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

BAKALÁŘSKÁ PRÁCE

Vliv porušení předpokladů při analýze rozptylu

Vedoucí bakalářské práce: **RNDr. et PhDr. Ivo Müller, Ph.D.**
Vypracoval: **PhDr. Daniel Dostál, Ph.D.**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2015

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: PhDr. Daniel Dostál, Ph.D.

Název práce: Vliv porušení předpokladů při analýze rozptylu

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: RNDr. et PhDr. Ivo Müller, Ph.D.

Rok obhajoby práce: 2015

Abstrakt: Cílem této práce je popsat předpoklady jednofaktorové analýzy rozptylu a následně ověřit dopad jejich porušení na celkovou přesnost testu. Důraz je kladen zejména na předpoklad homoskedasticity a předpoklad normálního rozdělení náhodné veličiny. Výsledky simulační studie potvrzují, že porušení libovolného z těchto předpokladů vede ke změně průběhu silofunkce. V případě heteroskedasticity se snižuje síla testu a zvyšuje pravděpodobnost chyby prvního druhu. Výrazné zešikmení náhodných veličin při zachování konstantního rozptylu naopak sílu testu zvyšuje a snižuje pravděpodobnost chyby prvního druhu. Efekt porušení podmínek je zkoumán ve vztahu k rozsahu výběru, jejich vyváženosti a počtu srovnávaných skupin.

Klíčová slova: analýza rozptylu, ANOVA, předpoklady, silofunkce

Počet stran: 56

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: PhDr. Daniel Dostál, Ph.D.

Title: The influence of assumption violation on analysis of variance

Type of thesis: Bachelor's

Department:
Department of Mathematical Analysis and Application of Mathematics

Supervisor: RNDr. et PhDr. Ivo Müller, Ph.D.

The year of presentation: 2015

Abstract: The aim of this thesis is to describe the assumptions of one-way analysis of variance and to verify the impact of the assumption violation on the test accuracy. The emphasis is on the assumption of homoscedasticity and the one of normal distribution. The simulation study confirms that violation of either condition leads to a change in power curve. In the case of heteroscedasticity, the power of the test is reduced and the probability of the type I error increases. On the contrary, significant skew of random variables, while a constant variance is maintained, leads to the increase of the power and reduction of the type I error probability. The effect of violations is examined in relation to the sample size, the balance and the number of compared groups.

Key words: analysis of variance, ANOVA, assumptions, power curve

Number of pages: 56

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením pana doktora Iva Müllera s použitím uvedené literatury.

V Olomouci dne 7. května 2015

Obsah

Poděkování

Touto prací uzavírám své studium aplikované statistiky, které nepřehlédnutelně ovlivňovalo náplň mých dnů s ročním přerušením v letech 2011 až 2015. Za to, že tento čas můžu označit nejen za přínosný, ale i příjemně strávený, patří díky mým spolužákům. Zejména Michalovi Polákovi, Adéle Vrtkové, Patrikovi Vidlářovi a později Tomovi Zdražilovi a Elišce Calábkové.

Bezmála veškeré znalosti statistiky, které jsem v těchto letech získal, mi poskytli čtyři učitelé – Ondřej Vencálek, Karel Hron, Eva Fišerová a Ivo Müller. Ač každý z nich poznatky předával svým osobitým způsobem, dohromady dali vzniknout pestré a vzájemně propojené mozaice poznání, za což jim jsem velmi vděčný. Poslední jmenovaný se zhostil také role mého vedoucího práce a byl mi průvodcem v tomto pozoruhodném světě statistiky. Za jeho podnětné nápady a připomínky patří největší díky právě jemu.

Úvod

Jeden z mých přátel z akademického prostředí formuloval – zřejmě na nějaké obzvláště nudné konferenci – originální hypotézu. Týká se toho, jak se zástupci nejrůznějších empirických vědních oborů staví ke statistickým testům, konkrétně k podmínkám jejich užití. Dle této hypotézy existují tři stádia, kterými si badatel prochází. V prvním, které obvykle proběhne ještě během studií, se dozví o existenci statistických testů a bezstarostně pak ověřuje statistickou významnost všech svých hypotéz bez ohledu na tvar rozdělení či další podmínky, kterými jsou testy svázány. Jednoho dne však pronikne ve svém poznání o něco hlouběji a zjistí, že statistické testy jsou odvozeny jen pro velmi specifické případy a že data, která získává nepřesnými nástroji, jen zřídkakdy těmto případům odpovídají. Takový badatel se pozná podle toho, že na konferencích se po každém příspěvku ptá, jestli byl proveden test normality, homoskedasticity atd., a pokud ne, tak označí výsledky za pochybné a nevěrohodné. Po letech praxe se badatel posune do třetího a posledního stádia. Zjistí, že řada statistických testů přináší výsledky zkreslené jen neznatelně bez ohledu na to, jestli testy byly použity v souladu se svou definicí, nebo jestli byla některá z podmínek jejich užití do nějaké míry porušena. Třetí stádium se tak na první pohled vypadá stejně jako to první.

Otázkou zůstává, který z těchto přístupů je správný. Který badatel se dopouští větší chyby – ten co při srovnání dvou souborů řekněme o stovce měření sáhne po t-testu, přestože jsou data nápadně zešikmená, nebo ten, který dá přednost neparametrickému postupu a připraví se tak nejspíš o část síly testu, možnost elegantně prezentovat popisné statistiky a tak dál? Odpověď se bude zřejmě lišit podle toho, komu tuto otázku položíme.

Úkolem této práce je se zapojit do takovéto diskuse. Na následujících stranách se pokusíme čtenáři přiblížit jeden z nejpoužívanějších statistických testů – analýzu rozptylu – a zmapovat, jak se tato metoda chová, když je aplikována na data, která do větší či menší míry nevyhovují jejím předpokladům.

Práce je rozdělena na čtyři kapitoly. V první popisujeme princip analýzy

rozptylu a její předpoklady a demonstrujeme její užití na konkrétních datech. V druhé kapitole čtenáři přiblížíme design prezentovaného výzkumu a definujeme parametry, s jejichž hodnotami budeme manipulovat. Použitou metodou jsou počítačové simulace – test budeme opakovaně aplikovat na posloupnosti pseudonáhodně generovaných čísel a sledovat v kolika procentech případů zamítá či přijímá nulovou hypotézu. Ve třetí kapitole představíme výsledky těchto simulací. Zaměříme se na chování testu za platnosti nulové hypotézy i na celý průběh silofunkce. V poslední kapitole tyto výsledky diskutujeme a reflektujeme silné i slabé stránky této studie.

1. Analýza rozptylu jednoduchého třídění

Idea analýzy rozptylu jednoduchého třídění[1] (zkráceně ANOVA z anglického analysis of variance) byla ve dvacátých letech minulého století navržena britským statistikem a genetikem sirem Ronaldem A. Fisherem (1921; 1925). Metoda si našla obrovské uplatnění v řadě vědních oborů a například v rámci psychologie bývá považována za nejhojněji užívaný statistický test (viz např. Howell, 2013).

Pomocí analýzy rozptylu testujeme hypotézu, zdali k nezávislých náhodných výběrů pochází z rozdělení pravděpodobnosti se stejnými středními hodnotami. Testovanou nulovou hypotézu tedy můžeme zapsat jako

$$H_0 : \mu_1 = ... = \mu_k. \tag{1}$$

Alternativní hypotéza naopak předpokládá nejméně jednu dvojici rozdílných středních hodnot:

$$H_A : \exists i, j \in \{1, ..., k\}, i \neq j : \mu_i \neq \mu_j. \tag{2}$$

Nezávislé výběry nemusí být nutně stejného rozsahu. Počty pozorování v jednotlivých výběrech značíme $n_1, ..., n_k$ a celkový počet měření jako n.

Princip analýzy rozptylu lze vysvětlit několika způsoby. Metoda je obvykle prezentována jako speciální případ lineárního modelu, což je užitečné zejména tehdy, chceme-li postup dále zobecňovat například pro vícefaktorové designy. Předtím, než představíme čtenáři toto vysvětlení, nastíníme postup na základě poněkud odlišné úvahy, která může být bližší zejména čtenářům bez hlubší znalosti práce s lineárními modely.

Předpokládejme, že rozdělení pravděpodobnosti, z nichž pochází naše náhodné výběry, mají stejný rozptyl $\sigma^2 = \sigma_1^2 = ... = \sigma_k^2$ a že velikost tohoto rozptylu chceme odhadnout. Tento odhad pak můžeme stanovit nejméně dvěma způsoby.

Jednak můžeme vypočítat odhad rozptylu v rámci každého výběru podle

[1] V dalším textu budeme vynechávat specifikaci toho, že se jedná o případ jednoduchého třídění. V této práci se totiž jinými obměnami, tedy vícefaktorvými designy, nezabýváme.

vzorce

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2, \quad j = 1, ..., k, \tag{3}$$

kde y_{ji} je i-té měření v rámci j-tého výběru a

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}. \tag{4}$$

Celkový odhad (označme jej jako rozptyl uvnitř skupin $\hat{\sigma}_e^2$) pak získáme jako průměr odhadů $\hat{\sigma}_j^2$. V případě různých rozsahů souborů se bude jednat o vážený průměr, kde jako váhy použijeme stupně volnosti $(n_j - 1)$. Tedy

$$\hat{\sigma}_e^2 = \frac{1}{n - k} \sum_{j=1}^{k} (n_j - 1)\hat{\sigma}_j^2 = \frac{1}{n - k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2. \tag{5}$$

K výpočtu druhého odhadu rozptylu využijeme naši znalost toho, že rozptyl průměru $\bar{y}_j$ je n_j-krát menší než rozptyl jednotlivých měření, z nichž průměr počítáme. Nejprve budeme předpokládat, že rozsahy všech skupin jsou stejné, tedy $n_1 = n_2 = ... = n_k$. Rozptyl měření (který tentokrát budeme označovat jako rozptyl mezi skupinami $\hat{\sigma}_A^2$) pak můžeme odhadnout jako

$$\hat{\sigma}_A^2 = n_1 \hat{\sigma}_{\bar{y}}^2 = n_1 \frac{1}{k - 1} \sum_{j=1}^{k} (\bar{y}_j - \bar{y})^2. \tag{6}$$

V případě rozdílných rozsahů skupin budeme příslušnou hodnotou n_j násobit každý sčítanec. Získáme tak vztah

$$\hat{\sigma}_A^2 = \frac{1}{k - 1} \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2. \tag{7}$$

Obě získané hodnoty, tedy $\hat{\sigma}_e^2$ i $\hat{\sigma}_A^2$, můžeme považovat za odhad rozptylu σ^2. Je mezi nimi jeden pro nás důležitý rozdíl: zatímco hodnota $\hat{\sigma}_e^2$ není závislá na platnosti nulové hypotézy, tak $\hat{\sigma}_A^2$ se zvětšuje s tím, jak moc se různí hodnoty μ_j.

Na základě rozdílu mezi těmito dvěma odhady pak můžeme usuzovat na platnost nulové hypotézy.

K provedení testu statistické významnosti musíme nicméně do našich úvah zahrnout ještě dva předpoklady. Budeme předpokládat nezávislost náhodných veličin, z nichž pochází naše výběry, a budeme přepokládat, že tyto náhodné veličiny mají normální rozdělení. Za těchto okolností pak víme, že pokud platí nulová hypotéza, tak

$$\frac{(k-1)\hat{\sigma}_A^2}{\sigma^2} \sim \chi^2_{k-1}, \tag{8}$$

$$\frac{(n-k)\hat{\sigma}_e^2}{\sigma^2} \sim \chi^2_{n-k}, \tag{9}$$

a tedy, pokud jsou veličny (8) a (9) nezávislé, platí i vztah

$$\frac{\hat{\sigma}_A^2}{\hat{\sigma}_e^2} = F \sim F_{k-1,n-k}. \tag{10}$$

Nalezené hodnoty F (značíme jako f), které jsou vyšší než kvantil Fisherova-Snedecorova rozdělení $F_{k-1,n-k,(1-\alpha)}$, nás vedou k zamítnutí platnosti nulové hypotézy na hladině významnosti α. Případně můžeme stanovit pravděpodobnost, s jakou (za platnosti nulové hypotézy) získáme hodnoty f rovny nebo větší pozorované hodnotě, pomocí vztahu

$$p = 1 - F^{-1}_{k-1,n-k,(f)}. \tag{11}$$

Nalezená pravděpodobnost p se označuje jako p-hodnota.

Druhý způsob, jakým můžeme chápat analýzu rozptylu, je popsat ji jako srovnání dvou lineárních modelů. Toto vysvětlení nicméně vyžaduje podrobnější znalost práce s lineárními modely, kterou čtenáři neposkytneme. Pro vynechané detaily jej můžeme odkázat na publikaci (Anděl, 2007) či skripta (Fišerová, 2013).

Modely, se kterými budeme pracovat, lze obecně vyjádřit vztahem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{12}$$

V případě analýzy rozptylu jej můžeme zapsat pomocí následující struktury:

$$\begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \\ \vdots \\ y_{k,1} \\ \vdots \\ y_{k,n_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{k,1} \\ \vdots \\ \epsilon_{k,n_k} \end{pmatrix}. \tag{13}$$

Po roznásobení matic dostaneme rovnice

$$y_{ji} = \mu_j + \epsilon_{ji}, \quad j = 1, ..., k, \quad i = 1, ..., n_j.$$

Za účelem testování statistických hypotéz předpokládáme, že $\epsilon_{ji} \sim N(0, \sigma^2), j = 1, ..., k, i = 1, ..., n_j$, a že $\epsilon_{1,1}, ..., \epsilon_{k,n_k}$ jsou nezávislé.

Pomocí metody nejmenších čtverců získáme vztah pro nalezení vektoru odhadů μ_j:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \frac{1}{n_1} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & \frac{1}{n_k} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n_1} y_{1,i} \\ \vdots \\ \sum_{i=1}^{n_k} y_{k,i} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix}.$$

Pro jednotlivé odhady regresních parametrů tedy platí, že $\hat{\mu}_j = \bar{y}_j$.

V případě platnosti nulové hypotézy lze nicméně navržený model zjednodušit. Jelikož $\mu_1 = ... = \mu_k$, tak vektor $\boldsymbol{\beta}$ lze zapsat pomocí jediného prvku μ a matice $\mathbf{X}$

se zredukuje na sloupcový vektor jedniček o délce k. Nový model tedy vysvětluje jednotlivá měření jako

$$y_{ji} = \mu + \epsilon_{ji}, \quad j = 1, ..., k, \quad i = 1, ..., n_j.$$

Odhad jediného parametru μ stanovíme opět pomocí metody nejmenších čtverců jako $\hat{\mu} = \bar{y}$.

Původní úlohu, kdy jsme ověřovali shodu středních hodnot k výběrů, jsme tímto postupem převedli na ekvivalentní problém, kdy budeme testovat, jestli první model dokáže vysvětlit více variability než jeho podmodel. Pro srovnání přesnosti dvou modelů využíváme vztah

$$F = \frac{\frac{S_{e0} - S_{e1}}{\varphi_0 - \varphi_1}}{\frac{S_{e1}}{\varphi_1}}, \tag{14}$$

kde S_{e1}, respektive S_{e0}, jsou součty čtverců reziduí příslušného modelu (obecně $S_e = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$) a φ je počet stupňů volnosti modelu, což odpovídá počtu pozorování n zmenšeného o počet odhadovaných parametrů k. Pro takto získanou statistiku platí $F \sim F_{\varphi_0 - \varphi_1, \varphi_1}$, což nám umožňuje testovat platnost nulové hypotézy.

V případě analýzy rozptylu tradičně využíváme jinou terminologii a hovoříme o součtech čtverců mezi skupinami S_A, uvnitř skupin S_e (tzv. reziduální s. č.) a celkovém součtu čtverců S_T. Pro jmenované statistiky platí

$$S_A = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2, \qquad S_e = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2,$$

$$S_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ji} - \bar{y})^2, \qquad S_T = S_A + S_e.$$

Úlohu řešíme dosazením nalezených náhodných veličin do tabulky 1.

Bez ohledu na to, kterou z nabízených úvah zvolíme, nalezená statistika F i p-hodnota jsou identické.

