigma^2}, \\ \frac{1}{2}\,\mathsf{var}\, S_A &= \sum_{i=1}^k \sigma_i^4 - \frac{2}{n}m\sum_{i=1}^k \sigma_i^4 + \left(\frac{m}{n}\sum_{i=1}^k \sigma_i^2\right)^2 \\ &= \sum_{i=1}^k \sigma_i^4 - \frac{2}{k}\sum_{i=1}^k \sigma_i^4 + \left(\frac{1}{k}\sum_{i=1}^k \sigma_i^2\right)^2 \\ &= k\overline{\sigma^4} - 2\overline{\sigma^4} + \overline{\sigma^2}^2. \end{aligned}$$

Střední hodnota vypadá, že pochází z rozdělení $\overline{\sigma^2}\chi^2_{k-1}$, avšak rozptyl takové známky nevykazuje. Muselo by platit

$$\left[(k-1)\overline{\sigma^2}\right]^2 = (k-1)\left(k\overline{\sigma^4} - 2\overline{\sigma^4} + \overline{\sigma^2}^2\right), \tag{3.4}$$

to jest

$$(k-1)\overline{\sigma^2}^2 = k\overline{\sigma^4} - 2\overline{\sigma^4} + \overline{\sigma^2}^2. \tag{3.5}$$

Platí ale pouze

$$(k-1)\overline{\sigma^2}^2 \leq k\overline{\sigma^4} - 2\overline{\sigma^4} + \overline{\sigma^2}^2 \tag{3.6}$$

Tato nerovnost je založena na nerovnosti mezi průměry: $(\overline{\sigma^2})^2 \leq \overline{\sigma^4}$, kterou lze ukázat třeba takto:

$$0 \leq \frac{1}{k}\sum_{i=1}^{k}(\sigma_i^2 - \overline{\sigma^2})^2 = \frac{1}{k}\left(\sum_{i=1}^{k}\sigma_i^4 - k(\overline{\sigma^2})^2\right) = \overline{\sigma^4} - (\overline{\sigma^2})^2.$$

Výraz (3.6) vznikne z této nerovnosti vynásobením číslem $(k-2)$ (pokud $k > 2$) a jednoduchou úpravou. Pro $k = 1$ a $k = 2$ je platnost nerovnosti zřejmá a v těchto dvou případech je v ní dokonce dosaženo rovnosti. Obecně pak v (3.6) nastává rovnost pouze pokud $\sigma_i^2 = \sigma^2$ pro všechna $i = 1, \ldots, k$, což už jsme vyřešili výše.

Stejně budeme postupovat i v případě rozdělení veličiny S_e. Opět lze výpočtem ověřit, že

$$S_e = \mathbf{Y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{Y} = \left\|(\mathbf{I} - \mathbf{H}_1)\mathbf{Y}\right\|^2,$$

kde $\mathbf{H}_1$ je jako v (3.2). Zde není zapotřebí brát ohled na hypotézu H_0 o rovnosti μ_i, protože $(\mathbf{I} - \mathbf{H}_1)(\mu_1\mathbf{1}_{n_1}, \ldots, \mu_k\mathbf{1}_{n_k})' = \mathbf{0}$. Proto můžeme i zde předpokládat, že $\boldsymbol{\mu} = \mathbf{0}$. Pak $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{V})$, $\mathbf{Z} = \mathbf{V}^{-1/2}\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$ a

$$\begin{aligned} S_e &= \mathbf{Y}'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}(\mathbf{I} - \mathbf{H}_1)\mathbf{V}^{1/2}\mathbf{V}^{-1/2}\mathbf{Y} \\ &= \mathbf{Z}'\mathbf{V}^{1/2}(\mathbf{I} - \mathbf{H}_1)\mathbf{V}^{1/2}\mathbf{Z} = \mathbf{Z}'\mathbf{B}\mathbf{Z}, \end{aligned} \tag{3.7}$$

kde $\mathbf{B} = \mathbf{V}^{1/2}(\mathbf{I} - \mathbf{H}_1)\mathbf{V}^{1/2}$.

Střední hodnotu a rozptyl spočítáme opět z vět 4 a 5.

$$\begin{aligned} \mathsf{E}\, S_e &= \mathsf{E}\, \mathbf{Z}'\mathbf{B}\mathbf{Z} = \mathsf{tr}\, \mathbf{B} = \mathsf{tr}[\mathbf{V}^{1/2}(\mathbf{I} - \mathbf{H}_1)\mathbf{V}^{1/2}] \\ &= \mathsf{tr}[\mathbf{V}(\mathbf{I} - \mathbf{H}_1)] = \mathsf{tr}(\mathbf{V}) - \mathsf{tr}(\mathbf{V}\mathbf{H}_1) \\ &= \sum_{i=1}^{k} n_i\sigma_i^2 - \sum_{i=1}^{k}\sigma_i^2 = n\overline{\sigma_n^2} - k\overline{\sigma^2}. \end{aligned}$$

Rozptyl splňuje

$$\begin{aligned}
\frac{1}{2}\,\mathsf{var}\,S_e &= \frac{1}{2}\,\mathsf{var}\,\mathbf{Z}'\mathbf{B}\mathbf{Z} = \mathsf{tr}\,\mathbf{B}^2 \\
&= \mathsf{tr}[\mathbf{V}^{1/2}(\mathbf{I}-\mathbf{H}_1)\mathbf{V}^{1/2}\mathbf{V}^{1/2}(\mathbf{I}-\mathbf{H}_1)\mathbf{V}^{1/2}] \\
&= \mathsf{tr}[\mathbf{V}(\mathbf{I}-\mathbf{H}_1)\mathbf{V}(\mathbf{I}-\mathbf{H}_1)] \\
&= \mathsf{tr}[\mathbf{V}(\mathbf{I}-\mathbf{H}_1)]^2 = \mathsf{tr}(\mathbf{V}-\mathbf{V}\mathbf{H}_1)^2 \\
&= \mathsf{tr}(\mathbf{V})^2 - 2\,\mathsf{tr}(\mathbf{V}^2\mathbf{H}_1) + \mathsf{tr}(\mathbf{V}\mathbf{H}_1)^2 \\
&= \sum_{i=1}^{k} n_i\sigma_i^4 - 2\sum_{i=1}^{k}\sigma_i^4 + \sum_{i=1}^{k}\sigma_i^4 \\
&= \sum_{i=1}^{k} n_i\sigma_i^4 - \sum_{i=1}^{k}\sigma_i^4 = n\overline{\sigma_n^4} - k\overline{\sigma^4}.
\end{aligned}$$

Z těchto hodnot je patrné, že pokud $\sigma_i^2 = \sigma^2$ pro $i = 1, \ldots, k$, pak

$$\begin{aligned}
\mathsf{E}\,S_e &= \sigma^2\sum_{i=1}^{k} n_i - \sigma^2 k = \sigma^2(n-k) \qquad \text{a} \\
\frac{1}{2}\,\mathsf{var}\,S_e &= \sigma^4\sum_{i=1}^{k} n_i - \sigma^4 k = \sigma^4(n-k),
\end{aligned}$$

což odpovídá rozdělení $\sigma^2\chi^2_{n-k}$.

