MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Matyáš Gulík

Testy shody rozptylů v jednoduchém třídění

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program: Finanční matematika
Studijní obor: MFMP

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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

Chtěl bych zde poděkovat vedoucímu bakalářské práce doc. RNDr. Arnoštu Komárkovi, Ph.D. za čas věnovaný konzultacím a veškeré poskytnuté rady, které vedly ke zkvalitnění této bakalářské práce.

Název práce: Testy shody rozptylů v jednoduchém třídění

Autor: Matyáš Gulík

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Bakalářská práce pojednává o testech shody rozptylů v kontextu jednoduchého třídění. Zaměřuje se na tři testy, které jsou běžně využívány v praxi. V práci je nejprve uveden přehled pojmů a poznatků z teorie pravděpodobnosti, které jsou využity během odvozování v dalších kapitolách. Dále je připomenuta analýza rozptylu jednoduchého třídění, která je pro testy shody rozptylů klíčová. V hlavní části práce je odvozen Leveneův test a dále je uveden Brown-Forsytheův test, který je jeho modifikací. Také je uveden Bartlettův test. Na závěr práce byly pomocí programu R provedeny simulace, jejichž cílem bylo zjistit, jak jsou testy schopny dodržet požadovanou hladinu.

Klíčová slova: shoda rozptylů, Levene, Brown-Forsythe, Bartlett

Title: Homoscedasticity tests in one-way classification

Author: Matyáš Gulík

Department: Department of Probability and Mathematical Statistics

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

Abstract: The bachelor thesis deals with tests of variance equality in the context of simple classification. It focuses on three tests commonly used in practice. The thesis first provides an overview of concepts and knowledge from probability theory, which are utilized in subsequent chapters. Additionally, a one-way analysis of variance is introduced, which is crucial for the tests of variance equality. In the main part of the thesis, the Levene test is derived, followed by the Brown-Forsythe test, which is its modification. The Bartlett test is also presented. Finally, simulations were conducted using the R program to determine the ability of the tests to maintain the desired significance level.

Keywords: equality of variances, Levene, Brown-Forsythe, Bartlett

Obsah

Seznam použitých zkratek

V rámci celé práce se budeme držet následujících značení:

$\boldsymbol{X}$	náhodný vektor $\boldsymbol{X}$
$\boldsymbol{X}^T$	transpozice náhodného vektoru $\boldsymbol{X}$
$\mathbf{0}$	nulový vektor
$\mathsf{E}(X)$	střední hodnota náhodné veličiny X
$\mathsf{var}(X)$	rozptyl náhodné veličiny X
$X \perp\!\!\!\perp Y$	X a Y jsou nezávislé
$\xrightarrow{D}$	konvergence v distribuci
$\xrightarrow{P}$	konvergence v pravděpodobnosti
$\mathrm{tr}(\mathbb{A})$	stopa matice $\mathbb{A}$
$\mathbb{A}^T$	transponovaná matice k matici $\mathbb{A}$
$f^{(n)}(0)$	hodnota n-té derivace funkce f v bodě 0
$\mathbb{R}$	množina reálných čísel
$\mathbb{I}_k$	jednotková matice řádu k
$X \sim \mathrm{N}(\mu, \sigma^2)$	X má normální rozdělení s parametry μ, σ^2
$\mathrm{N}_n(\boldsymbol{\mu}, \Sigma)$	n-rozměrné normální rozdělení s vektorem středních hodnot $\boldsymbol{\mu}$ a varianční maticí Σ

Úvod

Úkolem této práce je pojednat o testech shody rozptylů v kontextu jednoduchého třídění. V situaci, kdy máme dva nezávislé náhodné výběry z normálního rozdělení a chtěli bychom testovat nulovou hypotézu, zda výběry pocházejí z rozdělení se shodným rozptylem, lze použít dvouvýběrový F-test na rozptyl. Pokud bychom měli $k \geq 3$ nezávislých náhodných výběrů z normálního rozdělení a chtěli bychom opět testovat nulovou hypotézu, zda všechny výběry pocházejí z rozdělení se shodným rozptylem, tak by nás mohlo napadnout provést dvouvýběrový F-test na rozptyl na všechny možné dvojice náhodných výběrů. Pokud některý test nulovou hypotézu zamítne, tak zamítneme původní nulovou hypotézu, že všechny výběry pocházejí z rozdělení se shodným rozptylem. Lze však nahlédnout, že tento postup zvyšuje pravděpodobnost falešného zamítnutí, tzv. chyby I. druhu. V této situaci je tedy potřeba použít jiný test. Mezi tyto testy například patří Leveneův, Brown-Forsytheův a Bartlettův test, o kterých tato práce pojednává. Zaměříme se zejména na Leveneův test, pro který lze v literatuře často nalézt, jak jej na data aplikovat, ale zřídka lze nalézt jeho odvození. Odvození Leveneova testu provedeme od prvotního nápadu, který stojí za jeho vznikem, až po rozdělení testové statistiky a určení kritického oboru.

V první kapitole uvedeme přehled pojmů a poznatků z teorie pravděpodobnosti, které budeme v rámci práce využívat během odvozování. Ve druhé kapitole připomeneme analýzu rozptylu jednoduchého třídění. Ve třetí kapitole, která je hlavní částí této práce, se seznámíme s testy shody rozptylů v jednoduchém třídění, kde pro nás poznatky z první a druhé kapitoly budou zásadní. Na závěr této práce provedeme ve čtvrté kapitole simulační studii pomocí programu R (R Core Team, 2024), kde budeme chtít ověřit, zda jsou testy schopny dodržet požadovanou hladinu v závislosti na tom, z jakého rozdělení náhodné výběry pocházejí.

1. Přehled z teorie pravděpodobnosti

V této kapitole pro přehlednost shrneme pojmy a poznatky z teorie pravděpodobnosti, které nám budou užitečné v dalších kapitolách.

1.1 Chí-kvadrát rozdělení a kvadratické formy

Nejprve si připomeneme definici chí-kvadrát rozdělení. Pro chí-kvadrát používáme značení χ^2.

Definice 1 (Chí-kvadrát rozdělení)**.** *Nechť náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé a mají normované normální rozdělení $\mathrm{N}(0,1)$. Pak říkáme, že náhodná veličina $Y = \sum_{i=1}^n X_i^2$ má χ^2 rozdělení o n stupních volnosti. Značíme $Y \sim \chi^2_n$.*

Nyní zformulujeme větu o rozdělení kvadratické formy, která pro nás bude důležitá u odvození rozdělení testové statistiky Leveneova testu.

