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

DIPLOMOVÁ PRÁCE

Normální rozdělení a jak testovat normalitu

Vedoucí diplomové práce:
Mgr. Ondřej Vencálek Ph.D.
Rok odevzdání: 2013

Vypracoval:
Bc. Michal Trubač
AME, II. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana Mgr. Ondřeje Vencálka Ph.D. s použitím uvedené literatury.

V Olomouci dne 3. dubna 2013

Poděkování

Na tomto místě bych chtěl poděkovat především svému vedoucímu diplomové práce panu Mgr. Ondřeji Vencálkovi Ph.D., že měl se mnou dostatek trpělivosti, aby mi pomohl dovést tuto práci ke zdárnému konci. Také bych rád poděkoval své rodině a přátelům, kteří mě po celou dobu studia podporovali.

Obsah

Úvod

Normální rozdělení má mezi všemi pravděpodobnostními rozděleními zvláštní, řekli bychom privilegované postavení. Při počtu pravděpodobnosti se na normální rozdělení díváme jako na jeden ze základních kamenů. Důležitý význam tohoto rozdělení spočívá v tom, že za určitých podmínek velmi dobře aproximuje mnoho dalších spojitých i diskrétních rozdělení pravděpodobnosti. S normálním rozdělením se můžeme setkat v mnoha vědních oborech od matematiky, fyziky, statistiky, astronomie až po psychologii, medicínu či ekonomii.

V praxi je mnoho statistických metod založeno na předpokladu normality. Tento je však nutno pro konkrétní data vždy ověřit. Předpoklad normality je poměrně častý a to z důvodu příjemných vlastností u odvozených řešení, se kterými se potom lépe pracuje. Je však dobré si uvědomit, že v mnoha oborech nemusí být jednoduché předpokladu normality vyhovět. Kromě základních grafických nástrojů nabízí různé statistické softwary k ověření předpokladu normality celou řadu testů.

Cílem této práce je pochopit alespoň některé myšlenky, ze kterých tyto testy vycházejí. Nejde nám přitom o to, aby počet námi zmíněných testů byl co největší – nechceme sestavovat přehled všech používaných testů. Jde nám o pochopení podstaty těchto testů. Součástí práce bude i srovnání různých testů normality. Především budeme určovat pravděpodobnost chyby prvního a druhého druhu u jednotlivých testů v závislosti na rozsahu výběru. Na otázku, který test je pro ověření normality „nejlepší“, existuje několik odpovědí. Tyto se však vzájemně různí. My bychom se touto otázkou zabývali pokud možno hlouběji. Diplomová práce je vytvořena za pomoci typografického systému pro sázení dokumentů TEX. Pro matematické a statistické výpočty je využíváno jazyka R.

Práce je rozdělena na dvě hlavní části. Úvodní část se zabývá připomenutím základních pojmů a zavedením grafických metod. Pro lepší vysvětlení metod uvádíme konkrétní řešené příklady. Druhá část textu je stěžejní částí diplomové práce. Definujeme zde některé z testů normality. Pomocí simulací odhadujeme pravděpodobnost chyby prvního druhu a druhého druhu u Shapirova-Wilkova,

Jarqueova-Berova, Andersonova-Darlingova, Cramér-von Misesova, Lillieforsova a Shapirova-Franciova testu normality v závislosti na velikosti výběru. Na základě odhadnutých pravděpodobností prvního a druhého druhu porovnáváme „kvalitu“ testů.

1. Základní pojmy

1.1. Značení

Jelikož budeme v diplomové práci uvádět výpočty pomocí softwaru R, dovolujeme si využívat desetinné tečky místo desetinné čárky. Důvodem tohoto značení je snaha o vytvoření stejnorodého textu a zamezení zmatení čtenáře.

1.2. Náhodný výběr

Následující text vychází z [1].

Nechť $X_1, ..., X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin s rozdělením Q. Pak říkáme, že $X_1, ..., X_n$ je *náhodný výběr* z rozdělení Q. Číslo n se nazývá *rozsah výběru*. Položme

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad M_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Veličina $\bar{X}$ se nazývá *výběrový průměr* a veličinu M_2 budeme nazývat *výběrový rozptyl*. Hodnotě $\sqrt{M_2}$ budeme říkat *výběrová směrodatná odchylka*. Veličina S^2 je definována jen pro $n \geq 2$ a často se používá místo M_2. Mnozí autoři právě S^2 nazývají výběrový rozptyl. Dále zavedeme

$$M_3 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^3, \qquad M_4 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^4, \qquad \alpha_3 = \frac{M_3}{M_2^{3/2}}, \qquad \alpha_4 = \frac{M_4}{M_2^2}.$$

Veličina α_3 je *výběrová šikmost* a α_4 je *výběrová špičatost*.

Věta 1.1 *Nechť $X_1, ..., X_n$ je náhodný výběr z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2. Pak platí $E\bar{X} = \mu$, $var\bar{X} = \sigma^2/n$ a v případě $n \geq 2$ dále platí $ES^2 = \sigma^2$.*

Důkaz: První dvě tvrzení se dokáží prostou úpravou.

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} n E X_i = \mu$$

$$var\bar{X} = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} var\left(\sum_{i=1}^{n} \mathbf{X}_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Poslední tvrzení se dokáže tak, že na identitu

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} \left(X_i - \mu\right)^2 - n\left(\bar{X} - \mu\right)$$

aplikujeme operátor střední hodnoty E.

Vzhledem k vlastnosti $E\bar{X} = \mu$ se $\bar{X}$ nazývá *nestranný odhad* parametru μ. Analogicky vidíme, že S^2 je nestranný odhad parametru σ^2. Právě kvůli této vlastnosti je veličina S^2 oblíbena jako odhad rozptylu. Její výpočet se někdy provádí podle vzorce

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

1.3. Normální rozdělení

Následující text vychází z [1]-[3].

1.3.1. Normální rozdělení - trocha historie

Normální rozdělení bývá nazýváno „*Gaussovo rozdělení*“ nebo někdy též „*Laplaceovo-Gaussovo rozdělení*“ poté, co jej tito průkopníci objevili na přelomu 18. a 19. století. Ve skutečnosti bylo normální rozdělení objeveno mnohem dříve. Roku 1733 napsal Abraham de Moivre unikátní brožuru nazvanou *Approximatio ad summam terminorum binomii (a+b)^n in seriem expansi*, ve které uvedl, že normální rozdělení je odvozeno jako vhodná aproximace binomického rozdělení.

Toto tvrzení je dokázáno v následujícím zdroji [5]. Tato brožura byla napsána nejprve pouze v latině, ale v roce 1738 byla doplněna a přeložena do angličtiny. Jsou známy pouze dva výtisky této brožury.

Normální rozdělení bývá někdy nazýváno „*Zákon chyb*“. V souvislosti s tímto rozdělením bývají často zmiňovány náhodné chyby, způsobené velkým počtem neznámých a vzájemně nezávislých příčin. V astronomii se toto rozdělení stalo základním stavebním kamenem statistické práce, jelikož právě astronomická pozorování bývají ovlivněna chybami měření. Důvodem těchto chyb bývá měřící přístroj či vnější vlivy (teplota, vítr, rozdílná hustota zemské atmosféry, osobní zkušenost/zaujatost pozorovatele atd.). Samotný název byl obecně přijat až více než sto let po Gaussově smrti.

1.3.2. Normální rozdělení - obecně

Nechť $\mu \in \mathbb{R}$ a $\sigma > 0$ jsou dané konstanty (parametry). *Normální rozdělení* je určeno hustotou (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \forall x \in R^1$$

a označuje se symbolem $N(\mu, \sigma^2)$. Tvar hustoty závisí na parametru σ^2.

Výpočtem se dostane $EX = \mu$ a

$$E(X-EX)^{2k-1} = 0, \qquad E(X-EX)^{2k} = \frac{(2k)!}{k!}\frac{\sigma^{2k}}{2^k} \qquad pro \qquad k = 1, 2, \ldots.$$

Speciálně tedy máme $E(X - EX)^2 = \sigma^2$, dostaneme šikmost $\alpha_3 = 0$ a špičatost $\alpha_4 = 3$.

Normální rozdělení s obecnými parametry μ, σ má distribuční funkci

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy, \qquad -\infty < x < \infty.$$

Velmi často se můžeme setkat se speciálním případem normálního rozdělení a to s rozdělením $N(0,1)$. Toto rozdělení nazýváme *normované* či *standardizované* normální rozdělení. V takovémto případě má hustota náhodné veličiny U následující tvar:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{u^2}{2}\right\}, \qquad -\infty < u < \infty.$$

Je zřejmé, že

$$\varphi(u) = \varphi(-u), \qquad -\infty < u < \infty.$$

Hustota $\varphi(u)$ je sudá funkce. Má-li náhodná veličina X rozdělení $N(\mu, \sigma^2)$, pak náhodná veličina

$$U = \frac{X - \mu}{\sigma}$$

má rozdělení $N(0,1)$. Standardizované normální rozdělení má následující distribuční funkci

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{-\frac{t^2}{2}\right\} dt, \qquad -\infty < u < \infty.$$

Hodnoty této funkce jsou tabelovány, neboť integrál ve vztahu pro $\Phi(u)$ obecně neumíme vyjádřit .

1.3.3. Normální rozdělení - jedno z nejdůležitějších rozdělení

Důležitost normálního rozdělení plyne nejen z poměrně častého předpokladu normality při použití různých metod jako ANOVA či t-test, ale také z možnosti aproximovat pomocí normálního rozdělení řadu jiných spojitých i diskrétních pravděpodobnostních rozdělení. Normální rozdělení aproximuje nejen v praxi rozšířené binomické rozdělení, ale také Poissonovo a hypergeometrické rozdělení.

Mnoho statistických analýz předpokládá normalitu, ne však u jednotlivých dat, ale u jejich aritmetických průměrů nebo součtů na základě obecné teorie známé jako *centrální limitní věta*. Jeden z výkladů této teorie může být následující:

Nechť $X_1, ..., X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s konečnou střední hodnotou μ a s konečnou nenulovou směrodatnou odchylkou $\sigma > 0$. Pak distribuční funkce součtů

$$Y = \sum_{i=1}^{n} X_i,$$

má pro velké n přibližně normální rozdělení se střední hodnotou $n\mu$ a směrodatnou odchylkou $\sigma\sqrt{n}$.

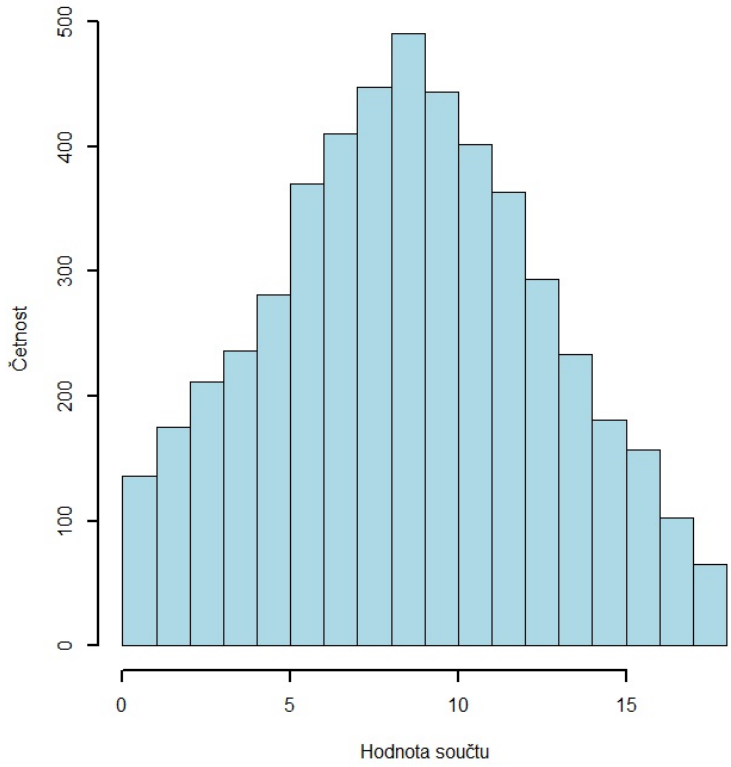
$$Y = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

má pro velké n přibližně normální rozdělení se střední hodnotou μ a směrodatnou odchylkou $\frac{\sigma}{\sqrt{n}}$.

Příklad 1: Představme si kotoučový zámek, na kterém je n kotoučů. Počet kotoučů může být libovolný. Na každém z kotoučů lze nastavit hodnoty od 0 do 9, přičemž hodnoty na jednotlivých kotoučích generujeme pomocí generátoru náhodných čísel statistického softwaru R. Hodnoty na různých kotoučích volíme nezávisle. Pro libovolné dva kotouče platí, že hodnota na jednom z nich nic nevypovídá o hodnotě na druhém. Pro danou n-tici hodnot spočteme jejich součet. S pomocí softwaru R vytvoříme 5000 fiktivních kotoučových zámků požadovaných vlastností a otestujeme normalitu u rozdělení součtů pomocí Shapirova-Wilkova[1] testu normality. Tuto simulaci provedeme 1000 krát a sledujeme počet zamítnutí normality při $\alpha = 0.05$.

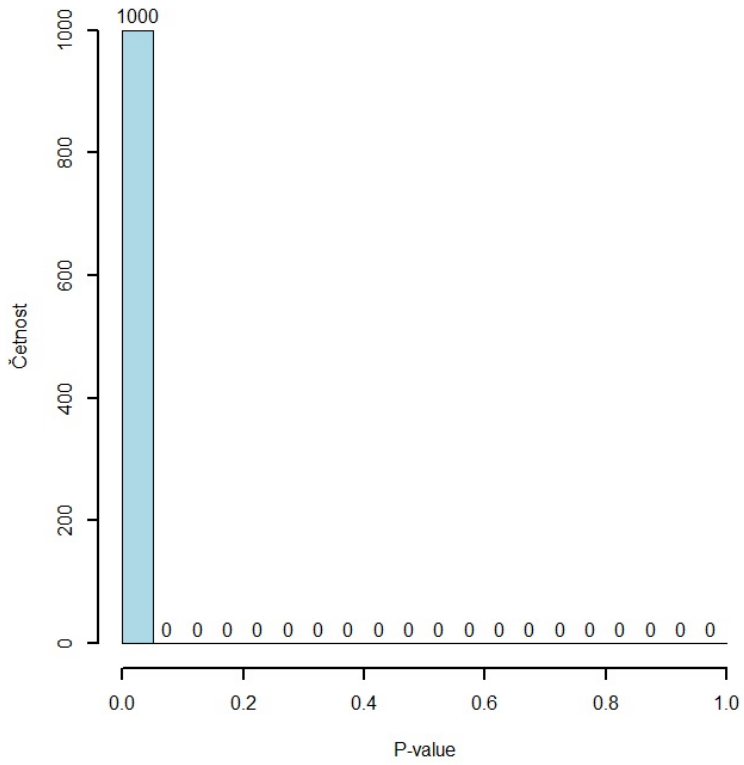
[1] Bližší informace o Shapirovu-Wilkovu testu jsou uvedeny v kapitole 3.

Krok 1.1. Nejprve volíme dva kotouče (tj. $n = 2$). Provedeme simulaci popsanou v Příkladě 1. Abychom si udělali lepší představu o tvaru rozdělení součtů, vykreslíme si tyto součty do histogramu (Obr. 1).