Tabulka 1: Tabulka analýzy rozptylu

zdroj variability	součet čtverců	stupně volnosti	podíl	F	p
skupiny	S_A	$\varphi_A = k - 1$	S_A/φ_A	$\frac{S_A/\varphi_A}{S_e/\varphi_e}$	$1 - F^{-1}_{k-1,n-k,(f)}$
rezidua	S_E	$\varphi_e = n - k$	S_e/φ_e		
celkový	S_T	$\varphi_T = n - 1$			

1.1. Příklad použití metody

Za příklad užití analýzy rozptylu použijeme hodnoty získané ve studii Plháková et al. (2013). Výzkumníci se pokoušeli ověřit hypotézu o tom, že cirkadiánní preference souvisí s výskytem subklinické depresivity. Cirkadiánní preferencí se rozumí to, jestli má jedinec sklony k nejvyšší aktivitě v ranních, respektive dopoledních hodinách, nebo zdali je aktivní až odpoledne, večer či v noci. Ač je patrné, že tato charakteristika je do značné míry modifikovaná nároky prostředí, existují doklady i o silné vrozené komponentě (Horne & Östberg, 1976).

Ke zjištění cirkadiánní preference byl využit český překlad Dotazníku ranních a večerních typů Horneho a Östberga (1976), dle jehož výsledků byli respondenti rozdělení na tři skupiny – ranní typ, nevyhraněný typ a večerní typ. Příznaky deprese byly sledovány pomocí Beckovy sebeposuzovací škály depresivity pro dospělé (BDI-II, Preiss & Vacíř, 1999). Tato metoda se dotazuje na 21 projevů chování a prožívání, které mohou indikovat depresi. Respondent vždy vybírá ze čtyř možností, které jsou skórovány 0, 1, 2 a 3 body. Výsledkem dotazníku je prostý součet bodů, a může se tedy pohybovat v rozmezí od 0 po 63.

Výzkum byl realizován v roce 2012 na rozsáhlém souboru studentů Univerzity Palackého, z didaktických účelů však do našeho příkladu zahrneme jen 30 pozorování, po deseti v každé skupině (viz. tabulka 2).

V souladu s výše uvedeným posupem vypočítáme jednotlivé statistiky a dosadíme do tabulky analýzy rozptylu. Nalezené hodnoty shrnuje tabulka 3.

Ač je z výsledků patrný určitý trend – zdá se, že večerní chronotyp je doprovázen častěji projevy depresivity – rozdíly mezi skupinami nejsou statisticky významné. Nalezená p-hodnota 0.059 těsně překračuje stanovenou hladinu $\alpha = 0.05$

Tabulka 2: Ukázka naměřených hodnot

Chronotyp	Naměřené hodnoty míry depresivity	Průměr	Rozptyl
ranní	2, 4, 5, 5, 8, 9, 10, 11, 12, 12	7.80	12.84
nevyhraněný	0, 3, 5, 7, 10, 10, 11, 19, 20, 27	11.20	71.07
večerní	0, 7, 8, 9, 11, 21, 23, 27, 32, 41	17.90	168.32
celkem		12.30	96.49

a nemůžeme tedy zamítnout nulovou hypotézu.

1.2. Předpoklady analýzy rozptylu

Analýza rozptylu se řadí mezi takzvané parametrické statistické testy. To obecně znamená, že tato metoda modeluje realitu pomocí náhodných veličin, které lze popsat konečným množstvím parametrů. V tomto konkrétním případě využíváme (jednorozměrné) normální rozdělení, které lze beze zbytku popsat pomocí parametrů μ a σ^2 reprezentujících jeho střední hodnotu a rozptyl. Výhodou parametrických testů je jejich velká síla – analýza rozptylu podobně jako t-test, který je jejím speciálním případem, je při normálním rozdělení sledované náhodné veličiny nejsilnějším statistickým testem (Anděl, 2007). Příznivého chování parametrických postupů je dosaženo za cenu poměrně přísných předpokladů, které na data klademe.

V literatuře jsou tradičně uváděny tři předpoklady: normalita rozdělení dat, homoskedasticita a nekorelovanost výběrů. Před tím, než tyto tři podmínky vysvětlíme, zmiňme ještě dva předpoklady, které bývají považovány za samozřejmé.

Tabulka 3: Výsledky analýzy rozptylu

zdroj variability	součet čtverců	stupně volnosti	podíl	F	p
skupiny	528.20	2	264.10	3.141	0.059
rezidua	2270.10	27	84.08		
celkový	2798.30	29			

Závisle proměnná[2] Y musí být měřena přinejmenším na intervalové úrovni[3]. To znamená, že výsledky ztrácí smysl, pokud by hodnoty sledované proměnné označovaly jen určité nominální kategorie nebo kdyby je sice bylo možné seřadit, ale neznali bychom vzdálenosti, jaké mezi sebou jednotlivé hodnoty mají (Stevens, 1946).

Za příklad nám můžou posloužit školní známky. Pokud bychom považovali známku za míru osvojení znalostí studentem, pak bychom měli tuto veličinu považovat nejspíše za pořadovou (ordinální). Těžko bychom totiž mohli tvrdit, že mezi stupněm 1 a 2 je stejně velký rozdíl jako mezi 3 a 4. Charakteristiky, jako je průměr či rozptyl, nejsou pak zcela smysluplné. Přísně vzato, pokud počítáme průměrné známky, jak je běžné, měli bychom připustit, že se naše závěry týkají zase jen známek, nikoli nějaké vlastnosti studentů, kterou tato čísla zastupují.

Druhou podmínkou je nezávislost jednotlivých měření. V literatuře nebývá zmiňována, jelikož je již implicitně obsažena v tvrzení, že pracujeme s náhodnými výběry, které se z definice skládají z vzájemně nezávislých realizací náhodné veličiny. V praxi ale platnost této podmínky nemusí být automaticky zajištěna.

Můžeme si položit otázku, jestli data, se kterými pracujeme v příkladu z kapitoly 1.1, tyto dvě přehlížené podmínky splňují. Co se týče nezávislosti, tak pokud účastníci výzkumu odpovídali samostatně (neopisovali) a pokud byli do souboru zařazeni dle vhodného klíče (ideálně náhodným výběrem), nemusíme mít obavy z vážného porušení této podmínky. Diskutabilnější je metrická povaha výsledků psychologických dotazníků. Již Likert (1932) nicméně poukazuje na to, že ač hodnoty získané z izolovaných položek tuto vlastnost mít nemusí, tak po jejich sečtení se vzdálenosti mezi jednotlivými body sobě začnou podobat a získanou proměnnou můžeme považovat za metrickou bez vážné ztráty přesnosti.

[2]V následujících odstavcích budeme uvažovat o Y jako o náhodném vektoru. Abychom zdůraznili, že jeho jednotlivé prvky chápeme jako náhodné veličiny, budeme je značit velkými písmeny (Y_{ji}) stejně jako z nich odvozené statistiky (např. $\bar{Y}$) místo původního značení (y_{ji}, $\bar{y}$).

[3]Intervalová a poměrová úroveň měření bývá často souhrnně označována jako metrická či kvantitativní úroveň. Můžeme tedy říci, že požadujeme, aby sledovaná proměnná byla metrická.

1.2.1. Nezávislost výběrů

Náhodné výběry Y a Z považujeme za nezávislé právě tehdy, když platí nezávislost v každé dvojici prvků jejich prvků. V praxi to znamená, že výsledek měření v rámci jedné skupiny není nijak ovlivněn hodnotami, které jsme získali v kterékoli jiné skupině.

S porušením této podmínky se v praxi nesetkáváme příliš často. Důvodem je to, že na rozdíl od zbývajících dvou podmínek zvolený výzkumný design jednoznačně určuje, jestli mezi náhodnými výběry může existovat závislost nebo ne. Její případné porušení je tedy metodologickým selháním a již před začátkem sběru dat dokážeme posoudit, jestli jsme provedli vhodná opatření, abychom závislosti výběrů zabránili.

Zůstaneme-li v oblasti psychologického výzkumu, představme si experimentální design, kdy n pokusných osob postupně vystavujeme k různým podmínkám a pokaždé měříme jejich výkon při určité činnosti. Ověřujeme pak hypotézu o středních hodnotách těchto k výběrů o stejném rozsahu. Pokud při volbě statistického testu sáhneme po jednofaktorové analýze rozptylu, dopustíme se porušení podmínky nezávislosti výběrů. Výsledky, kterých jedinec dosáhne za určitých podmínek, pravděpodobně souvisí s jeho výsledky za jiných podmínek, jelikož můžeme očekávat, že část variability výsledků je způsobena nějakými charakteristikami jedince, které jsou alespoň po dobu experimentu stabilní. Analýza rozptylu zde selhává v tom smyslu, že nevyužije celou informaci, kterou máme k dispozici, a je tedy zbytečně konzervativní. Řešením by mohlo být například zařazení druhého faktoru o n úrovních (jednu pro každou osobu) nebo využití Hotellingova testu, který může sloužit jako zobecnění párového t-testu.

Podmínka nezávislosti výběrů není při realizaci výzkumného designu obvykle z výše uvedených důvodů příliš palčivá a nebudeme jí v této práci věnovat další prostor.

1.2.2. Normalita rozdělení

Zřejmě nejčastěji ze všech podmínek analýzy rozptylu je v učebních textech připomínaný požadavek normálního rozdělení sledované náhodné veličiny. Tato podmínka vyplývá ze vztahů (8) a (9). Aby platilo, že výsledná statistika F (viz. roznice 10) má Fisherovo-Snedocorovo rozdělní, požadujeme, aby každý z odhadů rozptylu $\hat{\sigma}_e^2$ i $\hat{\sigma}_A^2$ vynásobený konstantou měl rozdělení χ^2.

Za platnosti nulové hypotézy tedy můžeme toto tvrzení přepsat jako

$$\frac{\varphi}{\sigma^2} \cdot \hat{\sigma}^2 \sim \chi^2_{\varphi}, \tag{15}$$

kde φ označuje počet stupňů volnosti. Prozkoumejme, jaké podmínky musí náhodná veličina Y splňovat, aby vztah (15) platil pro $\hat{\sigma}_e^2$ i $\hat{\sigma}_A^2$.

Zaměříme-li se na $\hat{\sigma}_A^2$, můžeme postupovat tak, že do vztahu (15) dosadíme vzorec pro výpočet tohoto odhadu rozptylu (7). Zanedbáme-li konstanty, které zajišťují zachování měřítka, získáme náhodou veličinu $\sum_{j=1}^{k}(\bar{Y}_j - \bar{Y})^2$. Obecně má náhodná veličina Z rozdělení χ^2 o n stupních volnosti tehdy, když je určena vztahem $Z = \sum_{i=1}^{n} X_i^2$, kde $X_i \sim N(0,1)$, vzájemně nezávislé, pro $i = 1, 2, ..., n$ (Anděl, 2007). Požadujeme tedy, aby každá z veličin $\bar{Y}_j - \bar{Y}$ měla normální rozdělení se středem v nule, respektive aby každá z veličin $\bar{Y}_j$ měla normální rozdělní se střední hodnotou μ_j. Toho dosáhneme tak, že vzneseme požadavek normality na jednotlivé veličiny Y_{ji}, jelikož libovolná lineární kombinace normálně rozdělených náhodných veličin má opět normální rozdělení (Anděl, 2007).

Z této úvahy je patrná jedna zajímavá skutečnost - při vysokých rozsazích výběru n není nezbytné, aby veličiny Y_{ji} měly normální rozdělení, a přesto se může rozdělní veličin $\bar{Y}_j$ normalitě blížit. Toto lze vyvodit z centrálních limitních vět. Například Lindebergova věta (viz přílohy publikace Anděl, 2007) říká, že pro posloupnost nezávislých náhodných veličin X_1, X_2,..., X_n se středními hodnotami μ a konečnými rozptyly σ^2 platí při n jdoucím do nekonečna

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2),$$

kde šipka s písmenem d označuje konvergenci v distribuci. Pro aritmetický průměr $\bar{X}$ těchto veličin X_i pak tedy lze při n jdoucím do nekonečna vyvodit vztah

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

S rozstoucími rozsahy výběru se tedy budou distribuce průměrů $\bar{Y}_j$ asymptoticky blížit k $N(\mu_j, \sigma^2/n_j)$.

Analogickou úvahu můžeme provést i pro odhad $\hat{\sigma}_e^2$. Do vzorce (15) dosadíme rovnici (5). Opět zanedbáme konstantu a získáme součet k náhodných veličin[4] ve tvaru $\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y}_j)^2$, z nichž u každé požadujeme rozdělení χ^2. Opět můžeme náš požadavek přenést na jednotlivé prvky sum a předpokládat, že veličiny $Y_{ji} - \bar{Y}_j$ mají normální rozdělení se středem v nule, respektive že veličiny Y_{ji} mají obecně normální rozdělení, ač nemusí mít napříč skupinami stejnou střední hodnotu. Případný vliv centrální limitní věty není v tomto případě zřejmý.

Přestože jsou teoretická východiska pro podmínku normality poměrně zřejmá, můžeme narazit na určité nejasnosti, jak tento předpoklad ověřovat. V praxi se často setkáme s názorem, že bychom měli testovat, jestli posloupnost hodnot y_{ji} pochází z normálního rozdělení. Z výše uvedeného ale vyplývá, že tento postup není správný. Hodnoty y_{ji} představují realizace náhodné veličiny s rozdělením $N(\mu, \sigma^2)$ pouze za platnosti nulové hypotézy. Pokud nulová hypotéza neplatí, budou naměřené hodnoty různě posunuté v závislosti na tom, ze kterého náhodného výběru pochází. Jsme-li tedy svědky toho, že histogram naměřených hodnot hustotu tohoto rozdělení nepřipomíná, nemusí to nutně znamenat porušení podmínky normality, ale neplatnost nulové hypotézy. Ve skutečnosti požadujeme, aby z normálního rozdělení pocházela rezidua $(y_{ji} - \bar{y}_j)$, k čemuž jsme došli i při zkoumání vlastností odhadu $\hat{\sigma}_e^2$.

Porušení podmínky normálního rozdělení reziduí můžeme testovat několika testy. Mezi nejčastěji používané patří Kolmogorovův-Smirnovův test (respektive jeho adaptace známá jako Lillieforsův test), test dobré shody nebo Shapirův-

[4]Připomeňme aditivitu rozdělení χ^2 – sečteme-li více nezávislých veličin s tímto rozdělením, bude mít výsledná náhodná veličina opět rozdělní χ^2 s počtem stupňů volnosti odpovídajícím součtu stupňů volnosti všech sčítanců.

Wilkův test, který je ze zmiňovaných postupů nejsilnější (Razali & Wah, 2011). Při vysokých rozsazích výběru dochází k zamítnutí nulové hypotézy i při velmi malé odchylce od normality, která přesnost výsledků neohrožuje, můžeme proto dát přednost kontrole histogramu reziduí tak říkajíc ,,od oka".

1.2.3. Homoskedasticita

Pojmem homoskedasticita se označuje shodnost rozptylů náhodných veličin. V případě analýzy rozptylu požadujeme, aby náhodné výběry, z nichž pochází jednotlivá měření, měly rozdělení $N(\mu_j, \sigma_j^2)$, kde $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. Tato podmínka opět vychází z definice rozdělní χ^2 jako součtu druhých mocnin náhodných veličin s normovaným normálním rozdělením. Porušení tohoto předpokladu by se opět odrazilo na vlastnostech odhadu chybového rozptylu $\hat{\sigma}_e^2$ i rozptylu mezi skupinami $\hat{\sigma}_A^2$, což si lze snadno představit, když se vrátíme ke vztahům (5) a (7).

Méně snadno si již představíme, jakým způsobem se heteroskdasticita promítá do vlastností testové statistiky F. Mohli bychom například spekulovat nad tím, že rozmanité hodnoty σ_j^2 povedou k jejímu vychýlení. Má-li náhodná veličina X rozdělení $N(0, \sigma^2)$, pak lze střední hodnotu její druhé mocniny popsat jako

$$\mathbb{E}[X^2] = \sigma^2.$$

Chování statistiky F v případě porušení podmínky homoskedasticity by tedy mohlo být odhadnutelné na základě toho, jak se mění střední hodnoty odhadů rozptylů $\hat{\sigma}_e^2$ a $\hat{\sigma}_A^2$ v závislosti na parametrech σ_1^2 až σ_k^2. Pokud by například odhad $\hat{\sigma}_e^2$ měl za určitých podmínek vyšší střední hodnotu než odhad $\hat{\sigma}_A^2$, pak bychom mohli očekávat, že za těchto okolností je analýza rozptylu příliš konzervativní.