Věta 1 (Rozdělení kvadratické formy)**.** *Nechť $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}, \Sigma)$ a nechť $\mathbb{A}_{n\times n}$ je symetrická pozitivně semidefinitní matice taková, že $\mathbb{A}\Sigma$ je nenulová a idempotentní (tj. $\mathbb{A}\Sigma\mathbb{A}\Sigma = \mathbb{A}\Sigma$). Pak $\boldsymbol{Y} = \boldsymbol{X}^{\boldsymbol{T}}\mathbb{A}\boldsymbol{X} \sim \chi^2_{\mathrm{tr}(\mathbb{A}\Sigma)}$.*

Důkaz. Důkaz lze najít např. v práci Anděl (2011, str. 68, věta 4.16).

□

Na závěr této podkapitoly uvedeme definici F-rozdělení.

Definice 2 (F-rozdělení)**.** *Nechť $X \sim \chi^2_n$ a $Y \sim \chi^2_m$ jsou nezávislé náhodné veličiny a nechť pro náhodnou veličinu F platí*

$$F = \frac{X/n}{Y/m}.$$

Pak řekneme, že F má F-rozdělení o n a m stupních volnosti. Značíme $F \sim F_{n,m}$.

1.2 Kumulanty

V této podkapitole se seznámíme s kumulanty, s kterými se setkáme během odvození Leveneova testu. Začneme definicí kumulativní vytvořující funkce.

Definice 3 (Kumulativní vytvořující funkce)**.** *Nechť X je náhodná veličina. Funkce $K_X(t) = \log \mathsf{E}(e^{tX})$, $t \in \mathbb{R}$ se nazývá kumulativní vytvořující funkce náhodné veličiny X.*

Poznámka. Kumulativní vytvořující funkce se dá ekvivalentně zapsat ve tvaru $K_X(t) = \log M_X(t)$, kde $M_X(t)$ je momentová vytvořující funkce náhodné veličiny X.

Dále si definujeme pojmy kumulant a normalizovaný kumulant.

Definice 4 (Kumulant)**.** *Nechť X je náhodná veličina a $n \in \mathbb{N}$, pak její n-tý kumulant definujeme vztahem $\kappa_n = K_X^{(n)}(0)$, pokud $K_X^{(n)}(0)$ existuje.*

Poznámka. Pro náhodnou veličinu X platí $\kappa_1 = \mathsf{E}(X)$, $\kappa_2 = \mathsf{var}(X)$, $\kappa_3 = \mathsf{E}(X - \mathsf{E}(X))^3$, $\kappa_4 = \mathsf{E}(X - \mathsf{E}(X))^4 - 3(\mathsf{var}(X))^2$.

Definice 5 (Normalizovaný kumulant)**.** *Nechť κ_n je n-tý kumulant náhodné veličiny X, $n \in \mathbb{N}$. Normalizovaný kumulant náhodné veličiny X definujeme vztahem*

$$\gamma_n = \frac{\kappa_n}{\sigma^n},$$

kde σ je směrodatná odchylka náhodné veličiny X.

Poznámka. Pro náhodnou veličinu je tedy γ_3 její šikmost, ale γ_4 už není její špičatost.

1.3 Polonormální rozdělení

V této podkapitole představíme polonormální rozdělení, jehož znalost pro nás bude u odvození Leveneova testu také důležitá. Nejprve si polonormální rozdělení definujeme.

Definice 6 (Polonormální rozdělení)**.** *Nechť X je náhodná veličina taková, že $X \sim \mathrm{N}(0, \sigma^2)$, kde $\sigma^2 > 0$. Pak řekneme, že náhodná veličina $Y = |X|$ má polonormální rozdělení s parametrem σ^2.*

Dále se zaměříme na momenty polonormálního rozdělení. Toto tvrzení dokážeme, jelikož pro nás bude stěžejní u odvození Leveneova testu.

Tvrzení 2. *Nechť náhodná veličina Y má polonormální rozdělení s parametrem σ^2. Pak platí*

$$\mathsf{E}(Y) = \sqrt{\frac{2}{\pi}}\sigma, \quad \mathsf{var}(Y) = \left(1 - \frac{2}{\pi}\right)\sigma^2.$$

Důkaz. Víme, že $Y = |X|$, kde $X \sim \mathrm{N}(0, \sigma^2)$. Hustotu náhodné veličiny X označme f_X a počítejme

$$\begin{aligned}\mathsf{E}(Y) = \mathsf{E}|X| &= \int_{-\infty}^{\infty} |x| f_X(x)\,dx = \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\,dx \\ &= 2\sigma \int_0^{\infty} \frac{x}{\sigma^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\,dx = \sqrt{\frac{2}{\pi}}\sigma \int_0^{\infty} e^{-y}\,dy \\ &= \sqrt{\frac{2}{\pi}}\sigma \Big[- e^{-y} \Big]_0^{\infty} = \sqrt{\frac{2}{\pi}}\sigma,\end{aligned}$$

kde v druhé rovnosti v druhém řádku jsme použili substituci $y = \frac{x^2}{2\sigma^2}$, $dy = \frac{x}{\sigma^2}dx$. Dále

$$\begin{aligned}\mathsf{var}(Y) &= \mathsf{E}(Y^2) - (\mathsf{E}(Y))^2 = \mathsf{E}|X|^2 - (\mathsf{E}(Y))^2 = \mathsf{E}(X^2) - (\mathsf{E}(Y))^2 \\ &= \mathsf{var}(X) + (\mathsf{E}(X))^2 - (\mathsf{E}(Y))^2 = \sigma^2 + 0^2 - \frac{2}{\pi}\sigma^2 = \left(1 - \frac{2}{\pi}\right)\sigma^2.\end{aligned}$$

□

2. Analýza rozptylu jednoduchého třídění

V této kapitole si připomeneme analýzu rozptylu jednoduchého třídění, jež hraje důležitou roli u testů shody rozptylů, kterými se budeme zabývat v nadcházející kapitole. Analýza rozptylu jednoduchého třídění je speciálním případem analýzy rozptylu. V rámci celé kapitoly budeme uvažovat následující model.