Obr. 1: Histogram součtů pro $n = 2$

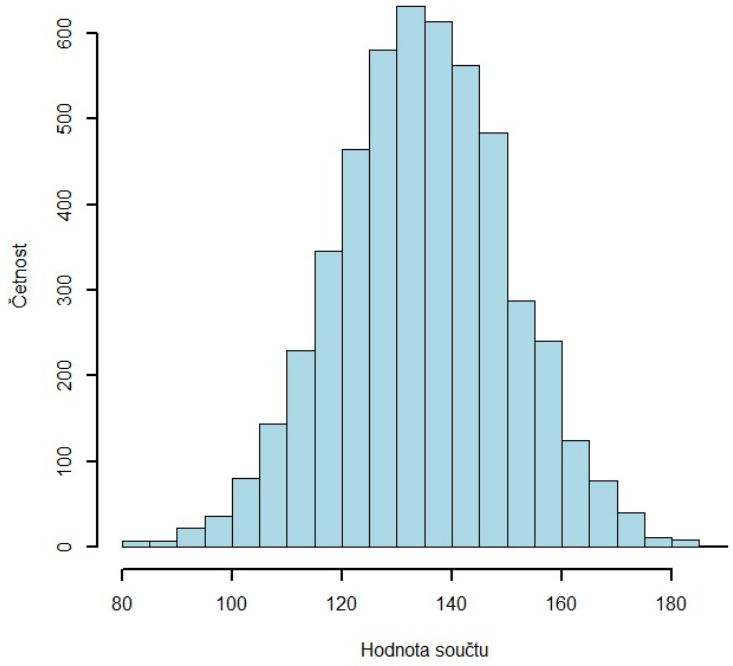
Krok 1.2. Pro zkoumání normality jsme využili Shapirova-Wilkova testu. Pro 1000 opakování jsme výsledné p-value vykreslili do (Obr. 2). Jelikož všechny p-value byly menší než námi zvolená hladina testu, ve všech 1000 případech jsme hypotézu normality zamítli.



Obr. 2: Histogram p-value Shapirova-Wilkova testu pro $n = 2$

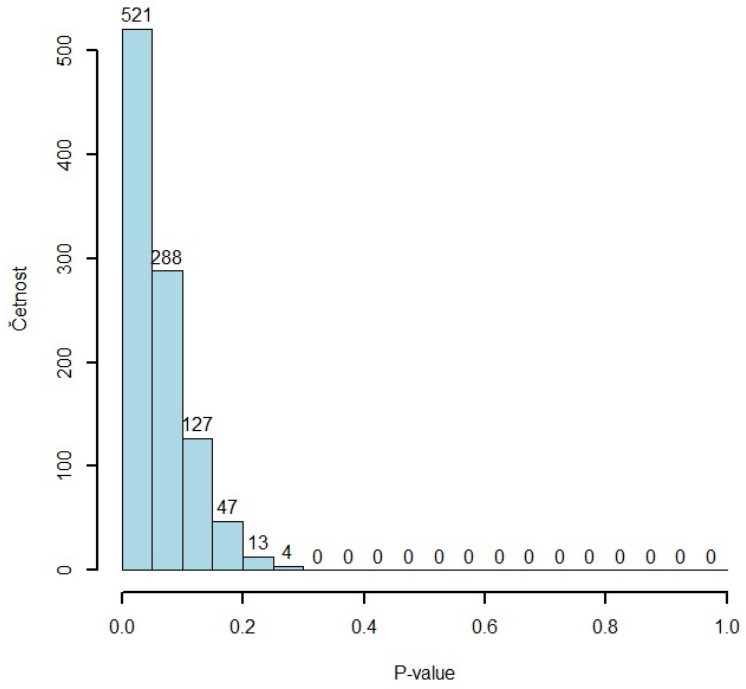
Krok 2.1. Volíme 30 kotoučů na zámku.[2] Rozdělení součtů zobrazíme pomocí histogramu (Obr. 3). Tvar histogramu začíná připomínat zvonovitý tvar normálního rozdělení, proto jsme provedli Krok 2.2.

[2]Experiment byl původně konstruován pro n = 2, 3, 5, 10, 30, 100, 1000, 5000, avšak pro názornost nám stačí uvést pouze $n = 2$, 30, 5000.



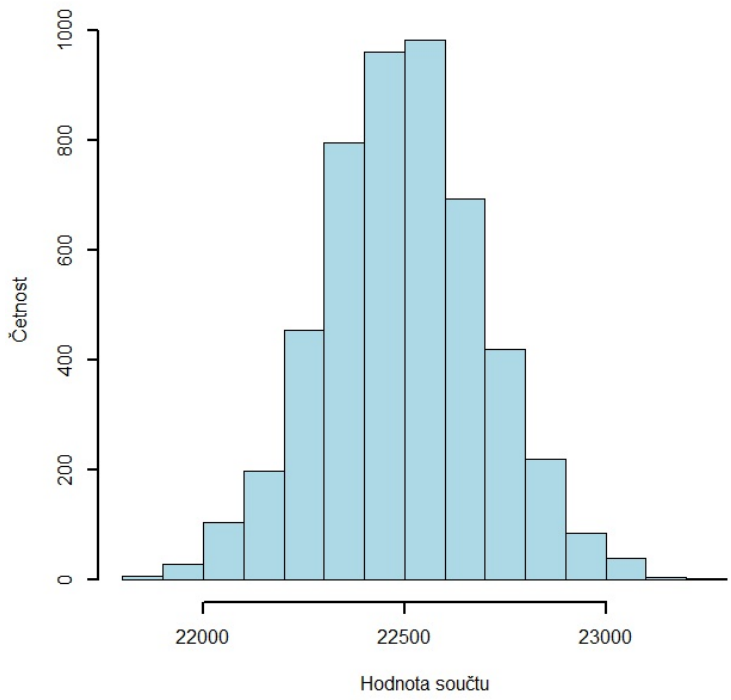
Obr. 3: Histogram součtů pro $n = 30$

Krok 2.2. Do histogramu (Obr. 4) vykreslíme p-value získané použitím Shapirova-Wilkova testu a vidíme, že již pro n v řádu několika desítek se začíná objevovat normalita. Konkrétně v 479 z 1000 případů nešlo hypotézu normality na 5% hladině významnosti zamítnout. Poznamenejme, že exaktní rozdělení není normální. Jde totiž o diskrétní rozdělení - součty mohou nabývat pouze celočíselných hodnot od 0 (v případě, že na všech 30 kotoučích je nastavena 0) do 270 (v případě, že na všech 30 kotoučích je nastavena 9).



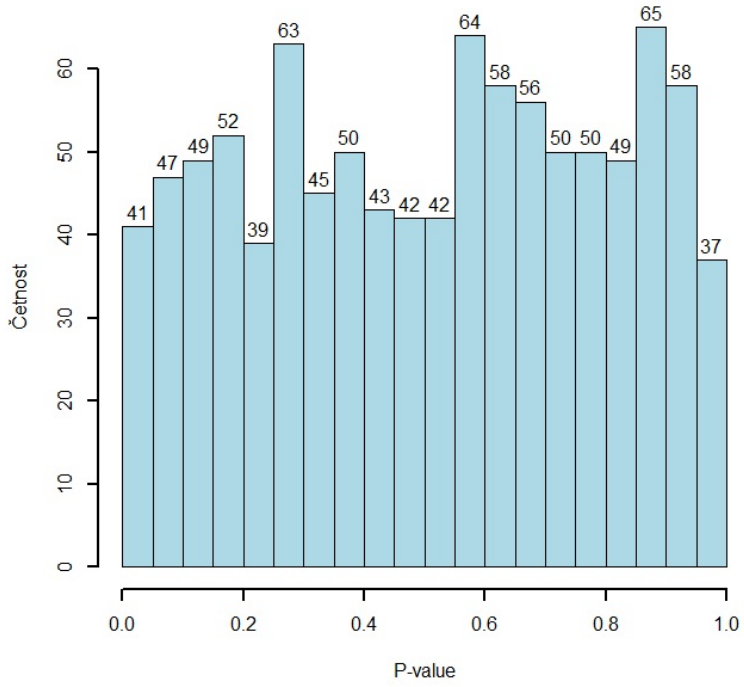
Obr. 4: Histogram p-value Shapirova-Wilkova testu pro $n = 30$

Krok 3.1. Jelikož jsme se chtěli co nejlépe přiblížit situaci popisované v centrální limitní větě, tedy situaci, kdy $n \to \infty$ a přitom obdržet výsledek v „rozumném čase“, zvolili jsme $n = 5000$. V histogramu (Obr. 5) můžeme vidět tvar velmi podobný zvonovitému tvaru hustoty normálního rozdělení. Ani zde exaktní rozdělení není normální, neboť je diskrétní.



Obr. 5: Histogram součtů pro $n = 5000$

Krok 3.2. V histogramu (Obr. 6) můžeme vidět, že pro $n = 5000$ jsme zamítli normalitu pouze 41 krát z 1000 při 0.05 hladině významnosti. Tímto jsme názorně demonstrovali centrální limitní větu.



Obr. 6: Histogram p-value Shapirova-Wilkova testu pro $n = 5000$

1.4. Jednoduchá a složená hypotéza H_0

Následující text vychází z [4].

Nechť máme náhodný výběr $x_1, \ldots, x_n$ o n nezávislých stejně rozdělených pozorováních náhodné veličiny X. Když zkoumáme, zda naše data pochází z normálního rozdělení nebo ne, můžeme nulovou (H_0) a alternativní (H_A) hypotézu pro test normality formulovat následovně:

$$H_0 : f_X(x) = f_{N(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \forall x \in R^1.$$

$$H_A : f_X(x) \neq f_{N(\mu,\sigma^2)}(x)$$

pro všechna $\mu \in \mathbb{R}$ a $\sigma > 0$.

V případě, že parametry μ a σ^2 jsou specifikovány a my testujeme hypotézu H_0 s $\mu = \mu_0$ a $\sigma^2 = \sigma_0^2$, kde μ_0 a σ_0^2 jsou známy, pak tuto hypotézu nazýváme *jednoduchá hypotéza*. V případě, kdy hypotéza neobsahuje žádná omezení na alespoň jeden z parametrů, nazýváme hypotézu H_0 *složená hypotéza*.

V této diplomové práci se budeme zabývat v praxi nejběžnějším případem, tedy případem, kdy jsou oba parametry neznámé.

1.5. Chyba prvního a druhého druhu

Následující text vychází z [1].

Může se stát, že hypotézu H_0 zamítneme, ačkoli je správná. V takovémto případě se dopouštíme *chyby prvního druhu*.

Když hypotézu H_0 nezamítneme, ačkoli není správná, dopustíme se *chyby druhého druhu.*

Hypotézu H_0 volíme zpravidla tak, aby chyba prvního druhu byla závažnější než chyba druhého druhu.

1.6. Knihovny v softwaru R

Pro matematické a statistické výpočty uvedené v této diplomové práci jsme využívali statistický software R. Jelikož při práci s tímto softwarem jsme pro použití některých příkazů museli načítat specializované knihovny, uvádíme na následující straně stručnou tabulku s těmito příkazy a knihovnami. Potřebnou knihovnu načteme pomocí příkazu *library (název knihovny)*.

příkaz	knihovna
jarque.test()	moments
jarque.bera.test()	tseries
ad.test()	nortest
cvm.test()	nortest
lillie.test()	nortest
sf.test()	nortest
rtriangle	triangle

1.7. Shape a omnibus test

Následující text vychází z [4].

V případě alternativní hypotézy H_A uvedené v kapitole 1.4., je možno rozlišit tři různé případy distribucí. První alternativou je tzv. asymetrické rozdělení se šikmostí různou od nuly. Druhou a třetí alternativou jsou rozdělení symetrická, která mají špičatost menší (resp. větší) než 3. Testy, které dokáží odhalit odchýlení pouze z hlediska šikmosti nebo špičatosti, nazýváme *shape (tvarové) testy*. Testy umožňující otestování šikmosti a špičatosti zároveň nazýváme *omnibus testy*.

2. Základní grafické nástroje

2.1. Histogram

Na histogram se můžeme dívat jako na grafický nástroj k znázornění distribuce dat pomocí sloupcového grafu[3] se sloupci stejné šířky, vyjadřující šířku intervalů (tříd), přičemž výška sloupců vyjadřuje četnost sledované veličiny v daném intervalu. V této diplomové práci využíváme histogram především k posouzení, zda zkoumaná data připomínají svým tvarem charakteristický symetrický tzv. zvonovitý tvar Gaussovy křivky.

Jeden z možných postupů sestrojení histogramu:

1. Data uspořádáme vzestupně.

2. Určíme minimální X_{min} a X_{max} maximální hodnotu a vypočteme variační rozpětí R:
$$R = X_{max} - X_{min}.$$

3. Určíme délku třídy C:
Jelikož pracujeme převážně se softwarem R, uvedeme vzorec (tzv. Sturges formula) [6], který využívá tento software.
$$C = \frac{R}{1 + 3.322 \log(n)},$$
Z důvodu jednoduššího vykreslování tříd histogramu nepoužijeme vypočtené C, ale použijeme místo něj C*, kde hodnota C* je číslo z níže uvedené množiny M, které je nejblíže vypočtenému C.
$$M_0 = \{1, 2, 5\}$$

[3] V této diplomové práci se přikláníme k histogramům, které jsou tvořeny sloupci (obdélníky). Existují však zdroje, které v histogramu běžně zakreslují např. trojúhelníky.

$$M = \left\{x \in \mathbb{R} : x = C \cdot 10^k, k \in \mathbb{Z}, C \in M_0\right\}$$

Tzn. prvky množiny M jsou celá čísla 1, 2, 5, 10, 20, 50, 100, 200, 500, ..., ale také čísla 0.5, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001, ...

4. Určení dolní hranice první třídy (označíme X_0), vykreslení histogramu:
   V tomto kroku určíme bod X_0, od kterého začneme vykreslovat sloupce, jejichž šířka je určena c* a výška je určena četností sledované veličiny v daném intervalu. První sloupec bude obsahovat hodnoty z intervalu $\langle X_0, X_0 + C^*)$, druhý sloupec $\langle X_0 + C^*, X_0 + 2 \cdot C^*)$. Dolní hranici první třídy určíme tak, aby hodnota X_{min} byla obsažena v první třídě. S vykreslováním sloupců končíme v okamžiku, kdy jsme vykreslili sloupec pro třídu, která obsahuje hodnotu X_{max}. V případě, že hodnota X_{max} bude ležet na některé z dolních hranic, zahrneme ji do předchozí třídy. Dolní hranici X_0 určíme jako nejbližší menší číslo z níže uvedené množiny N k číslu X_{min}. Pokud se X_{min} rovná některému číslu z množiny N, určíme toto číslo jako X_0.

$$N = \{x \in \mathbb{R} : x = C^* \cdot k, k \in \mathbb{Z}\}$$

Příklad 2: Hráč bowlingu naházel v 36 hrách tyto body: 183, 198, 172, 150, 128, 192, 204, 148, 184, 237, 216, 260, 170, 120, 126, 198, 128, 157, 246, 162, 168, 207, 242, 225, 192, 199, 202, 141, 159, 200, 232, 178, 160, 203, 255, 215. Jak může vypadat histogram jeho hry?

Postup: Víme, že $n = 36$.

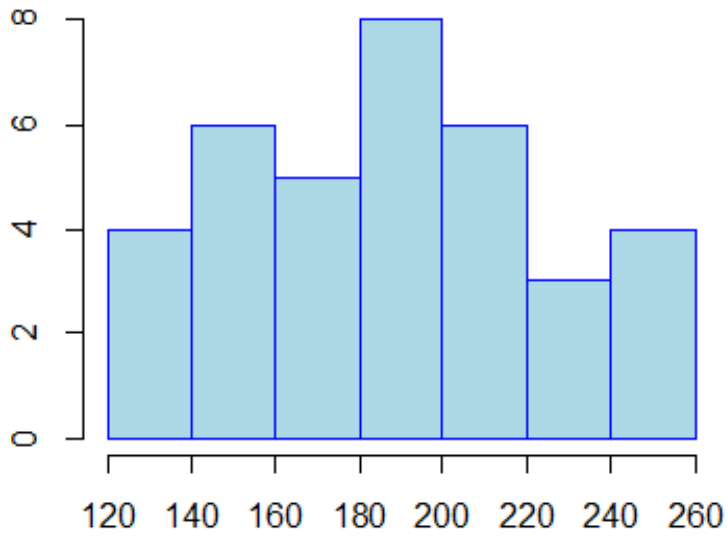
1. Data seřadíme vzestupně:
   120, 126, 128, 128, 141, 148, 150, 157, 159, 160, 162, 168, 170, 172, 178, 183, 184, 192, 192, 198, 198, 199, 200, 202, 203, 204, 207, 215, 216, 225, 232, 237, 242, 246, 255, 260.