Zkusme tyto podmínky popsat. Odvoďme, jakou střední hodnotu budou mít statistiky $\hat{\sigma}_e^2$ a $\hat{\sigma}_A^2$, pokud máme k dispozici k náhodných výběrů o rozsazích $n_1 = ... = n_k$. Pro výpočet nejprve uveďme rozptyly a střední hodnoty veličin $\bar{Y}_j$ a $\bar{Y}$

$$\bar{Y}_j \sim N\left(\mu_j, \frac{\sigma_j^2}{n_j}\right), \qquad \bar{Y} \sim N\left(\frac{1}{k}\sum_{j=1}^{k}\mu_j, \frac{1}{k^2}\sum_{j=1}^{k}\frac{\sigma_j^2}{n_j}\right),$$

a jejich kovarivnaci, víme-li, že veličiny $\bar{Y}_j$ jsou pro $j = 1, 2, ..., k$ nezávislé:

$$cov(\bar{Y}, \bar{Y}_j) = cov\left(\frac{1}{k}\sum_{l=1}^{k}\bar{Y}_l, \bar{Y}_j\right) = \frac{1}{k}\cdot\frac{\sigma_j^2}{n_j}.$$

Střední hodnotu odhadu $\hat{\sigma}_e^2$ můžeme vyjádřit jako

$$\mathbb{E}(\hat{\sigma}_e^2) = \frac{1}{n-k}\sum_{j=1}^{k}\mathbb{E}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2 = \frac{1}{n-k}\sum_{j=1}^{k}\sigma_j^2(n_j - 1).$$

Za podmínky stejných rozsahů souborů využijeme vztahu $n_1 k = n$ a nalezené řešení zjednodušíme

$$\mathbb{E}(\hat{\sigma}_e^2) = \frac{n_1 - 1}{n-k}\cdot\frac{k}{k}\sum_{j=1}^{k}\sigma_j^2 = \frac{1}{k}\sum_{j=1}^{k}\sigma_j^2.$$

Střední hodnotu odhadu $\hat{\sigma}_A^2$ vyjádříme následovně. Tentokrát budeme mimojiné předpokládat platnost nulové hypotézy:

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}_A^2) =& \frac{1}{k-1}\sum_{j=1}^{k} n_j\,\mathbb{E}\left[(\bar{Y}_j - \mu) - (\bar{Y} - \mu)\right]^2 = \\
& \frac{1}{k-1}\sum_{j=1}^{k} n_j\,\mathbb{E}\left[(\bar{Y}_j - \mu)^2 - 2(\bar{Y}_j - \mu)(\bar{Y} - \mu) + (\bar{Y} - \mu)^2\right] = \\
& \frac{1}{k-1}\sum_{j=1}^{k} n_j\,\mathbb{E}\left[\frac{\sigma_j^2}{n_j} - 2\frac{1}{k}\frac{\sigma_j^2}{n_j} + \frac{1}{k^2}\sum_{l=1}^{k}\frac{\sigma_l^2}{n_l}\right] = \\
& \frac{1}{k-1}\left[\sum_{j=1}^{k}\sigma_j^2 - \frac{2}{k}\sum_{j=1}^{k}\sigma_j^2 + \frac{n}{k^2}\frac{1}{n_1}\sum_{j=1}^{k}\sigma_j^2\right] = \\
& \frac{1}{k-1}\left(1 - \frac{2}{k} + \frac{1}{k}\right)\sum_{j=1}^{k}\sigma_j^2 = \frac{1}{k}\sum_{j=1}^{k}\sigma_j^2.
\end{aligned}$$

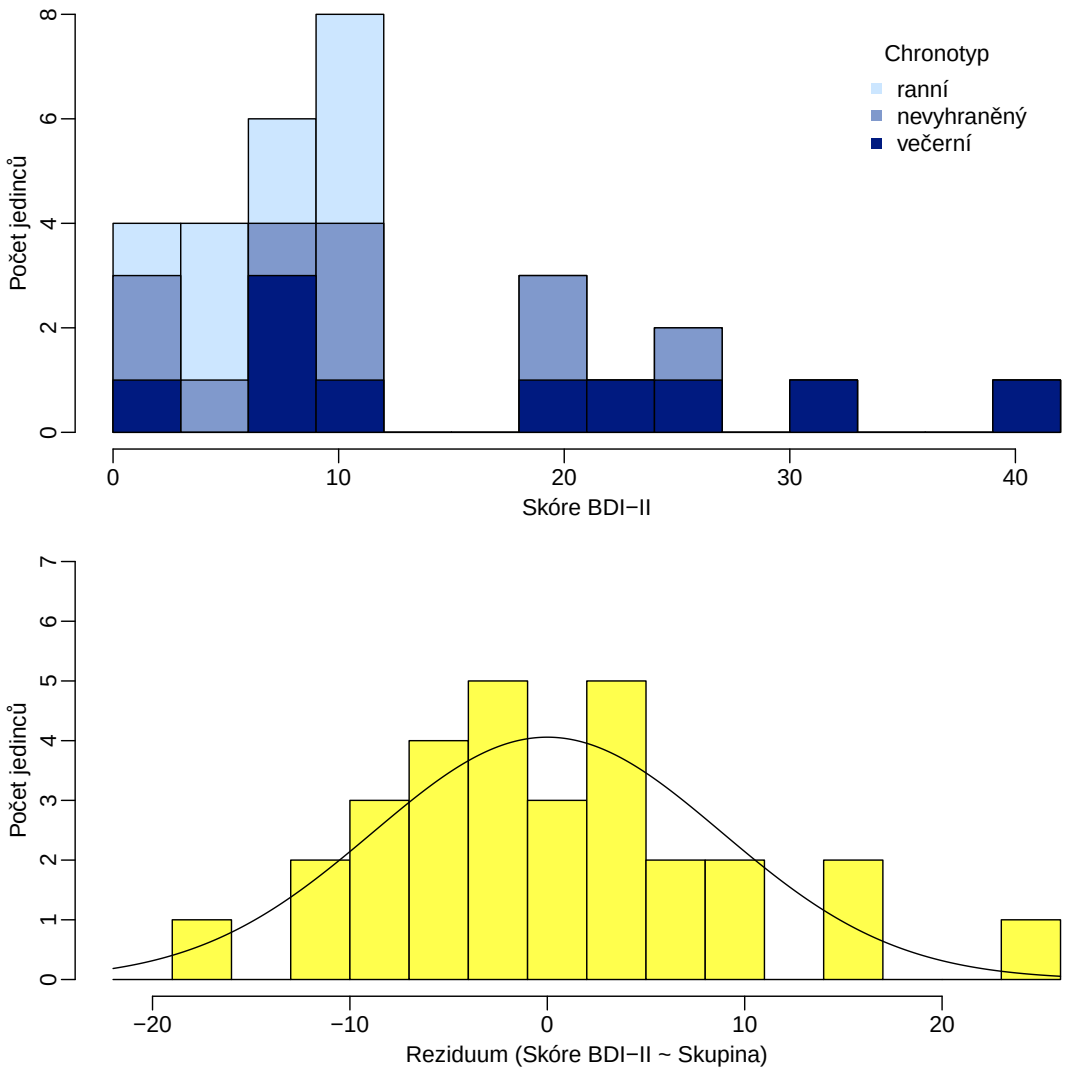
Naše úvaha vede k poměrně překvapivému zjištění: bez ohledu na to, jakým způsobem porušujeme podmínku shody rozptylů, jsou si střední hodnoty odhadů rozptylů $\hat{\sigma}_e^2$ i $\hat{\sigma}_A^2$ rovny. To platí přinejmenším v případě, kdy jsou náhodné

výběry stejného rozsahu. Ač jsme selhání testu plynoucí z různých středních hodnot odhadů rozptylu neodhalili, nedokazuje to, že by podmínka homoskedasticity byla zbytečná. Jednak nevíme, jestli zůstává zachována nezávislost obou odhadů rozptylu, a pak samotná střední hodnota neposkytuje dostatek informací o celém tvaru rozdělení testové statistiky F. Je možné, že při porušení podmínky homoskedasticity je její distribuce nějak deformovaná, ač si třeba zachovává nezměněnou střední hodnotu.

Porušení podmínky homoskedasticity lze testovat například Bartlettovým či Leveneovým testem. Ač je druhý jmenovaný pouze přibližnou metodou, z důvodu obstojné robustnosti bývá považován za metodu první volby. Leveneův test vychází z teoretického rámce analýzy rotpylu. Při jeho výpočtu v prvním kroku transformujeme naměřené hodnoty y_{ji} na hodnoty z_{ji} dle vztahu $z_{ji} = |y_{ji} - \bar{y}_j|$. V druhém kroku pak pomocí jednofaktorové analýzy rozptylu ověříme nulovou hypotézu o shodných středních hodnotách náhodných veličin $\bar{z}_j$.

Vraťme se k ukázkovému příkladu z kapitoly 1.1. Obrázek 1 znázorňuje distribuci měřené veličiny a jejích reziduí po analýze rozptylu. Je patrné, že data jsou značně zešikmená ke kladným hodnotám a normální rozdělení připomínají jen vzdáleně (koeficient šikmosti je roven hodnotě 1.20 místo očekávané nuly). Rozdělení reziduí se již nicméně k normalitě blíží (šikmost = 0.49, špičatost = 3.28). Shapirův-Wilkův test provedený na reziduích nulovou hypotézu nezamítá ($W = 0.98$, $p = 0.82$) a náš předpoklad normality nezpochybňuje. Problematičtější výsledky přináší srovnání rozptylů skupin. Porovnáme-li krajní skupiny, zjistíme, že třetí skupina má více než třináctkrát větší rozptyl než ta první. Leveneův test zamítá nulovou hypotézu o shodě rozptylů na hladině významnosti $\alpha = 0.001$ ($F = 15.14$). Data tedy nesplňují předpoklad homoskedasticity a můžeme mít pochybnosti, jestli naše rozhodnutí přijmout nulovou hypotézu, učiněné na základě výsledků analýzy rozptylu, bylo správné.

Obrázek 1: Histogram naměřených hodnot z ukázkové úlohy a jejich reziduí



2. Design simulační studie

Vlastnosti analýzy rozptylu při jednoduchém třídění budeme ověřovat pomocí simulační studie. Obecně tedy budeme postupovat tak, že na základě určitých parametrů vygenerujeme k posloupností náhodných čísel. Každá z těchto posloupností představuje množinu n_j realizací (značíme y_{ji}) náhodné veličiny Y_j. Pomocí analýzy rozptylu otestujeme, zdali tyto náhodné veličiny mají shodné střední hodnoty. Proceduru budeme opakovat a zaznamenáme, v kolika procentech případů dochází k zamítnutí nulové hypotézy na pětiprocentní hladině významnosti při daných hodnotách parametrů. Získaná hodnota je odhadem parametru $\pi_{\boldsymbol{\phi}}$, který definujeme jako

$$\pi_{\boldsymbol{\phi}} = \mathbb{E}[\Pi(\boldsymbol{\phi})], \tag{16}$$

kde $\mathbb{E}$ je operátor střední hodnoty a Π je náhodná veličina závislá na vektoru parametrů $\boldsymbol{\phi}$. Veličina Π je kritickou funkcí a má alternativní rozdělení. Hodnota 1 reprezentuje zamítnutí nulové hypotézy a 0 její nezamítnutí.

2.1. Definice parametrů $\boldsymbol{\phi}$

Abychom mohli výsledky jednotlivých simulací srovnávat a snadno popsat, je žádoucí, aby počet parametrů, které specifikují zadání libovolné námi zvolené simulace, byl rozumně nízký. Je zjevné, že pro dosažení tohoto cíle musíme pestrou škálu všech možností, se kterými se můžeme setkat, poněkud zjednodušit. Problémy plynou zejména z proměnlivého počtu náhodných výběrů, jež zkoumáme – pro přesný popis bychom totiž museli každý z k výběrů popsat jeho rozsahem, střední hodnotou jeho distribuční funkce, jejím rozptylem a případně dalšími charakteristikami. Konečný počet (skalárních) parametrů by tedy byl různě vysoký v závislosti na počtu srovnávaných výběrů. Zvolili jsme proto následující řešení, které citelně neredukuje komplexnost problému, a zároveň se opírá o nízký a neproměnlivý počet parametrů. Simulace jsme specifikovali pomocí následujících parametrů.

2.1.1. Počet a rozsahy výběrů (parametry n, k, u)

V předchozím textu jsme značili písmenem n celkový rozsah souboru a písmenem k počet náhodných výběrů. Obě tyto hodnoty budeme využívat jako parametry při generování dat pro simulaci. Vztah mezi n a k je zřejmý:

$$n = \sum_{j=1}^{k} n_j.$$

Parametry n a k nicméně samy jednoznačně neurčují rozsahy jednotlivých výběrů n_j. Pro jejich přesné vymezení použijeme třetí parametr u, který kvantifikuje nevyváženost rozsahů výběrů.

Aby byl parametr u intuitivně pochopitelný, definovali jsme jej na základě těchto požadavků:

$$u = \frac{n_k}{n_1}, \qquad n_j = n_{j-1} + \frac{u}{k-1}, \quad j = 2, ..., k. \tag{17}$$

Lze jej tedy interpretovat jako poměr rozsahů největší a nejmenší skupiny a předpokládáme, že rozsahy zbývajících skupin jsou rovnoměrně rozprostřeny mezi těmito extrémy.

Z rovnic (17) odvodíme vztah pro jednotlivé hodnoty n_j jako

$$n_j = 2n\frac{(u-1)j + k - u}{k(k-1)(u+1)}, \qquad j = 1, ..., k. \tag{18}$$

Tedy například pro $n = 120$, $k = 5$ a $u = 2$ dostaneme rozsahy $(n_1, n_2, n_3, n_4, n_5) = (16, 20, 24, 28, 32)$.

Parametr u může teoreticky nabývat libovolné hodnoty z $\mathbb{R}^+$. V praxi zejména pro nízké hodnoty n a vysoké k, můžou vysoké hodnoty u (nebo naopak ty blízké nule) generovat rozsahy menší než 2, což znemožňuje výpočet analýzy rozptylu. Nalezené hodnoty n_j navíc nejsou obecně celá čísla a je potřeba je zaokrouhlit. Kvůli zaokrouhlovací chybě se tak může stát, že součet rozsahů všech skupin nemusí přesně odpovídat původnímu n. V případě, kdy je u rovno jedné, jsou rozsahy všech výběrů shodné. Platí také, že pro u a $1/u$, získáme stejný vektor rozsahů výběrů s obráceným pořadím prvků.

2.1.2. Posunutí středních hodnot (d)

Parametr d vyjadřuje rozmanitost středních hodnot distribučních funkcí Y_j, z nichž byly generovány náhodné výběry[5]. Přímo tedy určuje, do jaké míry je porušena platnost nulové hypotézy o shodě středných hodnot. Při jeho stanovování jsme postupovali analogicky jako v případě rozmanitosti rozsahů souborů. Vztah mezi d a středními hodnotami μ_j jsme omezili těmito podmínkami:

$$\sum_{j=1}^{k} \mu_j = 0, \quad j = 1, ..., k, \qquad \mu_j = \mu_{j-1} + \frac{d}{k-1}, \quad j = 2, ..., k. \tag{19}$$

Za předpokladu jednotkových rozptylů můžeme vyjádřit $d = \mu_k - \mu_1$, a parametr d lze tedy interpretovat jako maximální rozdíl středních hodnot distribučních funkcí, z nichž jsou generovány výběry. Pokud rozptyl není roven jedné, bylo by vhodné parametr standardizovat vydělením výrazu směrodatnou odchylkou, a získat tak předpis $d = (\mu_k - \mu_1)/\sigma$. Vyjádřením μ_j z rovnic (19) získáme vztah mezi d a μ_j ve tvaru

$$\mu_j = \left(\frac{j-1}{k-1} - \frac{1}{2} \right) d. \tag{20}$$

Pokud je $d = 0$, pak jsou všechny střední hodnoty rovny nule. A například pro případ $k = 5$ a $d = 4$ jsou nalezené hodnoty $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-2, -1, 0, 1, 2)$. Hodnoty parameru d a $-d$ generují shodné střední hodnoty avšak s opačným pořadím.