Uvažujme $k \geq 2$ nezávislých náhodných výběrů s rozsahy $N_1, \ldots, N_k$, kde i-tý náhodný výběr pochází z rozdělení $\mathrm{N}(\mu_i, \sigma^2)$, $\mu_i \in \mathbb{R}$, $\sigma^2 > 0$, $i = 1, \ldots, k$, tj.

$$\begin{aligned} X_{1,1}, \ldots, X_{1,N_1} &\sim \mathrm{N}(\mu_1, \sigma^2), \\ &\vdots \\ X_{k,1}, \ldots, X_{k,N_k} &\sim \mathrm{N}(\mu_k, \sigma^2). \end{aligned} \tag{2.1}$$

Symbolem $X_{i,j}$ budeme značit j-té pozorování i-tého náhodného výběru, kde $j = 1, \ldots, N_i$. Zajímá nás, jestli uvažované náhodné výběry pocházejí z rozdělení se stejnou střední hodnotou. Chceme testovat nulovou hypotézu

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

proti alternativní hypotéze

$$H_1 : \exists i \neq j : \mu_i \neq \mu_j.$$

Pro přehlednost nejprve zavedeme značení.

Značení. Označme $N = \sum_{i=1}^{k} N_i$, $\overline{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{i,j}$, $\overline{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} X_{i,j}$. Symbol N značí celkový počet všech pozorování. Symbol $\overline{X}_i$ značí výběrový průměr i-tého náhodného výběru. Symbol $\overline{X}$ značí celkový průměr všech pozorování.

Nyní si definujeme součty čtverců.

Definice 7 (Součty čtverců)**.**
$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{i,j} - \overline{X})^2$ *nazýváme celkový součet čtverců.*
$SS_A = \sum_{i=1}^{k} N_i (\overline{X}_i - \overline{X})^2$ *nazýváme součet čtverců skupin.*
$SS_e = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{i,j} - \overline{X}_i)^2$ *nazýváme reziduální součet čtverců.*

Dále se zaměříme na rozdělení součtů čtverců.

Věta 3. *Nechť platí výše uvedený model* (2.1)*, pak*

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{N-k}.$$

Důkaz. Důkaz lze najít např. v práci Kulich (2022, str. 161).

□

Věta 4. *Nechť platí výše uvedený model* (2.1) *a navíc platí i nulová hypotéza H_0, pak*

1. $$\frac{SS_A}{\sigma^2} \sim \chi^2_{k-1},$$

2. $$SS_A \perp\!\!\!\perp SS_e.$$

Důkaz. Důkaz lze najít např. v práci Kulich (2022, str. 162).

□

Zvolíme testovou statistiku

$$F_A = \frac{SS_A/(k-1)}{SS_e/(N-k)} = \frac{\sum_{i=1}^{k} N_i(\overline{X}_i - \overline{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{i,j} - \overline{X}_i)^2/(N-k)}, \tag{2.2}$$

která má za platnosti nulové hypotézy rozdělení $F_{k-1,N-k}$. Tento závěr plyne z předchozích dvou vět a definice F-rozdělení (viz definice 2). Nyní nás zajímá, jaký bychom měli zvolit kritický obor. Víme, že $\overline{X}_i$ je odhadem střední hodnoty i-tého výběru a $\overline{X}$ je odhadem celkové střední hodnoty. Za platnosti nulové hypotézy by tedy $\overline{X}_i$ a $\overline{X}$ měla být podobná, a tedy SS_A by mělo být malé vzhledem k SS_e. Nulovou hypotézu budeme zamítat pro příliš vysoké hodnoty testové statistiky F_A.
Kritický obor: H_0 zamítáme na hladině $\alpha \Leftrightarrow F_A \geq F_{k-1,N-k}(1-\alpha)$,
kde $F_{k-1,N-k}(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdělení $F_{k-1,N-k}(1-\alpha)$.

Poznámka. Tento test bývá nazýván jako F-test analýzy rozptylu.

Poznámka. Všimněme si, že by také bylo postačující, pokud by v modelu (2.1) byl splněn předpoklad shody rozptylů pouze za platnosti nulové hypotézy. Museli bychom však poté k větě 3 přidat předpoklad, že nulová hypotéza platí. Tento poznatek nám bude užitečný v následující kapitole.

3. Testy shody rozptylů

V této kapitole se budeme zabývat testy shody rozptylů v jednoduchém třídění. V matematické statistice existuje nezanedbatelné množství modelů, které ve svých předpokladech mají zahrnut i předpoklad shody rozptylů. Příkladem je výše zmíněná analýza rozptylu jednoduchého třídění. Testy shody rozptylů nám mohou pomoci ověřit, zda tento předpoklad není porušen. Zaměříme se na Leveneův test. Brown-Forsytheův a Bartlettův test představíme pouze stručněji. V celé kapitole budeme používat stejné značení jako v kapitole 2.

3.1 Leveneův test

Uvažujme $k \geq 2$ nezávislých náhodných výběrů s rozsahy $N_1, \ldots, N_k$, kde i-tý náhodný výběr pochází z rozdělení $\mathrm{N}(\mu_i, \sigma_i^2)$, $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, \ldots, k$, tj.

$$\begin{gathered} X_{1,1}, \ldots, X_{1,N_1} \sim \mathrm{N}(\mu_1, \sigma_1^2), \\ \vdots \\ X_{k,1}, \ldots, X_{k,N_k} \sim \mathrm{N}(\mu_k, \sigma_k^2). \end{gathered}$$

Zajímá nás, jestli uvažované náhodné výběry pocházejí z rozdělení se shodným rozptylem. Chceme testovat nulovou hypotézu

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

proti alternativní hypotéze

$$H_1 : \exists i \neq j : \sigma_i^2 \neq \sigma_j^2.$$

Nejprve uvedeme prvotní nápad, který motivoval vznik samotného testu. Uvažujme náhodné veličiny $A_{i,j} = (X_{i,j} - \mu_i)^2$, $i = 1, \ldots, k$, $j = 1, \ldots, N_i$. Náhodné veličiny $A_{i,j}$ jsou nezávislé. Dále platí $\mathsf{E}(A_{i,j}) = \sigma_i^2$ a $\mathsf{var}(A_{i,j}) = 2\sigma_i^4$. Získali jsme tedy k nových nezávislých náhodných výběrů. Chtěli bychom na tyto výběry aplikovat analýzu rozptylu jednoduchého třídění. Náhodné veličiny $A_{i,j}$ však nejsou normálně rozdělené.