Obecně ale nic takového neplatí. Musela by být opět splněna rovnost

$$(n\overline{\sigma_n^2} - k\overline{\sigma^2})^2 = (n-k)(n\overline{\sigma_n^4} - k\overline{\sigma^4}).$$

Platí ale pouze

$$(n\overline{\sigma_n^2} - k\overline{\sigma^2})^2 \geq (n-k)(n\overline{\sigma_n^4} - k\overline{\sigma^4}).$$

Tuto nerovnost lze snadno získat z následující nerovnosti: pro kladná čísla $c_1, \ldots, c_k$ platí:

$$c\sum_{i=1}^{k} c_i x_i^2 \geq (\sum_{i=1}^{k} c_i x_i)^2, \tag{3.8}$$

kde $c = \sum_{i=1}^{k} c_i$, přičemž rovnost nastává jen tehdy, jsou-li všechna čísla x_i stejná.

Důkaz: Označme nejprve $\overline{x_c} = \frac{1}{c}\sum_{i=1}^{k} c_i x_i$. Pak

$$\begin{aligned} 0 \le c\sum_{i=1}^{k} c_i(x_i - \overline{x_c})^2 &= c\sum_{i=1}^{k} c_i x_i^2 - 2c\overline{x_c}\sum_{i=1}^{k} c_i x_i + c^2\overline{x_c}^2 \\ &= c\sum_{i=1}^{k} c_i x_i^2 - \left(\sum_{i=1}^{k} c_i x_i\right)^2. \end{aligned}$$

Tedy $c\sum_{i=1}^{k} c_i x_i^2 \ge \left(\sum_{i=1}^{k} c_i x_i\right)^2$, přičemž tvrzení o rovnosti je zřejmé. □

V této nerovnosti stačí nyní vzít $c_i = n_i - 1$ a $x_i = \sigma_i^2$. Potom

$$\begin{aligned} c &= n - k, \\ c\sum_{i=1}^{k} c_i x_i^2 &= (n-k)\sum_{i=1}^{k}(n_i - 1)\sigma_i^4 = (n-k)\left(\sum_{i=1}^{k} n_i\sigma_i^4 - \sum_{i=1}^{k}\sigma_i^4\right) \\ &= (n-k)(n\overline{\sigma_n^4} - k\overline{\sigma^4}), \\ \left(\sum_{i=1}^{k} c_i x_i\right)^2 &= \left[\sum_{i=1}^{k}(n_i-1)\sigma_i^2\right]^2 = \left(\sum_{i=1}^{k} n_i\sigma_i^2 - \sum_{i=1}^{k}\sigma_i^2\right)^2 \\ &= (n\overline{\sigma_n^2} - k\overline{\sigma^2})^2. \end{aligned}$$

Tedy $(n\overline{\sigma_n^2} - k\overline{\sigma^2})^2 \ge (n-k)(n\overline{\sigma_n^4} - k\overline{\sigma^4})$. Navíc rovnost nastává jen tehdy, jsou-li $\sigma_i^2 = \sigma^2$ pro všechna $i = 1, \ldots, k$, což je případ diskutovaný výše.

Pokud mají všechny výběry stejný rozsah, tj. $n_i = m$ pro všechna $i = 1, \ldots, k$, potom

$$\begin{aligned} \mathsf{E}\, S_e &= m\sum_{i=1}^{k}\sigma_i^2 - \sum_{i=1}^{k}\sigma_i^2 = mk\overline{\sigma^2} - k\overline{\sigma^2} = (n-k)\overline{\sigma^2} \qquad \text{a} \\ \frac{1}{2}\,\mathsf{var}\, S_e &= \sum_{i=1}^{k} m\sigma_i^4 - \sum_{i=1}^{k}\sigma_i^4 = (m-1)\sum_{i=1}^{k}\sigma_i^4 \\ &= (n-k)\frac{1}{k}\sum_{i=1}^{k}\sigma_i^4 = (n-k)\overline{\sigma^4}. \end{aligned}$$

Vypadá to jako střední hodnota a rozptyl rozdělení $\overline{\sigma^2}\chi^2_{n-k}$, ale i zde narážíme na to, že $(\overline{\sigma^2})^2 \leq \overline{\sigma^4}$. Rovnost nastává pouze v případě $\sigma_i^2 = \sigma^2$ pro $i = 1, \ldots, k$.

Z těchto výsledků je patrné, že veličiny S_A a S_e už nemají šanci vytvořit náhodnou veličinu s F-rozdělením. Přesto může být zajímavé zjistit, jak je to s jejich nezávislostí. Ta se snadno ukáže z následující věty.

Věta 6 Nechť $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{V})$ a nechť $\mathbf{F}_{n\times n} \geq 0$, $\mathbf{G}_{n\times n} \geq 0$. Nechť $\mathbf{c} \in \mathbb{R}_n, \mathbf{d} \in \mathbb{R}_n$. Platí-li $\mathbf{GVF} = \mathbf{0}$, jsou náhodné veličiny $(\mathbf{Y}-\mathbf{c})'\mathbf{F}(\mathbf{Y}-\mathbf{c})$ a $(\mathbf{Y}-\mathbf{d})'\mathbf{G}(\mathbf{Y}-\mathbf{d})$ nezávislé. [2, věta 4.20]

Z předchozího víme, že $S_A = \mathbf{Y}'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{Y}$ a $S_e = \mathbf{Y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{Y}$. Tedy stačí ve větě 5 vzít $\mathbf{F} = (\mathbf{H}_1 - \mathbf{H}_0)$, $\mathbf{G} = (\mathbf{I} - \mathbf{H}_1)$ a $\mathbf{c} = \mathbf{d} = \mathbf{0}$.

Je třeba jen ukázat, že matice $\mathbf{F}$ a $\mathbf{G}$ jsou pozitivně semidefinitní. To však dle [2, věta A.14] plyne triviálně z toho, že jsou symetrické a idempotentní.

