MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Jaromír Macoun

Lillieforsův test normality

Katedra pravděpodobnosti a matematické statistiky

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

Praha 2021

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

Tímto bych rád poděkoval panu doc. RNDr. Arnoštu Komárkovi, Ph.D., vedoucímu mé bakalářské práce, za trpělivost, odbornou pomoc a čas, který mi věnoval. Dále velký dík patří mé rodině za trpělivost a pochopení během celého mého studia.

Název práce: Lillieforsův test normality

Autor: Jaromír Macoun

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: V práci představíme jednovýběrový Kolmogorovův-Smirnovův test ověřující, zdali náhodná veličina pochází z rozdělení určeného známou spojitou distribuční funkcí. Nejdříve zavedeme značení a dokážeme některé základní vlastnosti testové statistiky a odvodíme asymptotické kritické hodnoty pro tento test. Závěrem první kapitoly ukážeme konzistenci testu. Dále zavedeme Lillieforsův test normality a budeme studovat jeho vlastnosti. Stěžejní výsledek práce bude, že rozdělení testové statistiky za určitých podmínek nezávisí na neznámých parametrech. Nakonec uvedeme aproximace kritických hodnot a porovnáme s již publikovanými.

Klíčová slova: Kolmogorovův-Smirnovův test, normální rozdělení, Lillieforsův test normality, test normality

Title: Lilliefors test of normality

Author: Jaromír Macoun

Department: Department of Probability and Mathematical Statistics

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

Abstract: In this bachelor thesis will be shown one-sample Kolmogorov-Smirnov test, which compares empirical distribution function with one specified distribution function. At first we introduce marking and prove some basic properties about the test statistics and derive asymptotic critical values for the test. At the end of the first chapter we show consistency of the test. In the next step we initiate Lilliefors test of normality. The crucial outcome of the thesis is that distribution of test statistics with some assumptions is independent of unknown parameters. Finally we show a table of approximated critical values and compare with already publicated.

Keywords: Kolmogorov-Smirnov test, normal distribution, Lilliefors test of normality, test of normality

Obsah

Úvod

Testování hypotéz je jedna ze základních statistických metod, jak zjistit informace ze souboru dat. Mnohokrát se při testování musí předpokládat určitá vlastnost náhodného výběru. Jeden z častých předpokladů bývá, že náhodný výběr pochází z normálního rozdělení. Přesný t-test, nebo F-test tuto vlastnost vyžaduje. K testování normality dat existuje celá řada metod, my se zde s jednou seznámíme.

Tato práce si klade za cíl představit Lillieforsův test normality a ukázat jeho souvislost s Kolmogorovým-Smirnovým testem. V první kapitole se seznámíme s Kolmogorovým-Smirnovým testem, ukážeme si jeho jednostrannou a oboustrannou verzi. Odvodíme si zajímavé vlastnosti a tvar kritického oboru. Nejdůležitějším výsledkem této kapitoly bude důkaz konzistence testu.

V druhé kapitole zavedeme Lillieforsův test. Stěžejním výsledkem v této kapitole bude důkaz, že testová statistika je v jistém smyslu univerzální. Tohoto faktu se následně využije v další kapitole.

Třetí kapitola představí výsledky dvou simulací. Uvedeme zde tabulku aproximací kritický hodnot Lillieforsova testu a porovnáme jí s již publikovanou. Nakonec budeme diskutovat výsledek druhé simulace, kde uvažujeme použití kritických hodnot z Kolmogorova-Smirnova testu pro test normality.

1. Jednovýběrový Kolmogorovův-Smirnovův test

Jednovýběrový Kolmogorovův-Smirnovův test zkoumá, zdali náhodná veličina má rozdělení určené danou pevně zvolenou spojitou distribuční funkcí. Testová statistika je založena na porovnání distribuční funkce a empirické distribuční funkce daného náhodného výběru.

1.1 Model, hypotéza, testová statistika

V následující přípravné sekci budeme vycházet z práce Omelka (2021)(kapitola 5.1).

Definice 1 (empirická distribuční funkce). *Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr. Potom funkci $\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$ označujeme* empirickou distribuční funkci.

Kolmogorovův-Smirnovův test uvažuje náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení s distribuční funkcí F_X z modelu $\mathcal{F}$, o rozsahu n.

Model:

$$\mathcal{F} = \{\text{Všechna spojitá rozdělení}\}.$$

Chceme zjistit, zdali se distribuční funkce náhodného výběru rovná naší hypotetické distribuční funkci F_0.

Hypotéza a alternativa:

$$H_0 : F_X(x) = F_0(x) \quad \forall x \in \mathbb{R}, \quad H_1 : \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x).$$

Testová statistika bude založena na vzdálenosti empirické distribuční funkce v supremální normě. Jinými slovy zaznamenává největší možný absolutní rozdíl mezi $F_0(x)$ a $\widehat{F_n}(x)$.

Testová statistika:

$$K_n = \sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F_0(x) \right|. \tag{1.1}$$

Z tvaru testové statistiky je vidět, že nabývá nezáporných hodnot. Pokud je F_0 skutečná distribuční funkce náhodného výběru $X_1, \ldots, X_n$, pak K_n bude blízké nule, tedy nabízí se zamítat pro vysoké hodnoty testové statistiky. Z toho plyne tvar kritického oboru: (C,∞) pro nějaké $C > 0$. Testovou statistiku lze ekvivalentně přepsat $K_n = \max(K_n^+; K_n^-)$, kde

$$K_n^+ = \sup_{x \in \mathbb{R}} \left(\widehat{F_n}(x) - F_0(x) \right),$$

$$K_n^- = \sup_{x \in \mathbb{R}} \left(F_0(x) - \widehat{F_n}(x) \right).$$

Toto vyjádření K_n dále využijeme pro odvození vlastností testové statistiky a také pomocí K_n^+ a K_n^- lze zformulovat jednostranná verze testu.

Pro jednostranný test budeme předpokládat stejný model jako u oboustranného testu. Lišit se bude pouze alternativa a testová statistika.