Uvažujme nyní náhodné veličiny $B_{i,j} = |X_{i,j} - \mu_i|$, $i = 1, \ldots, k$, $j = 1, \ldots, N_i$. Všimněme si, že $B_{i,j}$ jsou nezávislé a dle definice 6 mají polonormální rozdělení s parametrem σ_i^2. Pak dle tvrzení 2 platí

$$\mathsf{E}(B_{i,j}) = \sqrt{\frac{2}{\pi}}\sigma_i, \quad \mathsf{var}(B_{i,j}) = \left(1 - \frac{2}{\pi}\right)\sigma_i^2.$$

Opět jsme tedy získali k nových nezávislých náhodných výběrů. Ani $B_{i,j}$ nejsou normálně rozdělené, ale jejich rozdělení bude mít na spolehlivost F-testu analýzy rozptylu méně negativní vliv, než jaký by mělo rozdělení náhodných veličin $A_{i,j}$. Důvodem je, že dle Levene (1960) lze při porušení předpokladu normality přibližně určit míru negativního vlivu, který má rozdělení náhodných veličin na spolehlivost F-testu analýzy rozptylu, pomocí jejich třetího a čtvrtého normalizovaného kumulantu, tj. γ_3 a γ_4. Pro náhodnou veličinu s rozdělením $\mathrm{N}(\mu_i, \sigma_i^2)$ platí

$\gamma_3 = 0$ a $\gamma_4 = 0$. Dále dle Levene (1960) pro $A_{i,j}$ platí $\gamma_3 = 2{,}83$, $\gamma_4 = 12$ a pro $B_{i,j}$ platí $\gamma_3 = 1$, $\gamma_4 = 0{,}39$. Tento fakt naznačuje, že F-test analýzy rozptylu by měl být více spolehlivý, pokud ho budeme aplikovat na pozorování $B_{i,j}$. V praxi však obvykle μ_i nejsou známé, proto budeme uvažovat náhodné veličiny $Y_{i,j} = |X_{i,j} - \overline{X}_i|$, $i = 1, \dots, k$, $j = 1, \dots, N_i$. Nyní nastává další komplikace, jelikož $Y_{i,j}$ a $Y_{i,l}$ nejsou nezávislé pro $j \neq l$, a tedy $Y_{i,1}, \dots, Y_{i,N_i}$ nejsou náhodné výběry. Fisher (1920) ukazuje, že korelace $Y_{i,j}$ a $Y_{i,l}$ je řádu N_i^{-2}. Zde nám však ani nulová korelace nezajistí nezávislost, jelikož $Y_{i,j}$ a $Y_{i,l}$ nejsou normálně rozdělené, tedy nemohou mít ani sdružené normální rozdělení. Avšak dle Levene (1960) by korelace takového řádu pravděpodobně neměla mít vážný vliv na rozdělení testové statistiky (2.2). Miller (1972) ale upozorňuje, že pokud jsou rozsahy náhodných výběrů velmi malé, pak korelace způsobí, že F-test analýzy rozptylu nebude spolehlivý. Nyní se zaměříme na momenty $Y_{i,j}$.

Tvrzení 5. *Nechť $X_{i,j} \sim \mathrm{N}(\mu_i, \sigma_i^2)$ jsou nezávislé náhodné veličiny, kde $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, \dots, k$, $j = 1, \dots, N_i$. Dále uvažujme náhodné veličiny $Y_{i,j} = |X_{i,j} - \overline{X}_i|$. Pak platí*

$$\mathsf{E}(Y_{i,j}) = \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_i}\right)}\sigma_i, \quad \mathsf{var}(Y_{i,j}) = \left(1 - \frac{2}{\pi}\right)\left(1 - \frac{1}{N_i}\right)\sigma_i^2.$$

Důkaz. Uvažujme nejprve $Z_{i,j} = X_{i,j} - \overline{X}_i$.

$$Z_{i,j} = X_{i,j} - \overline{X}_i = X_{i,j} - \frac{1}{N_i}\sum_{l=1}^{N_i} X_{i,l} = X_{i,j}\left(1 - \frac{1}{N_i}\right) - \frac{1}{N_i}\sum_{l \neq j} X_{i,l}.$$

Víme, že $X_{i,j} \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Tedy $Z_{i,j}$ je lineární kombinací nezávislých náhodných veličin s normálním rozdělením. Z toho plyne, že $Z_{i,j}$ má také normální rozdělení. Určíme střední hodnotu a rozptyl $Z_{i,j}$.

$$\begin{aligned}
\mathsf{E}(Z_{i,j}) &= \mathsf{E}(X_{i,j}) - \mathsf{E}(\overline{X}_i) = \mathsf{E}(X_{i,j}) - \mathsf{E}(X_{i,j}) = 0.\\
\mathsf{var}(Z_{i,j}) &= \mathsf{var}(X_{i,j} - \overline{X}_i) = \mathsf{var}\left(X_{i,j}\left(1 - \frac{1}{N_i}\right) - \frac{1}{N_i}\sum_{l \neq j} X_{i,l}\right)\\
&= \mathsf{var}\left(X_{i,j}\left(1 - \frac{1}{N_i}\right)\right) + \mathsf{var}\left(\frac{1}{N_i}\sum_{l \neq j} X_{i,l}\right)\\
&= \left(1 - \frac{1}{N_i}\right)^2 \mathsf{var}(X_{i,j}) + \frac{1}{N_i^2}\sum_{l \neq j}\mathsf{var}(X_{i,l}) = \left(\frac{N_i - 1}{N_i}\right)^2 \sigma_i^2 + \frac{N_i - 1}{N_i^2}\sigma_i^2\\
&= \frac{N_i - 1}{N_i}\sigma_i^2 = \left(1 - \frac{1}{N_i}\right)\sigma_i^2.
\end{aligned}$$

Tedy $Z_{i,j} \sim \mathrm{N}\left(0, (1 - \frac{1}{N_i})\sigma_i^2\right)$. Víme, že $Y_{i,j} = |Z_{i,j}|$. Dle definice 6 má $Y_{i,j}$ polonormální rozdělení s parametrem $(1 - \frac{1}{N_i})\sigma_i^2$. Pak dle tvrzení 2 platí

$$\begin{aligned}
\mathsf{E}(Y_{i,j}) &= \sqrt{\frac{2}{\pi}}\sqrt{1 - \frac{1}{N_i}}\sigma_i = \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_i}\right)}\sigma_i,\\
\mathsf{var}(Y_{i,j}) &= \left(1 - \frac{2}{\pi}\right)\left(1 - \frac{1}{N_i}\right)\sigma_i^2.
\end{aligned}$$

□

Budeme tedy uvažovat k nezávislých skupin pozorování

$$\begin{gathered} Y_{1,1}, \ldots, Y_{1,N_1}, \\ \vdots \\ Y_{k,1}, \ldots, Y_{k,N_k}, \end{gathered} \tag{3.1}$$

na které aplikujeme analýzu rozptylu jednoduchého třídění. Testujeme nulovou hypotézu

$$H_0 : \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_1}\right)}\sigma_1 = \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_2}\right)}\sigma_2 = \cdots = \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_k}\right)}\sigma_k \tag{3.2}$$

proti alternativní hypotéze

$$H_1 : \exists i \neq j : \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_i}\right)}\sigma_i \neq \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{N_j}\right)}\sigma_j.$$