Dále je třeba ještě ověřit vztah: $\mathbf{GVF} = \mathbf{0}$. Roznásobíme:

$$\begin{aligned}\mathbf{GVF} &= (\mathbf{I} - \mathbf{H}_1)\,\mathbf{V}\,(\mathbf{H}_1 - \mathbf{H}_0) \\ &= \mathbf{VH}_1 - \mathbf{VH}_0 - \mathbf{H}_1\mathbf{VH}_1 + \mathbf{H}_1\mathbf{VH}_0 \\ &= \mathbf{0}.\end{aligned}$$

Platí totiž $\mathbf{VH}_1 = \mathbf{H}_1\mathbf{VH}_1$ a $\mathbf{VH}_0 = \mathbf{H}_1\mathbf{VH}_0$. Veličiny S_A a S_e jsou tedy nezávislé.

3.2 Testování hypotézy o rovnosti středních hodnot postupem B. L. Welche

V této podkapitole uvedeme modifikaci jednoduchého třídění v případě nestejných rozptylů uvedenou v [4].
Nechť Y_{ip} jsou jako v (3.1), tj. $Y_{ip} \sim N(\mu_i, \sigma_i^2)$. Připomeňme si, že

$$y_{i\cdot} = \frac{1}{n_i}\sum_{p=1}^{n_i} Y_{ip}.$$

Dle [2, věta 4.21 a)] má $y_{i\cdot}$ rozdělení $N(\mu_i, \frac{\sigma_i^2}{n_i})$. Dále uvažujme statistiku

$$v^2 = \sum_{i=1}^{k} w_i(y_{i\cdot} - \hat{y})^2,$$

kde

$$w_i = \frac{n_i}{S_i^2}\,, \qquad \text{(podobně budeme značit } \omega_i = \frac{n_i}{\sigma_i^2}\text{)}$$
$$\hat{y} = \frac{\sum_{i=1}^{k} w_i y_{i\cdot}}{\sum_{i=1}^{k} w_i}.$$

Veličiny S_i^2 jsou definované v (2.6) a dle (2.5) mají rozdělení $\frac{\sigma_i^2}{n_i-1}\chi^2_{n_i-1}$. Příslušné stupně volnosti budeme značit f_i (tj. $f_i = n_i - 1$).

Testujeme hypotézu $H_0 : \mu_1 = \ldots = \mu_k$ a bez újmy na obecnosti budeme předpokládat, že $\mu_i = 0$. V [4] je popsáno podrobné odvození aproximace momentové vytvořující funkce $M(u)$ a kumulantové vytvořující funkce $K(u)$ veličiny v^2, obojí do řádu $1/f_i$. Nám bude stačit jen jeho výsledek:

$$M(u) = (1-2u)^{-\frac{1}{2}(k-1)} \left\{ 1 + \left[2u(1-2u)^{-1} + 3u^2(1-2u)^{-2}\right] \left[\sum_{i=1}^{k} \frac{1}{f_i} \left(1 - \frac{\omega_i}{\sum_{i=1}^{k} \omega_i} \right)^2 \right] \right\} \quad (3.9)$$

$$K(u) = -\frac{1}{2}(k-1)\log(1-2u) + \left[2u(1-2u)^{-1} + 3u^2(1-2u)^{-2}\right] \left[\sum_{i=1}^{k} \frac{1}{f_i} \left(1 - \frac{\omega_i}{\sum_{i=1}^{k} \omega_i} \right)^2 \right]. \quad (3.10)$$

Připomeňme ještě, že momentová vytvořující funkce nějaké náhodné veličiny T je definovaná předpisem $M(u) = \mathsf{E}\, e^{uT}$, zatímco kumulantová vytvořující funkce je definovaná jako $K(u) = \log M(u)$. [5, str. 114, 115]

Dále budeme potřebovat veličinu

$$F = \frac{\frac{\chi_1^2}{\hat{f}_1}}{\frac{\chi_2^2}{\hat{f}_2}},$$

kde χ_1^2, χ_2^2 jsou nezávislé a mají stupně volnosti $\hat{f}_1$, resp. $\hat{f}_2$. Její momentová vytvořující funkce je opět odvozena v [4].

Ve svém článku si Welch na tomto místě pokládá $\hat{f}_1 = k-1$ a dále pracuje s veličinou $G = [(k-1)+\frac{A}{\hat{f}_2}]F$, kde A je zatím blíže neurčený výraz. Pro veličinu G obdržíme

$$M_G(u) = (1-2u)^{-\frac{1}{2}(k-1)}\left[1+\frac{A+2(k-1)}{\hat{f}_2}u(1-2u)^{-1}+\frac{k^2-1}{\hat{f}_2}u^2(1-2u)^{-2}\right] \tag{3.11}$$

$$K_G(u) = -\frac{1}{2}(k-1)log(1-2u)+\frac{1}{\hat{f}_2}[A+2(k-1)]u(1-2u)^{-1}+\frac{k^2-1}{\hat{f}_2}u^2(1-2u)^{-2}. \tag{3.12}$$

Zdůrazněme, že funkce $M_G(u)$ a $K_G(u)$ jsou opět pouze aproximace momentové a kumulantové vytvořující funkce veličiny G, a to do řádu $1/\hat{f}_2$.

Porovnáme-li kumulantové funkce $K(u)$ a $K_G(u)$, zjistíme, že je lze upravit tak, aby si byly rovny. Musí platit

$$\frac{1}{\hat{f}_2} = \frac{3}{k^2-1}\sum_{i=1}^{k}\frac{1}{f_i}\left(1-\frac{\omega_i}{\sum_{i=1}^{k}\omega_i}\right)^2 \quad \text{a} \tag{3.13}$$

$$\frac{A}{\hat{f}_2} = \frac{2(k-2)}{k+1}\sum_{i=1}^{k}\frac{1}{f_i}\left(1-\frac{\omega_i}{\sum_{i=1}^{k}\omega_i}\right)^2. \tag{3.14}$$

Z toho vyplývá, že veličina v^2 má rozdělení $\left[(k-1)+\frac{A}{\hat{f}_2}\right]F$, kde $\hat{f}_1 = k-1$ a $\hat{f}_2$ a $\frac{A}{\hat{f}_2}$ jsou jako v (3.13) a (3.14). Ze vztahu

$$v^2 \doteq \left[(k-1)+\frac{A}{\hat{f}_2}\right]F$$

dostáváme

$$F \doteq \frac{\dfrac{v^2}{k-1}}{1+\dfrac{A}{\hat{f}_2}\dfrac{1}{k-1}}. \tag{3.15}$$

Test hypotézy H_0 založený na tomto výsledku bude vypadat následovně. Do výrazu (3.15) dosadíme za v^2, $\hat{f}_1$ a $\frac{A}{\hat{f}_2}$. Tak získáme veličinu

$$F = \frac{\dfrac{\sum_{i=1}^{k}w_i(y_{i.}-\hat{y})^2}{k-1}}{1+\dfrac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\dfrac{1}{f_i}\left(1-\dfrac{\omega_i}{\sum_{i=1}^{k}\omega_i}\right)^2},$$

do které ještě za neznámé veličiny ω_i dosadíme jejich nestranné odhady w_i. Dále si spočteme stupně volnosti

$$\begin{aligned} \hat{f}_1 &= k-1, \\ \hat{f}_2 &= \left[\frac{3}{k^2-1} \sum_{i=1}^{k} \frac{1}{f_i} \left(1 - \frac{w_i}{\sum_{i=1}^{k} w_i} \right)^2 \right]^{-1}. \end{aligned}$$

Kritický obor pak bude mít tvar $\{F > F_{\hat{f}_1,\hat{f}_2}(\alpha)\}$, kde $F_{\hat{f}_1,\hat{f}_2}(\alpha)$ je kritická hodnota F-rozdělení o stupních volnosti $\hat{f}_1$ a $\hat{f}_2$ na hladině testu α. Pokud tedy bude platit $F > F_{\hat{f}_1,\hat{f}_2}(\alpha)$, pak zamítáme hypotézu H_0 a rovnosti středních hodnot na hladině α.