2. $x_{max} = 260$, $x_{min} = 120$, $r = 140$,

3. c = 22.6903, $c^* = 20$.

4. Dolní hranice první třídy $x_0 = 120$.

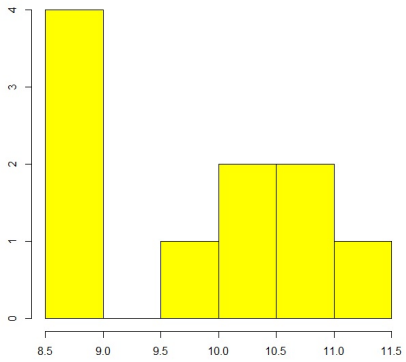
Histogram bude tedy vypadat:



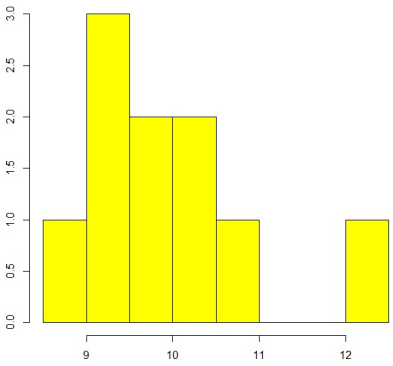
Obr. 7: Výsledný histogram bodových zisků v 36 hrách bowlingu

Náš pohled na histogram: Zkoumáme-li normalitu, pak budeme předpokládat, že data jsou rozložena přibližně symetricky (šikmost blízká nule), hodnoty se kumulují kolem výběrového průměru, výběrový průměr je blízký mediánu a předpokládáme minimální počet odlehlých hodnot.

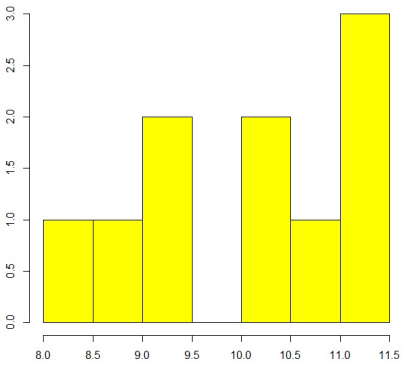
Otázka k zamyšlení: Jakou roli hraje u histogramu počet pozorování?
Na tuto otázku se pokusíme odpovědět následujícími pěti skupinami histogramů. Postupně jsme volili rozsahy $n = 10, 30, 50, 100, 200$ z $N(10, 1)$ vždy po třech výběrech ke každému z rozsahů.



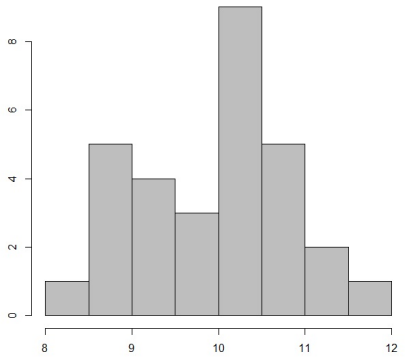
$N(10,1), n = 10$



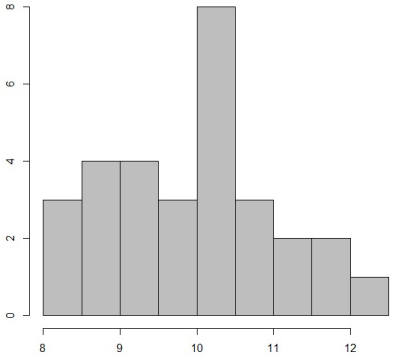
$N(10,1), n = 10$



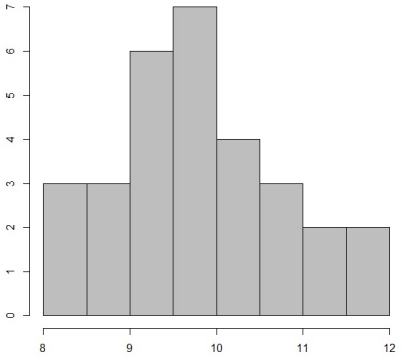
$N(10,1), n = 10$



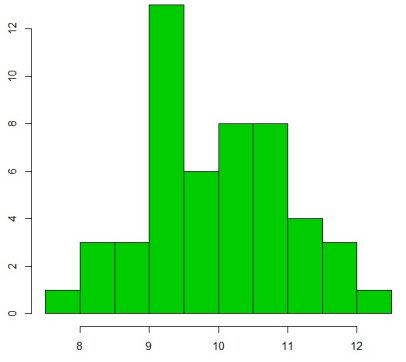
$N(10,1), n = 30$



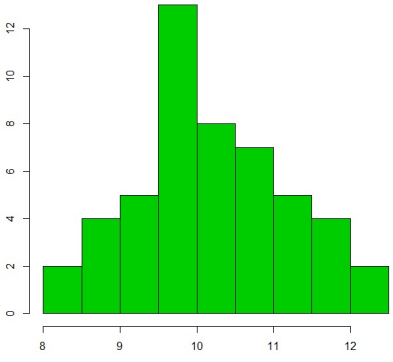
$N(10,1), n = 30$



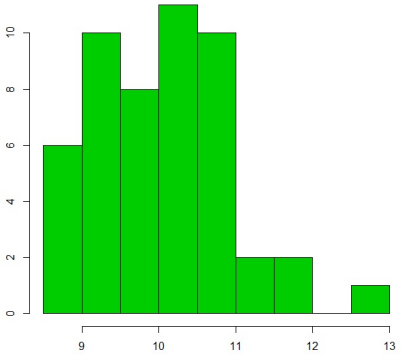
$N(10,1), n = 30$



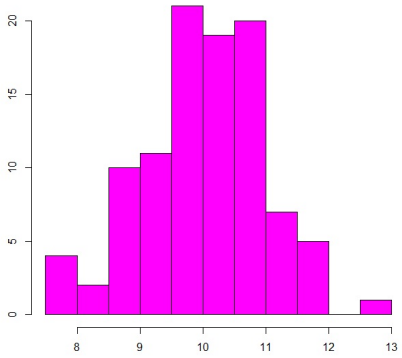
$N(10,1), n = 50$



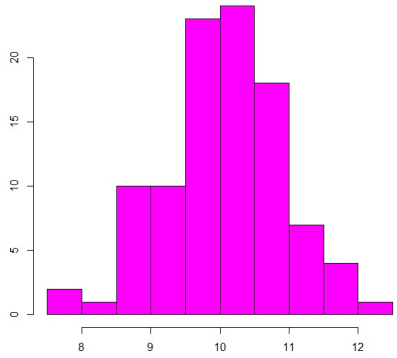
$N(10,1), n = 50$



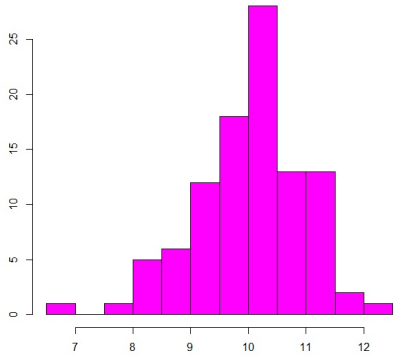
$N(10,1), n = 50$



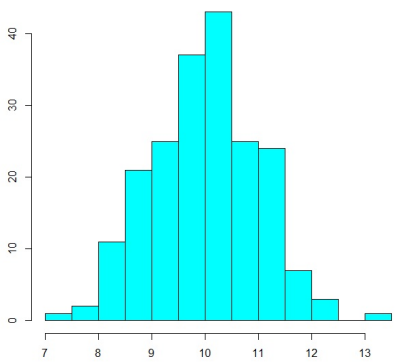
$N(10,1), n = 100$



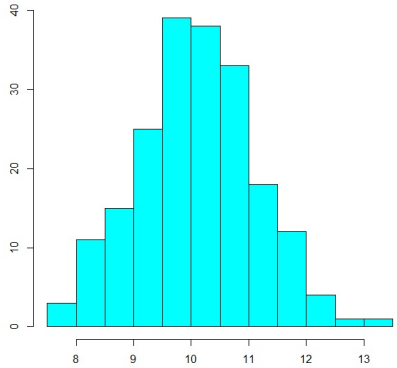
$N(10,1), n = 100$



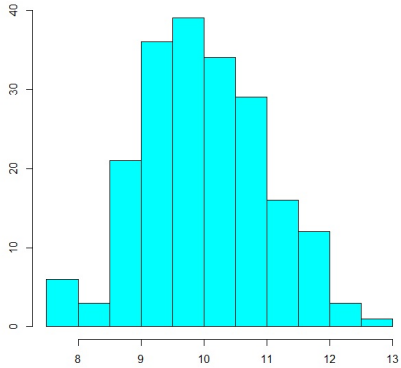
$N(10,1), n = 100$



$N(10,1), n = 200$



$N(10,1), n = 200$



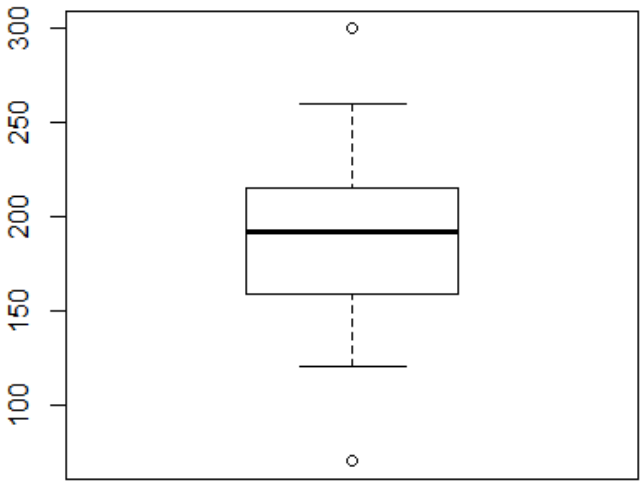
$N(10,1), n = 200$

Pozn.: *Počet tříd histogramu bývá často volen intuitivně na základě zkušeností. Zvláště u výběrů s nízkým počtem hodnot může nesprávně zvolený počet tříd resp. nesprávně zvolená šířka intervalu snížit vypovídací hodnotu grafu.*

2.2. Krabicový graf (Boxplot)

Krabicový graf je grafická metoda, která umožňuje znázornit:

1. Odhad mediánu,
2. hodnoty dolního a horního kvartilu,
3. minimální a maximální hodnotu,
4. odlehlé hodnoty (v případě že existují),
5. symetrii rozdělení.



Obr. 8: Ukázka krabicového grafu

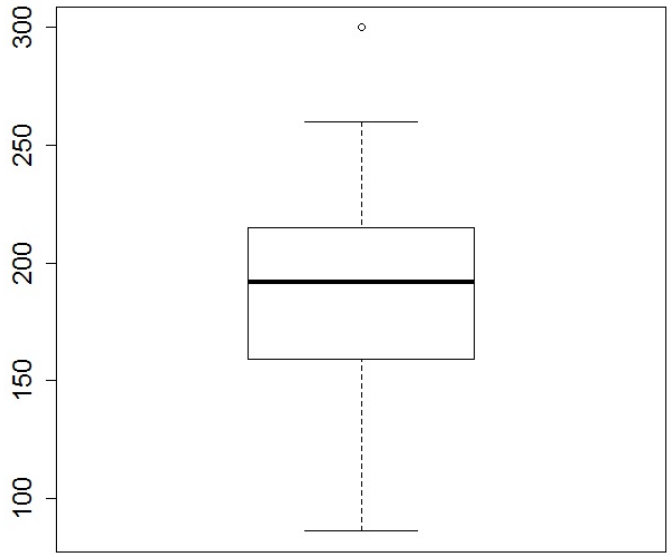
V obdélníku je vyznačen medián $x_{0.5}$ (tlustý předěl), dolní kvartil $x_{0.25}$ (dolní okraj vnitřního obdélníka) a horní kvartil $x_{0.75}$ (horní okraj vnitřního obdélníka) daného souboru dat. Hranice pro velmi vysokou hodnotu se určí jako $x_{0.75} + 1.5(x_{0.75} - x_{0.25})$. Hranice pro velmi nízkou hodnotu se určí jako $x_{0.25} - 1.5(x_{0.75} - x_{0.25})$. Software R tyto hranice vykresluje pouze v případě, že jim odpovídá konkrétní hodnota ze souboru dat. V případě, že náš soubor dat neobsahuje přímo hodnotu odpovídající $x_{0.75} + 1.5(x_{0.75} - x_{0.25})$, bude se vykreslovat nejbližší menší hodnota z daného souboru. Jestliže náš soubor dat neobsahuje přímo

hodnotu $x_{0.25} - 1.5(x_{0.75} - x_{0.25})$ vykreslí se nejbližší větší hodnota ze souboru dat. Vyskytují-li se v souboru hodnoty, které jsou menší než hranice velmi nízké hodnoty resp. větší než hranice velmi vysoké hodnoty, znázorňujeme je kroužky a jedná se o tzv. odlehlé hodnoty.

Příklad 3: Hráč bowlingu naházel v 38 hrách tyto body: 183, 198, 172, 150, 86, 128, 192, 204, 148, 184, 237, 216, 260, 170, 120, 126, 198, 128, 157, 300, 246, 162, 168, 207, 242, 225, 192, 199, 202, 141, 159, 200, 232, 178, 160, 203, 255, 215.

Pomocí softwaru jsme určili:

1. Medián $x_{0.5} = 192$,
2. dolní kvartil $x_{0.25} = 159.25$, horní kvartil, $x_{0.75} = 213$,
3. min $x_{min} = 86$, max $x_{max} = 300$.
4. Koeficienty: Šikmost $\alpha_3 = 0.11$, špičatost $\alpha_4 = 2.99$.



Obr. 9: Krabicový graf bodových zisků pro Příklad 3

Z obrázku (Obr. 9) lze poměrně snadno vyčíst, že krabicový graf může velmi dobře posloužit při určování minima, maxima, mediánu, dolního a horního kvartilu. Určování šikmosti a špičatosti je o něco obtížnější. *V naší diplomové práci využíváme krabicový graf především k určování odlehlých hodnot.*

2.3. Kvantil - kvantil graf (QQnorm)

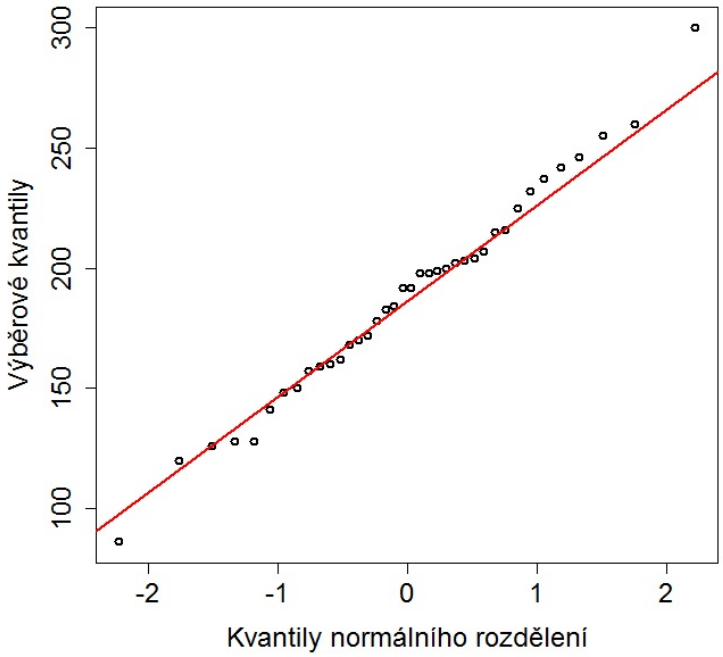
Umožňuje graficky posoudit, zda se sledované hodnoty „přibližují k normálnímu rozdělení".