Za platnosti nulové hypotézy (3.2) je pro aplikaci analýzy rozptylu jednoduchého třídění splněn předpoklad, že uvažované skupiny pozorování (3.1) pocházejí z rozdělení se shodným rozptylem. Dále pokud rozsahy všech náhodných výběrů jsou stejné, tj. $N_1 = N_2 = \cdots = N_k$, pak je nulová hypotéza (3.2) ekvivalentní nulové hypotéze, kterou jsme chtěli původně testovat na začátku kapitoly, tj.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2.$$

Budou-li však rozsahy náhodných výběrů dostatečně velké, pak $\sqrt{1 - \frac{1}{N_i}}$ budou blízko k číslu jedna. Je důležité podotknout, že pokud rozsahy všech náhodných výběrů nebudou stejné a zároveň nebudou dostatečně velké, tak může nastat situace, kdy nulová hypotéza (3.2) platí, ale přitom se rozptyly $\sigma_1^2, \ldots, \sigma_k^2$ budou podstatně lišit. Podobně může nastat situace, kdy $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, ale nulová hypotéza (3.2) neplatí. Tedy testování nulové hypotézy (3.2) nám v tomto případě neříká nic o tom, zda $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

Dříve jsme zmínili, že $Y_{i,j}$ nesplňují předpoklad normality. Zajímá nás tedy rozdělení testové statistiky (2.2), pokud náhodné výběry nepocházejí z normálního rozdělení. Za tímto účelem zformulujeme následující tvrzení.

Tvrzení 6. *Uvažujme $k \geq 2$ nezávislých náhodných výběrů s rozsahy $N_1, \ldots, N_k$, kde i-tý náhodný výběr pochází z rozdělení se střední hodnotou $\mu \in \mathbb{R}$ a konečným rozptylem $\sigma^2 > 0$, $i = 1, \ldots, k$. Dále nechť $N = \sum_{i=1}^k N_i \to \infty$ a $\frac{N_i}{N} \to M_i > 0$ pro každé $i \in \{1, \ldots, k\}$, když $N \to \infty$. Pak pro testovou statistiku F_A definovanou vzorcem (2.2) platí*

$$(k-1)F_A \xrightarrow[N\to\infty]{D} \chi_{k-1}^2.$$

Důkaz. Upravíme $(k-1)F_A$ do tvaru

$$(k-1)F_A = \frac{SS_A}{SS_e/(N-k)} = \frac{SS_A/\sigma^2}{SS_e/(\sigma^2(N-k))}.$$

1) Nejprve ukážeme, že

$$\frac{SS_e}{\sigma^2(N-k)} \xrightarrow[N\to\infty]{P} 1.$$

Označme $Z_{i,j}$ j-té pozorování i-tého náhodného výběru a upravujme

$$\begin{aligned}\frac{SS_e}{N-k} &= \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{i,j}-\overline{Z}_i)^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{i,j}^2 - 2Z_{i,j}\overline{Z}_i + \overline{Z}_i^2)\\ &= \frac{1}{N-k}\sum_{i=1}^{k}\Big(\sum_{j=1}^{N_i}Z_{i,j}^2 - 2N_i\overline{Z}_i^2 + N_i\overline{Z}_i^2\Big) = \frac{1}{N-k}\sum_{i=1}^{k}\Big(\sum_{j=1}^{N_i}Z_{i,j}^2 - N_i\overline{Z}_i^2\Big)\\ &= \frac{1}{N-k}\sum_{i=1}^{k}N_i\Big(\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}^2 - \overline{Z}_i^2\Big).\end{aligned}$$

Nyní jelikož $\mathsf{E}(Z_{i,j}) = \mu$ a $\mathsf{var}(Z_{i,j}) = \sigma^2$, tak z Chinčinova slabého zákona velkých čísel (viz Anděl, 2011, str. 330, věta B.3) a věty o spojité transformaci (viz Anděl, 2011, str. 332, věta B.9) plyne

$$\begin{aligned}\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}^2 &\xrightarrow[N_i\to\infty]{P} \mathsf{E}(Z_{i,j}^2) = (\mathsf{E}(Z_{i,j}))^2 + \mathsf{var}(Z_{i,j}) = \mu^2+\sigma^2,\\ \overline{Z}_i^2 = \Big(\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}\Big)^2 &\xrightarrow[N_i\to\infty]{P} (\mathsf{E}(Z_{i,j}))^2 = \mu^2.\end{aligned}$$

Dohromady tedy máme

$$\Big(\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}^2 - \overline{Z}_i^2\Big) \xrightarrow[N_i\to\infty]{P} \sigma^2.$$

Počítejme dále

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N-k}\sum_{i=1}^{k}N_i\Big(\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}^2 - \overline{Z}_i^2\Big) &= \sum_{i=1}^{k}\lim_{N\to\infty}\frac{N_i}{N-k}\Big(\frac{1}{N_i}\sum_{j=1}^{N_i}Z_{i,j}^2 - \overline{Z}_i^2\Big)\\ &= \sum_{i=1}^{k}M_i\sigma^2 = \sigma^2.\end{aligned}$$

Z výše uvedeného celkem dostáváme

$$\frac{SS_e}{N-k} \xrightarrow[N\to\infty]{P} \sigma^2,$$

z čehož plyne

$$\frac{SS_e}{\sigma^2(N-k)} \xrightarrow[N\to\infty]{P} 1.$$

2) Nyní ukážeme, že

$$\frac{SS_A}{\sigma^2} \xrightarrow[N\to\infty]{D} \chi^2_{k-1}.$$

Provedeme sérii úprav

$$\begin{aligned}\frac{SS_A}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{k}N_i(\overline{Z}_i-\overline{Z})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{k}N_i\Big((\overline{Z}_i-\mu)-(\overline{Z}-\mu)\Big)^2\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{k}\Big(\sqrt{N_i}(\overline{Z}_i-\mu)-\sqrt{N_i}(\overline{Z}-\mu)\Big)^2.\end{aligned}$$

Uvažujme $W_i = \sqrt{N_i}(\overline{Z}_i - \mu)$ a pomocí W_i vyjádřeme $\overline{Z} - \mu$

$$\overline{Z} - \mu = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{i,j} - \mu) = \frac{1}{N}\sum_{i=1}^{k} N_i(\overline{Z}_i - \mu) = \frac{1}{N}\sum_{i=1}^{k}\sqrt{N_i}W_i.$$