Hypotéza a alternativa:

$$H_0^{'} : F_X(x) = F_0(x) \quad \forall x \in \mathbb{R},$$

$$H_1^{'} : \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x) \quad \text{a} \quad F_X(x) \geq F_0(x) \quad \forall x \in \mathbb{R}.$$

Testová statistika:

$$K_n^+ = \sup_{x \in \mathbb{R}} \left(\widehat{F_n}(x) - F_0(x) \right). \tag{1.2}$$

Respektive symetricky pro druhou verzi jednostranného testu.

Hypotéza a alternativa:

$$H_0^{''} : F_X(x) = F_0(x) \quad \forall x \in \mathbb{R},$$

$$H_1^{''} : \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x) \quad \text{a} \quad F_X(x) \leq F_0(x) \quad \forall x \in \mathbb{R}$$

Testová statistika:

$$K_n^- = \sup_{x \in \mathbb{R}} \left(F_0(x) - \widehat{F_n}(x) \right). \tag{1.3}$$

Na první pohled je vidět, že nulová hypotéza je u všech tří případů stejná, liší se pouze alternativa, vůči které testujeme.

1.2 Vlastnosti testové statistiky

V této části budeme pracovat s uspořádaným náhodným výběrem, proto si na úvod uvedeme jeho definici.

Definice 2 (Uspořádaný náhodný výběr)**.** *Nechť $X_1, \ldots, X_n$ je náhodný výběr z jednorozměrného spojitého rozdělení. Seřadíme-li náhodný výběr $X_1, \ldots, X_n$ od nejmenší po největší, získáme* uspořádaný náhodný výběr
$X_{(1)} < X_{(2)} \cdots < X_{(n-1)} < X_{(n)}$.

Symbolem $X_{(k)}$ rozumíme k-tou nejmenší hodnotu mezi pozorováními $X_1, \ldots, X_n$; nazýváme ji k-tá pořádková statistika.

Lemma 1. *Nechť F_0 je spojitá distribuční funkce a $X_1, \ldots, X_n$ náhodný výběr z jednorozměrného spojitého rozdělení, potom platí*

$$K_n^+ = \max_{1 \leq i \leq n} \left(\frac{i}{n} - F_0(X_{(i)}) \right), \quad K_n^- = \max_{1 \leq i \leq n} \left(F_0(X_{(i)}) - \frac{i-1}{n} \right).$$

Důkaz. Podrobný důkaz je uveden v práci Omelka (2021)(Lemma 5.1). □

Důsledkem tohoto lemmatu je, že za nulové hypotézy má $F_0(X_{(i)})$ beta rozdělení viz Omelka (2021)(Věta 2.13). Z toho vyplývá, že rozdělení K_n, K_n^+ a K_n^- za platnosti nulové hypotézy (u všech třech typů testů je nulová hypotéza stejná) nezávisí na F_0.

Lemma 2. *Nechť $U_{(i)}$ je i-tá pořádková statistika náhodného výběru o rozsahu n z rovnoměrného rozdělení na intervalu (0,1) pro $i \in \{1,\ldots,n\}$. Buď F spojitá rostoucí distribuční funkce náhodného výběru $X_1,\ldots,X_n$, potom náhodná veličina $F\left(X_{(i)}\right)$ má stejné rozdělení jako $U_{(i)}$.*

Důkaz. Zvolme $i \in \{1,\ldots,n\}$. Z předpokladu víme, že existuje inverzní funkce F^{-1} k F . Spočtěme distribuční funkci náhodné veličiny $F\left(X_{(i)}\right)$.

$$P\left(F\left(X_{(i)}\right) \leq t\right) = P\left(X_{(i)} \leq F^{-1}(t)\right) = F_{X_{(i)}}(F^{-1}(t)) = \star$$

$F_{X_{(i)}}(F^{-1}(t))$ je distribuční funkce náhodné veličiny $X_{(i)}$. Již z prvního výrazu je vidět, že pro $t \geq 1$ bude pravděpodobnost 1. Pro $t \leq 0$ bude pravděpodobnost 0. Proto se nadále budeme zabývat pro případ $0 < t < 1$.

$$\star = \frac{1}{B(i,n-i+1)} \int_0^{F(F^{-1}(t))=t} z^{i-1}(1-z)^{n-i} dz$$

Tato rovnost platí viz Omelka (2021)(Věta 2.13). Lemma plyne z toho, že funkce $W(t) := t$ pro $0 < t < 1$ je distribuční funkce rovnoměrného rozdělení na (0,1).

□

Poznámka. Předchozí lemma by platilo i bez předpokladu, že funkce F je rostoucí.

Tvrzení 3. *Nechť F_0 je spojitá distribuční funkce náhodného výběru $X_1,\ldots,X_n$ a K_n^+ je jako v rovnosti (1.2), potom*

$$P\left(K_n^+ \leq t\right) = \begin{cases} 0, & t \leq 0, \\ n! \prod_{i=1}^n \int_{\frac{i}{n}-t}^{u_{i+1}} du, & 0 < t < 1, \\ 1, & t \geq 1, \end{cases}$$

kde $du = du_1 \ldots du_n \quad (u_1 < \ldots u_n)$ a $u_{n+1} = 1$.

Důkaz.

$$P\left(K_n^+ \leq t\right) = P\left(\sup_{x\in\mathbb{R}} \left[\widehat{F_n}(x) - F_0(x)\right] \leq t\right) \overset{(2)}{=} P\left(\max_{1\leq i\leq n}\left(\frac{i}{n} - F_0(X_{(i)})\right) \leq t\right) =$$

$$\overset{(3)}{=} P\left(\bigcap_{i=1}^n \left(\frac{i}{n} - F_0(X_{(i)})\right) \leq t\right) = P\left(\bigcap_{i=1}^n \left(F_0(X_{(i)}) \geq \frac{i}{n} - t\right)\right) =$$

$$\overset{(5)}{=} P\left(\bigcap_{i=1}^n \left(U_{(i)} \geq \frac{i}{n} - t\right)\right) = \star$$

V rovnosti (2) jsme použili lemma 1. V rovnosti (3) jsme využili, že jestliže je maximum menší než t, pak to platí i pro průnik přes všechny $i = 1,2,\ldots,n$. Dále v rovnosti (5) používáme lemma 2 ($U_{(i)}$ je i-tá pořádková statistika náhodného výběru z rovnoměrného rozdělení na intervalu (0,1)). Z práce Omelka (2021)(Věta 2.12) známe sdruženou hustotu uspořádaného náhodného výběru $U_{(1)},\ldots,U_{(n)}$.