Způsob konstrukce: Kvantil - kvantil graf konstruujeme tak, že na svislou osu vynášíme uspořádané hodnoty $x_1, \ldots, x_n$ a na vodorovnou osu kvantily $p[k]$ normálního rozdělení, kde

$$p[k] = \frac{k - r}{n + s},$$

přičemž r a s jsou korigující faktory. V naší diplomové práci využíváme $r = \frac{3}{8}$ a $s = \frac{1}{4}$, což je v softwaru R označováno jako „Type 9". [4] V softwaru R je možné vykreslit do kvantil - kvantil grafu pomocí příkazu *qqline* přímku, která prochází prvním a třetím kvartilem. Čím méně se body kvantil - kvantil grafu odchylují od této přímky, tím blíže je svým tvarem empirické rozdělení k rozdělení normálnímu.

[4] Software R umožňuje výpočet z 9 předprogramovaných vzorců pro p[k]. Vzorce se liší v nastavení korigujících faktorů r a s. Ve standardním nastavení je využíván Type 7, u kterého jsou korigující faktory $r = 1$ $s =$ -1.



Obr. 10: Kvantil - kvantil graf bodových zisků pro Příklad 3

Pozn.: *V případě normálně rozděleného souboru dat se hodnoty ve výše uvedeném grafu soustředí kolem červené (qqline) přímky.*

2.4. Empirická distribuční funkce (ECDF)

Následující text vychází z [7], [8].

Empirická distribuční funkce slouží jako grafický nástroj zobrazující uspořádaná data.

Způsob konstrukce:

1. Data upořádáme podle velikosti $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$,
2. na osu x vyneseme $x_{(i)}$ kde $i = 1, 2, \ldots, n$,
3. na osu y vyneseme ke každé hodnotě $x_{(i)}$ hodnotu $\frac{i}{n}$,

4. dvojice $\left[x_{(i)}, \frac{i}{n}\right]$ tvoří body, ze kterých tvoříme graf empirické distribuční funkce.

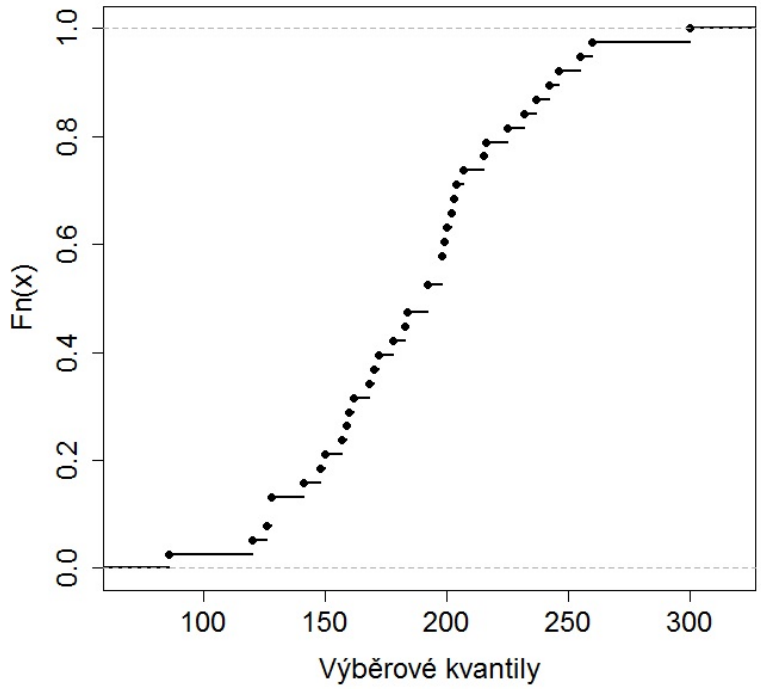
V případě dat seskupených do tříd konstruujeme empirickou distribuční funkci následovně:

1. Na osu x vyneseme meze třídních intervalů,

2. na osu y vyneseme kumulované relativní třídní četnosti,

$$F_j = \frac{1}{n} \sum_{j=1}^{k} n_j,$$

3. dvojice $\left[z_{(j)} + \frac{h}{2}, F_j\right]$ tvoří body, ze kterých tvoříme graf empirické distribuční funkce. Číslo h představuje délku třídních intervalů, $z_{(j)}$ představuje dolní meze třídních intervalů.



Obr.11: Empirická distribuční funkce bodových zisků pro Příklad 3

Pozn.: *V případě normálního rozdělení by data ve výše uvedeném grafu kopírovala tvar charakteristické esovité křivky.*

Pozn.: *Sestavení grafu empirické distribuční funkce bývá usnadněno normováním dat, které nemá vliv na vzhled grafu.*

3. Testy normality v softwaru R

Vyšetřujeme-li normalitu, grafické metody nám poskytnou alespoň hrubou představu o našem souboru dat. Takovéto vizuální zkoumání normality souboru dat je určitě dobrou startovací metodou, avšak různí pozorovatelé mohou vyvodit různé závěry u téhož souboru dat. Je tedy zřejmé, že u grafických metod hraje svou roli zkušenost statistika. Na grafické metody se díváme jako na velmi subjektivní nástroje zkoumání normality. Grafické metody nedokáží přesně určit okamžik, kdy se ještě jedná o normálně rozdělená data a kdy už ne. Abychom co nejlépe dokázali vyšetřit normalitu, vytvořili jsme následující kapitolu nazvanou *Testy normality v softwaru R*. Testy normality (stejně jako grafické metody) nedokáží přesně určit, kdy se ještě o normalitu jedná a kdy už ne. Testy normality bereme jako objektivnější nástroj k vyšetření normality, jelikož u nich dokážeme říci, že za daných testových podmínek normalitu lze/nelze zamítnout.

3.1. Shapirův-Wilkův test (SW test)

Následující kapitola vychází z [9]-[18].

V roce 1965 zveřejnili Samuel Shapiro a Martin Wilk testovou statistiku W. Shapirův-Wilkův test byl původně navržen pro rozsahy $3 \leq n \leq 50$. Royston [9] rozšířil tento test na rozsahy $3 \leq n \leq 2000$. Shapirův-Wilkův test lze použít i pro rozsahy větší něž $n = 2000$, avšak test již není schopen dosáhnout požadované hladiny významnosti. Tuto skutečnost budeme názorně demonstrovat v kapitole 4. Pro rozsahy do $n = 5000$ můžeme využít modifikaci SW testu nazvanou Shapirův-Franciův test. Vhodnost této modifikace budeme názorně demonstrovat v kapitole 4.2. Testová procedura *shapiro.test* v softwaru R povoluje chybějící hodnoty. SW test je nejčastěji využívaným testem normality u výběrů rozsahu $n \leq 2000$ [16].

Shapirův-Wilkův test využíváme k testování hypotézy, že náhodný výběr $x_1, ..., x_n$ pochází z normálního rozdělení. Test navazuje na grafickou metodu kvantil - kvantil grafu. Zjišťujeme, zda body sestrojeného kvantil - kvantil grafu

se významně odchylují od přímky proložené prvním a třetím kvartilem. Testová statistika může nabývat hodnot od 0 do 1, kde hodnota 1 představuje perfektní shodu dat s normálním rozdělením.

Pozn.: *Obecně však nemůžeme říci, že čím více se s W blížíme k 1, tím blíže jsme s naším empirickým rozdělením k rozdělení teoretickému, jelikož zde hraje roli rozsah výběru n. Vyjde-li nám u výběru např. n = 3 a w = 0.8, tak normalitu nezamítneme. Vyjde-li však u výběru n = 50 a w = 0.9, normalitu zamítneme, i když hodnota testové statistiky W byla u druhého testu blíže 1 než u testu prvního.*

Testovou statistiku W můžeme definovat:

$$W = \frac{\left(\sum_{i=1}^{m} a_{i,n}(x_{(n-i+1)} - x_{(i)})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

kde $m = \frac{n}{2}$ pro sudá n, pro n lichá $m = \frac{n-1}{2}$. Koeficienty $a_{i,n}$ jsou tabelovány [17].

Statistiku W lze vyjádřit i následovně:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

kde $x_{(i)}$ značí i-tou hodnotu uspořádaného výběru a a_i váhy, které jsou odvozeny ze středních hodnot a varianční matice uspořádaného výběru z $N(0,1)$ rozsahu n. Tyto váhy jsou tabelovány.

Jeden z možných postupů průběhu SW testu:

1. Data uspořádáme vzestupně (označíme $x_{(1)}, \ldots, x_{(n)}$).

2. Vypočteme $\sum_{i=1}^{n}(x_i - \bar{x})^2$.

3. Vypočteme $\left(\sum_{i=1}^{m} a_{i,n}(x_{(n-i+1)} - x_{(i)})\right)^2$.

4. Vypočteme hodnotu testové statistiky W.

5. Vypočtenou hodnotu testové statistiky porovnáme s tabelovanou kritickou hodnotou SW testu.

6. Zamítneme/nezamítneme nulovou hypotézu na dané hladině významnosti.

Příklad 4: Otestujme pomocí SW testu hypotézu, že náhodný výběr x = (7, 0, -3, 10, -8, 4, 6) pochází z normálního rozdělení. Pracujme na hladině významnosti $\alpha = 0.05$.

Víme: $n = 7$, $m = \frac{n-1}{2} = 3$

1. Data uspořádáme vzestupně (-8, -3, 0, 4, 6, 7, 10).

2. Vypočteme $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 237.4286$.

3. Vypočteme $\left(\sum_{i=1}^{m} a_{i,n}(x_{(n-i+1)} - x_{(i)})\right)^2 =$
   $\left(a_{1,7}(x_{(7)} - x_{(1)}) + a_{2,7}(x_{(6)} - x_{(2)}) + a_{3,7}(x_{(5)} - x_{(3)})\right)^2 =$
   $(0.6233(10 + 8) + 0.3031(7 + 3) + 0.1401(6 - 0))^2 = 15.091^2 = 227.7383.$
   Koeficienty $a_{i,n}$ jsou uvedeny v [17].

4. Vypočteme hodnotu testové statistiky $w = \frac{227.7383}{237.4286} = 0.9592$

5. Porovnáme w s kritickou hodnotou Shapirova-Wilkova testu [18]: $0.9592 > 0.803$.

6. Normalitu nelze zamítnout na dané hladině významnosti.

7. Pro kontrolu jsme využili statistický software R. Hodnota testové statistiky nám vyšla w = 0.9594, p-value= 0.8138. Jde vidět, že p-value je větší než námi zvolená hladina významnosti, proto normalitu nelze zamítnout.
   Pozn.: Při kontrole softwarem R se hodnota testové statistiky W lišila o 0.0002.

3.2. Jarqueův-Berův test (JB test)

Následující kapitola vychází z [19]-[20].

Tento test vynalezli Carlos Jarque a Anil K. Bera. Jde o test dobré shody, který porovnává výběrovou šikmost a špičatost se šikmostí a špičatostí normálního rozdělení. Testová hypotéza H_0 říká, že šikmost je nulová a špičatost rovna třem. Čím více se výběrová šikmost (resp. špičatost) odchyluje od šikmosti (resp. špičatosti) normálního rozdělení, tím více narůstá statistika JB. Testová procedura *jarque.bera.test* v softwaru R umožňuje testovat soubory dat, jejichž rozsah je větší než $n = 1$. Tato procedura nepovoluje chybějící hodnoty. Testovou statistiku JB můžeme definovat:

$$JB = \frac{n}{6}\left(\alpha_3^2 + \frac{(\alpha_4 - 3)^2}{4}\right),$$

kde α_3 je výběrová šikmost a α_4 je výběrová špičatost. JB má asymptoticky chí-kvadrát χ_2^2 rozdělení se dvěma stupni volnosti.

Jeden z možných postupů průběhu JB testu:

1. Vypočteme $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. Vypočteme $M_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$.

3. Vypočteme $M_3 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^3$.

4. Vypočteme $M_4 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^4$.

5. Vypočteme $\alpha_3 = \frac{M_3}{M_2^{3/2}}$.

6. Vypočteme $\alpha_4 = \frac{M_4}{M_2^2}$.

7. Vypočteme testovou statistiku JB.

8. Vypočtenou hodnotu testové statistiky porovnáme s kritickou hodnotou χ^2_2.

9. Zamítneme/nezamítneme nulovou hypotézu na dané hladině významnosti.

Příklad 5: Otestujme pomocí JB testu hypotézu, že náhodný výběr x = (7, 0, -3, 10, -8, 4, 6) pochází z normálního rozdělení. Pracujme na hladině významnosti $\alpha = 0.05$.

Víme : $n = 7$

1. Vypočteme $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2.2857$.

2. Vypočteme $m_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 33.9183$.

3. Vypočteme $m_3 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3 = -89.6676$.

4. Vypočteme $m_4 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^4 = 2319.286$.

5. Vypočteme $\alpha_3 = \frac{m_3}{m_2^{3/2}} = -0.4539$.

6. Vypočteme $\alpha_4 = \frac{m_4}{m_2^2} = 2.0159$.

7. Vypočteme testovou statistiku JB = 0.5228.

8. Vypočtenou hodnotu testové statistiky porovnáme s kritickou hodnotou $\chi^2_2 = 5.99$.

9. Jelikož $0.5228 < 5.99$ nezamítneme nulovou hypotézu na dané hladině významnosti.

3.3. Andersonův-Darlingův test (AD test)

Následující kapitola vychází z [21], [22].

V roce 1952 Theodore Wilbur Anderson a Donald A. Darling společně publikovali tento test. Andersonův-Darlingův test je omnibus test viz. kapitola 1.7., který využíváme k testování složené hypotézy, že náhodný výběr $x_1, \ldots, x_n$ pochází z normálního rozdělení. Testová procedura *ad.test* softwaru R umožňuje testovat soubory dat, jejichž rozsah je větší než $n = 7$.

Testová statistika je definována následovně:

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(p_{(i)}) + \ln(1 - p_{(n-i+1)})),$$

kde $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$. Φ je kumulativní distribuční funkce *N(0,1)* standardizovaného normálního rozdělení, $\overline{X}$ výběrový průměr a S směrodatná odchylka souboru dat. Hodnoty $p_{(i)}$ můžeme nalézt v [24]. P-value se počítá pomocí modifikované statistiky:

$$Z = A(1 + \frac{0.75}{n} + \frac{2.25}{n^2}).$$

Jeden z možných postupů průběhu AD testu:

1. Data uspořádáme vzestupně (označíme $X_{(1)}, \ldots, X_{(n)}$).
2. Vypočteme $(X_{(i)} - \overline{X})/S$ a určíme $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$.
3. Vypočteme $A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(p_{(i)}) + \ln(1 - p_{(n-i+1)}))$.
4. Vypočteme $Z = A(1 + \frac{0.75}{n} + \frac{2.25}{n^2})$.
5. Vypočteme p-value.
   Když je $Z < 0.2$ p-value určíme jako $p = 1 - e^{-13.436+101.14Z-223.73Z^2}$.

Když je $0.2 \leq Z < 0.34$ určíme p-value jako $p = 1 - e^{-8.318+42.796Z-59.938Z^2}$.
Když je $0.34 \leq Z < 0.6$ určíme p-value jako $p = e^{0.9177-4.279Z-1.38Z^2}$.
Když je $Z \geq 0.6$ určíme p-value jako $p = e^{1.2937-5.709Z+0.0186Z^2}$.