Za zmínku stojí vztah mezi parametrem d a klasickými ukazateli míry účinku pro analýzu rozptylu. Cohen (1988) definuje ukazatel míry účinku pro analýzu rozptylu jako

$$f = \sqrt{\frac{\sum_{j=1}^{k} \frac{n_j}{n} (\mu_j - \mu)^2}{\sigma^2}}. \tag{21}$$

Vztah mezi f a d pak můžeme vyjádřit z rovnic (20) a (21)[6]. Budeme před-

[5] Ač u všech parametrů používáme pro indexaci skupin stejné označení j, tak hodnoty indexu pro každý parametr permutujeme.

[6] Preferuje-li čtenář obecný ukazatel míry účinku pro lineární model R^2, k převodu může použít vztah $f = \sqrt{R^2/(1 - R^2)}$

pokládat, že $\mu = 0$, $\sigma = 1$ a že rozsahy všech skupin jsou shodné.

$$f = \sqrt{\frac{\sum_{j=1}^{k} \frac{n_j}{n}(\mu_j - \mu)^2}{\sigma^2}} = \sqrt{\frac{1}{k}\sum_{j=1}^{k} \mu_j^2} = d \cdot \sqrt{\frac{k+1}{12(k-1)}} \tag{22}$$

Pozoruhodné je to, že námi odvozený parametr d můžeme vedle standardních ukazatelů míry účinku najít v Cohenově práci z roku 1988 (str. 276), ba co víc, je zde diskutována i situace, kdy jsou průměry výběrů rovnoměrně rozmístěny mezi extrémy, což zcela přesně odpovídá námi popsanému uspořádání. Cohen (1988) tento ukazatel označuje jako *standardizovaný rozsah populačních průměrů* a používá pro něj stejně jako autor této práce písmeno d.

Výhodou tohoto ukazatele je především jeho intuitivní povaha a snadná pochopitelnost i pro jedince bez hlubšího matematického vzdělání. Jeho hodnotu při pohledu na naměřená data můžeme odhadnout jako

$$d \approx \frac{\max\limits_{j=1,\ldots,k} \hat{\mu}_j - \min\limits_{j=1,\ldots,k} \hat{\mu}_j}{\hat{\sigma}_e}. \tag{23}$$

2.1.3. Parametr heteroskedasticity (v)

Abychom mohli modelovat situace, kdy je porušena podmínka homoskedascity, potřebujeme zařadit parametry, které vyjadřují rozmanitost rozptylů σ^2. Situaci opět zjednodušíme tím, že velikost rozptylů v k skupinách vyjádříme jediným parametrem v. Podobně jako v předešlých situacích stanovíme vztah mezi v a σ_j^2 na základě určitých požadavků, které na tento ukazatel klademe.

Opět budeme požadovat snadnou interpretovatelnost parametru. Parametr v proto definujeme jako poměr nejmenšího a největšího rozptylu:

$$v = \frac{\sigma_k^2}{\sigma_1^2}. \tag{24}$$

Dále požadujeme, aby změny parametru v neovlivňovaly velikost chybového rozptylu σ_e^2. Pro stejné rozsahy výběrů to zajistíme podmínkou

$$\sum_{j=1}^{k} \sigma_j^2 = k. \tag{25}$$

Nakonec požadujeme, aby v limitním případě, kdy se hodnota v blíží k nekonečnu, byl veškerý rozptyl soustředěn v jediném výběru a ostatní výběry aby měly nulovou varianci. To lze zajistit tak, že jednotlivé hodnoty σ_j^2 budeme chápat jako členy geometrické posloupnosti posunuté o konstantu, která umožní platnost rovnice (24). Tedy

$$\sigma_j^2 = C_1 v^{j-1} + C_2. \tag{26}$$

Využijeme-li znalost o součtu konečné geometrické posloupnosti, tak po dosazení do vztahu (25) získáme rovnici

$$C_1 \frac{1 - v^k}{1 - v} + kC_2 = k, \tag{27}$$

a dosazením do vztahu (24) rovnici

$$\frac{C_1 v^{k-1} + C_2}{C_1 + C_2} = v. \tag{28}$$

Ze soustavy rovnic (27) a (28) vyjádříme konstanty C_1 a C_2 jako

$$C_1 = \frac{k(v-1)}{v^{k-1}k - vk + v^k - 1}, \tag{29}$$

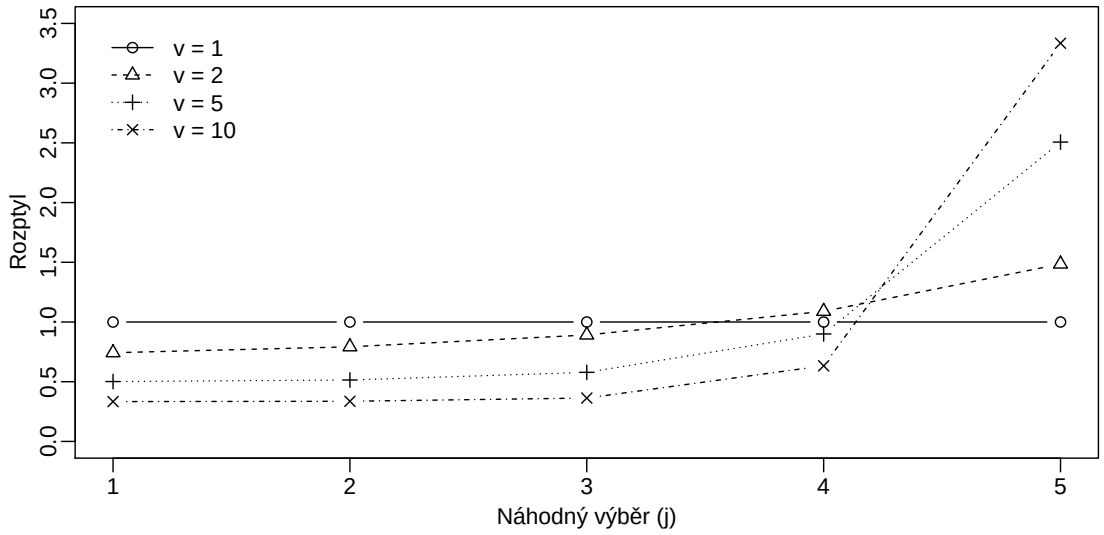
$$C_2 = \frac{v^{k-1} - v}{v^{k-1}k - vk + v^k - 1}. \tag{30}$$

Po dosazení zpět do rovnice (26) a úpravě získáme konečně vztah mezi v a σ_j^2 ve tvaru

$$\sigma_j^2 = \frac{k[(v-1)v^j + v^k - v^2]}{v^k(k+v) - v(kv+1)}. \tag{31}$$

Nalezená rovnice nicméně nemá řešení, pokud $v = 1$, jelikož výsledkem je $0/0$. Pro tyto případy dodefinujeme řešení v souladu s našimi podmínkami jako $\sigma_j^2 = 1$, pro $j = 1, ..., k$. Podobně jako v případě parametru u generuje hodnota v i $1/v$ stejné rozptyly, jen v opačném pořadí. V obrázku 2 jsou znázorněny nalezené rozptyly pro $k = 5$ a v rovno hodnotám 1, 2, 5 a 10.

Obrázek 2: Rozptyly výběrů pro různé hodnoty parametru v



2.1.4. Parametr zešikmení (b)

Poslední parametr b udává míru porušení předpokladu normálního rozdělení náhodné veličiny Y. Způsobů, jakým se může náhodná veličina odchylovat od normálního rozdělení, existuje nekonečné množství. Pro účely této práce jsme zvolili odchýlení prostřednictvím zešikmení. Hledali jsme takový parametr zešikmení, který splňuje následující tři vlastnosti:

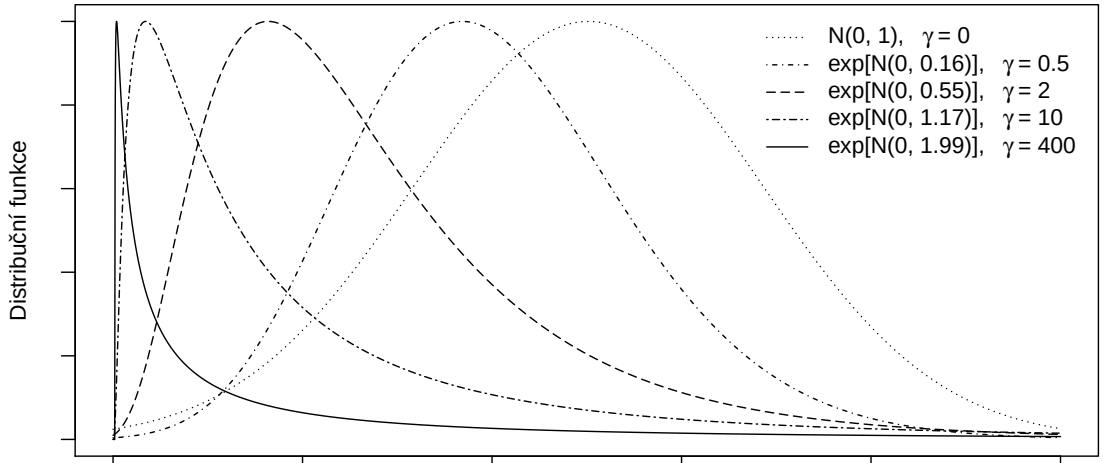
(i) Hodnota parametru b odpovídá třetímu centrálnímu momentu – tedy koeficientu šikmosti – náhodných veličin Y_j .

(ii) Je-li hodnota parametru zešikmení rovna nule, mají veličiny Y_j normální rozdělení.

(iii) Střední hodnota ani rozptyl náhodných veličin Y_j není závislý na hodnotě parametru b.

Jako poměrně schůdná cesta se zdá být generování hodnot z normálního rozdělení a jejich následné transformování do požadované podoby. Můžeme například vynásobit všechny kladné hodnoty nějakou konstantou odvozenou od pa-

Obrázek 3: Tvary hustot log-normálního rozdělení



Pozn.: Osy grafu jsou bezrozměrné a byly pro každou funkci nastaveny zvlášť, tak aby bylo možné srovnat jejich tvar. Plochy pod křivkami se tedy různí.

rametru zešikmení a následně všechna měření vynásobit jinou, normovací, konstantou, která datům vrátí původní rozptyl. Ukázalo se však, že tato metoda velmi rychle vede k limitnímu rozdělení, kdy se záporné hodnoty transformují na čísla v těsné blízkosti nuly a kladné hodnoty přestanou dále růst bez ohledu na parametr šikmosti, aby zachovávaly původní rozptyl.

Vyhovující řešení bylo nakonec nalezeno pomocí log-normálního rozdělení. Připomeňme, že náhodná veličina Z má log-normální rozdělení, pokud pro ni platí vztah $Z = e^X$, kde $X \sim N(\mu, \sigma^2)$. Využijeme toho, že známe střední hodnotu, rozptyl i šikmost (označena γ) takto definovaného rozdělení (Johnson et al., 1995):

$$\mathbb{E}(Z) = e^{\mu+\sigma^2/2}, \qquad var(Z) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2},$$
$$\gamma = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}.$$

Položíme-li pak hodnotu parametru $\mu = 0$, můžeme pomocí parametru σ^2 měnit zešikmení veličiny. Vysoké hodnoty parametru σ^2 povedou k distribuci nápadně zešikmené ke kladným hodnotám. Naopak, bude-li se σ^2 blížit nule, bude distribuce připomínat normální rozdělení, což ilustruje obrázek 3.

Abychom zajistili nezávislost střední hodnoty, rozptylu a šikmosti, využijeme toho, že střední hodnota i rozptyl jsou ovlivněny lineárními transformacemi, zatímco koeficient šikmosti zůstává (až na případné znaménko) invariantní. To nám umožní vytvořit náhodnou veličinu

$$Z = \frac{e^X - e^{\sigma^2/2}}{\sqrt{(e^{\sigma^2} - 1)e^{\sigma^2}}}, \tag{32}$$

kde (opomene-li řešení, která obsahují komplexně sdružená čísla) lze σ^2 jednoznačně vyjádřit pomocí b jako

$$\sigma^2 = log\left(\frac{2^{2/3}}{(b + \sqrt{b^2 + 4})^{2/3}} + \frac{(b + \sqrt{b^2 + 4})^{2/3}}{2^{2/3}} - 1\right), \tag{33}$$

a pro kterou při $X \sim N(0, \sigma^2)$ platí, $E(Z) = 0$, $var(Z) = 1$ a b je jejím koeficientem šikmosti.

Při $b = 0$ vzniká degenerovaná náhodná veličina s nulovým rozptylem. Pro tuto situaci proto dodefinujeme $Z \sim N(0, 1)$ ve shodě s předchozím vývojem křivky. Parametr b může nabývat libovolné hodnoty z $\mathbb{R}_0^+$. Pomocí lineární transformace lze následně převést veličinu Z na Y v souladu s dříve definovanými parametry d a v.

2.1.5. Vynechané parametry

Chování analýzy rozptylu v případě porušení podmínky nezávislosti náhodných výběrů v této práci zkoumat nebudeme. Jak jsme zmínili v kapitole 1.2.1, závislost výběrů obvykle nepředstavuje pro výzkumníka palčivý problém, jelikož jí lze často předejít, případně můžeme použít nástroje, které z ní naopak můžou těžit.

Za zmínku stojí také technický problém, kterému bychom museli čelit, pokud bychom měli ambice zkoumat chování testu při porušení této podmínky: jaké vlastnosti by tento parametr měl mít. Mohl by například nějak určovat korelovanost výběrů, nicméně otázkou je, které výběry by mezi sebou byly korelované a které ne, jaký by byl směr těchto korelací a v neposlední řadě, jak nastavit tento

parametr, aby se jeho interpretace neměnila ani v případě výběrů nestejných rozsahů. Řešení tohoto problému by nás zřejmě stálo hned několik dalších parametrů, které bychom museli do procedury zařadit. A také velkou část její elegance.

2.2. Postup ověření hladiny testu

Jedním z projevů selhání statistického testu může být to, že nerespektuje zadanou hladinu α. V praxi se to projeví tak, že při opakovaném provádění testu na datových souborech, které byly náhodně generované za platnosti nulové hypotézy, bude docházet k chybě prvního druhu (tedy zamítnutí H_0) ve více než α procentech případů. V našem případě platnost H_0 zajistíme tak, že parametr d nastavíme na hodnotu 0. Za hladinu významnosti α zvolíme hodnotu 0.05, jelikož tato hladina je ve výzkumných studiích používána nejčastěji.

Překračuje-li test v důsledku porušení svých předpokladů hladinu α pouze nepatrně, nemusí to mít v praxi žádný vliv na validitu našich závěrů, zejména pak ve společenských vědách, kde je jakékoli měření zatíženo velkou chybou (často dosahující až desítek procent rozptylu sledované proměnné) a badatel se i při dodržení všech předpokladů statistického testu musí potýkat se značnou mírou nejistoty. Stanovme proto určitou mez, kterou budeme považovat za nejvyšší nepřesnost, kterou můžeme u testu akceptovat a jeho výsledky ještě považovat za směrodatné. Tuto hladinu stanovíme poměrně benevolentně na 0.10. Je-li relativní četnost zamítnutých hypotéz π_ϕ větší než 0.10, pak budeme hovořit o selhání statistického testu.

2.2.1. Stanovení počtu cyklů

Přesnost, s jakou odhadujeme pravděpodobnost zamítnutí nulové hypotézy (π_ϕ), je závislá na tom, kolikrát simulaci pro dané hodnoty parametru ϕ opakujeme. Dostačující přesnost ($\sqrt{var(\hat{\pi}_\phi)}$) pro naše účely je 0.001. Minimální počet opakování simulace (budeme jej značit C), který garantuje tuto přesnost, odvodíme z vlastností alternativního rozdělení a vlastností rozptylu náhodné veličiny obecně (dle Anděl, 2007).

Pro součet k vzájemně nezávislých náhodných veličin X_j platí vztah

$$var\left(\sum_{j=1}^{k} X_j\right) = \sum_{j=1}^{k} var(X_j).$$

Dále pro libovolnou náhodnou veličinu X s konečným rozptylem σ^2 a konstantu a platí, že

$$var(aX) = a^2\sigma^2.$$

Také připomeňme, že rozptyl náhodné veličiny s alternativním rozdělením se střední hodnotou p je roven $p(1-p)$.