3.3 Testování hypotézy o rovnosti středních hodnot postupem G. E. P. Boxe

Věta 7 Nechť Q', resp. Q jsou nezávislé kvadratické formy s rozdělením $\sum_{j=1}^{r'} \lambda'_j \chi^2_{\nu'_j}$, resp. $\sum_{j=1}^{r} \lambda_j \chi^2_{\nu_j}$. Potom veličina Q'/Q má aproximativně rozdělení $bF_{h',h}$, kde

$$b = \frac{\mathsf{E}(Q')}{\mathsf{E}(Q)} \tag{3.16}$$

$$h' = \frac{2[\mathsf{E}(Q')]^2}{\mathsf{var}(Q')} \tag{3.17}$$

$$h = \frac{2[\mathsf{E}(Q)]^2}{\mathsf{var}(Q)}. \tag{3.18}$$

[3, věta 3.1 a 6.1]

Použijeme model (3.1) a vyšetříme opět rozdělení S_A a S_e. Za platnosti hypotézy H_0 o rovnosti středních hodnot můžeme podle (3.3) kvadratickou formu S_A zapsat ve tvaru

$$S_A = \mathbf{Z}'\mathbf{A}\mathbf{Z},$$

kde $\mathbf{A} = \boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}$ a $\mathbf{Z} = \boldsymbol{V}^{-1/2}\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$. Z lemmatu 5 k matici $\mathbf{A}$ existuje matice $\mathbf{Q}$ tak, že $\mathbf{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ a $\mathbf{I} = \mathbf{Q}\mathbf{Q}'$, přičemž $\boldsymbol{\Lambda} = \mathsf{diag}\,(\lambda_1, \ldots, \lambda_n)$, kde $\lambda_1 \geq \ldots \geq \lambda_n$ jsou vlastní čísla matice $\mathbf{A}$.

Potom

$$S_A = \mathbf{Z}'\mathbf{A}\mathbf{Z} = \mathbf{Z}'\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'\mathbf{Z} = \mathbf{U}'\boldsymbol{\Lambda}\boldsymbol{U} = \sum_{i=1}^{k-1} \lambda_i U_i^2, \tag{3.19}$$

kde $\mathbf{U} = \mathbf{Q}'\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$. Suma ve výrazu (3.19) má pouze $k-1$ sčítanců, neboť nenulových čísel matice $\mathbf{A}$ je pouze $k-1$. To si můžeme ověřit tak, že si uvědomíme, že počet nenulových vlastních čísel každé matice je roven její hodnosti. Matice $\mathbf{V}^{1/2}$ je navíc regulární, tedy

$$h(\mathbf{A}) = h[\boldsymbol{V}^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\boldsymbol{V}^{1/2}] = h(\mathbf{H}_1 - \mathbf{H}_0).$$

Jenže matice $(\mathbf{H}_1 - \mathbf{H}_0)$ je idempotentní, a tudíž

$$h(\mathbf{A}) = h(\mathbf{H}_1 - \mathbf{H}_0) = \mathsf{tr}(\mathbf{H}_1 - \mathbf{H}_0) = k - 1.$$

Zřejmě každé U_i^2 má rozdělení χ_1^2, takže součet čtverců S_A je nezápornou lineární kombinací $k-1$ nezávislých náhodných veličin s rozdělením χ_1^2, kde koeficienty lineární kombinace jsou vlastní čísla λ_i.

Rozdělení S_e nahlédneme trochu jiným způsobem. Vyjádření veličiny S_e si lze takto upravit:

$$S_e = \sum_{i=1}^{k}\sum_{p=1}^{n_i}(Y_{ip} - y_{i\cdot})^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2,$$

kde $S_i^2 = \frac{1}{n_i - 1}\sum_{p=1}^{n_i}(Y_{ip} - y_{i\cdot})^2$. Z vlastností náhodného výběru z $N(\mu_i, \sigma_i^2)$ [2, věta 4.21 b)] víme, že

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi_{n_i - 1}^2,$$

a tudíž

$$S_e = \sum_{i=1}^{k}(n_i - 1)S_i^2 \sim \sum_{i=1}^{k}\sigma_i^2\chi_{n_i-1}^2.$$

Reziduální součet čtverců S_e je tedy součtem k nezávislých náhodných veličin s rozdělením $\sigma_i^2\chi_{n_i-1}^2$.

Aproximaci rozdělení podílu S_A/S_e dostaneme, vezmeme-li ve větě 9 $Q' = S_A$ a $Q = S_e$. Stačí jen spočítat koeficient b a stupně volnosti h a h'.

K tomu je zapotřebí znát střední hodnoty a rozptyly obou těchto veličin. Ty už máme spočítané v podkapitole 3.1. Máme

$$\begin{aligned}
\mathsf{E}\, S_A &= k\overline{\sigma^2} - \overline{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{k}\sigma_i^2(n-n_i) \\
\mathsf{var}\, S_A &= 2(k\overline{\sigma^4} - 2\overline{\sigma_n^4} + \overline{\sigma_n^2}^2) = 2\left[\frac{1}{n^2}\left(\sum_{i=1}^{k}\sigma_i^2 n_i\right)^2 + \frac{1}{n}\sum_{i=1}^{k}\sigma_i^4(n-2n_i)\right] \\
\mathsf{E}\, S_e &= n\overline{\sigma_n^2} - k\overline{\sigma^2} = \sum_{i=1}^{k}\sigma_i^2(n_i-1) \\
\mathsf{var}\, S_e &= 2(n\overline{\sigma_n^4} - k\overline{\sigma^4}) = 2\sum_{i=1}^{k}\sigma_i^4(n_i-1)
\end{aligned}$$