Pro $0 < u_1 < \cdots < u_n < 1$ a $t \in (0,1)$ platí

$$\star = n! \int_{1-t}^1 \int_{\frac{n-1}{n}-t}^1 \cdots \int_{\frac{2}{n}-t}^1 \int_{\frac{1}{n}-t}^1 1 du_1 du_2 \ldots du_{n-1} du_n =$$

$$= n! \int_{1-t}^{1} \int_{\frac{n-1}{n}-t}^{u_n} \cdots \int_{\frac{2}{n}-t}^{u_3} \int_{\frac{1}{n}-t}^{u_2} 1 du_1 du_2 \dots du_{n-1} du_n$$

Pro $t \geq 1$ platí $0 \geq 1-t \geq \frac{n-1}{n} - t \geq \cdots \geq \frac{1}{n} - t$, tedy integrujeme $\int_0^1 \int_0^1 \cdots \int_0^1 \int_0^1 1 du_1 du_2 \dots du_{n-1} du_n$, což je rovno 1.

Pro $t \leq 0$ platí $u_n > \cdots > u_1 \geq 1-t \geq 1$. Integruje se přes nulovou množinu, tedy výraz je roven 0.

Celkem jsme dostali požadovaný vzorec.

□

Poznámka. Pro K_n^- lze odvodit podobný vzorec, s obdobným důkazem.

Důsledek. Jestliže uvažujeme jednostranný statistický test s testovou statistikou K_n^+(respektive K_n^-) jako bylo uvažováno v sekci 1.1, pak za nulové hypotézy $H_0^{'}$ (respektive $H_0^{''}$) známe distribuční funkci testové statistiky a tedy můžeme vypočítat přesný kritický obor.

Výsledek získaný z tvrzení 3 je v praxi použitelný pro malé rozsahy náhodných výběrů. Pro velké n je výraz značně složitý, proto se používá asymptotické rozdělení K_n^+.

Věta 4. *Nechť F_0 je spojitá distribuční funkce náhodného výběru $X_1, \dots, X_n$ o rozsahu n a K_n^+ je jako v rovnosti 1.2, potom pro $t > 0$ platí:*

$$\lim_{n \to \infty} P\left(\sqrt{n} K_n^+ \leq t\right) = 1 - e^{-2t^2}.$$

Důkaz. Důkaz viz Smirnov (1944)

□

Nyní se zaměříme více na vlastnosti K_n definovaném v rovnosti (1.1) v úvodu kapitoly. Podobně jako v tvrzení 3 lze z lemmatu 1 odvodit přesné rozdělení náhodné veličiny K_n, neboť platí : $K_n = \max\{K_n^+, K_n^-\}$. My se však spíše zaměříme na její asymptotické vlastnosti.

Dále budeme uvažovat model, hypotézu a alternativu oboustranného testu, jako jsme již definovali dříve v sekci 1.1.

Tvrzení 5. 1. *Nechť platí hypotéza H_0, potom $K_n \xrightarrow{p} 0$.*

2. *Nechť platí alternativa H_1, potom $K_n \xrightarrow{p} \sup_{x \in \mathbb{R}} |F_X(x) - F_0(x)|$, kde F_X je skutečná distribuční funkce náhodného výběru.*

Důkaz.

1. Z předpokladu F_0 je spojitá distribuční funkce náhodného výběru $X_1, \dots, X_n$. Z vlastností empirické distribuční funkce (respektive z Glivenko-Cantelliho věty) plyne, že $K_n \xrightarrow{p} 0$.

2. Nechť platí H_1. Nejdříve provedeme vhodné odhady.

$$K_n = \sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F_0(x) \right| = \sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F_X(x) + F_X(x) - F_0(x) \right| \leq$$

$$\overset{(1)}{\leq} \sup_{x\in\mathbb{R}} \left|\widehat{F_n}(x) - F_X(x)\right| + \sup_{x\in\mathbb{R}} |F_X(x) - F_0(x)|$$

V nerovnosti (1) jsme využili trojúhelníkovou nerovnost ze základní vlastnosti suprema.

Obdobně lze odvodit následující odhad.

$$\sup_{x\in\mathbb{R}} |F_X(x) - F_0(x)| \leq K_n + \sup_{x\in\mathbb{R}} \left|\widehat{F_n}(x) - F_X(x)\right|$$

Zvolme $\varepsilon > 0$, označme si $\sup_{x\in\mathbb{R}} |F_X(x) - F_0(x)|$ jako S.

$$P\left(|K_n - S| > \varepsilon\right) = P\left((K_n - S) > \varepsilon\right) + P\left(S > (\varepsilon + K_n)\right) \leq$$

$$\overset{(2)}{\leq} P\left(\sup_{x\in\mathbb{R}} \left|\widehat{F_n}(x) - F_X(x)\right| > \varepsilon\right) + P\left(\sup_{x\in\mathbb{R}} \left|\widehat{F_n}(x) - F_X(x)\right| > \varepsilon\right)$$

V nerovnosti (2) jsme použili odvozené nerovnosti z první části důkazu. Protože

$$P\left(\sup_{x\in\mathbb{R}} \left|\widehat{F_n}(x) - F_X(x)\right| > \varepsilon\right) \overset{n\to\infty}{\to} 0,$$

což plyne opět z Glivenko-Cantelliho věty, tak

$$P\left(|K_n - S| > \varepsilon\right) \overset{n\to\infty}{\to} 0.$$

Z toho již plyne tvrzení.

□

Uvedeme si větu podobnou větě 4, která je stěžejní pro určení asymptotického kritického oboru pro oboustranný Kolmogorovův-Smirnovův test. Věta se bude hodit i pro odvození některých vlastností K_n.