Jelikož jsme definovali parametr π_ϕ jako střední hodnotu kritické funkce Π (viz. 16), který lze odhadnout zprůměrováním jejich realizací, můžeme rozptyl (respektive směrodatnou odchylku) tohoto odhadu vajádřít jako

$$\sqrt{var(\hat{\pi}_\phi)} = \sqrt{\frac{1}{C^2}\sum_{l=1}^{C} \pi_\phi(1-\pi_\phi)} = \sqrt{\frac{\pi_\phi(1-\pi_\phi)}{C}}. \tag{34}$$

Výraz $\pi_\phi(1-\pi_\phi)$ může nabývat maximální (tedy pro nás nejméně příznivé) hodnoty $1/4$ v případě, že $\pi = 1/2$. Tuto hodnotu spolu s požadovanou přesností dosadíme do vztahu (34):

$$\sqrt{var(\hat{\pi}_\phi)} = \sqrt{\frac{1/4}{C}} = \frac{1}{2\sqrt{C}}$$

a z něj vyjádříme minimální počet cyklů C:

$$C = \frac{1}{4var(\hat{\pi}_\phi)}. \tag{35}$$

Abychom mohli garantovat přesnost na desetiny procenta, musíme tedy provést 250 000 opakování simulace.

2.3. Postup ověření síly testu

Kromě nedodržení hladiny α může statistický test selhat ještě jedním způsobem. V případech, kdy H_0 neplatí, může být test příliš konzervativní a zamítat

nulovou hypotézu v příliš málo případech. Abychom odhalili tento nedostatek, budeme vyšetřovat celý průběh silofunkce a srovnávat ji se silofunkcí analýzy rozptylu bez porušení předpokladů.

Budeme tedy provádět série simulací pro různé hodnoty parametru d, který určuje míru porušení nulové hypotézy. Začneme s $d = 0$ a postupně budeme pokračovat v krocích po 0.01 až 0.1[7]. Jelikož víme, že silofunkce je na takto vymezeném intervalu neklesající funkcí omezenou shora hodnotou jedna, budeme pokračovat do té doby, než se k tomuto maximu přiblížíme. Jednotlivé body pro nás nemají takovou důležitost jako měření při $d = 0$, snížíme proto počet opakování simulací pro každou hodnotu d na 2 500. Budeme tedy měřit s přesností 0.01.

2.3.1. Odhad průběhu silofunkce

Pomocí výše uvedeného postupu můžeme vyjádřit silofunkci jako množinu desítek diskrétních bodů zaměřených s určitou chybou. Tato reprezentace není pro další práci příliš praktická. Abychom mohli silofunkce zobrazit a dále s mini pracovat, proložíme jednotlivé body spojitou křivkou.

Z možných alternativ jsme dali přednost proložení polynomem, přestože tato metoda má určité nevýhody – například to, že v krajních hodnotách ztrácí přesnost a nijak nevyužívá známé vlastnosti silofunkce. Tyto nevýhody lze nicméně korigovat pomocí restrikcí, kterými model omezíme.

Budeme požadovat, aby nalezená funkce byla při maximální hodnotě d rovna jedné. Dále požadujeme, aby v tomto bodě byla její první derivace rovna nule – při vyšších hodnotách ji totiž bez závažné ztráty přesnosti můžeme nahradit konstantní funkcí. Nakonec budeme těžit ze znalosti toho, že silofunkce má své minimum při $d = 0$ a budeme tedy v bodě 0 požadovat nulovou první derivaci[8].

[7]První simulace ukázaly, že pro zmapování průběhu silofunkce stačí obvykle volit krok o délce 0.1. V případě, že její průběh je velmi strmý, bylo nutné měřítko zjemnit, aby ji bylo možné spolehlivě proložit křivkou.

[8]Zde kromě minima v nule předpokládáme také to, že silofunkce je hladká. O platnosti prvního předpokladu nám napovídá fakt, že silofunkce je symetrická (jelikož d a $-d$ vede ke stejnému posunutí). Předpoklad hladkosti (zejména v bodě $d = 0$) se zdá také intuitivně správný a hodnoty pozorované v této studii jej podporují, důkaz pro něj však nemáme.

Otázkou je stanovení vhodného stupně polynomu, kterým budeme naměřené hodnoty prokládat. Polynomy nízkého stupně nedokážou silofunkci kopírovat s dostatečnou věrností. A naopak polynomy příliš vysokého stupně jsou citlivé na odchylky jednotlivých měření a můžou být také problematické při numerickém výpočtu kvůli vysokému číslu podmíněnosti matice $\mathbf{X}'\mathbf{X}$. Za těchto podmínek se ukázal pro popis silofunkce jako nejvhodnější kompromis polynom šestého stupně.

Jelikož v bodě 0 měříme s vyšší přesností než v ostatních bodech, stanovíme hodnoty regresních parametrů pomocí zobecněné metody nejmenších čtverců. Pro jejich odhad využijeme následující vztah:

$$\hat{\boldsymbol{\beta}} = [\mathbf{I} - \mathbf{C}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}\mathbf{B}]\mathbf{C}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \mathbf{C}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}\mathbf{b}, \qquad (36)$$

kde

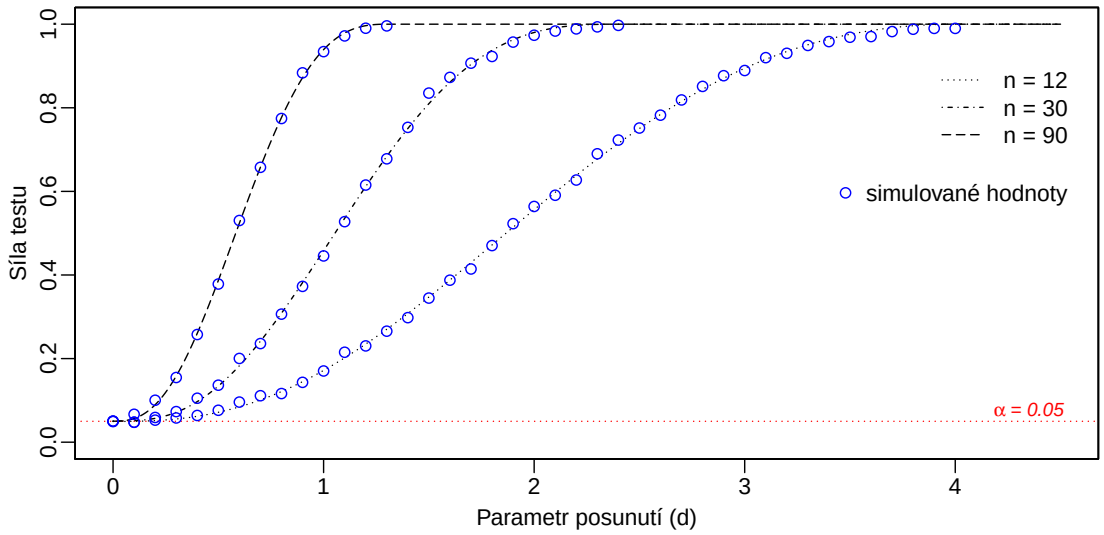
$$\mathbf{C}^{-1} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}.$$

Sloupcový vektor Y obsahuje odhadnuté hodnoty π_ϕ, $\mathbf{X}$ je matice regresorů, $\boldsymbol{\Sigma}^{-1}$ je diagonální matice obsahující informace o přesnosti měření pro jednotlivé body a vektor $\mathbf{b}$ s maticí $\mathbf{B}$ specifikuje tři výše uvedené podmínky, které požadujeme. Pro podrobnější popis lineárních modelů s podmínkami čtenáře odkazujeme například na skripta (Fišerová, 2013). Konkrétní podoba jednotlivých vektorů a matic je uvedena v kapitole 2.4 této práce v syntaxi jazyka R.

Pro maximální přesnost proložení bodů polynomem se ukázalo důležité pomocí vhodně zvoleného zastavovacího kritéria vymezit interval, na kterém měření provádíme – tedy stanovit hodnotu d, ve které měření ukončíme. V případě, že zastavíme příliš pozdě, bude část křivky téměř rovnoběžná s osou x, což je při modelování pomocí polynomu poměrně problematické. Často v tomto případě dochází k nežádoucímu ,,zvlnění" průběhu funkce. Testování proto zastavujeme již tehdy, když dvakrát po sobě nalezneme hodnotu $\hat{\pi}_\phi$ vyšší než 0.99.

Modely, ke kterým vede popsaná metoda, vystihují naměřená data velmi přesně – hodnota R^2 dle našich pozorování nikdy neklesá 0.98. Několik sad simulovaných dat a modelovaných silofunkcí znázorňuje obrázek 4.

Obrázek 4: Simulované hodnoty a odhadnutý průběh silofunkce
$(k = 3, u = 1, v = 1, b = 0)$



2.4. Implementace do prostředí R

Při realizaci navrhovaného postupu je nezbytná pokud možno maximální automatizace procesu. Manuální přenastavování hodnot parametrů a spouštění cyklu simulací by vedlo k neadekvátnímu nárůstu potřebného času. Pro realizaci simulací bylo proto zvoleno prostředí jazyka R, který umožňuje dosažení potřebné míry automatizace a zároveň obsahuje široké spektrum nástrojů pro statistickou analýzu.

Zásadní otázkou pro nás byla také optimalizace samotného výpočtu. Abychom dodrželi stanovenou přesnost, je nutné opakovaně generovat velké množství pseudonáhodných čísel a mnohokrát vyhodnotit statistický test. Počet opakování tohoto procesu bude dosahovat řádu statisíců až milionů. Problém optimalizace je v jazyce R obzvláště palčivý, jelikož ten ve srovnání s jinými jazyky patří k velmi pomalým.

Jazyk R je nadstavbou jazyka C, který bývá sám o sobě považován za mimořádně rychlý. Komunikace mezi oběma vrstvami nicméně celou proceduru zpomaluje (Ligges & Fox, 2008). Při přípravě skriptu jsme se proto pokoušeli přenést

co nejvíce úkonů na úroveň, která je realizována v jazyce C. K tomu lze využít zejména předdefinované funkce, které nám často umožní se vyhnout zařazení cyklů na úrovni jazyka R. Pro představu uveďme rozdíl mezi rychlostí výpočtu průměru řady celých čísel od 1 do 10^8 pomocí funkce *mean()* a pomocí námi definované funkce *slowMean()*:

```
slowMean = function(x)
{
  suma = 0
  pocet = 0
  for(i in x)
  {
    suma = suma + i
    pocet = pocet + 1
  }
  return(suma/pocet)
}

x = 1:(10^8)
system.time(slowMean(x))[1]/system.time(mean(x))[1]
```

Poměr rychlostí získaný na počítači, který byl užit k simulacím v této práci, se pohybuje mezi překvapivě vysokými hodnotami 350 až 500 ku jedné.

Pro provedení simulací jsme využili následující funkci. Jejími parametry jsou hodnoty n, k, u, v, d, b představené výše a počet cyklů. Získanou hodnotou je pak odhad veličiny π_ϕ.

```
anova_simulace = function(N,k,u,v,b,d,pocet_cyklu)
{
  n = round( 2*N*((u-1)*(1:k)+k-u) / (k*(k-1)*(u+1)) )
  if(v==1){s=rep(1,k)}else{s = (k*((v-1)*v^(1:k)+v^k-v^2)) / (v^k*(k
      +v)-v*(k*v+1))}
  m = (((1:k)-1)/(k-1)-0.5)*d

  N = sum(n) #korekce po chybě ze zaokrohlení

  p = numeric(pocet_cyklu)
  skupiny = factor(rep(1:k,n))

  for(i in 1:pocet_cyklu)
  {
```

```
    mereni = rSkewNorm(N,rep(sample(m),n),rep(sample(s),n),b)
    prum = tapply(mereni, skupiny, mean)
    rozp = tapply(mereni, skupiny, var)
    F = ( (sum(n*(prum - mean(mereni))^2)/(k-1)) / (sum((n-1) * rozp
        )/(N-k)) )
    p[i] = pf(F, k-1, N-k, lower.tail = FALSE)
  }

  return(mean(p<0.05))
}
```

Pozn.: Funkce rSkewNorm(), jejíž plné znění neuvádíme, generuje pseudonáhodná čísla ze zešikmeného rozdělení s šikmostí určenou jejím čtvrtým parametrem. Pokud ten je roven nule, pak přesně odpovídá funkci rnorm(). V případech, kdy nebylo potřeba generovat zešikmené rozdělení, byla použita přímo funkce rnorm() kvůli úspoře výpočetní kapacity.

Další funkce slouží k nalezení polynomu, který popisuje průběh silofunkce. Vstupní hodnotou je řádek tabulky (dataframu), který obsahuje hodnoty všech parametrů vyjma d. Dále je zde uveden počet cyklů pro $d \neq 0$ a počet cyklů pro $d = 0$. Nakonec je zde údaj o délce kroku – tedy hodnotě, která se po každé sadě simulací přičítá k hodnotě d při vyšetřování průběhu silofunkce. Funkce vrací opět řádek tabulky doplněný o odhadnuté koeficienty polynomu $\hat{\beta}$, hodnotu $\hat{\pi}$ pro $d = 0$ a maximální hodnotu d, pro kterou byla simulace prováděna.

```
najdi_silofunkci = function(zadani)
{
  list2env(lapply(as.list(zadani[c("N","k","u","v","b","krok_d","
      pocet_cyklu","pocet_cyklu_pri_d0")]), as.numeric),envir=parent.
      frame())

  sila = anova_simulace(N,k,u,v,b,0,pocet_cyklu_pri_d0)
  if(pocet_cyklu==0){zadani["alfa0"] = sila[1]; return(zadani);}
  sila[2] = anova_simulace(N,k,u,v,b,krok_d,pocet_cyklu)

  j = 2
  while(sila[j]<0.99 | sila[j-1]<0.99)
  {
    j = j+1
    sila[j] = anova_simulace(N,k,u,v,b,krok_d*(j-1),pocet_cyklu)
  }
```

```
  d = (((1:j)-1)*krok_d)
  h = 7
  X = matrix(numeric(h*length(d)),ncol=h)
  for(i in 1:h){X[,i] = d^(i-1)}
  # V je varianční matice na mínus prvou
  V = diag(c(pocet_cyklu_pri_d0/pocet_cyklu,rep(1,length(d)-1)))

  a = max(d)
  B = matrix(c(c(0,1,rep(0,h-2)),(1:h-1)*a^(1:h-2),a^(1:h-1)), byrow
      =TRUE, nrow=3)
  b = c(0,0,-1)
  C = solve(t(X) %*% V %*% X) #pomocná matice
  beta = (diag(h) - C %*% t(B) %*% solve(B %*% C %*% t(B)) %*% B) %*
      % C %*% t(X) %*% V %*% sila - C %*% t(B) %*% solve(B %*% C %*%
      t(B)) %*% b
  zadani[1:h+which(colnames(tabulka)=="beta0")-1] = beta
  zadani["max_d"] = a
  zadani["alfa0"] = sila[1]

  return(zadani)
}
```

Poslední část skriptu načítá tabulku se zadáním simulací ze souboru ve formátu *.csv* a postupně prochází její řádky. Pro každý řádek pak volá funkci pro nalezení silofunkce a její výsledky ukládá zpět do tabulky. Skript je ukončen tehdy, když jsou všechny chybějící hodnoty v tabulce doplněny.

```
soubor = winDialogString("Tabulka se vstupními údaji: ","tab.csv")
repeat
{
  tabulka = read.table(soubor, header = TRUE, sep = ";", dec = ",")
  if(min(tabulka["vyplneno"])>0){break}
  cislo_radku = min(tabulka[tabulka["vyplneno"]==0,"id"])
  tabulka[cislo_radku,] = najdi_silofunkci(tabulka[cislo_radku,])
  tabulka[cislo_radku,"vyplneno"] = 1;
  print(Sys.time())
  print(tabulka[cislo_radku,])
  write.table(tabulka, soubor, col.names = TRUE, row.names = FALSE,
      sep = ";" dec = ",")
}
```

Výhodou tohoto postupu je to, že skript může pracovat bez průběžných zásahů uživatele. Ten pouze připraví tabulku, která obsahuje všechny kombinace para-

metrů, pro něž máme v plánu simulace provádět, a spustí program. Dle rozsahu uživatelovy zakázky jsou pak do souboru doplněny hledané hodnoty.