Dle věty 7 lze tedy statistiku

$$\frac{\dfrac{S_A}{k-1}}{\dfrac{S_e}{n-k}}$$

aproximovat pomocí rozdělení $bF_{h',h}$, kde koeficienty jsou dány vzorci:

$$\begin{aligned}
b &= \frac{n-k}{k-1}\frac{\mathsf{E}(S_A)}{\mathsf{E}(S_e)} = \frac{n-k}{n(k-1)}\frac{\sum_{i=1}^{k}(n-n_i)\sigma_i^2}{\sum_{i=1}^{k}(n_i-1)\sigma_i^2} \\
h' &= \frac{2[\mathsf{E}(S_A)]^2}{\mathsf{var}(S_A)} = \frac{\left[\sum_{i=1}^{k}(n-n_i)\sigma_i^2\right]^2}{\left[\sum_{i=1}^{k} n_i\sigma_i^2\right]^2 + n\sum_{i=1}^{k}(n-2n_i)\sigma_i^4} \\
h &= \frac{2[\mathsf{E}(S_e)]^2}{\mathsf{var}(S_e)} = \frac{\left[\sum_{i=1}^{k}(n_i-1)\sigma_i^2\right]^2}{\sum_{i=1}^{k}(n_i-1)\sigma_i^4}
\end{aligned}$$

Test hypotézy pak bude vypadat tak, že spočteme veličinu $\frac{S_A/(k-1)}{S_e/(n-k)}$ a porovnáme ji s kritickou hodnotou $bF_{h',h}(\alpha)$. Bude-li

$$\frac{S_A/(k-1)}{S_e/(n-k)} > bF_{h',h}(\alpha),$$

pak na hladině α zamítneme hypotézu H_0 a rovnosti středních hodnot.

3.4 Kruskalův-Wallisův test

Kruskalův-Wallisův test je neparametrickou obdobou jednoduchého třídění a používá se zejména tehdy, jde-li o výběry z rozdělení značně se lišících od normálního. Jeho předpokladem je, že výběry $Y_{i1}, \ldots, Y_{in_i}$ jsou na sobě nezávislé a pocházejí z rozdělení se spojitou distribuční funkcí F_i, kde $i = 1, \ldots, k$.

Pomocí tohoto testu lze testovat hypotézu o rovnosti distribučních funkcí, tj. $H_0 : F_1(x) = \ldots = F_k(x)$ pro všechna x.

Samotný test pak probíhá tak, že každé veličině Y_{ip} přiřadíme její pořadí R_{ip} ve sdruženém výběru. Jako testovou statistiku použijeme

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1),$$

kde $T_i = \sum_{p=1}^{n_i} R_{ip}$ je součet pořadí v i-tém výběru.

Význam této statistiky si můžeme odůvodnit následující úvahou. Připomeňme, že řádkový součet čtverců je definován jako

$$S_A = \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2,$$

kde $y_{i.} = \frac{1}{n_i} \sum_{p=1}^{n_i} Y_{ip}$ a $y_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{p=1}^{n_i} Y_{ip}$. V těchto výrazech nahradíme veličiny Y_{ip} pořadími R_{ip} a vytvoříme tak obdobu veličiny S_A, kterou nazveme S'_A. Veličina S'_A bude mít tvar

$$S'_A = \sum_{i=1}^{k} n_i \left(\frac{T_i}{n_i} - \overline{T} \right)^2,$$

kde T_i je jako v definici statistiky H a $\overline{T} = \frac{1}{n} \sum_{i=1}^{k} T_i = \frac{1+2+\ldots+n}{n} = \frac{n+1}{2}$. Upravíme-li si tento vzorec, dostaneme

$$S'_A = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - n\overline{T}^2 = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{n(n+1)^2}{4}.$$

Místo výrazu $S_e/(n-k)$, který odhaduje σ^2 (bez ohledu na H_0), použijeme obdobný odhad pro rozptyl (tentokráte za hypotézy H_0) založený na pořadí:

$$\hat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{p=1}^{n_i} (R_{ip} - \overline{R})^2.$$

Tento výraz si dále upravíme, přičemž si uvědomíme, že $\overline{R} = \overline{T} = \frac{n+1}{2}$. Budeme mít

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n-1}\sum_{i=1}^{n} i^2 - \frac{n}{n-1}\overline{R}^2 \\ &= \frac{n(n+1)(2n+1)}{6(n-1)} - \frac{n}{n-1}\frac{(n+1)^2}{4} \\ &= \frac{n(n+1)}{12}.\end{aligned}$$

Celkově pak dostaneme

$$\begin{aligned}\frac{S'_A}{\hat{\sigma^2}} &= \frac{12}{n(n+1)}S'_A = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - \frac{n(n+1)^2}{4}\frac{12}{n(n+1)} \\ &= \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n+1),\end{aligned}$$

což je přesně statistika H.

Navíc z důkazu věty 3 víme, že veličina $\frac{S_A}{\sigma^2}$ má rozdělení χ^2_{k-1}. Něco podobného bychom tedy očekávali i od naší veličiny $\frac{S'_A}{\hat{\sigma^2}}$. A skutečně, dle [1, str. 115] lze dokázat, že za platnosti H_0 má statistika H asymptoticky χ^2 rozdělení. Dle tvrzení [1, věta 9.11] o tvaru statistik tohoto typu je $\mathsf{E}\,H = k-1$, tedy půjde o asymptotické χ^2 rozdělení s $k-1$ stupni volnosti. Proto v našem testu hypotézu H_0 zamítneme, pokud $H \geq \chi^2_{k-1}(\alpha)$.

Pokud hypotézu H_0 nezamítneme, pak můžeme předpokládat, že střední hodnoty (pokud existují) výběrů $Y_{i1}, \ldots, Y_{in_i}$ jsou si rovny, neboť rovnají-li se distribuční funkce, rovnají se také všechny momenty. Je třeba si jen uvědomit, že obrácená implikace neplatí a zamítnutí H_0 nedává o středních hodnotách žádnou informaci.

Kapitola 4

Simulace

4.1 Ověření hladiny a síly testů

V této kapitole budeme zkoumat všechny dříve zmíněné testy jednoduchého třídění: klasický F-test, Welchovu metodu, Kruskalův-Wallisův test a Boxovu metodu. Podíváme se, jak tyto testy udržují hladinu a jaká je jejich síla v různých situacích, pro různé rozsahy výběrů a pro různá rozdělení.

Simulace provedeme pomocí programu R (zdrojový kód je na přiloženém CD). Nejprve si zvolíme rozsahy jednotlivých výběrů $n_1, \ldots, n_k$ a jejich počet k. Dále vektor středních hodnot $(\mu_1, \ldots, \mu_k)$, vektor směrodatných odchylek $(\sigma_1, \ldots, \sigma_k)$, hladinu testů α a počet simulací, které budeme chtít provést. My jsme zvolili vždy $k = 3$, $\alpha = 0{,}05$ a počet simulací jako 1000. Počítač tedy 1000 krát nageneruje data zadaných parametrů a otestuje na nich hypotézu H_0 o rovnosti středních hodnot, a to všemi čtyřmi testy. Relativní počty zamítnutí hypotézy každým testem jsou pak zaneseny do tabulky.