Věta 6. *Nechť F_0 je spojitá distribuční funkce náhodného výběru $X_1, \ldots, X_n$ o rozsahu n, potom*

$$\sqrt{n}K_n \overset{d}{\to} Z,$$

kde Z je náhodná veličina s distribuční funkcí G definovanou:

$$G(y) = \begin{cases} 0, & y \leq 0, \\ 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}, & y > 0. \end{cases}$$

Důkaz. Důkaz viz Smirnov (1944)

□

Definice 3. *Nechť $Y_1, \ldots, Y_n$ je posloupnost náhodných veličin definovaných na jednom pravděpodobnostním prostoru. Řekneme, že $\{Y_i\}_{i=1}^{\infty}$ je omezená v pravděpodobnosti, jestliže platí*

$$\forall\varepsilon > 0 \quad \exists M \in \mathbb{R}, M > 0 \quad \exists n_0 \in \mathbb{N}, \quad \forall n \in \mathbb{N}, n \geq n_0 : P(|Y_n| > M) < \varepsilon.$$

Tvrzení 7. *Nechť platí H_0, potom je náhodná veličina $\sqrt{n}K_n$ omezená v pravděpodobnosti.*

Důkaz. Zvolme $\varepsilon \in \mathbb{R}, \quad \varepsilon > 0$. Potřebujeme nalézt $M > 0$ a n_0, tak aby platilo pro $\forall n \geq n_0$:

$$P(\left|\sqrt{n}K_n\right| > M) < \varepsilon.$$

Toto je ekvivalentní výrazu

$$P(\sqrt{n}K_n < M) > 1 - \varepsilon.$$

K_n nabývá skoro jistě nezáporných hodnot, proto můžeme vynechat absolutní hodnotu a vhodně upravit nerovnosti. Označme G distribuční funkci $\sqrt{n}K_n$ jako ve větě 6. Dále označme $G_n(y) := P(\sqrt{n}K_n < y)$, pro nějaké $y > 0$. Protože je řada $1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}$ pro $y > 0$ lokálně stejnoměrně konvergentní, tak je funkce $G(y)$ spojitá. Z věty 6 $G_n(y)$ konverguje k $G(y)$ pro $y > 0$. Máme tedy $\varepsilon > 0$, k němu nalezněme $M > 0$ takové, že

$$1 - \frac{\varepsilon}{2} < G(M). \tag{1.4}$$

To lze ze spojitosti a vlastnosti distribuční funkce ($\lim_{m\to\infty} G(m) = 1$). Dále nalezněme $n_0 \in \mathbb{N}$ takové, že $\forall n \geq n_0$ platí:

$$G(M) - \frac{\varepsilon}{2} < G_n(M) < G(M) + \frac{\varepsilon}{2}. \tag{1.5}$$

Toto opět lze, protože $G_n(M) \overset{n\to\infty}{\to} G(M)$. Dohromady z nerovností (1.4) a (1.5) získáváme:

$$P(\sqrt{n}K_n < M) = G_n(M) > G(M) - \frac{\varepsilon}{2} > 1 - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = 1 - \varepsilon.$$

Z této nerovnosti a z diskuze v úvodu důkazu plyne tvrzení.

□

1.3 Kritický obor oboustranného testu

Z sekce 1.1 již máme zformulovaný model, hypotézu a testovou statistiku oboustranného testu, avšak nám chybí přesně určený kritický obor. Zde vyjdeme z věty 6. Za nulové hypotézy má náhodná veličina $\sqrt{n}K_n$ asymptotické rozdělení dané distribuční funkcí G definovanou ve větě 6. Nechť $\alpha \in (0,1)$ je požadovaná hladina testu. Označme k_α jako α kvantil příslušný distribuční funkci G, pak test s testovou statistikou $\sqrt{n}K_n$ a kritickým oborem $(k_{1-\alpha},\infty)$ dodržuje asymptoticky hladinu α. Na základě vlastností odvozených v sekci 1.2 si odvodíme konzistenci tohoto testu.

Poznámka. Označme si

$\mathcal{F}_1 = \{F \in \mathcal{F};$ takové, že splňují $H_1\}$,

$C(\alpha)$ kritický obor testu příslušný hladině alfa,

$\mathbb{X}$ náhodný vektor sestavený z náhodného výběru $X_1 \dots X_n$.

Poznámka. V následujících dvou definicích uvažujeme obecný model, hypotézu, testovou statistiku. Dále v lemmatu 8 a větě 9 opět uvažujeme model, hypotézu a testovou statistiku definovanou na začátku sekce 1.1.

Definice 4. *Nechť* $X_1, \ldots, X_n$ *je náhodný výběr z modelu* $\mathcal{F}$*. Mějme* $\alpha \in (0,1)$*, hypotézu* H_0*, alternativu* H_1*, testovou statistiku* $S_n(\mathbb{X})$ *a kritický obor* $C(\alpha)$*. Pro* $F \in \mathcal{F}$ *definujeme silofunkci testu jako*

$$\beta_n(F) := \int \mathbf{1}\{S_n(\mathbf{x}) \in C(\alpha)\} dF_{(x_1)} \ldots dF_{(x_n)}.$$

Definice 5. *Nechť* $X_1, \ldots, X_n$ *je náhodný výběr z modelu* $\mathcal{F}$*.*

Dále nechť $S_n(\mathbb{X})$ *je testová statistika a* $C(\alpha)$ *je kritický obor příslušný hypotéze a alternativě. Řekneme, že test* $(S_n(\mathbb{X}), C(\alpha))$ *je konzistentní, jestliže platí:*

$$\forall F \in \mathcal{F}_1, \quad \lim_{n \to \infty} \beta_n(F) = 1.$$

Lemma 8. *Nechť* $X_1, \ldots, X_n$ *je náhodný výběr z modelu* $\mathcal{F}$ *a platí* H_1*, pak* $\sqrt{n} K_n \xrightarrow{p} \infty$

Důkaz. Za platnosti H_1 $K_n \xrightarrow{p} \sup_{x \in \mathbb{R}} |F_X(x) - F_0(x)| := K > 0$, kde F_X je skutečná distribuční funkce výběru a F_0 je funkce z hypotézy. Z toho plyne, že $\forall \varepsilon \in \mathbb{R}, \quad \varepsilon > 0$ platí:

$$P(K_n - K > \varepsilon) + P(K_n - K < -\varepsilon) = P(|K_n - K| > \varepsilon) \overset{n \to \infty}{\rightarrow} 0 \tag{1.6}$$

Oba výrazy na levé straně jsou nezáporné a shora omezené, tedy oba definované. Z výrazu (1.6) plyne, že $\forall \varepsilon > 0 : P(K_n - K < -\varepsilon) \overset{n \to \infty}{\rightarrow} 0$.