3. Výsledky

Celkově jsme provedli přibližně $3.4 \cdot 10^8$ statistických testů na více než $3.5 \cdot 10^{10}$ pseudonáhodných číslech. Zkoumali jsme zvlášť chování analýzy rozptylu při porušení podmínky homoskedasticity a pro zešikmění rozdělení sledované náhodné veličiny. V obou případech jsme zjišťovali chování testu za platnosti nulové hypotézy i v případě jejího porušení. Pozornost byla věnována také tomu, jakou roli hraje rozsah náhodného výběru, počet skupin, do kterých je rozdělen, a vyváženost jejich rozsahů.

3.1. Dopad porušení podmínky homoskedasticity

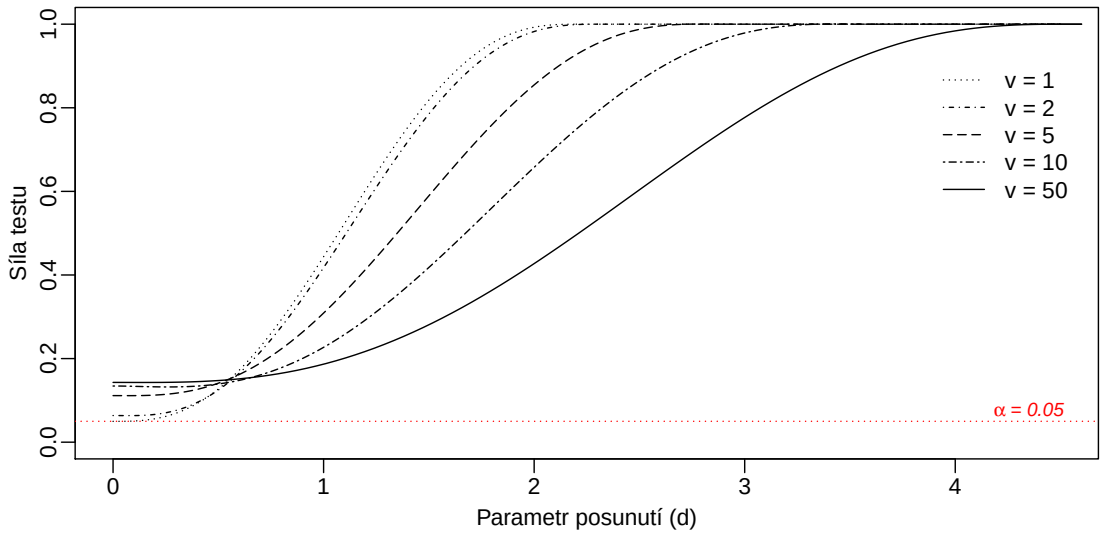
První otázka, kterou si položíme, je, jestli chování analýzy rozptylu vůbec souvisí s hodnotou parametru v. Odpověď jsme hledali tak, že jsme stanovili pevné hodnoty parametru k a n a sledovali jsme změny ve tvaru silofunkce v závislosti na změnách parametru v. Za tyto pevné hodnoty jsme zvolili pět náhodných výběrů ($k = 5$), každý o deseti měřeních ($n = 50$). Obrázek 5 znázorňuje průběhy silofunkce pro hodnoty parametru v od 1 (tedy shodné rozptyly ve všech skupinách) po 50 (tedy nejheterogennější skupina má padesátkrát větší rozptyl než ta nejhomogennější).

Je patrné, že porušení podmínky homoskedasticity má na výsledky analýzy rozptylu značný vliv. Při parametru $v = 2$ se průběh silofunkce téměř neliší od situace s vyrovnanými rozptyly, nicméně již při $v = 5$ je silofunkce znatelně deformovaná. Její počátek neleží na hladině α a ve srovnání s $v = 1$ neroste dostatečně strmě. Vyšší hodnoty parametru v toto chování ještě zdůrazňují.

3.1.1. Vliv rozsahu výběru

Vliv rozsahu výběrů na chování testu při porušení podmínky homoskedasticity jsme zjišťovali srovnáním průběhu silofunkce při hodnotě parametru $v = 5$ a $v = 1$

Obrázek 5: Průběh silofunkce v závislosti na porušení předpokladu homoskedasticity
$(k = 5, n = 50, u = 1, b = 0)$



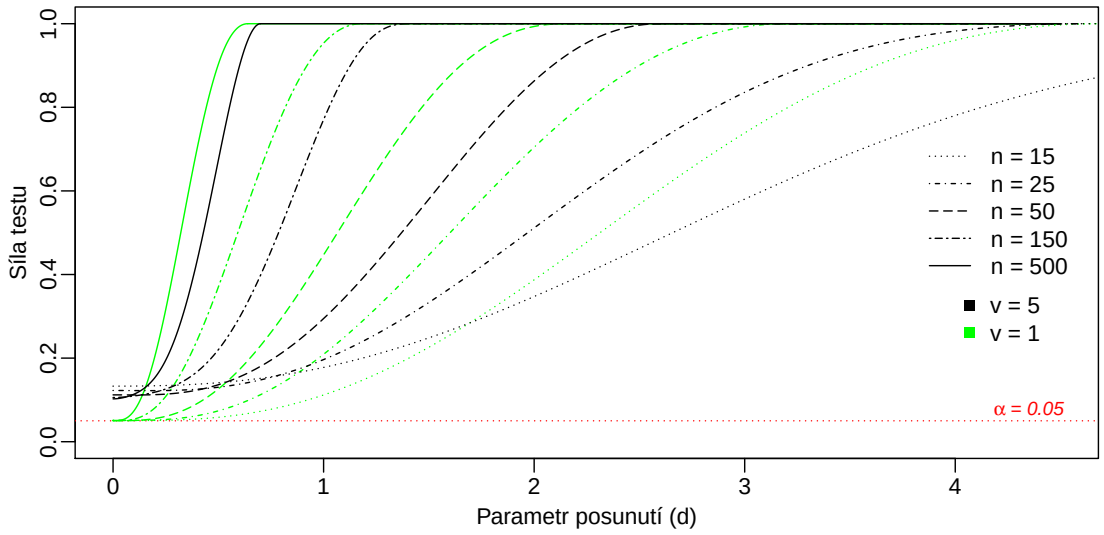
při pěti rozsazích výběrů – 15, 25, 50, 150 a 500 měření. Ve všech případech byly pozorování rovnoměrně rozděleny do pěti skupin (tedy $k = 5$, $u = 1$). Výsledky srovnání ukazuje obrázek 6.

Výsledky naznačují, že míra deformace silofunkce je ovlivněna rozsahem výběru n, ač je tento vliv spíše malý. I při nejvyšším rozsahu výběru $n = 500$ je zde patrný méně strmý nárůst hodnoty silofunkce, test je tedy méně citlivý. Při menších rozsazích výběrů je tento efekt ještě o něco silnější. Ani jedna z nalezených silofunkcí nerespektuje stanovenou hladinu α. Toto přehledněji znázorňuje obrázek 7. Je z něj patrné, že již při malém porušení podmínky homoskedasticity dochází k nárůstu pravděpodobnosti chyby prvního druhu, a při $k = 5$ a $v = 5$ již ani datové soubory o rozsazích několika set měření nezabrání selhání testu.

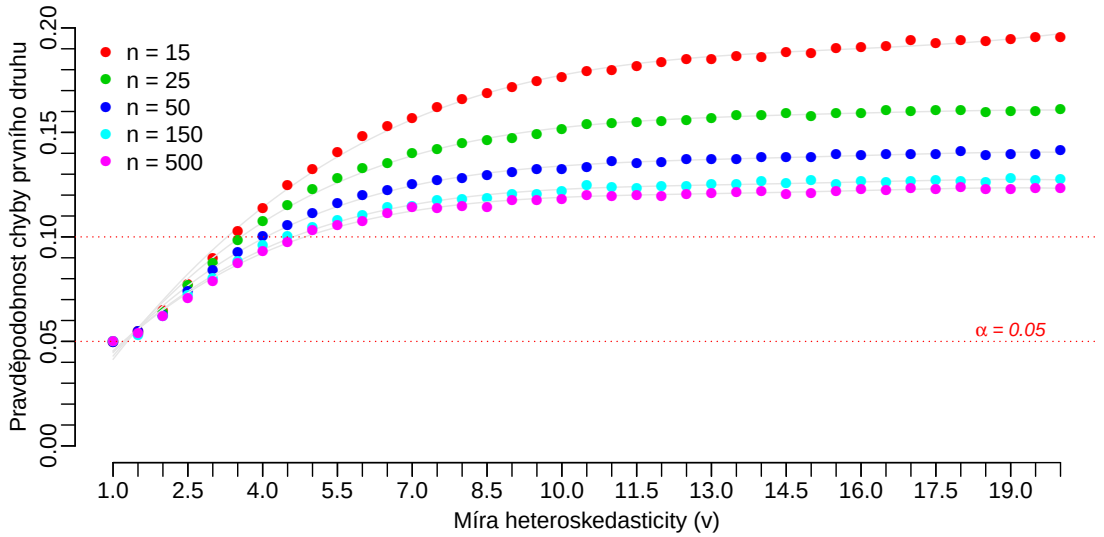
3.1.2. Vliv počtu výběrů

Z obrázku 8 je patrné, že silofunkce je nejvíce deformovaná, srovnáváme-li velké množství náhodných výběrů. Malé oslabení testu je patrné i při $k = 3$, nicméně citelný úbytek síly testu pozorujeme až při extrémních případech, kdy

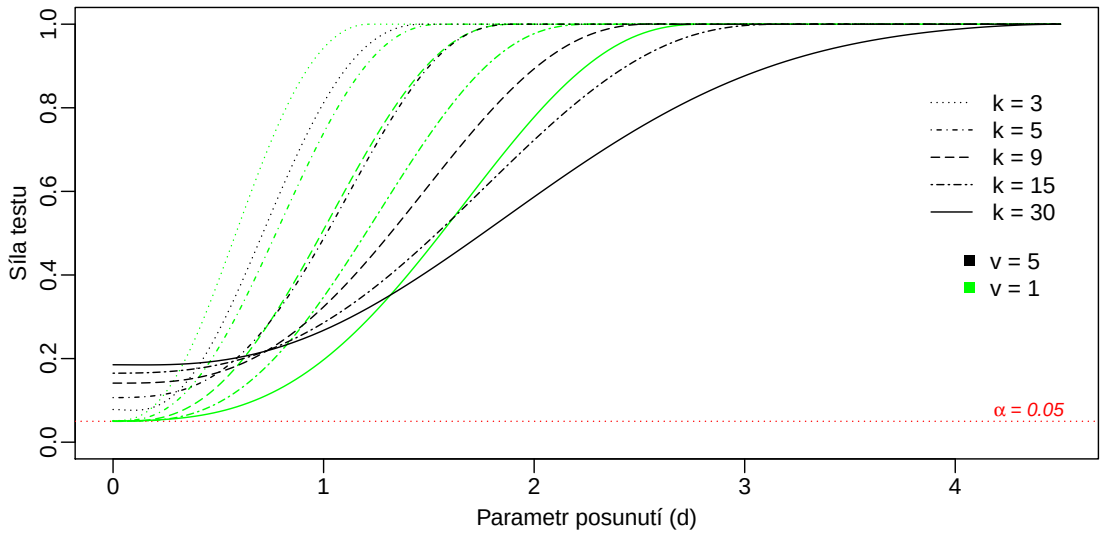
Obrázek 6: Změna tvaru silofunkce při porušení předpokladu homoskedascitiy v závislosti na rozsahu výběru
$(k = 5, u = 1, b = 0)$



Obrázek 7: Pravděpodobnost chyby prvního druhu při porušení předpokladu homoskedascitiy v závislosti na rozsahu výběru
$(k = 5, u = 1, b = 0)$



Obrázek 8: Změna tvaru silofunkce při porušení předpokladu homoskedascitiy v závislosti na počtu výběrů
$(n = 90, u = 1, b = 0)$



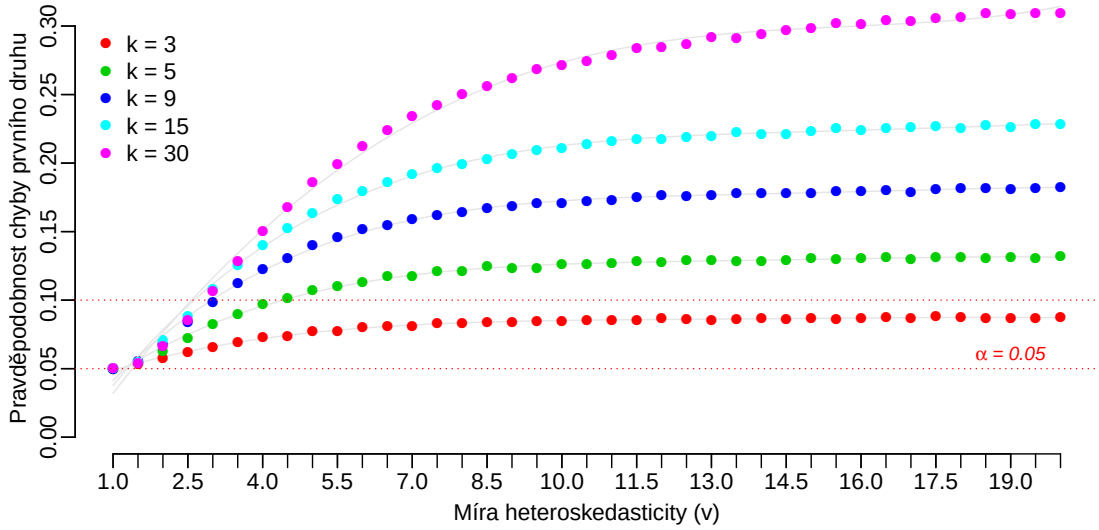
jsme srovnávali 30 skupin po 3 měření nebo 15 skupin po 6 měřeních. Větší pozornost budí chování testu za platnosti nulové hypotézy, což dokládá obrázek 9. Zdá se, že při malém počtu výběrů je test poměrně robustní vůči porušení podmínky homoskedasticity, alespoň co se týče dodržení hladiny α. V našem případě při $n = 90$ a $k = 3$ nevedla ani extrémní hodnota $v = 20$ k selhání statistického testu – pravděpodobnost chyby prvního druhu se ustálila na přibližně 8 procentech. Při srovnání desítek malých skupin toto číslo vystoupalo až do desítek procent.

3.1.3. Vliv vyváženosti rozsahů výběrů

Posledním parametrem, jehož vliv na chování analýzy rozptylu při porušení jejich předpokladů jsme zkoumali, je parametr nevyváženosti rozsahů výběrů u. Simulovali jsme situaci, kdy srovnáváme 80 měření rozdělených do pěti skupin. Za hodnotu parametru u jsme postupně volili hodnoty 1, 2, 3, 5 a 10. Při $u = 1$ byl tedy rozsah každé skupiny 16 měření, při $u = 10$ to byly rozsahy 3,10,15,18 a 20 měření.

Tvary silofunkcí zachycené v obrázku 10 naznačují jen velmi malý vliv vyvá-

Obrázek 9: Pravděpodobnost chyby prvního druhu při porušení předpokladu homoskedascitiy v závislosti na počtu výběrů
$(n = 90, u = 1, b = 0)$



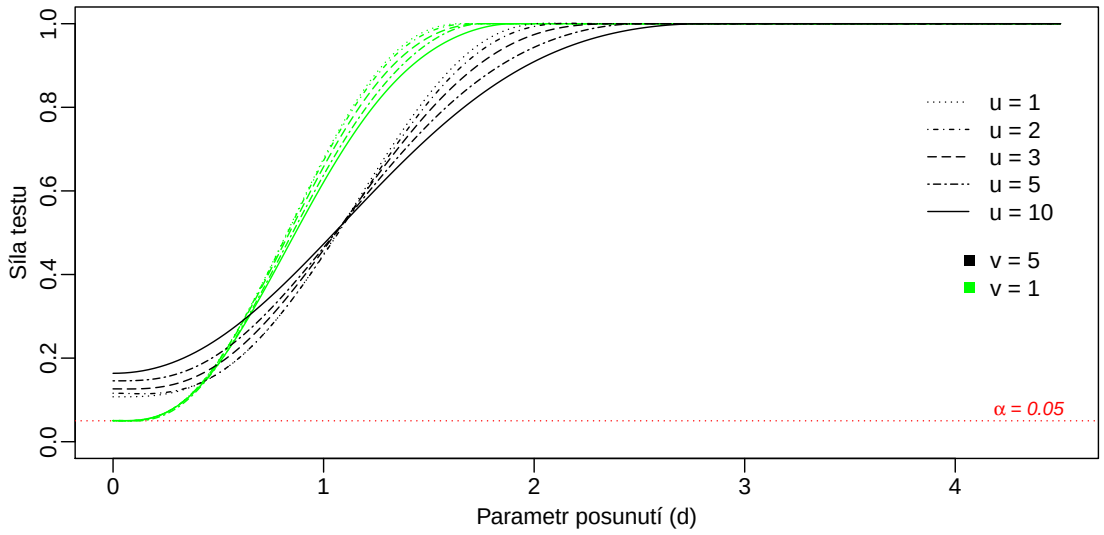
ženosti skupin na chování analýzy rozptylu při porušení podmínky homoskedasticity. Zdá se, že test je nejsilnější v případě vyvážených rozsahů, rozdíl nicméně není velký, ani když srovnáme krajní hodnoty. Ze stejného obrázku je také patrné, jak souvisí síla testu s vyvážeností rozsahů výběrů v případě, když žádnou podmínku neporušujeme – vliv parametru u je pak téměř neznatelný.