Pokračujme v úpravách

$$\begin{aligned}
\frac{SS_A}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{k}\Big(\sqrt{N_i}(\overline{Z}_i - \mu) - \sqrt{N_i}(\overline{Z} - \mu)\Big)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{k}\Big(W_i - \frac{\sqrt{N_i}}{N}\sum_{i=1}^{k}\sqrt{N_i}W_i\Big)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{k}\Big(W_i^2 - 2W_i\frac{\sqrt{N_i}}{N}\sum_{i=1}^{k}\sqrt{N_i}W_i + \frac{N_i}{N^2}\big(\sum_{i=1}^{k}\sqrt{N_i}W_i\big)^2\Big) \\
&= \frac{1}{\sigma^2}\Big(\sum_{i=1}^{k}W_i^2 - \frac{2}{N}\big(\sum_{i=1}^{k}\sqrt{N_i}W_i\big)^2 + \frac{1}{N}\big(\sum_{i=1}^{k}\sqrt{N_i}W_i\big)^2\Big) \\
&= \frac{1}{\sigma^2}\Big(\sum_{i=1}^{k}W_i^2 - \frac{1}{N}\big(\sum_{i=1}^{k}\sqrt{N_i}W_i\big)^2\Big).
\end{aligned}$$

Označme $\boldsymbol{W} = (W_1, \ldots, W_k)^T$, $\sqrt{\boldsymbol{N}} = (\sqrt{N_1}, \ldots, \sqrt{N_k})^T$, $\mathbb{A} = \mathbb{I}_k - \frac{1}{N}\sqrt{\boldsymbol{N}}\sqrt{\boldsymbol{N}}^T$ a pokračujme v úpravách

$$\begin{aligned}
\frac{SS_A}{\sigma^2} &= \frac{1}{\sigma^2}\Big(\sum_{i=1}^{k}W_i^2 - \frac{1}{N}\big(\sum_{i=1}^{k}\sqrt{N_i}W_i\big)^2\Big) \\
&= \frac{1}{\sigma^2}\Big(\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{N}\boldsymbol{W}^T\sqrt{\boldsymbol{N}}\sqrt{\boldsymbol{N}}^T\boldsymbol{W}\Big) \\
&= \frac{\boldsymbol{W}^T}{\sigma}\Big(\mathbb{I}_k - \frac{1}{N}\sqrt{\boldsymbol{N}}\sqrt{\boldsymbol{N}}^T\Big)\frac{\boldsymbol{W}}{\sigma} = \frac{\boldsymbol{W}^T}{\sigma}\mathbb{A}\frac{\boldsymbol{W}}{\sigma}.
\end{aligned}$$

Uvažujme nyní také nezávislé náhodné veličiny $U_i \sim \mathrm{N}(0,1)$ a označme $\boldsymbol{U} = (U_1, \ldots, U_k)^T$, tedy $\boldsymbol{U} \sim \mathrm{N}_k(\boldsymbol{0}, \mathbb{I}_k)$. Dle Lindebergovy centrální limitní věty (viz Anděl, 2011, str. 331, věta B.5) platí

$$W_i = \sqrt{N_i}(\overline{Z}_i - \mu) \xrightarrow[N_i \to \infty]{D} \mathrm{N}(0, \sigma^2),$$

pak tedy také

$$\frac{W_i}{\sigma} \xrightarrow[N_i \to \infty]{D} U_i.$$

Zároveň si všimněme, že W_i jsou nezávislé. Celkem tedy dostáváme

$$\frac{\boldsymbol{W}}{\sigma} \xrightarrow[N \to \infty]{D} \boldsymbol{U}.$$

Označme $\mathbb{B} = \mathbb{I}_k - \sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T$, kde $\sqrt{\boldsymbol{M}} = (\sqrt{M_1}, \ldots, \sqrt{M_k})^T$. Potom platí $\mathbb{A} \to \mathbb{B}$, když $N \to \infty$. Matice $\mathbb{B}\mathbb{I}_k = \mathbb{B}$ je nenulová a idempotentní

$$\begin{aligned}
\mathbb{B}\mathbb{B} &= \Big(\mathbb{I}_k - \sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T\Big)\Big(\mathbb{I}_k - \sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T\Big) \\
&= \mathbb{I}_k - 2\sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T + \sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T\sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T \\
&= \mathbb{I}_k - \sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^T = \mathbb{B}.
\end{aligned}$$

Lze také snadno nahlédnout, že matice $\mathbb{B}$ je symetrická. Z toho vcelku plyne, že $\mathbb{B}$ je pozitivně semidefinitní, jelikož $\boldsymbol{x}^T\mathbb{B}\boldsymbol{x} = \boldsymbol{x}^T\mathbb{B}^T\mathbb{B}\boldsymbol{x} = (\mathbb{B}\boldsymbol{x})^T(\mathbb{B}\boldsymbol{x}) \geq 0$, kde $\boldsymbol{x} \in \mathbb{R}^k$. Dále $\mathrm{tr}(\mathbb{B}\mathbb{I}_k) = \mathrm{tr}(\mathbb{B}) = k-1$. Ve výsledku spojením Crámerovo-Sluckého věty (viz Anděl, 2011, str. 333, věta B.10) a věty o rozdělení kvadratické formy (viz věta 1) dostáváme

$$\frac{SS_A}{\sigma^2} = \frac{\boldsymbol{W}^T}{\sigma}\mathbb{A}\frac{\boldsymbol{W}}{\sigma} \xrightarrow[N\to\infty]{D} \boldsymbol{U}^T\mathbb{B}\boldsymbol{U} \sim \chi^2_{k-1}.$$

3) Nyní použijeme Crámerovo-Sluckého větu na výsledky obdržené z 1) a 2), tedy

$$(k-1)F_A = \frac{SS_A/\sigma^2}{SS_e/(\sigma^2(N-k))} \xrightarrow[N\to\infty]{D} \chi^2_{k-1},$$

čímž je důkaz dokončen.

□

Poznámka. Za platnosti nulové hypotézy (3.2) pochází všechny skupiny pozorování (3.1) z rozdělení se stejnou střední hodnotou a stejným rozptylem, jak bylo řečeno výše.