Dodržování hladiny ověříme počáteční volbou parametrů, při níž bude H_0 platit, tj. zvolíme $\mu_1 = \mu_2 = \mu_3$. Relativní četnost zamítnutí pak bude odhadem pravděpodobnosti, že daný test zamítne nulovou hypotézu, která ale platí, tedy půjde o odhad pravděpodobnosti chyby prvního druhu, tedy o odhad hladiny. Nás bude zajímat, pro které testy a které situace se bude tento odhad blížit skutečné hodnotě hladiny α. Počet zamítnutí S je náhodná veličina s rozdělením $Bi(m, p)$, kde m je počet simulací (tedy $m = 1000$) a p je skutečná pravděpodobnost zamítnutí H_0, když H_0 platí. Testujeme tedy vlastně hypotézu, že $p = \alpha$. Dle centrální limitní věty [2, věta B.5] má za

platnosti H_0 testová statistika

$$\frac{S - \mathsf{E}\, S}{\sqrt{\mathsf{var}\, S}} = \frac{S - m\alpha}{\sqrt{m\alpha(1-\alpha)}}$$

asymptoticky rozdělení $N(0,1)$. Kritický obor tedy bude mít tvar

$$\left\{ \frac{|S - m\alpha|}{\sqrt{m\alpha(1-\alpha)}} \geq u\left(\frac{\alpha}{2}\right) \right\},$$

kde $u(\alpha/2)$ je kritická hodnota rozdělení $N(0,1)$. Pro relativní četnosti bude kritický obor vypadat takto

$$\left\{ \frac{\left|\frac{S}{m} - \alpha\right|}{\sqrt{\frac{\alpha(1-\alpha)}{m}}} \geq u\left(\frac{\alpha}{2}\right) \right\}.$$

Tedy hypotézu o tom, že daný test dodržuje hladinu, zamítáme, pokud bude platit

$$\left|\frac{S}{m} - \alpha\right| \geq u\left(\frac{\alpha}{2}\right) \sqrt{\frac{\alpha(1-\alpha)}{m}}.$$

Dosadíme-li do tohoto vzorce námi zvolené hodnoty (tedy $m = 1000$ a $\alpha = 0{,}05$) dostáváme, že hypotézu zamítáme pokud

$$\left|\frac{S}{1000} - 0{,}05\right| \geq 0{,}013.$$

Hodnoty, které tuto nerovnost nesplňují, tedy podporují hypotézu o dodržení hladiny testem, budou vždy v příslušné tabulce uvedeny tučně.

Dodržení hladiny je pro správné fungování testu nutné, ale nikoli postačující kritérium. Je třeba ještě zjistit jeho sílu. Tu budeme ověřovat tak, že zvolíme za parametry pro data hodnoty, při kterých hypotéza o rovnosti středních hodnot neplatí. Relativní četnost zamítnutí pak bude odhadem pro pravděpodobnost, že test zamítl hypotézu H_0, která neplatila, tudíž půjde o odhad síly testu. Test, který se v dané situaci ukáže jako nejsilnější, bude v tabulkách uveden tučným písmem.

Je třeba ještě podotknout, že samotná velikost síly testu v dané situaci je dána především velikostí zvolených směrodatných odchylek. Volba větších

hodnot nutně znejistí rozhodování o nulové hypotéze a povede k nižší hodnotě síly. Abychom tedy mohli porovnávat testy mezi sebou i v rámci různých situací (řádků tabulky), budeme se snažit udržovat konstantní hodnotu K rozptylu celkového průměru. Pro každou zvolenou kombinaci rozsahů výběrů n_1, n_2, n_3 a směrodatných odchylek $\sigma_1, \sigma_2, \sigma_3$ tedy bude muset platit

$$K = \mathsf{var}(y_{..}) = \mathsf{var}\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{p=1}^{n_i} Y_{ip}\right) = \frac{1}{n^2}\sum_{i=1}^{k}\sum_{p=1}^{n_i}\sigma_i^2$$
$$= \frac{1}{n^2}\sum_{i=1}^{k} n_i\sigma_i^2.$$

My budeme volit všechny výběry tak, aby součet jejich rozsahů n byl ve všech situacích stejný a byl roven hodnotě 30. Pak tedy stačí, aby byl pro všechny řádky konstantní výraz: $\sum_{i=1}^{k} n_i\sigma_i^2$. Budeme požadovat

$$\sum_{i=1}^{3} n_i\sigma_i^2 = Kn^2 = 140.$$

Pro přehlednost si do tabulky navolíme celočíselné hodnoty směrodatných odchylek $\sigma_1, \sigma_2, \sigma_3$, a pak najdeme konstantu c tak, aby platilo

$$\sum_{i=1}^{3} n_i(c\,\sigma_i)^2 = 140.$$

Konstanta c bude stejná pro všechna σ_i v rámci jednoho řádku a v tabulce bude uvedena vždy ve sloupečku před příslušným vektorem směrodatných odchylek. Tak bude možné porovnávat nasimulované hodnoty nejen v rámci stejných rozsahů výběrů a stejných rozptylů, ale i mezi sebou.

Simulace provedeme pro normální, logistické a gamma rozdělení. Pro každé rozdělení budeme zkoumat stejné rozsahy výběrů a stejné kombinace středních hodnot a směrodatných odchylek. Pak budeme u každého testu sledovat, zda dodržuje hladinu a jak je v dané situaci silný.