My chceme dokázat, že $P(|\sqrt{n} K_n| < C) \overset{n \to \infty}{\rightarrow} 0$. Nejprve si výraz upravíme, buď $C > 0$.

$$P\left(\left|\sqrt{n} K_n\right| < C\right) \overset{(1)}{=} P\left(\sqrt{n} K_n < C\right) = P\left(K_n < \frac{C}{\sqrt{n}}\right) =$$

$$P\left(K_n - K < \frac{C}{\sqrt{n}} - K\right) = \star$$

V rovnosti (1) jsme využili nezápornost K_n a v ostatních rovnostech ekvivalentní úpravy. Víme, že $K > 0$ a zřejmě

$$\frac{C}{\sqrt{n}} \overset{n \to \infty}{\rightarrow} 0.$$

Zvolme si $\delta \in \mathbb{R}, \delta > 0$, tak aby $\delta < K$. Dále pro δ nalezněme $n_0 \in \mathbb{N}$ takové, že $\forall n \geq n_0 : \frac{C}{\sqrt{n}} < \delta$. Zřejmě $\delta - K < 0$, tedy z úvodu důkazu stačí zvolit ε tak, aby $\delta - K < -\varepsilon$. Z toho dostáváme $\forall n \geq n_0$:

$$\star \leq P(K_n - K < \delta - K) \overset{n \to \infty}{\rightarrow} 0.$$

Z toho již plyne důkaz.

□

Věta 9. *Výše definovaný asymptotický test* $(\sqrt{n}K_n, C(\alpha))$ *je konzistentní.*

Důkaz. Z lemmatu 8 již víme, že $\sqrt{n}K_n \xrightarrow{p} \infty$ za platnosti alternativy. Toto je ekvivalentní $\lim_{n\to\infty} P\left(\sqrt{n}K_n > C\right) = 1$, pro jakékoliv $C > 0$. Existuje $C > 0$ takové, že $k_{1-\alpha} < C$. Celkem dostáváme:

$$1 \geq P\left(\sqrt{n}K_n > k_{1-\alpha}\right) \geq P\left(\sqrt{n}K_n > C\right) \overset{n\to\infty}{\to} 1.$$

Z toho již plyne, že $P\left(\sqrt{n}K_n > k_{1-\alpha}\right) \overset{n\to\infty}{\to} 1$, tedy test je konzistentní.

$\square$

V kapitole 1 jsme se podrobně seznámili s Kolmogorovým-Smirnovým testem. Ukázali jsme chování testové statistiky za určitých podmínek a zároveň konzistenci asymptotického testu. Některé získané výsledky využijeme pro odvozování vlastností Lillieforsova testu.

2. Test normality

V této kapitole si na úvod připomeneme základní definici a vlastnosti normálního rozdělení, protože tyto poznatky budeme používat v sekci 2.2. Dále z Kolmogorova-Smirnova testu odvodíme test normality a zjistíme o tomto testu dílčí vlastnosti.

2.1 Shrnutí poznatků o normálním rozdělení

V Kolmogorově-Smirnově testu se porovnávají distribuční funkce, proto si uvedeme definici distribuční funkce normálního rozdělení a její značení.

Definice 6. *Distribuční funkce normálního rozdělení s parametry* $\mu \in \mathbb{R}$ *a* $\sigma^2 \in \mathbb{R}, \quad \sigma^2 > 0$ *(značíme* $N(\mu,\sigma^2)$*) je definována:*

$$\phi(t;\mu,\sigma^2) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Normované normální rozdělení je speciální případ pro $\mu = 0$ *a* $\sigma = 1$ *a distribuční funkci tohoto rozdělení značíme* Φ.

Mezi distribuční funkcí normálního rozdělení a normovaného normálního rozdělení platí jednoduchý vztah.

Lemma 10. *Buď* ϕ *a* Φ *jako v definici 6 pro* $\mu \in \mathbb{R}$ *a* $\sigma^2 \in \mathbb{R}, \quad \sigma^2 > 0$*, pak*

$$\phi(x;\mu,\sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

.

Důkaz. Důkaz plyne jednoduchou substitucí v integrálu $\frac{x-\mu}{\sigma} = z$, $\frac{1}{\sigma}dx = dz$ □

2.2 Lillieforsův test

Mějme náhodný výběr $X_1,\ldots,X_n$ z nějakého rozdělení. Nyní nechceme vědět z jakého rozdělení výběr pochází, ale jestli rozdělení je normální. Nabízí se možnost použít klasický Kolmogorovův-Smirnovův test, kde testovaná distribuční funkce bude distribuční funkce normálního rozdělení. Ta je spojitá, a je jednoznačně určená dvěma parametry, střední hodnotou a rozptylem (respektive směrodatnou odchylkou). Tyto parametry lze jednoduše odhadnout výběrovým průměrem

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a výběrovým rozptylem

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Do distribuční funkce normálního rozdělení dosadíme místo parametrů tyto odhady a dostáváme hlavní ideu Lillieforsova testu.

V odvození vlastností nelze jednoduše převzít všechny již zjištěné informace o Kolmogorově-Smirnově testu, protože v testové statistice nahrazujeme neznámé parametry odhady. Situace tedy není tak jednoduchá, jak by se mohlo na první pohled zdát.

Model pro tento test:

$$\mathcal{F} = \{\text{Všechna spojitá rozdělení}\}.$$

Nechť F_X je distribuční funkce rozdělení, z kterého pochází náhodný výběr.