Hodnota parametru u má o něco významnější dopad na chování testu v případě platnosti nulové hypotézy, což znázorňuje obrázek 11. Vyvážené rozsahy skupin dávají testu o něco větší robustnost, nicméně ani zde není vliv příliš patrný. V našem případě, kdy $n = 80$ a $k = 5$, je test více méně spolehlivý při $v < 4$, pokud jsou skupiny vyvážené, a při $v < 2.5$, pokud jsou výrazně nevyvážené $(u = 10)$.

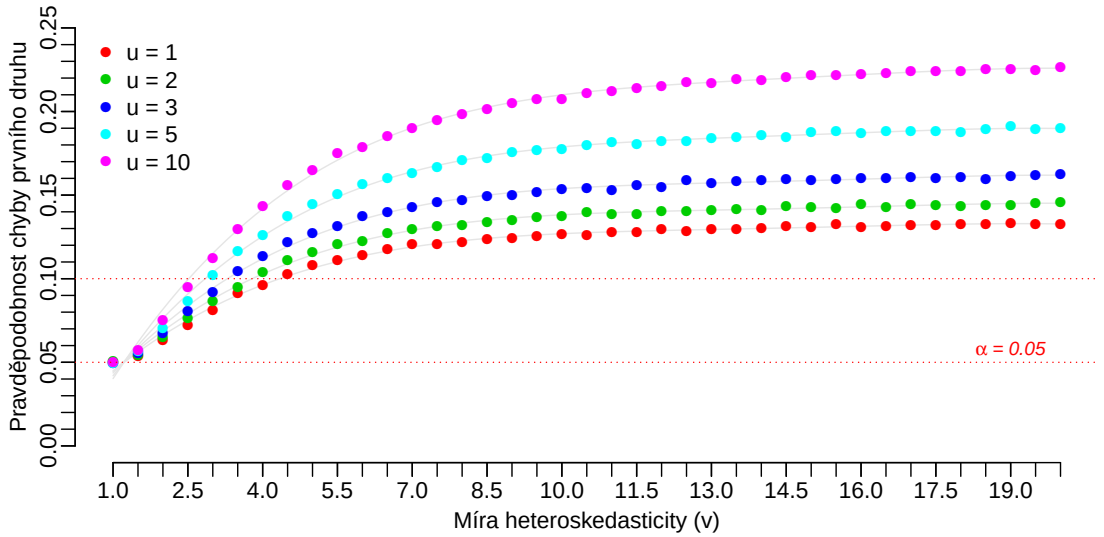
3.2. Dopad porušení podmínky normality rozdělení

Chování analýzy rozptylu při zešikmeném rozdělení jsme zkoumali stejným způsobem jako vliv homoskedasticity. Opět jsme vždy zvolili pevný rozsah výběru a počet skupin a odhadovali sílu testu v závislosti na míře porušení nulové hy-

Obrázek 10: Změna tvaru silofunkce při porušení předpokladu homoskedascitiy v závislosti na vyváženosti rozsahů výběrů $(k = 5, n = 80, b = 0)$



Obrázek 11: Pravděpodobnost chyby prvního druhu při porušení předpokladu homoskedascitiy v závislosti na vyváženosti výběrů $(k = 5, n = 80, b = 0)$



potézy. Všechny výsledky v této kapitole byly získány při dodržení podmínky homoskedasticity, tedy $v = 1$. Obrázek 12 znázorňuje průběh silofunkce při $k = 5$ a $n = 50$ pro různě zešikmená rozdělní pravděpodobnosti, z nichž byly generovány náhodné výběry.

Výsledek je poměrně překvapivý. Silofunkce stoupá nejstrměji při maximální šikmosti ($b = 400$). Srovnáme-li tvar křivky se situací, kdy podmínka normality rozdělení porušena není ($b = 0$), zjistíme, že šikmost rozdělení zvyšuje sílu statistického testu. Tento efekt je poměrně markantní, vyskytuje se však až u vysoce zešikmených rozdělení – hodnoty parametru $b \leq 2$ vedly k téměř identickým křivkám.

Stejně překvapivé je i chování testu za platnosti nulové hypotézy. Všech pět silofunkcí respektuje stanovenou hladinu α a zdá se, že některé mají počátek v ještě nižších hodnotách, což doložíme v následující podkapitole.

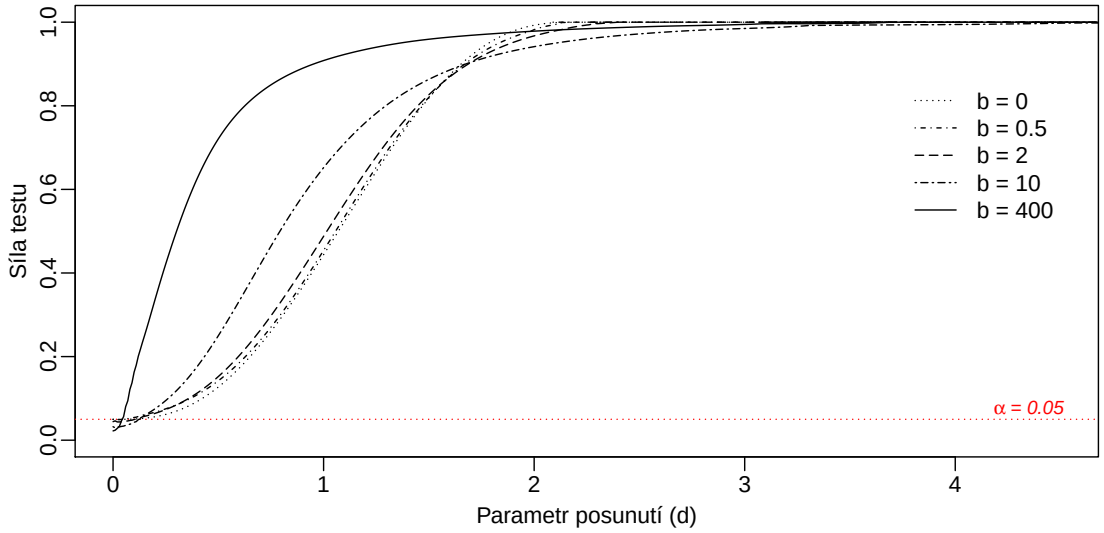
Nutno také zmínit ještě jeden důsledek porušení podmínky normality dat. V případě zešikmené náhodné veličiny má silofunkce jiný průběh nejen než ve všech doposud prezentovaných případech. Nejde jen o rozdíl kvantitativní, ale nově nalezené silofunkce zřejmě spadají do jiné rodiny funkcí než ty dřívější. Nepříjemným důsledkem je pro nás to, že v mnoha případech nelze věrně body proložit polynomem, který jsme definovali dříve. V případech, kdy se nám nepodařilo dostatečně přesné proložení, vynášíme do grafu samotná měření spojená lomenou čarou. Abychom dosáhli hladšího průběhu, vyrovnali jsme před vynesením body pomocí klouzavého průměru s velikostí okna 7 měření[9]. Při zjemnění kroku na 0.01 bodu tento postup vede k uspokojivým výsledkům a kolísání je vidět jen na několika místech.

3.2.1. Vliv rozsahu výběru

V kapitole 1.2.2 jsme naznačili, že očekáváme rozdílný dopad porušení podmínky normality v závislosti na rozsahu výběru. Chování náhodných veličin popisované centrálními limitními větami by mělo u velkých rozsahů výběru zmírňovat

[9]Vyhlazení bylo provedeno tak, že pro všechny body vyjma prvních a posledních třech byly vypočteny vyrovnané hodnoty jako $\pi^*_{d_0} = \frac{1}{7}\sum_{l=-3}^{3} \pi_{d_l}$.

Obrázek 12: Změna tvaru silofunkce při v závislosti na zešikmení rozdělení $(k = 5, n = 50, u = 1, v = 1)$



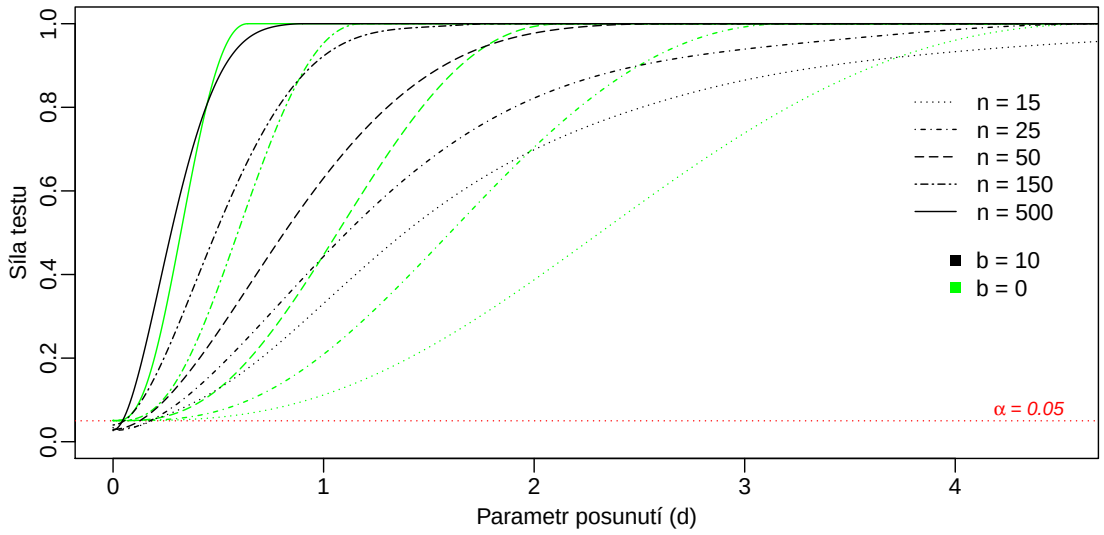
dopad zešikmení na výsledek statistického testu. Obrázek 13 naznačuje, že naše domněnka byla správná. Při srovnání silofunkce při pěti různých rozsazích výběru ($n = 15, 25, 50, 150$ a 500), vždy při $k = 5, u = 1, v = 1$, byl vliv porušení podmínky normality zřetelně nejmenší při $n = 500$. Pro malé výběry jsme pozorovali výrazné rozdíly mezi silofunkcemi – při srovnávání pěti skupin po třech měřeních byl rozdíl při d kolem hodnoty 2 více než 0.25. Ve všech případech vedlo zešikmení k nárůstu síly testu na většině délky průběhu silofunkce.

Potvrdilo se i naše dřívější pozorování, že i přes porušení podmínky normality test respektuje stanovenou hladinu α. Z obrázku 14 je patrné, že se rostoucím zešikmením pravděpodobnost chyby prvního druhu klesá hlouběji pod stanovenou mez. U malých souborů je tento vliv výraznější, ač zde má rozsah výběru jen velmi malý vliv.

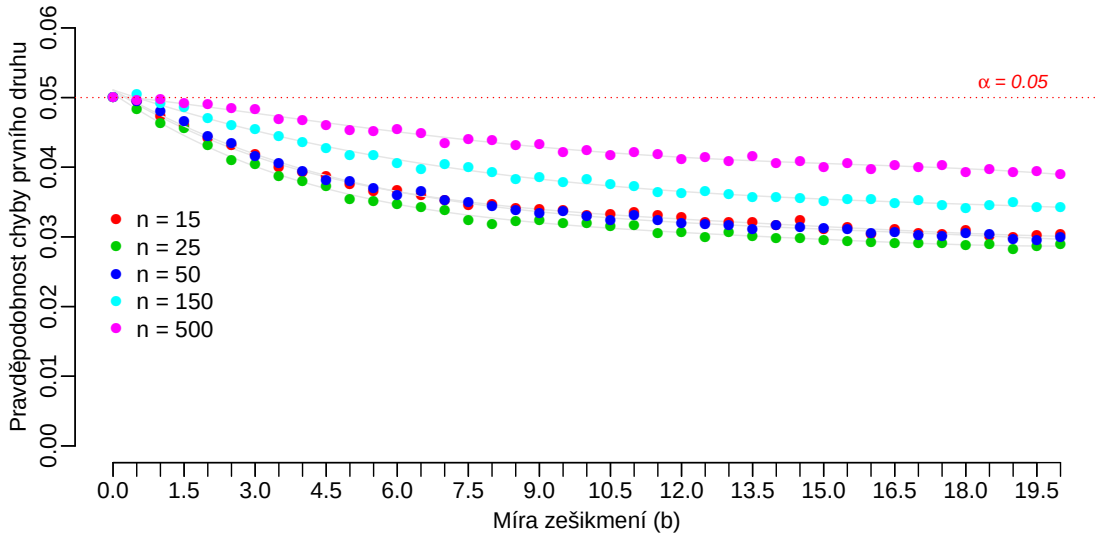
3.2.2. Vliv počtu výběrů

Našim dalším cílem bylo prozkoumat, jestli se dopad porušení předpokladu normality různí podle počtu náhodných výběrů, se kterými pracujeme. Simulovali jsme proto situaci, kdy máme 90 měření rozdělených do $3, 5, 9, 15$ a 30 skupin. Ve

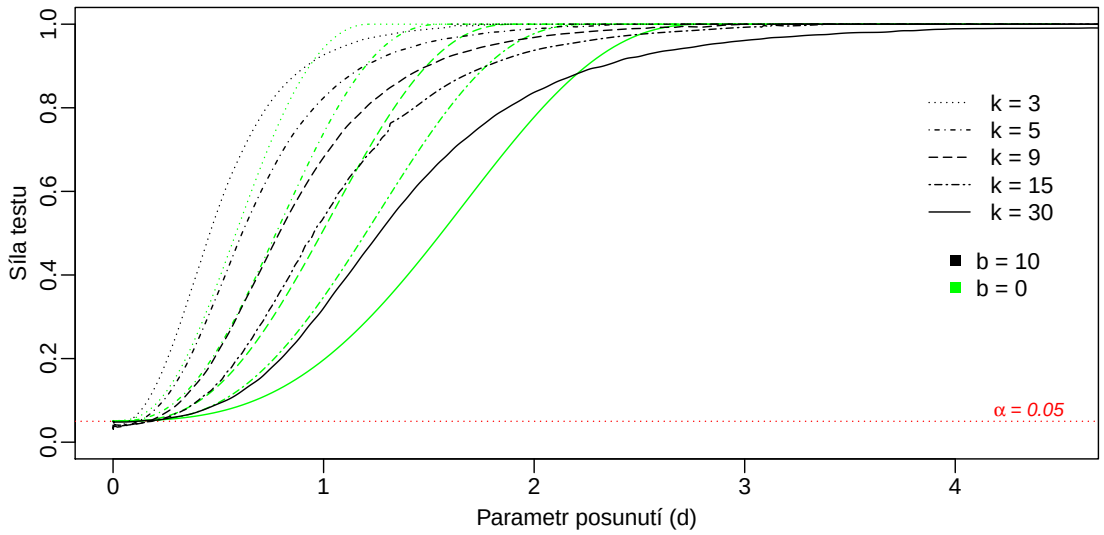
Obrázek 13: Změna tvaru silofunkce při zešikmeném rozdělení v závislosti na rozsahu výběru
$(k = 5, u = 1, v = 1)$



Obrázek 14: Pravděpodobnost chyby prvního druhu při zešikmeném rozdělení v závislosti na rozsahu výběru
$(k = 5, u = 1, v = 1)$