4.2 Normální rozdělení

Nejprve se podíváme na situaci, kdy všechny výběry pocházejí z normálního rozdělení. Pro toto rozdělení je za platnosti předpokladu rovnosti rozptylů

možno jako pro jediné spočítat skutečnou hodnotu síly F-testu. Tato hodnota je uvedena v tabulce 4.2 v pátém sloupečku pod názvem "F-test (teor.)". Simulace dopadly takto:

Tabulka 4.1: Empirická hladina testů pro normální rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test	Welch	Kruskal -Wallis	Box
10,10,10	0,0,0	2,16	1,1,1	**0,060**	**0,061**	**0,058**	**0,058**
9,10,11	0,0,0	2,16	1,1,1	**0,063**	**0,054**	**0,060**	**0,055**
7,10,13	0,0,0	2,16	1,1,1	**0,044**	**0,051**	**0,045**	**0,044**
6,7,17	0,0,0	2,16	1,1,1	**0,050**	**0,048**	**0,038**	**0,051**
10,10,10	0,0,0	1,00	1,2,3	**0,063**	**0,048**	**0,048**	**0,044**
9,10,11	0,0,0	0,97	1,2,3	**0,044**	**0,048**	**0,052**	**0,039**
7,10,13	0,0,0	0,92	1,2,3	0,023	**0,048**	0,031	**0,037**
6,7,17	0,0,0	0,92	1,1,3	0,003	0,031	0,014	0,032
6,7,17	0,0,0	1,34	3,1,1	0,188	**0,056**	0,095	0,065

Tabulka 4.2: Empirická síla testů pro normální rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test (teor.)	F-test	Welch	Kruskal -Wallis	Box
10,10,10	1,2,3	2,16	1,1,1	0,400	**0,382**	0,354	0,341	0,371
9,10,11	1,2,3	2,16	1,1,1	0,398	**0,410**	0,396	0,390	0,398
7,10,13	1,2,3	2,16	1,1,1	0,379	**0,375**	0,358	0,357	0,367
6,7,17	1,2,3	2,16	1,1,1	0,382	**0,391**	0,345	0,352	**0,391**
10,10,10	1,2,3	1,00	1,2,3	—	0,393	**0,479**	0,410	0,335
9,10,11	1,2,3	0,97	1,2,3	—	0,396	**0,508**	0,405	0,374
7,10,13	1,2,3	0,92	1,2,3	—	0,327	**0,531**	0,357	0,378
6,7,17	1,2,3	0,92	1,1,3	—	0,298	**0,612**	0,404	0,584
6,7,17	1,2,3	1,34	3,1,1	—	**0,427**	0,294	0,365	0,214
6,7,17	3,2,1	0,92	1,1,3	—	0,261	**0,606**	0,368	0,527
6,7,17	3,2,1	1,34	3,1,1	—	**0,400**	0,270	0,342	0,188

Z tabulky 4.1 vidíme, že jsou-li směrodatné odchylky (a tudíž i rozptyly) stejné, dodržují hladinu všechny čtyři testy, bez ohledu na rozsahy jednotlivých výběrů. Co se týče jejich síly (viz tabulka 4.2), tak ta má v těchto případech také velmi vyrovnané hodnoty. Porovnáme-li navíc nasimulované hladiny F-testu s jejich teoretickými hodnotami, zpozorujeme značnou shodu.

Jsou-li směrodatné odchylky různé, ale rozsahy výběrů jsou stejné, nebo se liší jen velmi málo, pak hladinu opět dodržují všechny testy. I s jejich sílou je to velmi podobné jako v předchozím případě. Výjimku tvoří jen

Welchův testu, jehož síla výrazně stoupla a můžeme ho proto za této situace doporučit.

Jsou-li směrodatné odchylky různé a navíc se značně liší i rozsahy jednotlivých výběrů, pak hladinu α ve všech případech nedodržel žádný test. Avšak velmi dobře si vedl Welchův test, který selhal (a to navíc celkem těsně) pouze v jediném případě. Hned za ním pak Boxova metoda, která hladinu nedodržela ve dvou případech, ale opět relativně těsně. Ostatní testy v této extrémní situaci daly zcela nepřípustné hodnoty. V případech, kdy největší rozptyl byl u výběru s největším rozsahem, se Welchův test ukázal i jako nejsilnější, opět bezprostředně následován Boxovým testem.

Není-li největší směrodatná odchylka u výběru s největším rozsahem, přičemž ostatní dva výběry jsou si rozsahově podobné, je těžké vybrat, který z testů doporučit. Hladinu v tomto případě dodržel pouze Welchův test (připustit by se dal i Boxův test), ale tyto dva se zase ukázaly jako nejslabší. Testy, které v této situaci měly jednoznačně největší sílu, zase hrubě nedodržely hladinu. Vezmeme-li ovšem v úvahu rozdíly v hladinách a rozdíly v sílách jednotlivých testů, je zřejmé, že Welchův test je nejlepším kompromisem.

4.3 Logistické rozdělení

Nyní vyzkoušíme vlastnosti testů pro výběry z logistického rozdělení. Toto rozdělení volíme z toho důvodu, že Kruskalův-Wallisův test je znám jakožto lokálně nejsilnější pořadový test v případě, že výběry pocházejí právě z logistického rozdělení [2, str. 246]. Střední hodnoty výběrů budeme i nadále značit μ_1, μ_2, μ_3 a směrodatné odchylky $\sigma_1, \sigma_2, \sigma_3$.

Výsledky simulací jsou v tabulkách 4.3 a 4.4.

Co se týče dodržování hladiny, jsou výsledky vesměs podobné jako u normálního rozdělení. Jsou-li směrodatné odchylky stejné, pak hladinu dodržely všechny testy. Jako nejsilnější se zde ukázaly být F-test a Kruskalův-Wallisův test.

V případě nestejných směrodatných odchylek je situace opět trochu složitější. Pokud nejsou rozdíly v rozsazích výběrů veliké, pak hladinu dodržely takřka všechny testy. Výjimku tvořil jen Welchův a Boxův test, které v jednom případě těsně hladinu nedoržely. Jsou-li rozsahy různé, pak se situace otočila a hladinu dodržely pouze Welchův a Boxův test. Se silou testů je to zde podobné jako u normálního rozdělení. V extrémním případě, kdy

Tabulka 4.3: Empirická hladina testů pro logistické rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test	Welch	Kruskal -Wallis	Box
10,10,10	0,0,0	2,16	1,1,1	**0,058**	**0,049**	**0,055**	**0,051**
9,10,11	0,0,0	2,16	1,1,1	**0,042**	**0,038**	**0,038**	**0,040**
7,10,13	0,0,0	2,16	1,1,1	**0,051**	**0,041**	**0,045**	**0,043**
6,7,17	0,0,0	2,16	1,1,1	**0,058**	**0,061**	**0,053**	**0,060**
10,10,10	0,0,0	1,00	1,2,3	**0,051**	**0,041**	**0,045**	**0,044**
9,10,11	0,0,0	0,97	1,2,3	**0,042**	0,033	**0,041**	0,036
7,10,13	0,0,0	0,92	1,2,3	0,033	**0,045**	**0,041**	**0,041**
6,7,17	0,0,0	0,92	1,1,3	0,008	**0,048**	0,024	**0,037**
6,7,17	0,0,0	1,34	3,1,1	0,173	**0,045**	0,077	**0,063**