Hypotéza a alternativa:

$$H_0 : \exists \mu \in \mathbb{R} \text{ a } \sigma \in \mathbb{R}, \sigma > 0, \quad \forall t \in \mathbb{R} : \quad F_X(t) = \phi(t; \mu, \sigma),$$

$$H_1 : \forall \mu \in \mathbb{R} \text{ a } \sigma \in \mathbb{R}, \sigma > 0, \quad \exists t \in \mathbb{R} : \quad F_X(t) \neq \phi(t; \mu, \sigma).$$

Testová statistika:

$$T_n := \sup_{t \in \mathbb{R}} \left| \widehat{F_n}(t) - \phi(t; \overline{X}_n, S_n^2) \right| \tag{2.1}$$

Poznámka. Jinými slovy hypotéza je, že náhodný výběr pochází z nějakého normálního rozdělení. Nadále v této sekci budeme značit T_n jako výraz definovaný v rovnosti (2.1), stejně H_0, H_1 a $\mathcal{F}$.

Pro tento případ nelze použít kritický obor, který byl odvozen v sekci 1.3. Neznámé parametry distribuční funkce jsou odhadnuté výběrovým průměrem a rozptylem, a proto nelze využít věty 6.

Pro sestavení kritického oboru by se nám hodilo znát, nebo alespoň vhodně aproximovat rozdělení testové statistiky za nulové hypotézy. Aproximace může být provedena na základě metody Monte-Carlo simulace. Lilliefors (1967) ve své práci ukázal s odkazem na článek David a Johnson (1948), že za nulové hypotézy H_0 rozdělení testové statistiky T_n nezávisí na neznámých parametrech. Na základě tohoto výsledku lze sestavit univerzální tabulku kritický hodnot pro tento test. My zde ukážeme, jak to platí pro případ, kdy známe rozptyl σ^2.

Uvažujme tedy *Model*:

$$\mathcal{F}' = \{\text{Všechna spojitá rozdělení s rozptylem } \sigma^2 \in \mathbb{R}, \quad \sigma^2 > 0\}.$$

Nechť F_X je distribuční funkce rozdělení, z kterého pochází náhodný výběr.

Hypotéza a alternativa:

$$H_0' : \exists \mu \in \mathbb{R}, \quad \forall t \in \mathbb{R} : \quad F_X(t) = \phi(t; \mu, \sigma^2),$$

$$H_1' : \forall \mu \in \mathbb{R}, \quad \exists t \in \mathbb{R} : \quad F_X(t) \neq \phi(t; \mu, \sigma^2).$$

Testová statistika:

$$D_n := \sup_{t \in \mathbb{R}} \left| \widehat{F_n}(t) - \phi(t; \overline{X}_n, \sigma^2) \right|. \tag{2.2}$$

Poznámka. Jinými slovy hypotéza je, že náhodný výběr pochází z nějakého normálního rozdělení s rozptylem σ^2. Nadále v této sekci 2.2 budeme značit D_n jako výraz definovaný v rovnosti (2.2), stejně H_0', H_1' a $\mathcal{F}'$.

Analogicky jako v sekci 1.1 definujeme

$$D_n^+ = \sup_{t\in\mathbb{R}} \left(\widehat{F_n}(t) - \phi(t;\overline{X}_n,\sigma^2)\right),$$

$$D_n^- = \sup_{t\in\mathbb{R}} \left(\phi(t;\overline{X}_n,\sigma^2) - \widehat{F_n}(t)\right).$$

D_n lze vypočítat jako $\max\{D_n^+; D_n^-\}$.

Tedy stačí studovat pouze vlastnosti D_n^+ a D_n^-. Následující lemma je pouze konkrétní použití lemmat 1 a 10.

Lemma 11. *Nechť $X_1,\ldots,X_n$ je náhodný výběr z jednorozměrného spojitého rozdělení, potom platí*

$$D_n^+ = \max_{1\le i\le n}\left(\frac{i}{n} - \phi(X_{(i)};\overline{X}_n,\sigma^2)\right) = \max_{1\le i\le n}\left(\frac{i}{n} - \Phi\left(\frac{X_{(i)}-\overline{X}_n}{\sigma}\right)\right),$$

$$D_n^- = \max_{1\le i\le n}\left(\phi(X_{(i)};\overline{X}_n,\sigma^2) - \frac{i-1}{n}\right) = \max_{1\le i\le n}\left(\Phi\left(\frac{X_{(i)}-\overline{X}_n}{\sigma}\right) - \frac{i-1}{n}\right).$$

Důkaz. Důkaz plyne z lemmatu 1, kde volíme $F_0(t) = \psi(t;\overline{X}_n,\sigma^2), t\in\mathbb{R}$. Dále se využívá lemma 10 pro převedení na distribuční funkci normovaného normálního rozdělení.

□

Pro přehlednost budeme nadále používat značení:

$$Z_i := \frac{X_i - \overline{X}_n}{\sigma},$$

$$Z_{(i)} := \frac{X_{(i)} - \overline{X}_n}{\sigma}.$$

Pro tento případ nelze použít kritický obor, který byl odvozený v sekci 1.2. Neznámý parametr střední hodnoty je odhadnutý výběrovým průměrem, tedy nelze použít větu 6. Pomocí následujícího lemmatu dokážeme požadovanou vlastnost D_n.

Lemma 12. *Nechť $X_1\ldots X_n$ je náhodný výběr s rozdělením z modelu $\mathcal{F}'$ a platí nulová hypotéza H_0', potom vektor $(Z_1,Z_2,\ldots,Z_{n-1},Z_n)^T$ má n-rozměrné normální rozdělení s nulovým vektorem střední hodnoty a varianční maticí*

$$\boldsymbol{I_n} - \frac{1}{n}\boldsymbol{1_n}\boldsymbol{1_n}^T,$$

kde $\boldsymbol{I_n}$ je n-rozměrná jednotková matice a $\boldsymbol{1_n}$ je n-rozměrný vektor $(1,\ldots,1)^T$.