Testová statistika Leveneova testu je tvaru

$$W_L = \frac{\sum_{i=1}^k N_i(\overline{Y}_i - \overline{Y})^2/(k-1)}{\sum_{i=1}^k \sum_{j=1}^{N_i}(Y_{i,j} - \overline{Y}_i)^2/(N-k)},$$

kde $Y_{i,j} = |X_{i,j} - \overline{X}_i|$. Nulovou hypotézu (3.2) budeme zamítat pro příliš vysoké hodnoty testové statistiky W_L.
Kritický obor: H_0 zamítáme přibližně na hladině $\alpha \Leftrightarrow W_L \geq \chi^2_{k-1}(1-\alpha)/(k-1)$, kde $\chi^2_{k-1}(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdělení χ^2_{k-1}.

Obvykle také bývá kritický obor formulovaný konzervativnějším způsobem, kde H_0 zamítáme přibližně na hladině α, pokud $W_L \geq F_{k-1,N-k}(1-\alpha)$. Důvodem je, že pro $N > k$ platí $F_{k-1,N-k}(1-\alpha) \geq \chi^2_{k-1}(1-\alpha)/(k-1)$. To znamená, že takto zvoleným kritickým oborem nezvyšujeme pravděpodobnost chyby I. druhu.

3.2 Brown-Forsytheův test

Brown-Forsytheův test představuje modifikaci Leveneova testu, který jsme odvodili v podkapitole 3.1. Modifikace spočívá v záměně střední hodnoty za medián. U Leveneova testu jsme nahradili střední hodnotu výběrovým průměrem, jelikož v praxi není běžné známá. Medián také v praxi není běžně známý, proto ho analogicky nahradíme výběrovým mediánem. Budeme tedy nyní místo náhodných veličin $Y_{i,j} = |X_{i,j} - \overline{X}_i|$ uvažovat náhodné veličiny $Z_{i,j} = |X_{i,j} - \widetilde{X}_i|$, kde $\widetilde{X}_i$ značí výběrový medián i-tého náhodného výběru, $i = 1, \ldots, k$, $j = 1, \ldots, N_i$. Brown-Forsytheův test probíhá stejně jako Leveneův test. Na náhodné veličiny $Z_{i,j}$ aplikujeme analýzu rozptylu jednoduchého třídění.

Testová statistika Brown-Forsytheova testu je tvaru

$$W_{BF} = \frac{\sum_{i=1}^k N_i(\overline{Z}_i - \overline{Z})^2/(k-1)}{\sum_{i=1}^k \sum_{j=1}^{N_i}(Z_{i,j} - \overline{Z}_i)^2/(N-k)},$$

kde $Z_{i,j} = |X_{i,j} - \widetilde{X}_i|$. Nulovou hypotézu o shodě rozptylů budeme zamítat pro příliš vysoké hodnoty testové statistiky W_{BF}.
Kritický obor: H_0 zamítáme přibližně na hladině $\alpha \Leftrightarrow W_{BF} \geq \chi^2_{k-1}(1-\alpha)/(k-1)$. Obvykle bývá opět používán konzervativnější přístup, kde H_0 zamítáme přibližně na hladině α, pokud $W_{BF} \geq F_{k-1,N-k}(1-\alpha)$.

3.3 Bartlettův test

Tato podkapitola vychází z práce Zvára (2008). Uvažujme $k \geq 2$ nezávislých náhodných výběrů z normálního rozdělení, které jsme uvedli na začátku podkapitoly 3.1. Odhad rozptylu i-tého náhodného výběru označme

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{i,j} - \overline{X}_i)^2, \quad i = 1, \dots, k.$$

Odhad celkového rozptylu všech pozorování označme

$$S^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{i,j} - \overline{X}_i)^2 = \sum_{i=1}^{k} \frac{N_i - 1}{N - k} S_i^2.$$

Dále uvažujme konstantu C, která je tvaru

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{N_i - 1} - \frac{1}{N-k} \right).$$

Testová statistika Bartlettova testu je

$$\begin{aligned} B &= \frac{1}{C} \left((N-k) \log S^2 - \sum_{i=1}^{k} (N_i - 1) \log S_i^2 \right) \\ &= \frac{N-k}{C} \left(\log S^2 - \sum_{i=1}^{k} \frac{N_i - 1}{N - k} \log S_i^2 \right). \end{aligned}$$

Je-li $N_i \geq 7$ pro $i = 1, \dots, k$, pak se udává, že testová statistika B má za platnosti nulové hypotézy o shodě rozptylů přibližně rozdělení χ^2_{k-1}. Vysoké hodnoty testové statistiky B svědčí proti platnosti nulové hypotézy. Nulovou hypotézu tedy zamítneme přibližně na hladině α, pokud $B \geq \chi^2_{k-1}(1-\alpha)$. Bartlettův test je velmi citlivý na porušení předpokladu normality. Lze ho dokonce považovat za test normality. Avšak v situaci, kdy je předpoklad normality splněn, se zdá být nejsilnější z dostupných testů.

4. Simulační studie

Cílem této kapitoly je pomocí simulační studie ověřit, zda jsou testy uvedené v předchozí kapitole schopny dodržet požadovanou hladinu.

4.1 Návrh simulační studie

Prostřednictvím programu R (R Core Team, 2024) budeme generovat trojice pseudonáhodných výběrů z vybraných rozdělení. Tato rozdělení jsou

1. Normální rozdělení $\mathrm{N}(\mu_i, \sigma^2)$,
2. Studentovo t-rozdělení t_5 posunuté o parametr μ_i,
3. Logaritmicko-normální rozdělení $\mathrm{LN}(0, \sigma^2)$ posunuté o parametr μ_i,

kde $i = 1, 2, 3$. Pracujeme s hodnotami $\sigma^2 = 1$ a $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 1$. Trojice pseudonáhodných výběrů tedy budeme generovat za platnosti nulové hypotézy shody rozptylů. Dále při generování trojic výběrů z daného rozdělení budeme i-tý výběr každé trojice generovat z rozdělení s parametrem μ_i. Pro výběry budeme uvažovat rozsahy (10, 10, 10), (50, 50, 50), (100, 100, 100), tj. uvažujeme vyvážené výběry. Pro každou z těchto voleb rozsahů vygenerujeme z každého rozdělení 10 000 trojic výběrů. Pro každou trojici budeme testovat nulovou hypotézu shody rozptylů pomocí testů uvedených v kapitole 3 a uvedeme, u kolika z nich ji zamítneme. Tyto testy jsou

1. Leveneův test s testovou statistikou W_L,
2. Brown-Forsytheův test s testovou statistikou W_{BF},
3. Bartlettův test s testovou statistikou B.