Pozn.: *V případě testování dat generovaných z t_{20} rozdělení jsme se setkali s problémem, kdy nebylo možné určit p-value. Když soubor dat obsahoval alespoň jednu „velmi“ odlehlou hodnotu, která po znormování byla mimo interval (-8.3, 8.3), testová statistika A vycházela ∞ resp($-\infty$). V takovémto případě software R vracel u p-value NaN hodnotu. Pravděpodobnost, že při generování z t_{20} rozdělení nagenerujeme hodnotu, která bude po znormování větší než $|8.3|$ jsme určili jako 2 $\cdot pt(-8.3, 20) = 6.6 \cdot 10^{-8}$.*

6. Vypočtenou p-value porovnáme se zvolenou hladinou testu.

7. Zamítneme/nezamítneme nulovou hypotézu na dané hladině významnosti.

Příklad 6: Otestujme pomocí AD testu hypotézu, že náhodný výběr x = (10, -3, -1, 4, 0, 6, 7, 8) pochází z normálního rozdělení. Pracujme na hladině významnosti $\alpha = 0.05$.

Víme: $n = 8$.

1. Data uspořádáme vzestupně (označíme $(-3, -1, 0, 4, 6, 7, 8, 10)$).

2. Vypočteme $(X_{(i)} - \overline{X})/S$ = (-1.4616085, -1.0364133, -0.8238157, 0.0265747, 0.4517699, 0.6643675, 0.8769651, 1.3021603).

3. Určíme $p_{(i)}$ = (0.07192426, 0.15000468, 0.20502214, 0.51060052, 0.67428262, 0.74677244, 0.80974720, 0.90356920).

4. Vypočteme $A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(p_{(i)}) + \ln(1 - p_{(n-i+1)})) = 0.2749$.

5. Vypočteme $Z = A(1 + \frac{0.75}{n} + \frac{2.25}{n^2}) = 0.3651$.
Když $0.34 \leq Z < 0.6$ určíme p-value jako $p = e^{0.9177-4.279Z-1.38Z^2} = 0.4366$.

6. Normalitu nelze zamítnout.

3.4. Cramér-von Misesův test (CVM test)

Následující kapitola vychází z [21], [22].

Cramér-von Misesův test je omnibus test viz. kapitola 1.7., který využíváme k testování složené hypotézy, že náhodný výběr $x_1, \ldots, x_n$ pochází z normálního rozdělení. Testová procedura *cvm.test* v softwaru R umožňuje testovat soubory dat, jejichž rozsah je větší než $n = 7$. Tento test v porovnáním s AD testem dává menší váhu odlehlým hodnotám. Testová statistika je definována následovně:

$$W = \frac{1}{12n} + \sum_{i=1}^{n} (p_{(i)} - \frac{2i-1}{2n}),$$

kde $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$. Φ je kumulativní distribuční funkce *N(0,1)* standardizovaného normálního rozdělení, $\overline{X}$ výběrový průměr a S výběrová směrodatná odchylka. P-value se počítá pomocí modifikované statistiky:

$$Z = W(1 + \frac{0.5}{n}).$$

Jeden z možných postupů průběhu CVM testu:

1. Data uspořádáme vzestupně (označíme $X_{(1)}, \ldots, X_{(n)}$).
2. Vypočteme $(X_{(i)} - \overline{X})/S$ a určíme $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$.
3. Vypočteme $W = \frac{1}{12n} + \sum_{i=1}^{n} (p_{(i)} - \frac{2i-1}{2n})$.
4. Vypočteme $Z = W(1 + \frac{0.5}{n})$.
5. Vypočteme p-value.
   Když je $Z < 0.0275$ p-value určíme jako $p = 1 - e^{-13.953+775.5 \cdot Z - 12542.61 \cdot Z^2}$.
   Když je $0.0275 \leq Z < 0.051$ určíme p-value jako $p = 1 - e^{-5.903+179.546 \cdot Z - 1515.29 \cdot Z^2}$.
   Když je $0.051 \leq Z < 0.092$ určíme p-value jako $p = e^{0.886 - 31.62 \cdot Z + 10.897 \cdot Z^2}$.
   Když je $0.092 \leq Z < 1.1$ určíme p-value jako $p = e^{1.111 - 34.242 \cdot Z + 12.832 \cdot Z^2}$.
   Když je $W \geq 1.1$ p-value je menší než 7.3710^{-10}.

6. Vypočtenou p-value porovnáme se zvolenou hladinou testu.

7. Zamítneme/nezamítneme nulovou hypotézu na dané hladině významnosti.

Příklad 7: Otestujme pomocí CVM testu hypotézu, že náhodný výběr x = (10, -3, -1, 4, 0, 6, 7, 8) pochází z normálního rozdělení. Pracujme na hladině významnosti $\alpha = 0.05$.

Víme: $n = 8$.

1. Data uspořádáme vzestupně (označíme $(-3, -1, 0, 4, 6, 7, 8, 10)$).

2. Vypočteme $(x_i - \overline{x})/s =$ (-1.4616085, -1.0364133, -0.8238157, 0.0265747, 0.4517699, 0.6643675, 0.8769651, 1.3021603).

3. Určíme $p_{(i)} =$ (0.07192426, 0.15000468, 0.20502214, 0.51060052, 0.67428262, 0.74677244, 0.80974720, 0.90356920).

4. Vypočteme $w = \frac{1}{12n} + \sum_{i=1}^{n}(p_{(i)} - \frac{2i-1}{2n}) = 0.0459$.

5. Vypočteme $z = w(1 + \frac{0.5}{n}) = 0.0488$.

6. Vypočteme p-value $p = 1 - e^{-5.903+179.546 \cdot z - 1515.29 \cdot z^2} = 0.526797$.

7. Normalitu nelze zamítnout.

3.5. Lillieforsův (Kolmogorovův-Smirnovův) test (LILLIE test)

Následující kapitola vychází z [21]-[23].

Testová statistika je maximální absolutní odchylka mezi empirickou a hypotetickou kumulativní distribuční funkcí. Testová procedura *lillie.test* softwaru R umožňuje testovat soubory dat, jejichž rozsah je větší než $n = 4$. Testovou statistiku můžeme počítat následovně :

$$D = max\left\{D^+, D^-\right\},$$

kde

$$D^+ = max\left\{\frac{i}{n} - p_{(i)}\right\}, D^- = max\left\{p_{(i)} - \frac{i-1}{n}\right\},$$

kde $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$. Φ je kumulativní distribuční funkce *N(0,1)* standardizovaného normálního rozdělení, $\overline{X}$ výběrový průměr a S výběrová směrodatná odchylka. P-value se počítá na základě Dallalova-Wilkinsonova vzorce [23]. Tento vzorec funguje spolehlivě pouze pro p-value menší než 0.1. Pokud nám dle Dallalova-Wilkinsovona vzorce vyjde p-value větší než 0.1, pak je p-value počítána pomocí modifikované statistiky:

$$DDK = D(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}).$$

Jeden z možných postupů průběhu LILLIE testu:

1. Data uspořádáme vzestupně (označíme $X_{(1)}, \ldots, X_{(n)}$).
2. Vypočteme $(X_{(i)} - \overline{X})/S$ a určíme $p_{(i)} = \Phi((X_{(i)} - \overline{X})/S)$.
3. Určíme $D^+ = max\left\{\frac{i}{n} - p_{(i)}\right\}$ a $D^- = max\left\{p_{(i)} - \frac{i-1}{n}\right\}$.
4. Určíme $D = max\left\{D^+, D^-\right\}$.

5. Pro $n \leq 100$ určíme Dd = D a nd = n.

6. Jinak je $Dd = D \cdot ((\frac{n}{100})^{0.49})$ a nd = 100.

7. Určíme $\mathrm{p} = e^{-7.01256 \cdot Dd^2 \cdot (nd+2.78019) + 2.99587 \cdot Dd \cdot \sqrt{nd+2.78019} - 0.122119 + \frac{0.974598}{\sqrt{nd}} + \frac{1.67997}{nd}}$.

   Když p ≤ 0.1, máme výsledek. Pokud nám vyjde větší p, pak je potřeba využít modifikované statistiky: DDK $= \sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}$.

Pro:	p-value určíme jako:
DDK ≤ 0.302	1
$0.302 <$ DDK ≤ 0.5	2.76773 - 19.828315 $\cdot DDK + 80.709644 \cdot DDK^2$ - 138.55152 $\cdot DDK^3 + 81.218052 \cdot DDK^4$
$0.5 <$ DDK ≤ 0.9	$-4.901232 + 40.662806 \cdot DDK - 97.490286 \cdot DDK^2$ + 94.029866 $\cdot DDK^3 - 32.355711 \cdot DDK^4$
$0.9 <$ DDK ≤ 1.31	$6.198765 - 19.558097 \cdot DDK + 23.186922 \cdot DDK^2$ - 12.234627 $\cdot DDK^3 + 2.423045 \cdot DDK^4$
Pro jiné DDK	0

8. Vypočtenou p-value porovnáme se zvolenou hladinou testu.

9. Zamítneme/nezamítneme nulovou hypotézu na dané hladině významnosti.

Příklad 8: Otestujme pomocí LILLIE testu hypotézu, že náhodný výběr x = (10, -3, -1, 4, 0, 6, 7, 8) pochází z normálního rozdělení. Pracujme na hladině významnosti $\alpha = 0.05$.

Víme: $n = 8$.

1. Data uspořádáme vzestupně (označíme $(-3, -1, 0, 4, 6, 7, 8, 10)$).

2. Vypočteme $(x_i - \overline{x})/s$ = (-1.4616085, -1.0364133, -0.8238157, 0.0265747, 0.4517699, 0.6643675, 0.8769651, 1.3021603).

3. Určíme $p_{(i)}$ = (0.07192426, 0.15000468, 0.20502214, 0.51060052, 0.67428262, 0.74677244, 0.80974720, 0.90356920).

4. Určíme $D^+ = max\left\{\frac{i}{n} - p_{(i)}\right\} = 0.1699$ a $D^- = max\left\{p_{(i)} - \frac{i-1}{n}\right\} = 0.1742$.

5. Určíme $D = max\{D^+, D^-\} = 0.1743$.

6. Pro $n \leq 100$ určíme $Dd = D$ a $nd = n$.

7. Určíme $\mathrm{p} = e^{-7.01256 \cdot Dd^2 \cdot (nd+2.78019) + 2.99587 \cdot Dd \cdot \sqrt{nd+2.78019} - 0.122119 + \frac{0.974598}{\sqrt{nd}} + \frac{1.67997}{nd}}$ $= 0.8612$.
   Jelikož p-value vyšla větší než 0.1, přepočítáme $DDK = \sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \cdot D$ $= 0.5436$
   $\mathrm{p} = -4.901232 + 40.662806 \cdot DDK - 97.490286 \cdot DDK^2 + 94.029866 \cdot DDK^3 - 32.355711 \cdot DK^4 = 0.6738$

8. Normalitu nelze zamítnout.

3.6. Shapirův-Franciův test (SF test)

Následující kapitola vychází z [21], [25], [26].

Test umožňuje testovat soubory, jejichž rozsah je v rozmezí 5 až 5000. Procedura v softwaru R, která se nazývá *sf.test*, povoluje chybějící hodnoty. Test slouží k testování složené hypotézy normality. SF testová statistika představuje korelaci mezi hodnotami uspořádaného náhodného výběru x a odhadovanými kvantily standardizovaného normálního rozdělení y. Kvantily standardizovaného normálního rozdělení odhadujeme v softwaru R pomocí příkazu *qnorm(ppoints(x, a = 3/8))*. Tato aproximace se v softwaru R využívá i při tvorbě kvantil-kvantil grafu u výběrů s rozsahem větším než 10, avšak parametr $a = 1/2$. Za platnosti nulové hypotézy je rozdělení testové statistiky neznámé a musí být odhadnuto [25]. Royston 1983 dokázal pomocí simulací Monte Carlo, že transformace log(1- W), kde $W = cor(x, y)^2$ je přibližně normální.

Testovou statistiku definujeme následovně :

$$Z = \frac{\log(1 - W) - \mu}{\sigma}$$

kde $W = cor(x, y)^2 = \left(\frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} \right)^2$, μ = -1.2725 + 1.0521 $\cdot(v - u)$, σ = 1.0308 - 0.26758 $\cdot(v + 2/u)$, $u = \log(n)$, $v = \log(u)$, y = *qnorm(ppoints(n, a = 3/8))*, kde příkaz *qnorm()* určuje kvantily standardizovaného normálního rozdělení a n je počet pozorování. Víme, že log(1-W) má přibližně $N(\mu, \sigma^2)$, kde μ = -1.2725 + 1.0521 $\cdot(v - u)$, σ = 1.0308 - 0.26758 $\cdot(v + 2/u)$, testová statistika Z má přibližně *N(0,1)* rozdělení. Mějme hodnotu testové statistiky $z_0 \in \mathbb{R}$, P-value určíme:

$p = P(Z > z_0)$, což v softwaru R určíme pomocí příkazu $pnorm(z_0, lower.tail = FALSE)$.

Příklad 9: Otestujte pomocí SF testu hypotézu, že náhodný výběr (10, -3, -1, 4, 0, 6, 7, 8) pochází z normálního rozdělení. Pracujte na hladině významnosti $\alpha = 0.05$.

Víme: $n = 8$.

1. Data uspořádáme vzestupně (označíme x= (-3, -1, 0, 4, 6, 7, 8, 10)).
2. Určíme y = (-1.4342002, -0.8524950, -0.4727891, -0.1525060, 0.1525060, 0.4727891, 0.8524950, 1.4342002).
3. Určíme w = $cor(x, y)^2 = 0.9544$.
4. Určíme u = $\log(n) = 2.0794$ a v = $\log(u) = 0.7321$.
5. Určíme $\mu = -1.2725 + 1.0521 \cdot (v - u) =$ -2.69.
6. Určíme $\sigma = 1.0308 - 0.26758 \cdot (v + 2/u) = 0.5775$.

7. Vypočteme z = $\frac{\log(1-W)-\mu}{\sigma}$ = -0.6895059.

8. Určíme p = pnorm(z, lower.tail = FALSE), p = 0.7547.

9. Normalitu nelze zamítnout.

4. Porovnání testů normality v softwaru R

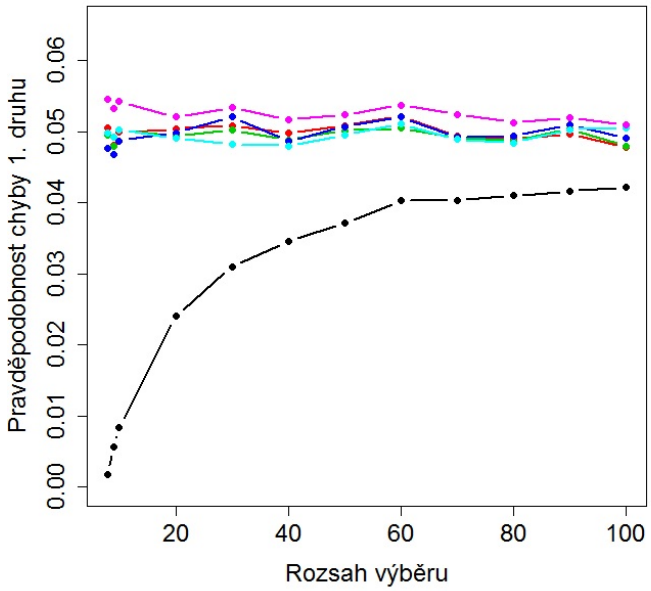
4.1. Vývoj pravděpodobnosti chyby prvního druhu

V této kapitole budeme určovat pravděpodobnost chyby prvního druhu u JB, SW, AD, CVM, LILLIE a SF testu v závislosti na rozsahu výběru n. Rozsahy výběru jsme volili (8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2200, 2400, 2600, 2800, 3000, 3200, 3400, 3600, 3800, 4000, 4200, 4400, 4600, 4800, 5000). Chybu prvního druhu jsme definovali v kapitole 1.5.