Důkaz. Za nulové hypotézy má X_i normální rozdělení se střední hodnotou μ a rozptylem σ^2 pro $i\in\{1,\ldots n\}$. X_i pro $i\in\{1,\ldots,n\}$ jsou vzájemně nezávislé náhodné veličiny, a tedy vektor $(X_1,\ldots,X_n)^T$ má n-rozměrné normální rozdělení s parametrem střední hodnoty $\mu\boldsymbol{1_n}$ a varianční matici $\sigma^2\boldsymbol{I_n}$. Jednoduchou lineární transformací lze získat vektor $(X_1,\ldots,X_n,\overline{X}_n)^T$. Matice B dimenze $(n+1) \times (n)$ určující lineární transformaci vypadá následovně:

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}.$$

Z toho plyne, že vektor $(X_1,\ldots,X_n,\overline{X}_n)^T$ má $(n+1)$-rozměrné normální rozdělení s parametrem střední hodnoty $\mu\mathbf{1_{n+1}}$ a varianční maticí Σ.

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 & \frac{\sigma^2}{n} \\ 0 & \sigma^2 & \ldots & 0 & \frac{\sigma^2}{n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \sigma^2 & \frac{\sigma^2}{n} \\ \frac{\sigma^2}{n} & \frac{\sigma^2}{n} & \ldots & \frac{\sigma^2}{n} & \frac{\sigma^2}{n} \end{pmatrix}$$

Při výpočtu jsme využili vlastnosti mnohorozměrného normálního rozdělení, a dále vztah $\Sigma = \sigma^2 B\boldsymbol{I_n}B^T$ a $B\mu\mathbf{1_n} = \mu\mathbf{1_{n+1}}$.

Vektor $\boldsymbol{Z} := (Z_1,Z_2,\ldots,Z_{n-1},Z_n)^T$ lze opět získat jednoduchou lineární transformací určenou maticí A a vektoru $\boldsymbol{X} := (X_1,\ldots,X_n,\overline{X}_n)^T$.

$$A := \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ 0 & 1 & \ldots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -1 \end{pmatrix}$$

$\boldsymbol{Z} = A\boldsymbol{X}$, tedy $\boldsymbol{Z}$ má n-rozměrné normální rozdělení, což opět plyne z vlastností mnohorozměrného normálního rozdělení.

Vektor $\boldsymbol{Z}$ má parametr střední hodnoty $A\mu\mathbf{1_{n+1}}$, který po vynásobení vychází roven nulovému vektoru. Varianční matice vektoru $\boldsymbol{Z}$ jest $A\Sigma A^T$, po výpočtu vychází následovně:

$$A\Sigma A^T = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \ldots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \ldots & 1-\frac{1}{n} \end{pmatrix} = \boldsymbol{I_n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n^T}.$$

□

Nyní pomocí lemmat 11 a 8 získáme požadovanou "univerzálnost"testové statistiky D_n.

Věta 13. *Rozdělení testové statistiky D_n za platnosti hypotézy $H_0^{'}$ nezávisí na neznámém parametru $\mu \in \mathbb{R}$, ani na známém parametru $\sigma^2 \in \mathbb{R}, \sigma^2 > 0$.*

Důkaz. Z lemmatu 11 a z faktu $D_n = \max\{D_n^+, D_n^-\}$ plyne, že rozdělení D_n bude odvozeno ze sdruženého rozdělení $\left(Z_{(1)},\ldots,Z_{(n)}\right)^T$, protože D_n je funkcí

$\left(Z_{(1)}, \ldots, Z_{(n)}\right)^T$. Dále vektor $\left(Z_{(1)}, \ldots, Z_{(n)}\right)^T$ vznikne seřazením náhodných veličin Z_i. Rozdělení vektoru $(Z_1, \ldots, Z_n)^T$ za platnosti nulové hypotézy nezávisí na parametrech μ a σ^2, což plyne z lemmatu 12. Celkem tedy dostáváme, že rozdělení D_n za nulové hypotézy nezávisí na μ a σ^2.

$\square$

Jak bylo již výše diskutováno, je možné sestavit tabulku kritických hodnot pro tento test na základě Monte-Carlo simulace. My provedeme simulaci pro obecný případ, kdy rozptyl je neznámý.

3. Simulace

3.1 Aproximace kritických hodnot

V této kapitole budeme uvažovat test definovaný v sekci 2.2. Ve výpočetním prostředí R-studio jsem implementoval program provádějící Monte-Carlo simulaci za účelem aproximace rozdělení testové statistiky T_n za nulové hypotézy. Kritické hodnoty budeme uvažovat pro rozsahy výběru $n \in \{30, 40, 100, 400\}$. Tento test bude zřejmě zamítat pro velké hodnoty testové statistiky a kritický obor bude tvaru $C(\alpha) = (q_{1-\alpha}, \infty)$, kde α bude hladina testu a $q_{1-\alpha}$ bude $(1-\alpha)-$kvantil rozdělení T_n za nulové hypotézy. V našem případě budeme uvažovat hladiny $\alpha \in \{0.01, 0.05, 0.1\}$.

V simulaci jsem uvažoval 100000 opakování náhodného výběru o rozsahu $30, 40, 100, 400$. Stejnou simulaci s počtem opakování až 1000000 provedl Leland Wilkinson a Gerard E. Dallal v roce 1986. Mé výsledky kritických hodnot (tabulka 3.1) se shodují s publikovanými v článku Dallal a Wilkinson (1986). Respektive maximální absolutní rozdíl mezi kritickými hodnotami napočítané mojí simulací a simulací z článku Dallal a Wilkinson (1986) je 0.001 a to pouze ve čtyřech případech, ve zbytku se v rámci zaokrouhlovací chyby shodují.