Simulace budeme provádět pro testy na hladině $\alpha = 0{,}05$ a také na hladině $\alpha = 0{,}01$.

4.2 Výsledky simulační studie

Výsledky simulační studie jsou prezentovány v níže uvedených tabulkách. Lze v nich nalézt poměr zamítnutí nulové hypotézy v procentech pro rozdělení uvedená v návrhu simulační studie. Rozsahy výběrů jsou v tabulkách značeny symbolem N. Tabulky jsou uvedeny pouze pro testy na hladině $\alpha = 0{,}05$, jelikož testy na hladině $\alpha = 0{,}01$ měly z hlediska dodržení hladiny analogické výsledky.

Pro normální rozdělení, které je symetrické a má lehké chvosty, byla hladina testu přibližně dodržena všemi testy pro téměř každou volbu rozsahů. Bartlettův test zamítal nulovou hypotézu téměř vždy tak, jak bychom očekávali. Leveneův test zamítal nulovou hypotézu přibližně o 1 % častěji v případě, kdy rozsahy všech výběrů byly rovny 10. Brown-Forsytheův test naopak nulovou hypotézu zamítal méně často, kdy pro rozsahy rovny 10 tomu tak dokonce bylo o téměř 2 %. V případě normálního rozdělení se zdá správnou volbou volit kterýkoliv z testů, ačkoliv pro malé rozsahy výběrů se zdá lepší zvolit Bartlettův či Brown-Forsytheův test.

N	Levene	Brown-Forsythe	Bartlett
10	5,70	3,28	4,74
50	5,23	4,25	5,15
100	5,11	4,77	5,09

Tabulka 4.1: Poměr zamítnutí nulové hypotézy v procentech pro normální rozdělení na hladině $\alpha = 0{,}05$.

Pro Studentovo t-rozdělení o 5 stupních volnosti, které je symetrické a má těžší chvosty než normální rozdělení, nedodržel hladinu testu zejména Bartlettův test. Projevila se zde jeho značná citlivost na porušení předpokladu normality. Leveneův test zamítal nulovou hypotézu přibližně o 2 % častěji, když rozsahy výběrů byly rovny 10. Brown-Forsytheův test měl velmi podobné výsledky jako pro normální rozdělení. V této situaci lze zvolit Leveneův i Brown-Forsytheův test, ačkoliv pro malé rozsahy výběrů se Leveneův test zdá být opět méně vhodný.

N	Levene	Brown-Forsythe	Bartlett
10	6,85	3,31	16,80
50	5,20	4,25	27,70
100	5,14	4,60	30,84

Tabulka 4.2: Poměr zamítnutí nulové hypotézy v procentech pro Studentovo t-rozdělení o 5 stupních volnosti na hladině $\alpha = 0{,}05$.

Pro logaritmicko-normální rozdělení, které má těžší chvosty než Studentovo t-rozdělení o 5 stupních volnosti a je jako jediné asymetrické, zcela selhává Leveneův i Bartlettův test. Leveneův test zamítal nulovou hypotézu přibližně o 20 % častěji, než bychom očekávali. U Bartlettova testu tomu bylo dokonce o 75 % častěji pro větší rozsahy výběrů, jelikož toto rozdělení z výše uvedených nejvíce porušuje předpoklad normality. Brown-Forsytheův test jako jediný byl schopen dodržet hladinu 5 %. Nulovou hypotézu zamítá opět méně často, než bychom očekávali. V této situaci je jediným testem, který je vhodné použít.

N	Levene	Brown-Forsythe	Bartlett
10	26,95	3,95	58,30
50	24,81	3,76	77,06
100	24,23	4,49	81,44

Tabulka 4.3: Poměr zamítnutí nulové hypotézy v procentech pro logaritmicko-normální rozdělení na hladině $\alpha = 0{,}05$.

Závěr

Práce představila tři testy na shodu rozptylů v jednoduchém třídění. Nejprve jsme uvedli důležité poznatky z teorie pravděpodobnosti a také jsme si připomněli analýzu rozptylu jednoduchého třídění. Na tyto kapitoly jsme plynule navázali odvozením Leveneova testu, jenž je na těchto znalostech založen. Vysvětlili jsme také, že Leveneův test je pro testování nulové hypotézy shody rozptylů vhodné použít pouze tehdy, když rozsahy všech výběrů jsou stejné nebo dostatečně velké. Dále jsme uvedli Brown-Forsytheův test, který je modifikací Leveneova testu. V neposlední řadě jsme představili také Bartlettův test. V závěru práce jsme provedli simulační studii, kde jsme zkoumali, zda jsou testy schopny dodržet požadovanou hladinou v závislosti na rozdělení, ze kterého výběry pocházely. V rámci simulací se potvrdila značná citlivost Bartlettova testu na porušení předpokladu normality, jelikož pro jiné než normální rozdělení nebyl schopen dodržet požadovanou hladinu. Dále jsme zjistili, že Leveneův test je na porušení předpokladu normality naopak méně citlivý. Selhává však tehdy, když výběry pocházejí z logaritmicko-normálního rozdělení, které je asymetrické a má těžké chvosty. Zároveň také zamítá častěji pro malé rozsahy výběrů. Brown-Forsytheův test byl jako jediný schopen dodržet požadovanou hladinu pro všechna uvedená rozdělení. Na základě obdržených výsledků tedy lze říci, že pokud si nejsme jisti, z jakého rozdělení data pocházejí, tak Brown-Forsytheův test se zdá být z uvedených testů nejvhodnější.

Seznam použité literatury

Anděl, J. (2011). *Základy matematické statistiky.* Třetí vydání. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Fisher, R. A. (1920). A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error. *Monthly Notices of the Royal Astronomical Society*, **80**(8), 758–770. ISSN 0035-8711.

Kulich, M. (2022). NMSA331 Matematická statistika 1: Poznámky k přednášce. URL `https://www.karlin.mff.cuni.cz/~komarek/vyuka/2023_24/nmsa331/ms1.pdf`. Přístup 5. 4. 2024.

Levene, H. (1960). Robust tests for equality of variances. In Olkin, I. a kol., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, 278–292. ISBN 978-0804705967.

Miller, A. J. (1972). Letters to the editor. *Technometrics*, **14**(2), 507. ISSN 00401706.

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Zvára, K. (2008). *Regrese.* První vydání. Matfyzpress, Praha. ISBN 978-80-7378-041-8.

Seznam tabulek

Přílohy

Skript psaný v programu R, pomocí kterého byla provedena simulační studie v kapitole 4, lze nalézt v příloze práce pod názvem: `skript-simulace.R`