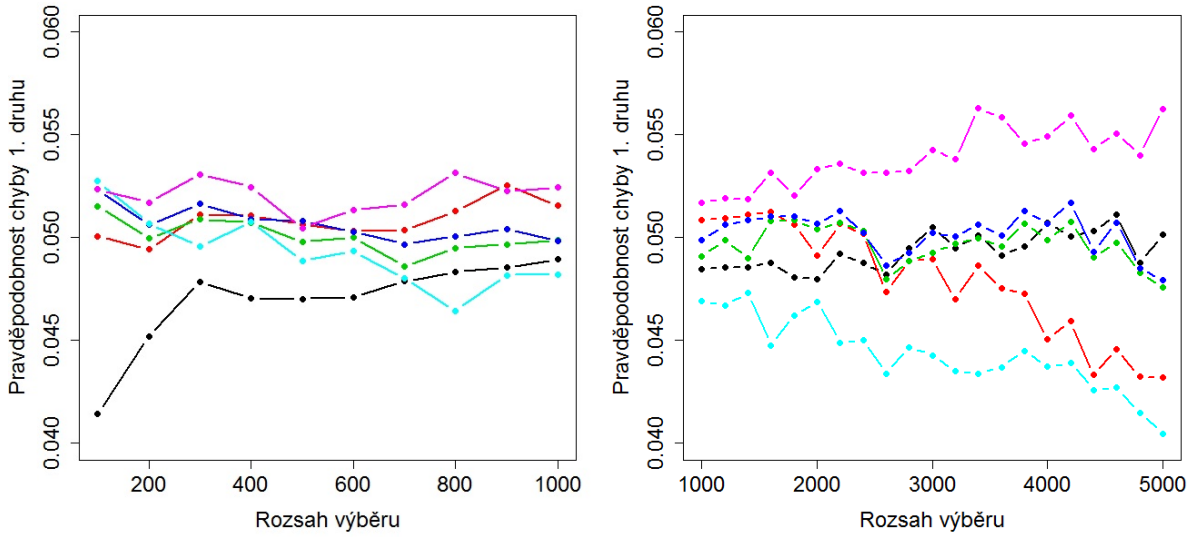
Pro každý výše zmíněný rozsah výběru jsme nagenerovali hodnoty z $N(0, 1)$ rozdělení a otestovali normalitu pomocí výše zmíněné šestice testů. U všech testů jsme zaznamenali, zda test zamítl nebo nezamítl složenou hypotézu H_0 (viz kapitola 1.4.) na hladině významnosti $\alpha = 0.05$. Tuto simulaci jsme provedli 50000 krát, aby naše odhady pravděpodobnosti chyby prvního druhu byly co nejpřesnější. Pravděpodobnost chyby prvního druhu pro konkrétní test normality a konkrétní rozsah výběru n jsme v případě dat nagenerovaných z $N(0, 1)$ odhadli jako podíl počtu zamítnutí hypotézy H_0 daným testem na daném rozsahu ku 50000. Pravděpodobnost chyby prvního druhu jsme určili u každého testu na každém rozsahu výběru. Výsledné pravděpodobnosti chyby prvního budou zaznamenány v Obr. 12 a Obr. 13.

Při zkoumání vývoje pravděpodobnosti chyby prvního druhu nejprve sledujeme, zda některý z testů normality překročil zvolenou hladinu významnosti $\alpha = 0.05$. V případě, že tento jev nastal, daný test na daném rozsahu nedoporučujeme používat, jelikož nevyhověl zvolenému požadavku na hladinu významnosti $\alpha = 0.05$. Jednodušeji řečeno „testu nelze věřit“. Dalším problémem (i když méně závažným) je, když se u některého testu pohybujeme s pravděpodobností chyby prvního druhu výrazně pod zvolenou hranicí $\alpha = 0.05$. Snížíme-li pravděpodobnost chyby prvního druhu, pak zároveň zvýšíme pravděpodobnost chyby druhu druhého. Vývoj pravděpodobnosti chyby druhého druhu budeme sledovat v následují kapitole. V tabulce Tabulka 1. jsme zaznamenali náš komentář

k pravděpodobnosti chyby prvního druhu u JB, SW, AD, CVM, LILLIE a SF testu normality u rozsahů $n = 10, 30, 50, 70, 200, 500, 800, 2000, 3000, 4000$. Podobným způsobem bychom mohli okomentovat pravděpodobnost chyby prvního druhu jednotlivých testů i u dalších rozsahů výběru na základě obrázků Obr. 12 a Obr. 13.



Obr. 12: Vývoj pravděpodobnosti chyby prvního druhu pro data generovaná z *N(0,1)* rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy.



Obr. 13: Vývoj pravděpodobnosti chyby prvního druhu pro data generovaná z *N(0,1)* rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro střední rozsahy(vlevo), velké rozsahy (vpravo).

Tabulka 1.: Pravděpodobnost chyby 1. druhu

test $\|\|rozsah$	10	30	50	70	200	500	800	2000	3000	4000
JB	-	-	-	-	-	-	-	-	0	0
SW	0	0	0	0	0	0	0	0	-	-
AD	0	0	0	0	0	0	0	0	0	0
CVM	0	+	0	0	0	0	0	0	0	0
LILLIE	0	0	0	0	0	0	-	-	-	-
SF	+	+	+	+	+	0	+	+	+	+

0/- test dosahuje/nedosahuje na hladinu významnosti

\+ test překračuje hladinu významnosti

V tabulce 1. můžeme vidět, že AD a CVM test si z hlediska pravděpodobnosti chyby prvního druhu na vybraných rozsazích vedl výborně. Velmi dobře si u zvolených rozsahů do 2000 vedl i SW test. Naopak SF test si nevedl dobře, jelikož u většiny rozsahů překračoval zvolenou hladinu významnosti. V případě, že se dopustíme chyby prvního druhu, jedná se o závažnější přečin, než když se dopustíme chyby druhého druhu.

4.2. Vývoj pravděpodobnosti chyby druhého druhu

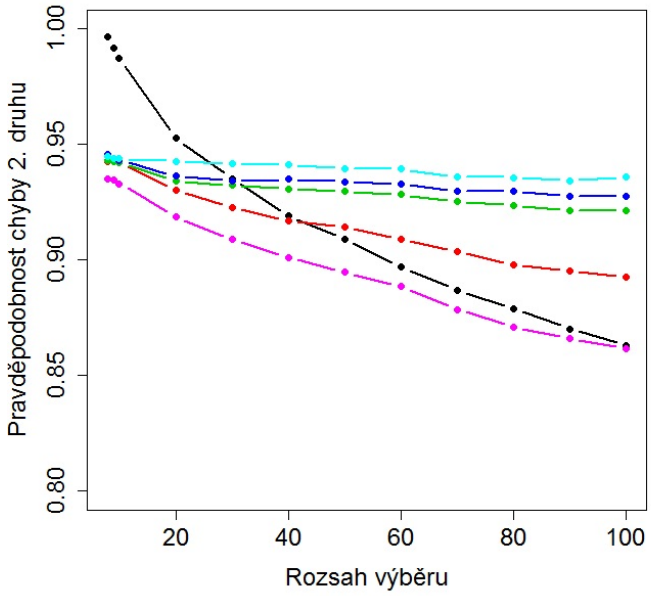
V této kapitole budeme určovat pravděpodobnost chyby druhého druhu u JB, SW, AD, CVM, LILLIE a SF testu v závislosti na rozsahu výběru n. Rozsahy výběru jsme volili stejně jako v kapitole 4.1. Chybu druhého druhu jsme definovali v kapitole 1.5.

Pro každý výše zmíněný rozsah výběru jsme postupně nagenerovali hodnoty ze studentova rozdělení o 20 stupních volnosti, trojúhelníkového rozdělení s krajními mezemi 0, 2 a modem 1, směsi standardizovaného normálního rozdělení a normálního rozdělení se střední hodnotou 2 a rozptylem 1, exponenciálního rozdělení se střední hodnotou 0.1, lognormálního rozdělení se střední hodnotou 0 a směrodatnou odchylkou 1, chí-kvadrát rozdělení o 10 stupních volnosti viz kapitola (4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5, 4.2.6, 4.2.7) a otestovali normalitu pomocí výše zmíněné šestice testů. U všech testů jsme zaznamenali, zda test zamítl nebo nezamítl složenou hypotézu H_0 (viz kapitola 1.4.) na hladině významnosti $\alpha = 0.05$. Tuto simulaci jsme provedli 50000 krát. Pravděpodobnost chyby druhého druhu pro konkrétní test normality a konkrétní rozsah výběru n jsme určili jako podíl počtu nezamítnutí hypotézy H_0 daným testem na daném rozsahu ku 50000. Pravděpodobnost chyby druhého druhu jsme určili u každého testu na každém uvedeném rozsahu výběru. Výsledné pravděpodobnosti chyby druhého druhu budou zaznamenány v Obr. 14 až Obr. 23.

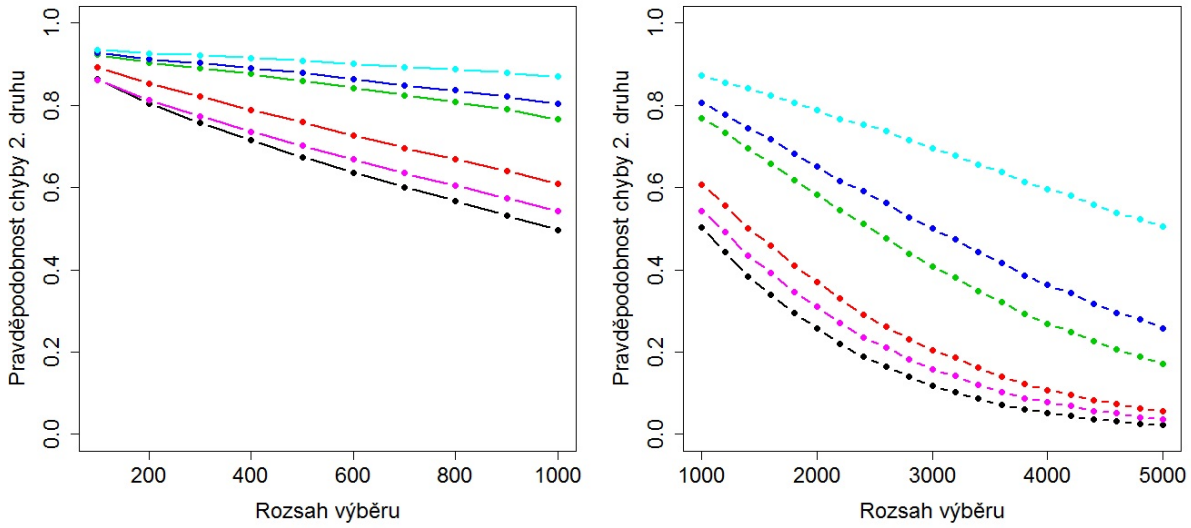
Při zkoumání vývoje pravděpodobnosti chyby druhého druhu sledujeme, jak velké tyto pravděpodobnosti byly u jednotlivých testů normality v závislosti na rozsahu výběru. Test s nejmenší pravděpodobností chyby druhéhu druhu bereme z hlediska pravděpodobnosti chyby druhého druhu jako nejlepší. Abychom dokázali vyvodit obstojné závěry o vývoji pravděpodobnosti chyby druhého druhu, zvolili jsme výše uvedenou sedmici rozdělení. Rozdělení byla volena na základě jejich tvarů. Do naší analýzy jsme se snažili zahrnout jak rozdělení velmi podobná normálnímu rozdělení (t_{20}, směs $N(0,1)$ a $N(2,1)$), tak rozdělení méně podobná (např. χ^2_{10}) až rozdělení výrazně odlišná ($Exp(10), Lnorm(0,1)$).

4.2.1. Studentovo rozdělení (t-rozdělení)

Studentovo rozdělení jsme volili z důvodu velké podobnosti k rozdělení normálnímu. Toto rozdělení je stejně jako normální rozdělení symetrické, liší se především tzv. těžkými chvosty. Z důvodu velké podobnosti mají testy normality problém se zamítáním normality, což je patrné na obrázcích Obr. 14, Obr. 15, Obr. 16. V tabulce Tabulka 2. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.). Již z této tabulky je zřejmé, že pokud bychom hodnotili kvalitu testu pouze na základě pravděpodobnosti chyby prvního druhu, nebyl by náš závěr úplně správný. Např.: CVM test funguje dle pravděpodobnosti chyby prvního druhu výborně, avšak z hlediska pravděpodobnosti chyby druhého druhu u t_{20} se drží až na 4. a především 5. pozici.



Obr. 14: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z t_{20} rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy.



Obr. 15: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z t_{20} rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro střední rozsahy (vlevo), velké rozsahy (vpravo).

Pozn.: *Z obrázků Obr. 14 a Obr.15 je vidět, že s rostoucím rozsahem výběru se u jednotlivých testů normality snižovala pravděpodobnost chyby druhého druhu. „Testy častěji poznaly, že se nejedná o data z normálního rozdělení.“*

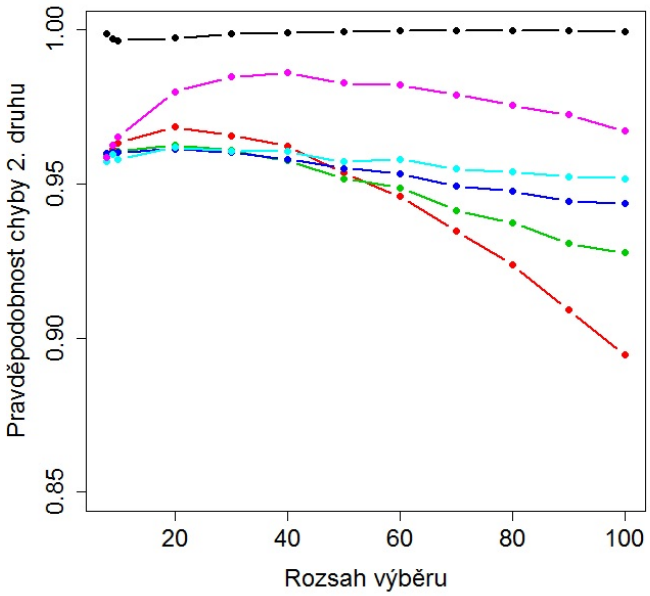
Tabulka 2.: Pravděpodobnost chyby 2. druhu

test $\|\|rozsah$	8	20	30	40	50	80	100	200 - 5000*
JB	6.	6.	5.	3.	2.	2.	2.	1.
SW	2.	2.	2.	2.	3.	3.	3.	3.
AD	3.	3.	3.	4.	4.	4.	4.	4.
CVM	5.	4.	4.	5.	5.	5.	5.	5.
LILLIE	4.	5.	6.	6.	6.	6.	6.	6.
SF	1.	1.	1.	1.	1.	1.	1.	2.