3.2 Použití chybného kritického oboru

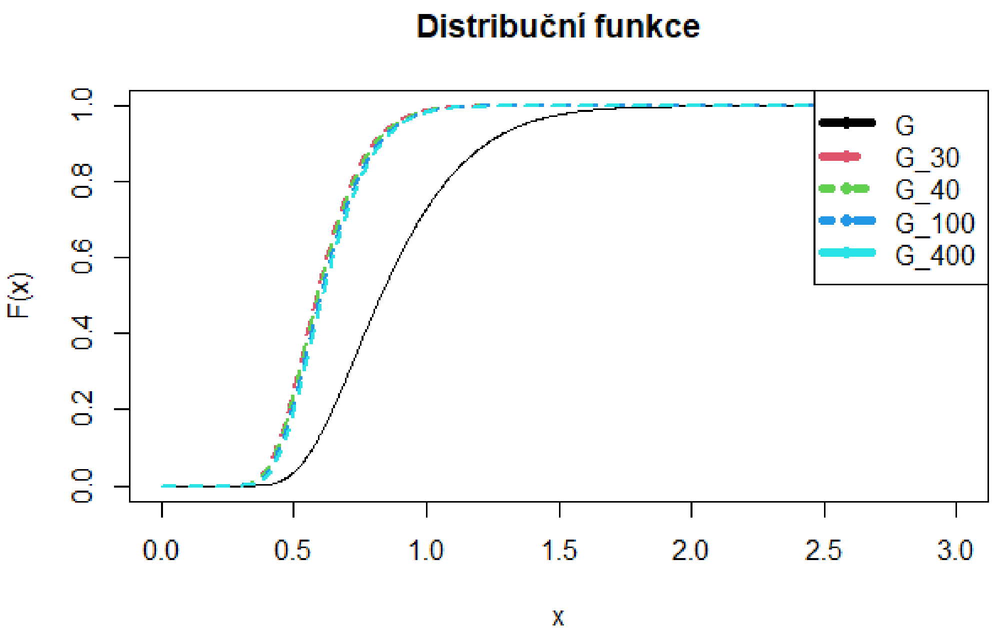
Z konstrukce Lillieforsova testu normality ze sekce 2.2 si můžeme myslet, že lze využít výsledek věty 6. Konkrétněji vzít si testovou statistiku T_n, vynásobit $\sqrt{n}$ a použít kritický obor odvozený v sekci 1.3. Dle článku Lilliefors (1967) se ukazuje, že by tento test měl být konzervativní. Ve shodě s tímto faktem je i výsledek simulace uvedený v obrázku 3.1. V obrázku uvažujeme distribuční funkci G definovanou ve větě 6 a porovnáváme je s distribučními funkcemi $G_{30}, G_{40}, G_{100}, G_{400}$. G_n je empirická distribuční funkce spočítaná ze simulace pro statistiku T_n vynásobenou $\sqrt{n}$ pro $n \in \{30, 40, 100, 400\}$. Z obrázku 3.1 je jasně vidět, že $G_n \geq G$, tedy z tohoto plyne již diskutovaná konzervativnost.

Pro test o rozsahu výběru 400 a hladině významnosti 0.05 jsem provedl detailnější simulaci, která má ukázat, jaká zhruba je hladina testu. Tedy použijeme-li kritický obor odvozený v sekci 1.3 a testovou statistiku $\sqrt{n}T_n$, dostáváme test se skutečnou hladinou zhruba okolo 0.00017, což je o dost nižší hladina, než která byla požadována.

Dále je na obrázku 3.1 vidět, že se funkce G_n značně překrývají. Zdá se, že G_n konvergují k nějaké spojité distribuční funkci. Můžeme se tedy domnívat, že $\sqrt{n}T_n$ konverguje v distribuci k nějaké neznámé veličině se spojitou distribuční funkcí.

rozsah výběru	hladina významnosti		
n	0.1	0.05	0.01
30	0.146	0.159	0.184
40	0.127	0.138	0.161
100	0.082	0.089	0.104
400	0.041	0.045	0.052

Tabulka 3.1: Tabulka kritických hodnot



Obrázek 3.1: Porovnání distribučních funkcí

Závěr

Práce se věnovala Kolmogorově-Smirnově testu, z kterého vychází Lillieforsův test normality. Konkrétněji přebírá jeho testovou statistiku.

Zadefinovali jsme si jednostranný Kolmogorův-Smirnovův test a dokázali tvar distribuční funkce testové statistiky za nulové hypotézy. Pomocí této statistiky by se odvodil kritický obor tohoto testu. Nicméně distribuční funkce má značně složitý tvar, proto jsme uvedli asymptotické rozdělení, z kterého se konstruuje kritický obor pro velké rozsahy výběru.

Následně jsme se věnovali klasickému oboustrannému Kolmogorově-Smirnově testu. Uvedli jsme základní vlastnosti jeho testové statistiky a podrobně popsali tvar kritického oboru. V návaznosti jsme ukázali, že tento test je konzistentní pro jakoukoliv alternativu.

Podrobně jsme popsali model, hypotézu a testovou statistiku pro Lillieforsův test. Pro verzi testu se známým rozptylem jsme dokázali, že rozdělení testové statistiky nezávisí na neznámém parametru střední hodnoty, ani na známém rozptylu. Toto nám dalo možnost aproximovat kritické hodnoty pro tento test. Alternativní vlastnost platí i pro obecný test normality.

V závěru jsme ukázali aproximace kritických hodnot pro Lillieforsův test. Tyto aproximace se shodovali s výsledky publikovanými jinými autory. Dále jsme uvažovali použití kritického oboru odvozeného v první kapitole pro test normality. Na základě výsledků simulace se ukázalo, že by tento test normality byl extrémně konzervativní.

Seznam použité literatury

DALLAL, G. E. a WILKINSON, L. (1986). An analytic approximation to the distribution of lilliefors's test statistic for normality. *The American Statistician*, **40**(4), 294–296.

DAVID, F. N. a JOHNSON, N. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, **35**(1/2), 182–190.

LILLIEFORS, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, **62**(318), 399–402.

OMELKA, M. (2021). NMSA331 poznámky k přednášce. `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf`. Accessed: 2021-02-27.

SMIRNOV, N. V. (1944). Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, **28**(10), 179–206.

Seznam obrázků

Seznam tabulek

A. Přílohy

Zdrojový kód programu R, který provádí simulační studii ze sekce 3.1

- simulace.R