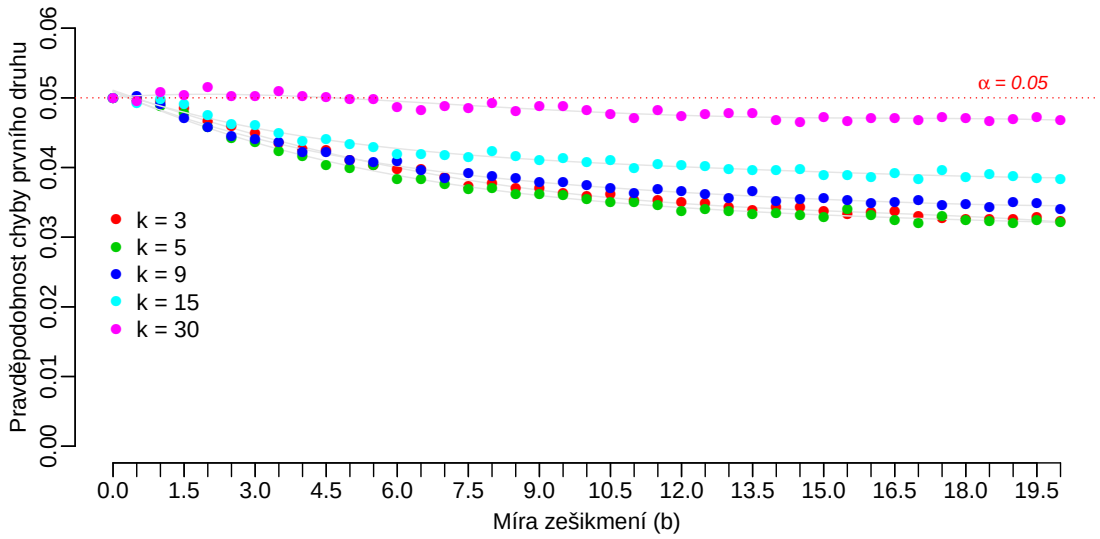
Obrázek 15: Změna tvaru silofunkce při zešikmeném rozdělení v závislosti na počtu výběrů
$(n = 90, u = 1, v = 1)$



všech případech jsme srovnávali silofunkci pro případ, kdy je koeficient šikmosti $b = 0$ a $b = 10$. Výsledky potvrzují již dříve pozorovaný jev, že test obecně ztrácí sílu rozdělením konstantního počtu měření do více skupin. Samotný fakt, jestli rozdělení pravděpodobnosti bylo zešikmené nebo ne, silofunkci sice ovlivňuje, ale nelze jednoznačně říct, že se velikost tohoto dopadu mění v závislosti na parametru k. Na obrázku 15, který výsledky zachycuje, je sice rozdíl nejpatrnější při nejvyšším počtu skupin, ale toto je částečně způsobeno tím, že zde má křivka celkově menší sklon.

Určité rozdíly můžeme pozorovat na chování testu v případě platnosti nulové hypotézy, viz obrázek 16. V případě srovnání 3 skupin po 30 měřeních se pravděpodobnost chyby prvního druhu držela těsně pod hladinou α pro koeficient šikmosti od 0 po 20. Pro $k \geq 5$ s rostoucím zešikmením klesal k hodnotám kolem 0.03.

Obrázek 16: Pravděpodobnost chyby prvního druhu při zešikmeném rozdělení v závislosti na počtu výběrů
$(n = 90, u = 1, v = 1)$

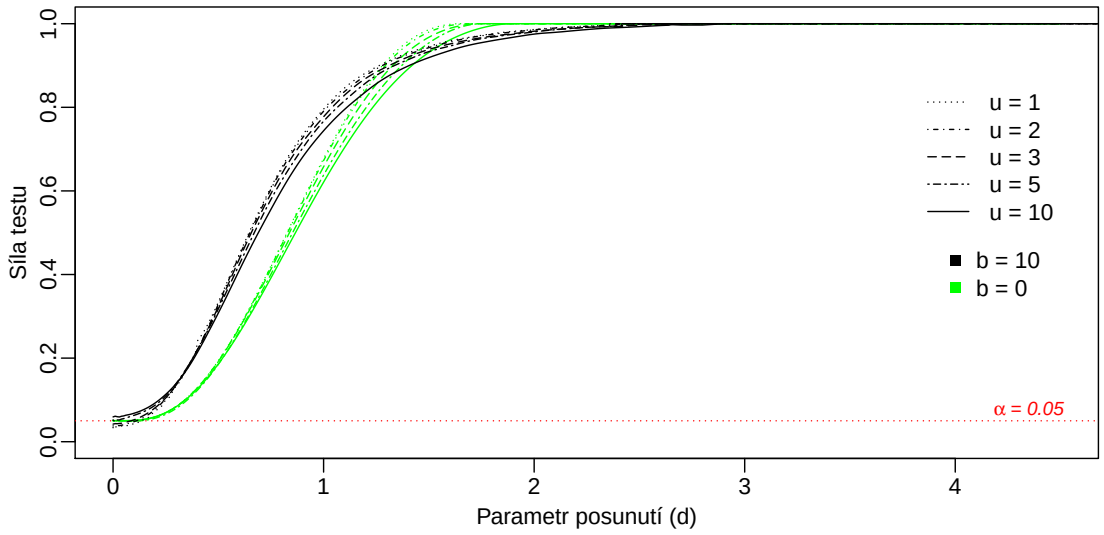


3.2.3. Vliv vyváženosti rozsahů výběrů

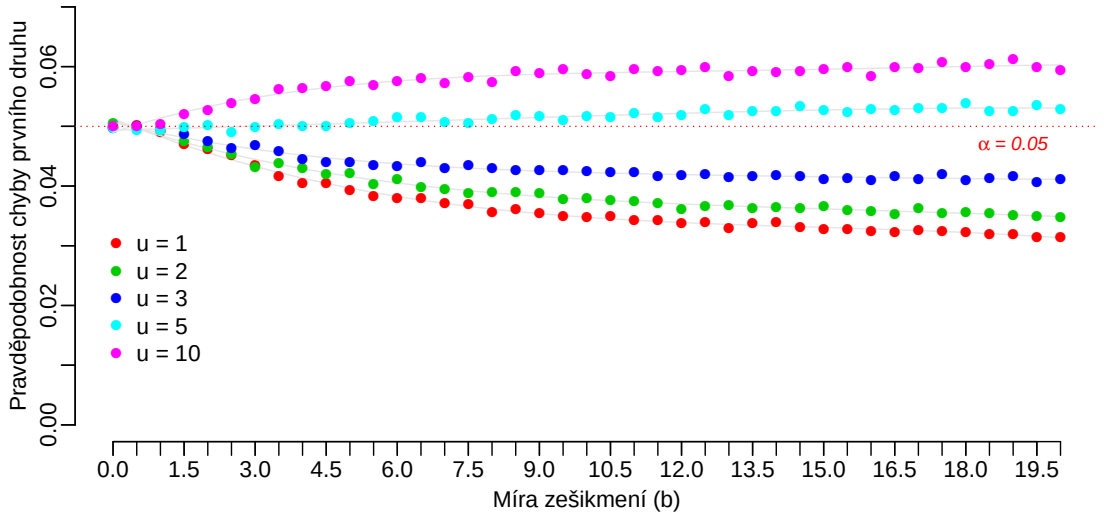
Poslední situací, kterou jsme simulovali, byl případ náhodných výběrů s nestejnými rozsahy. Opět jsme stanovili parametry n, k za konstantní, konkrétně 5 skupin o 80 měřeních, a sledovali jsme dopad porušení podmínky normality při hodnotách parametru $u = 1, 2, 3, 5$ a 10. Obrázek 17 ukazuje, že nevyváženost skupin nehraje prakticky žádnou roli, co se týče průběhu silofunkce testu.

Zajímavější výsledky jsme nalezli, když jsme zkoumali chování testu při platnosti nulové hypotézy, viz obrázek 18. Se stoupající hodnotou parametru u narůstá pravděpodobnost chyby prvního druhu. Pro $u = 5$ bez ohledu na míru zešikmení test dodržuje téměř přesně stanovenou hladinu α. Při $u = 10$ tuto hladinu přesahuje, ač jen o málo – při $b = 20$ dosahuje hodnoty přibližně 0.06.

Obrázek 17: Změna tvaru silofunkce při zešikmeném rozdělení v závislosti na vyváženosti rozsahů výběrů
$(k = 5, n = 80, v = 1)$



Obrázek 18: Pravděpodobnost chyby prvního druhu při zešikmeném rozdělení v závislosti na vyváženosti rozsahů výběrů
$(k = 5, u = 1, v = 1)$



4. Diskuse nalezených poznatků

Výsledky prezentované studie potvrzují důležitost dodržení předpokladů analýzy rozptylu. Ať už se jedná o přítomnost heteroskedasticity nebo zešikmení rozdělení zkoumané veličiny, silofunkce i pravděpodobnost chyby prvního druhu se mění v závislosti na míře porušení podmínky. Překvapivým zjištěním je to, že porušení každé z těchto podmínek se odráží na chování testu jiným způsobem.

Porušení podmínky homoskedasticity se na výsledcích projevilo tak, jak bychom intuitivně očekávali – test ztratil část své síly a přestal dodržovat stanovenou hladinu α. Toto nepříznivé chování testu lze do jisté míry zmírnit vysokým rozsahem souboru a tím, že budeme srovnávat spíše menší množství náhodných výběrů, které mají v ideálním případě stejný rozsah. Test nicméně ani v tomto případě příliš robustní není. S výjimkou extrémních hodnot, například když srovnáváme jen tři skupiny o rozsazích v řádech stovek či tisíců měření, doporučujeme striktně trvat na dodržení podmínky homoskedasticity, případně sáhnout po jiné metodě, která tímto předpokladem zatížena není. Nabízí se zobecnění testu popsaného B. L. Welchem (1947), které, ač se jedná pouze o přibližnou metodu, si zachovává dobré vlastnosti analýzy rozptylu, včetně srovnatelné síly (Ruxton, 2006).

Je určitým paradoxem, že podmínka shody rozptylů bývá diskutována méně často, než podmínka normality rozdělení. Dle našich výsledků porušení druhé jmenované podmínky představuje výrazně menší problém než první jmenované. Naopak se ukázalo, že pokud pracujeme s výrazně zešikmenými daty (bez změny rozptylu), tak test dokáže citlivěji odhalit porušení nulové hypotézy, než v případě normálního rozdělení, a přitom ani nevyčerpá stanovenou hladinu α. Tento výsledek je natolik překvapivý, že jej autor této práce zprvu považoval za důsledek nějaké skryté chyby v proceduře simulací. Pozorovaný jev se dá nicméně alespoň částečně vysvětlit následujícím způsobem.

Náhodná veličina s rozdělením pravděpodobnosti s parametrem b dosahujícím vysokých hodnot se nejčatěji realizuje na velmi úzkém intervalu. Například při $b = 400$ je rozdíl mezi mediánem a nejmenší hodnotou jen přibližně 0.02. Rozptyl

přitom zůstává roven 1. Zvláště u malých rozsahů výběru je tedy poměrně velká šance, že všechna pozorování budou spadat do velmi úzkého intervalu, a dojde tak k radikálnímu podhodnocení odhadu rozptylu. Například při náhodném výběru o rozsahu 10 pozorování má odhad rozptylu medián roven přibližně hodnotě 0.015 a devadesátý percentil hodnotě 0.37, při střední hodnotě odhadu rovné 1[10]. Často tedy nastává situace, kdy jsou některé nebo všechny srovnávané skupiny velmi homogenní, a test tak získá vysokou citlivost na porušení nulové hypotézy určené parametrem b.

Je nicméně nutné brát zřetel na to, že situace, kterou jsme modelovali, je velmi specifická. V reálných podmínkách jsou obvykle střední hodnota, rozptyl a šikmost rozdělení provázané. Nejčastěji pak pozorujeme situaci, kdy se střední hodnota veličiny mění nikoli posunutím celé její hustoty pravděpodobnosti, ale spíše současným zvětšením její šikmosti a rozptylu tak, jak jsme demonstrovali v naší ukázkové úloze (viz obrázek 1).

Do praxe si můžeme odnést poznatek, že s rostoucím počtem měření naléhavost podmínky normality rozdělení klesá, a to jen s malým vlivem vyváženosti skupin a jejich počtem. Pokud nemáme dostatek měření nebo vyžadujeme vysokou spolehlivost statistického testu, mezi parametrickými metodami vhodnou alternativu analýzy rozptylu nenalezneme. Spolehlivou neparametrickou náhradou je pak Kruskalův-Wallisův test, který je ovšem zatížen problémy typickými pro neparametrické metody (např. obtížná prezentace výsledků, omezená nabídka post-hoc testů, v některých případech nižší síla).

Co se týče limitů prezentované studie, nelze přehlédnout její hlavní nedostatek, kterým je jen nedostatečné pokrytí nabízeného tématu. Bylo by žádoucí nezkoumat jen izolované případy vzniklé manipulací s jednotlivými parametry, ale i jejich vzájemné vztahy. V ideálním případě bychom mohli vytvořit prostor o počtu dimenzí odpovídajícím počtu parametrů a každému jeho bodu přiřadit hodnotu rovnou pravděpodobnosti zamítnutí nulové hypotézy. Mohli bychom pak

[10]Získáno simulací s 250 000 opakováními.

zmapovat oblasti, kde je test ještě spolehlivý a kde již není. Takového úsilí by ale zřejmě poněkud překračovalo očekávaný rozsah bakalářské práce, nehledě na to, že vzhledem k možnému přínosu by se zřejmě nejednalo o příliš hospodárně vynaložený čas.

Další nedostatky se týkají spíše technického provedení. Příliš se neosvědčilo prokládání silofunkce polynomem. Tuto metodu jsme zvolili pro její eleganci – celou silofunkci jsme byli schopni věrně popsat pomocí sedmi hodnot. Ukázalo se však, že některé tvary, které vznikly při práci se zešikmeným rozdělením, se takto vystihnout nedají, což nás přinutilo k těžkopádné práci s jednotlivými body.

Naopak jako velmi praktická se ukázala plná automatizace procesu simulací včetně průběžného ukládání výsledků.

Závěry

Poznatky, které jsme získali pomocí simulační studie o chování analýzy rozptylu v závislosti na porušení jejich podmínek, lze shrnout do těchto bodů:

- Porušení předpokladu homoskedasticity i porušení předpokladu normálního rozdělení ovlivňují průběh silofunkce včetně chování testu za platnosti nulové hypotézy.

- V případě heteroskedasticity se test stává slabším a nedodržuje stanovenou hladinu α. Tyto důsledky se nápadně projevují již při poměrně malém rozdílu mezi rozptyly - například již při poměru rozptylů nejheterogennější a nejhomogennější skupiny rovnému pěti ku jedné.

- Velké soubory rozdělené do malého množství skupin jsou vůči tomuto zkreslení robustnější, důsledek porušení této podmínky však nelze zcela odstranit ani volbou velkých rozsahů výběrů.

- Zešikmení rozdělení samo o sobě nevede k oslabení síly testu, ani porušení stanovené hladiny α. V určitých případech může mít i pozitivní důsledky, což jsme doložili v případě silně zešikmených dat s nízkým rozptylem.

- Dopad porušení podmínky normality lze do velké míry odbourat vysokým rozsahem výběrů, zejména pak při malém množství srovnávaných skupin.

- Vyváženost rozsahů srovnávaných výběrů prakticky neovlivňuje robustnost testu vůči porušení normality.

Literatura

Anděl, J. (2007). *Základy matematické statistiky*. Praha: Matfyzpress.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Routledge.

Fišerová, E. (2013). *Lineární statistické modely*. Olomouc: Univerzita Palackého.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, *1*.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Horne, J. A. & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*.

Howell, D. (2013). *Statistical Methods for Psychology*. New York: Wadsworth.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*. New York: Wiley.

Ligges, U. & Fox, J. (2008). R Help Desk: How can I avoid this loop or make it faster? *R News*, *8*(1), 46–50.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*.

Plháková, A., Dostál, D., & Janečková, D. (2013). Cirkadiánní preference ve vztahu k depresivitě, subjektivní kvalitě spánku a cloningerovým dimenzím osobnosti. *Česká a slovenská psychiatrie*, *3*(109), 577–583.

Preiss, M. & Vacíř, K. (1999). *Beckova sebeposuzovací stupnice deprese. Příručka*. Brno: Psychodiagnostika.

Razali, N. & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, *17*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, *1-2*(34), 28–35.