Tabulka 4.4: Empirická síla testů pro logistické rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test	Welch	Kruskal -Wallis	Box
10,10,10	1,2,3	2,16	1,1,1	**0,397**	0,391	0,391	0,380
9,10,11	1,2,3	2,16	1,1,1	0,412	0,399	**0,416**	0,391
7,10,13	1,2,3	2,16	1,1,1	**0,388**	0,370	0,376	0,365
6,7,17	1,2,3	2,16	1,1,1	**0,403**	0,360	0,400	0,379
10,10,10	1,2,3	1,00	1,2,3	0,422	**0,477**	0,467	0,362
9,10,11	1,2,3	0,97	1,2,3	0,393	**0,513**	0,454	0,371
7,10,13	1,2,3	0,92	1,2,3	0,352	**0,569**	0,429	0,406
6,7,17	1,2,3	0,92	1,1,3	0,311	**0,661**	0,477	0,552
6,7,17	1,2,3	1,34	3,1,1	**0,451**	0,343	0,417	0,233
6,7,17	3,2,1	0,92	1,1,3	0,333	**0,650**	0,475	0,579
6,7,17	3,2,1	1,34	3,1,1	**0,496**	0,334	0,451	0,241

největší rozptyl není u výběru s největším rozsahem, byly nejsilnější F-test a Kruskalův-Wallisův test, v ostatních případech Welchův test a Boxův test.

Souhrnně lze říci, že jsou-li rozptyly stejné, můžeme F-test doporučit i pro logistické rozdělení. Stejně tak Welchův test pro nestejné rozptyly. Kruskalův-Wallisův test by byl pro toto rozdělení také použitelný, obzvláště v případě, kdy si rovností rozptylů nejsme jisti. Ne vždy sice dodržel hladinu, ale téměř ve všech případech byl (s malým rozdílem) druhý nejsilnější.

4.4 Gamma rozdělení

Toto rozdělení zde uvádíme coby zástupce nesymetrických rozdělení. Střední hodnota tohoto rozdělení je vždy kladná, tedy bude třeba tomuto požadavku

přizpůsobit volbu středních hodnot jednotlivých výběrů. Proto u ověřování hladiny položíme všechny střední hodnoty rovny 1. To nikterak nevadí, protože hladina testů na faktické velikosti střední hodnoty nezávisí. Značení ponecháme stejné jako v předchozích tabulkách, tedy μ_1, μ_2, μ_3 pro střední hodnoty a $\sigma_1, \sigma_2, \sigma_3$ pro směrodatné odchylky.

Simulace dopadly takto:

Tabulka 4.5: Empirická hladina testů pro gamma rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test	Welch	Kruskal -Wallis	Box
10,10,10	1,1,1	2,16	1,1,1	0,029	0,028	**0,039**	0,012
9,10,11	1,1,1	2,16	1,1,1	0,028	0,036	**0,047**	0,012
7,10,13	1,1,1	2,16	1,1,1	0,034	**0,041**	**0,049**	0,007
6,7,17	1,1,1	2,16	1,1,1	0,032	**0,047**	**0,044**	0,021
10,10,10	1,1,1	1,00	1,2,3	0,112	0,206	0,477	0,077
9,10,11	1,1,1	0,97	1,2,3	0,103	0,157	0,464	0,066
7,10,13	1,1,1	0,92	1,2,3	0,081	0,116	0,403	**0,040**
6,7,17	1,1,1	0,92	1,1,3	0,111	0,084	0,515	0,085
6,7,17	1,1,1	1,34	3,1,1	0,082	0,390	0,563	0,121

Tabulka 4.6: Empirická síla testů pro gamma rozdělení

n_1, n_2, n_3	μ_1, μ_2, μ_3	c	$\sigma_1, \sigma_2, \sigma_3$	F-test	Welch	Kruskal -Wallis	Box
10,10,10	1,2,3	2,16	1,1,1	0,518	0,546	**0,804**	0,473
9,10,11	1,2,3	2,16	1,1,1	0,525	0,580	**0,825**	0,498
7,10,13	1,2,3	2,16	1,1,1	0,483	0,613	**0,750**	0,493
6,7,17	1,2,3	2,16	1,1,1	0,480	0,633	**0,742**	0,512
10,10,10	1,2,3	1,00	1,2,3	0,393	**0,404**	0,394	0,300
9,10,11	1,2,3	0,97	1,2,3	0,370	**0,494**	0,453	0,326
7,10,13	1,2,3	0,92	1,2,3	0,291	**0,556**	0,400	0,368
6,7,17	1,2,3	0,92	1,1,3	0,264	**0,651**	0,447	0,589
6,7,17	1,2,3	1,34	3,1,1	0,753	0,761	**0,918**	0,720
6,7,17	3,2,1	0,92	1,1,3	0,592	0,767	**0,958**	0,679
6,7,17	3,2,1	1,34	3,1,1	0,442	0,236	**0,498**	0,139

Z výsledků vidíme, že je-li splněna rovnost směrodatných odchylek, je jednoznačnou volbou Kruskalův-Wallisův test. Nejenže vždy dodržel hladinu, ale jednoznačně měl i největší sílu. I Welchův test ve dvou případech uvedl hodnoty odpovídající předepsané hladině a jeho síla také nebyl špatná.

Co se týče situací s nestejným rozptylem, je vidět, že žádný z testů není pro gamma rozdělení příliš vhodný. Zcela lze zavrhnout F-test, Welchův

test a Kruskalův-Wallisův test, které ve všech případech značně nedodržely hladinu. Boxův test na tom byl zdánlivě podobně, byl však ze všech testů jednoznačně nejblíže intervalu pro udržení hladiny $(0{,}037; 0{,}063)$ a jeho síla se také držela v mezích únosnosti. Pokud bychom tedy přeci jen chtěli některý z testů použít, pak právě Boxův test.

4.5 Závěr

Závěrem můžeme jen potvrdit, že F-test je nejlepším testem pro testování hypotézy o rovnosti středních hodnot v případě stejných rozptylů, a to nejen pro normální rozdělení, ale i pro logistické. Totéž platí pro Welchův test v případě nestejných rozptylů. Kruskalův-Wallisův test prokázal svou sílu pro výběry pocházející z logistického rozdělení a ani v ostatních situacích neměl špatné výsledky. Zejména překvapil svou silou a dodržováním hladiny u gamma rozdělení při stejných rozptylech. Boxova metoda má pro normální a logistické rozdělení co do hladiny a síly podobně dobré vlastnosti jako Welchův postup a ukázala se i jako použitelná pro gamma rozdělení v případě nestejných rozptylů.

Literatura

[1] Anděl J.: *Statistické metody*, Matfyzpress, Praha, 1998.

[2] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2007.

[3] Box G. P. E.: Some theorems on qudratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, The Annals of Mathematical Statistics **25** (1954) 290-302.

[4] Welch B. L.: On the comparison of several mean values: an alternative approach, Biometrika **38** (1951) 330-336.

[5] Wilks S. S.: *Mathematical Statistics*, John Wiley & Sons, Inc., New York, 1962.