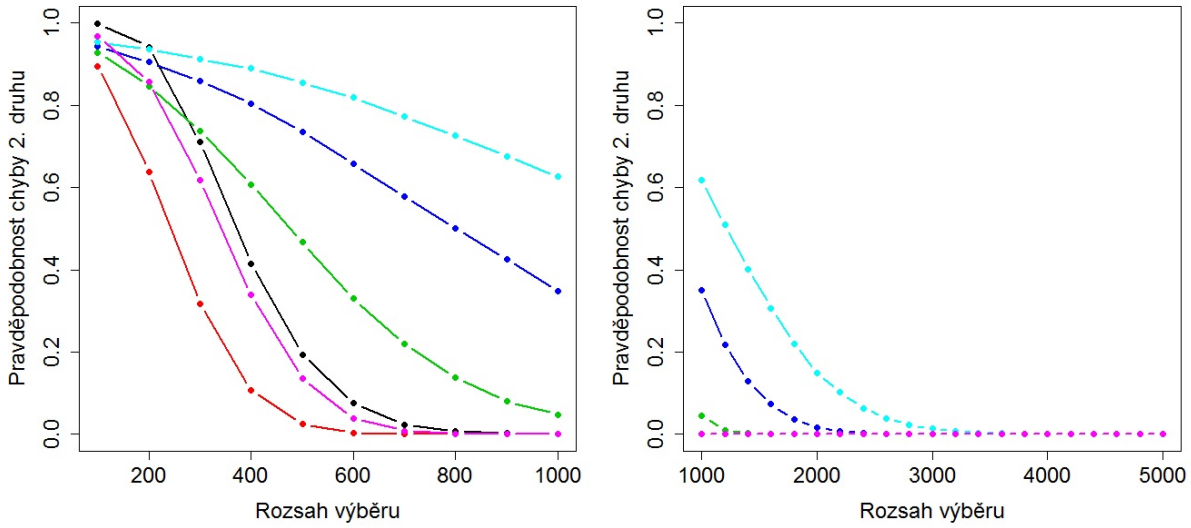
* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (29 rozsahů)

4.2.2. Trojúhelníkové rozdělení

Trojúhelníkové rozdělení obecně nemusí být symetrické. Abychom byli tvarem trojúhelníkového rozdělení blíže normálnímu rozdělení, zvolili jsme symetrické trojúhelníkové rozdělení Tri(0,2,1). Symetrické trojúhelníkové rozdělení bereme jako rozdělení poměrně podobné rozdělení normálnímu, jelikož testy normality vykazovaly vysokou pravděpodobnost chyby druhého druhu (převážně u malých rozsahů) viz Obr. 16. V tabulce Tabulka 3. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 16: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z Tri(0,2,1) rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy.



Obr. 17: Vývoj pravděpodobnosti chyby 2. druhu pro data generovaná z Tri(0,2,1) rozdělení u JB(černě), SW(červeně), AD(zeleně),CVM(modře), LILLIE(světle modře), SF(růžově) testu pro střední rozsahy(vlevo), velké rozsahy(vpravo).

Tabulka 3.: Pravděpodobnost chyby 2. druhu

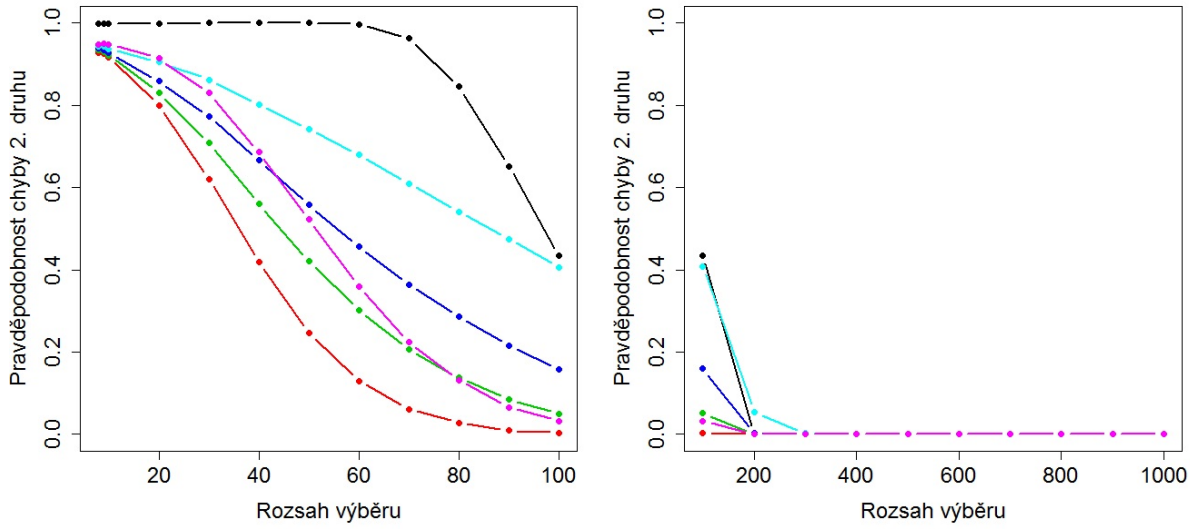
test $\|\|rozsah$	10	20	30	50	60	100	200	300 - 3600*	3800 - 5000**
JB	6.	6.	6.	6.	6.	6.	6.	3.	1.
SW	4.	4.	4.	2.	1.	1.	1.	1.	1.
AD	3.	3.	3.	1.	2.	2.	2.	4.	1.
CVM	2.	1.	1.	3.	3.	3.	4.	5.	1.
LILLIE	1.	2.	2.	4.	4.	4.	5.	6.	1.
SF	5.	5.	5.	5.	5.	5.	3.	2.	1.

* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (21 rozsahů)

** Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (7 rozsahů)

4.2.3. Rovnoměrné rozdělení

Jako další symetrické rozdělení uvádíme rovnoměrné rozdělení $Ro(0, 1)$. Toto rozdělení je sice symetrické, avšak svou špičatostí se výrazně liší od rozdělení normálního. V tabulce Tabulka 4. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 18: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z Ro(0,1) rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy(vlevo), střední rozsahy(vpravo).

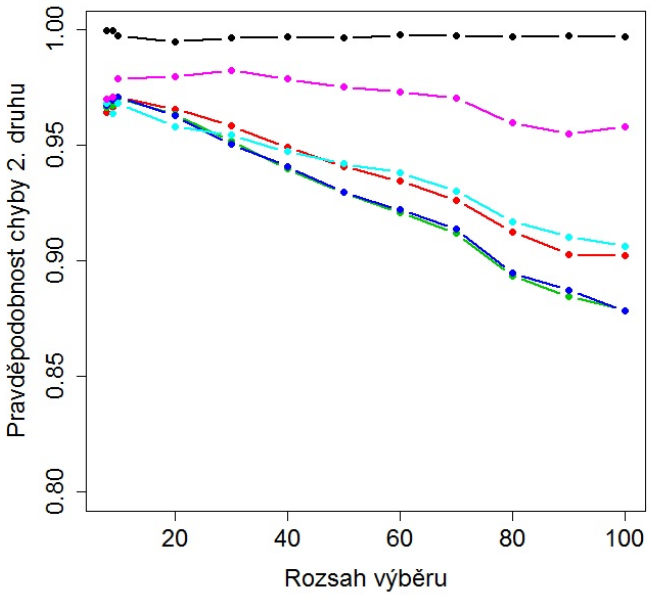
Tabulka 4.: Pravděpodobnost chyby 2. druhu

test $\|\|rozsah$	10	20	40	50	70	80	100	300 - 5000*
JB	6.	6.	6.	6.	6.	6.	6.	1.
SW	1.	1.	1.	1.	1.	1.	1.	1.
AD	2.	2.	2.	2.	2.	3.	3.	1.
CVM	3.	3.	3.	4.	4.	4.	4.	1.
LILLIE	4.	4.	5.	5.	5.	5.	5.	1.
SF	5.	5.	4.	3.	3.	2.	2.	1.

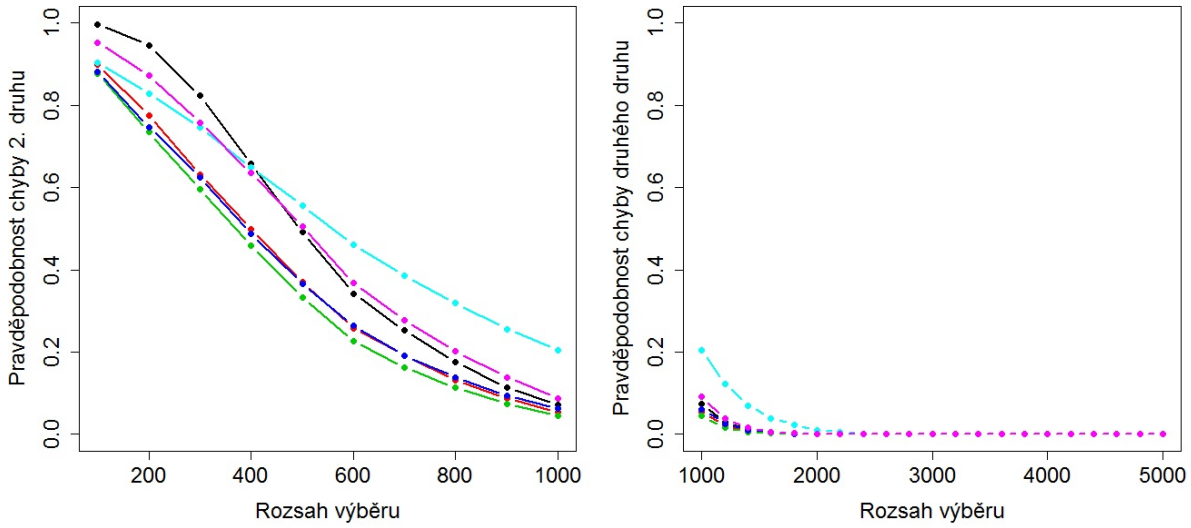
* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (28 rozsahů)

4.2.4. Směs dvou normálních rozdělení

Jako další rozdělení uvádíme směs rozdělení $N(0,1)$ a $N(2,1)$. Toto rozdělení je symetrické a také svou špičatostí velmi podobné rozdělení normálnímu. Testy normality vykazovaly velkou pravděpodobnost chyby druhého druhu u malých a středních rozsahů viz Obr. 19 a Obr. 20. Pokud by naše směs byla tvořena dvěma rozděleními, která by se *překrývala* méně (např. $N(0,1)$ a $N(4,1)$), testy normality by vykazovaly menší pravděpodobnost chyby druhého druhu. V tabulce Tabulka 5. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 19: Vývoj chyby druhého druhu pro data generovaná ze směsi N(0,1) a N(2,1) rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy .



Obr. 20: Vývoj chyby druhého druhu pro data generovaná ze směsi N(0,1) a N(2,1) rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro střední rozsahy(vlevo), velké rozsahy(vpravo).

Tabulka 5.: Pravděpodobnost chyby 2. druhu

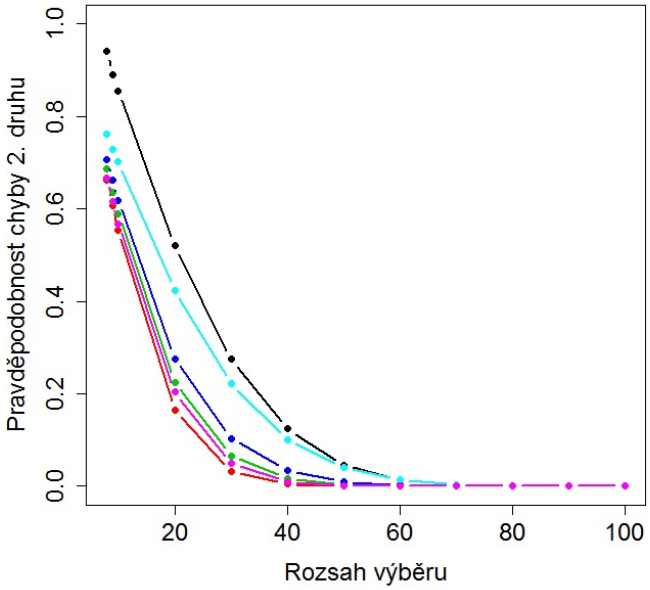
test $\|\|rozsah$	10	30	50	80	100	300	600	800 - 2400*	2600 - 5000**
JB	6.	6.	6.	6.	6.	6.	4.	4.	1.
SW	3.	4.	3.	3.	3.	3.	2.	2.	1.
AD	2.	3.	1.	1.	1.	1.	1.	1.	1.
CVM	4.	2.	1.	2.	1.	2.	3.	3.	1.
LILLIE	1.	1.	4.	4.	4.	5.	6.	6.	1.
SF	5.	5.	5.	5.	5.	4.	5.	5.	1.

* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (10 rozsahů)

** Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (13 rozsahů)

4.2.5. Exponenciální rozdělení

Exponenciální rozdělení $Exp(10)$ jsme volili, jelikož jeho tvar se výrazně liší od tvaru normálního rozdělení. Na Obr. 21 můžeme vidět, že testy normality poměrně brzy vykazují velmi malou pravděpodobnost chyby druhého druhu a od rozsahu $n = 70$ všechny testy mají nulovou pravděpodobnost chyby druhého druhu. V tabulce Tabulka 6. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 21: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z Exp(10) u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy.

Tabulka 6.: Pravděpodobnost chyby 2. druhu

test $\|rozsah$	8 - 50*	60	80 - 5000**
JB	6.	5.	1.
SW	1.	1.	1.
AD	3.	1.	1.
CVM	4.	4.	1.
LILLIE	5.	5.	1.
SF	2.	1.	1.

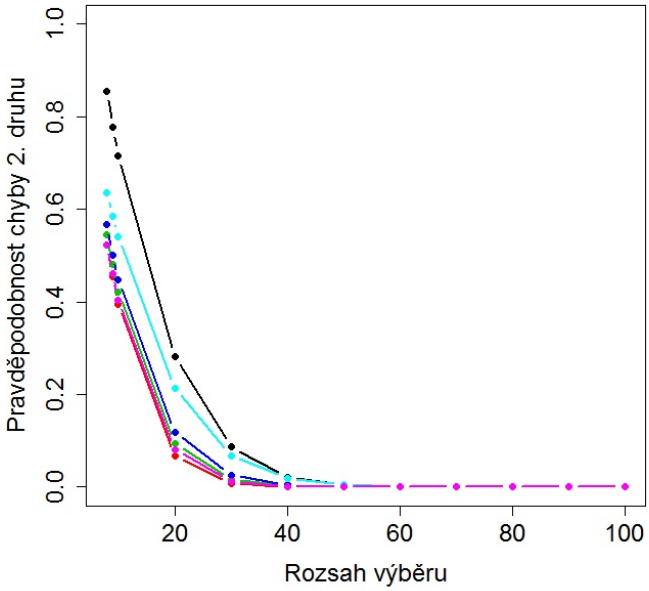
* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (7 rozsahů)

** Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (32 rozsahů)

4.2.6. Lognormální rozdělení

Lognormální rozdělení jsme volili, jelikož jeho tvar se výrazně liší od tvaru normálního rozdělení. Na Obr. 21 můžeme vidět, že testy normality poměrně brzy vykazují velmi malou pravděpodobnost chyby druhého druhu a od rozsahu $n = 60$ všechny testy mají nulovou pravděpodobnost chyby druhého druhu.

V tabulce Tabulka 7. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 22: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z Lnorm(0,1) rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro malé rozsahy.

Tabulka 7.: Pravděpodobnost chyby 2. druhu

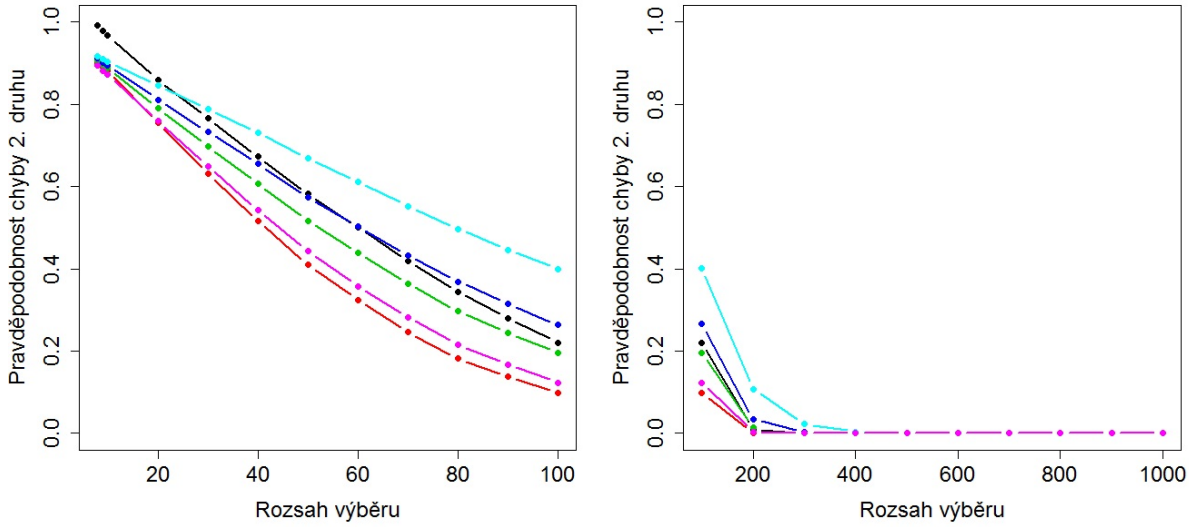
test $\|rozsah$	8	9 - 50*	60 - 5000**
JB	6.	6.	1.
SW	1.	1.	1.
AD	3.	3.	1.
CVM	4.	4.	1.
LILLIE	5.	5.	1.
SF	1.	2.	1.

* Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (6 rozsahů)

** Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (34 rozsahů)

4.2.7. Chí-kvadrát rozdělení

Chí-kvadrát rozdělení χ^2_{10} jsme volili, jelikož jeho tvar je méně podobný tvaru normálního rozdělení. Na Obr. 23 můžeme vidět, že pravděpodobnost chyby druhého druhu u všech šesti testů poměrně rychle klesala a už od rozsahu $n = 400$ testy mají nulovou pravděpodobnost chyby druhého druhu. V tabulce Tabulka 8. uvádíme seřazení námi zkoumaných testů od nejlepšího po nejhorší u vybraných rozsahů výběru na základě pravděpodobnosti chyby druhého druhu. Kde nejlepšímu testu přiřazujeme (1.), nejhoršímu (6.).



Obr. 23: Vývoj pravděpodobnosti chyby druhého druhu pro data generovaná z χ^2_{10} rozdělení u JB(černě), SW(červeně), AD(zeleně), CVM(modře), LILLIE(světle modře), SF(růžově) testu pro střední rozsahy(vlevo), velké rozsahy(vpravo).

Tabulka 8.: Pravděpodobnost chyby 2. druhu

test $\|rozsah$	8 - 10*	20	30 - 50**	60	70 - 200***	400 - 5000****
JB	6.	6.	5.	4.	4.	1.
SW	2.	1.	1.	1.	1.	1.
AD	3.	3.	3.	3.	3.	1.
CVM	4.	4.	4.	5.	5.	1.
LILLIE	5.	5.	6.	6.	6.	1.
SF	1.	2.	2.	2.	2.	1.

* - **** Pouze rozsahy uvedené v prvním odstavci kapitoly 4.1. (3, 3, 5, 27 rozsahů)

Závěr

V této práci jsme uvedli dva základní přístupy zkoumání normality. Prvním přístupem byly grafické metody, které doporučujeme použít jako startovací metody. Tyto poměrně jednoduché metody nám umožní vytvořit si alespoň hrubou grafickou představu o tvaru rozdělení našich dat. S použitím grafických metod můžeme zjistit například tvar rozdělení (jak jsou data šikmá, špičatá), zda soubor dat obsahuje odlehlé hodnoty, ... Druhým základním přístupem zkoumání normality jsou testy normality. Testy normality jsou stěžejní částí této diplomové práce. Zabývali jsme se šesticí testů v softwaru R : Shapirův-Wilkův test, Jarqueův-Berův test, Andersonův-Darlingův test, Cramér-von Misesův test, Lillieforsův test a Shapirův-Franciův test. Tyto testy jsou založeny na následujících myšlenkách: porovnání na základě kvantil-kvantil grafu, porovnání výběrové šikmosti a špičatosti s šikmostí a špičatostí normálního rozdělení, porovnání rozdílu (odchýlení) mezi empirickou a hypotetickou kumulativní distribuční funkcí. Jde tedy vidět, že testy normality často navazují na grafické metody. Při tvorbě této diplomové práce jsme došli k závěru, že *při zkoumání normality je vhodné využít obou základních přístupů*, tedy jak grafických nástrojů, tak i testů normality.

Dále jsme se zabývali vývojem pravděpodobnosti chyby prvního a druhého druhu u výše zmíněných testů normality. Na otázku „Který z testů normality je nelepší?“ nelze jednoznačně odpovědět. Během naší studie uvedené v kapitole 4. jsme ukázali, že neexistuje jeden konkrétní test, který by byl ve všech případech nejlepší. To je také důvodem, proč nemáme pouze jeden test normality, ale existuje jich více.

Na základě naší studie uvedené v kapitole 4. jsme vytvořili závěrečné porovnání zkoumaných šesti testů normality z hlediska pravděpodobnosti chyby prvního a druhého druhu. Závěrečné porovnání uvádíme v tabulce na následujích dvou stranách. Tato tabulka obsahuje pouze stěžejní rozsahy výběrů, vývoj u ostatních zkoumaných rozsahů je možno sledovat na obrázcích Obr.12 až Obr. 23 v kapitole 4.

Na základě studie uvedené v kapitole 4. jsme došli k závěru, že pro testování

normality bychom *nedoporučili* Shapirův-Franciův test. Tento test téměř na všech námi zvolených rozsazích překračoval zvolenou hladinu významnosti $\alpha = 0.05$. V případě, že se dopustíme chyby prvního druhu, jedná se o závažnější přečin, než když se dopustíme chyby druhého druhu. Nutno podotknout, že na úkor pravděpodobnosti chyby prvního druhu tento test dosahoval velmi dobrých výsledků u pravděpodobnosti chyby druhého druhu například u t-rozdělení, exponenciálního rozdělení, lognormálního rozdělení a chí-kvadrát rozdělení.

K testování normality bychom *doporučili* Shapirův-Wilkův test a Andersonův-Darlingův test normality. Shapirův-Wilkův test dopadl velmi dobře při zkoumání pravděpodobnosti chyby prvního druhu pro rozsahy do 2000. U větších rozsahů test nedosáhl na zvolenou hladinu významnosti viz tabulka na následující straně. Při studii pravděpodobnosti chyby druhého druhu se však SW test držel velmi často na prvním místě a v žádném případě nebyl velmi špatný.

Andersonův-Darlingův test dopadl výborně při zkoumání pravděpodobnosti chyby prvního druhu viz tabulka na následující straně. Při zkoumání pravděpodobnosti chyby druhého druhu se AD test držel především na 1.,2. a 3. pozici.

Jarqův-Berův test u rozsahů do 2000 nedosahoval s pravděpodobonstí chyby prvního druhu na zvolenou hladinu významnosti. U větších rozsahů to již bylo v pořádku. S pravděpodobností chyby druhého druhu se tento test velmi často držel na poslední pozici. Především u malých rozsahů tento test vykazoval velkou pravděpodobnost chyby druhého druhu. U rozsahů větších než 2000 byl na tom velmi dobře, jelikož tento test je konstruován jako asymptoticky optimální, tedy „nejlepší“ mezi testy při velkých výběrech.

Cramér-von Misesův test dopadl při zkoumání pravděpodobnosti chyby prvního druhu velmi dobře. S pravděpodobností chyby druhého druhu byl velmi často na 3. a 4. pozici, avšak několikrát se vyskytl i na 1. a 2. pozici. Z tohoto důvodu jej řadíme jako třetí nejlepší test v naší studii.

Lillieforsův test dosáhl na danou hladinu významnosti především u rozsahů $n \leq 600$. U větších rozsahů nedosahoval na zvolenou hladinu významnosti. S pravděpodobností chyby druhého druhu se držel především na posledních pozicích.

test $\|rozsah$	10	30	50	70	200	500	800	2000	3000	4000
SW $N(0,1)$	0	0	0	0	0	0	0	0	-	-
SW t_{20}	3.	2.	3.	3.	3.	3.	3.	3.	3.	3.
SW $Tri(0,2,1)$	4.	4.	2.	1.	1.	1.	1.	1.	1.	1.
SW směs	3.	4.	3.	3.	3.	3.	2.	2.	1.	1.
SW χ^2_{10}	2.	1.	1.	1.	1.	1.	1.	1.	1.	1.
SW $Ro(0,1)$	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
SW $Exp(10)$	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
SW $Lnorm(0,1)$	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
AD $N(0,1)$	0	0	0	0	0	0	0	0	0	0
AD t_{20}	2.	3.	4.	4.	4.	4.	4.	4.	4.	4.
AD $Tri(0,2,1)$	3.	3.	1.	2.	2.	4.	4.	4.	4.	1.
AD směs	2.	3.	1.	1.	1.	1.	1.	1.	1.	1.
AD χ^2_{10}	3.	3.	3.	3.	3.	1.	1.	1.	1.	1.
AD $Ro(0,1)$	2.	2.	2.	2.	1.	1.	1.	1.	1.	1.
AD $Exp(10)$	3.	3.	3.	1.	1.	1.	1.	1.	1.	1.
AD $Lnorm(0,1)$	3.	3.	1.	1.	1.	1.	1.	1.	1.	1.
CVM $N(0,1)$	0	+	0	0	0	0	0	0	0	0
CVM t_{20}	4.	4.	5.	5.	5.	5.	5.	5.	5.	5.
CVM $Tri(0,2,1)$	2.	1.	3.	3.	4.	5.	5.	5.	5.	1.
CVM směs	4.	2.	1.	2.	2.	2.	3.	3.	1.	1.
CVM χ^2_{10}	4.	4.	4.	5.	5.	1.	1.	1.	1.	1.
CVM $Ro(0,1)$	3.	3.	4.	4.	4.	1.	1.	1.	1.	1.
CVM $Exp(10)$	4.	4.	4.	5.	1.	1.	1.	1.	1.	1.
CVM $Lnorm(0,1)$	4.	4.	5.	1.	1.	1.	1.	1.	1.	1.
LILLIE $N(0,1)$	0	0	0	0	0	0	-	-	-	-
LILLIE t_{20}	5.	6.	6.	6.	6.	6.	6.	6.	6.	6.
LILLIE $Tri(0,2,1)$	1.	2.	4.	4.	5.	6.	6.	6.	6.	1.
LILLIE směs	1.	1.	4.	4.	4.	6.	6.	6.	1.	1.
LILLIE χ^2_{10}	5.	6.	6.	6.	6.	1.	1.	1.	1.	1.
LILLIE $Ro(0,1)$	4.	5.	5.	5.	6.	1.	1.	1.	1.	1.
LILLIE $Exp(10)$	5.	5.	5.	6.	1.	1.	1.	1.	1.	1.
LILLIE $Lnorm(0,1)$	5.	5.	6.	1.	1.	1.	1.	1.	1.	1.
JB $N(0,1)$	-	-	-	-	-	-	-	-	0	0
JB t_{20}	6.	5.	2.	2.	1.	1.	1.	1.	1.	1.
JB $Tri(0,2,1)$	6.	6.	6.	6.	6.	3.	3.	3.	3.	1.
JB směs	6.	6.	6.	6.	6.	4.	4.	4.	1.	1.
JB χ^2_{10}	6.	5.	5.	4.	4.	1.	1.	1.	1.	1.
JB $Ro(0,1)$	6.	6.	6.	6.	1.	1.	1.	1.	1.	1.
JB $Exp(10)$	6.	6.	6.	1.	1.	1.	1.	1.	1.	1.
JB $Lnorm(0,1)$	6.	6.	1.	1.	1.	1.	1.	1.	1.	1.

test $\|rozsah$	10	30	50	70	200	500	800	2000	3000	4000
SF $N(0,1)$	+	+	+	+	+	0	+	+	+	+
SF t_{20}	1.	1.	1.	1.	2.	2.	2.	2.	2.	2.
SF $Tri(0,2,1)$	5.	5.	5.	5.	3.	2.	2.	2.	2.	1.
SF směs	5.	5.	5.	5.	5.	5.	5.	5.	1.	1.
SF χ^2_{10}	1.	2.	2.	2.	2.	1.	1.	1.	1.	1.
SF $Ro(0,1)$	5.	4.	3.	3.	1.	1.	1.	1.	1.	1.
SF $Exp(10)$	2.	2.	1.	1.	1.	1.	1.	1.	1.	1.
SF $Lnorm(0,1)$	2.	2.	1.	1.	1.	1.	1.	1.	1.	1.

0 test dosahuje na hladinu významnosti

\+ test překračuje hladinu významnosti

\- test nedosahuje na hladinu významnosti

Uvádíme hodnocení 1.-6., kde 1. je nejlepší a 6. nejhorší výsledek.

Literatura

[1] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha 2005.

[2] Daw, R.H.: *Why the normal distribution*, JSS 18 (1) (1966) 2-15, citováno dne 6.1.2013

[3] Normam L. Johnson, Samuel Kotz, N. Balakrishnan: *Continuous univariate distributions*, 2. vydání

[4] Falk M.: *Comparison of Common Tests for Normality*, Julius-Maximilians-Universität Würzburg, August 2010

[5] Aproximace binomického rozdělení: http://mathworld.wolfram.com/BinomialDistribution.html, citováno dne: 6.1.2013

[6] Herbert A. Sturges: *The Choice of a Class Interval*, Journal of the American Statistical Association, Vol. 21, No. 153 (Mar., 1926), pp. 65-66., JSTOR, citováno dne: 13.1. 2013

[7] ECDF v softwaru R: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/ecdf.html, citováno dne: 6.1. 2013

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

[9] Royston, J.P., 1982b. *An extension of the Shapiro and Wilk's W test for normality to large samples.* Applied Statistics, 31, 115-124.

[10] Royston, J.P., 1982a. Algorithm AS 181: *The W test for normality.* Applied Statistics, 31, 176- 180.Royston, J.P., 1983a. Correction: Algorithm AS 181: The W test for normality. Applied Statistics,32, 224.

[11] Royston, J.P., 1983b. *Some techniques for assessing multivariate normality based on the Shapiro-Wilk W.* Applied Statistics, 32, 121-133.

[12] Royston, J.P., 1992. *Approximating the Shapiro-Wilk W-Test for non-normality.* Statistics and Computing, 2, 117-119.

[13] Royston, J.P., 1993. *A Toolkit for testing for non-normality in complete and censored samples*. The Statistician, 42, 37-43.

[14] Royston, J.P., 1995. Remark AS R94: A remark on Algorithm AS 181: *The W test for normality*. Applied Statistics, 44, 547-551.

[15] Shapiro, S.S., Wilk M.B.: *An analysis of Variance Test for Normality (Complete Samples)*, Biometrika, Vol. 52, No 3/4 (Dec.,1965), 591-611. JSTOR

[16] Rost, M., 2007. *Statistické metody v ekonomii*. Jihočeská univerzita v Českých Budějovicích.

[17] Koeficienty Shapirova-Wilkova testu: *Coefficients de Shapiro-Wilk*, www.math.unicaen.fr/.../tables-valuers.pdf, citováno dne: 10.2. 2013

[18] Kritické hodnoty Shapirova-Wilkova testu: http://www.kmt.zcu.cz/person/Kohout/infosoubory/letnisem/tabulky.htm citováno dne: 17.3. 2013

[19] Anil K. Bera, Pin T. Ng: *A large sample normality test*, Department of Economics, University of Illinois, Champaign, IL 31820, Department of Economics, University of Houston, TX 77204-5882, November 22, 1993

[20] Thorsten Thadewald, Herbert Bning: *Jargue-Bera Test and its Competitors for Testing Normality – A Power Comparison*, Journal of Applied Statistics, Vol. 34, No. 1, 87-105, January 2007, Institute for Statistics and Econometrics, Free University Berlin, Germany

[21] Gross, J.: *Tests for Normality*, Version 1.0-2, Date-05-30, citováno dne: 17.3. 2013

[22] Stephens, M.A. (1974): *EDF statistics for goodness of fit and some comparisons*. Journal of the American Statistical Association, 69, 730–737.

[23] Dallal, G.E. and Wilkinson, L. (1986): *An analytic approximation to the distribution of Lilliefors' test for normality.* The American Statistician, 40, 294–296.

[24] Budíková, M. (2007): *Teorie pravděpodobnosti a matematická statistika*, Masarykova univerzita, 3. vydání, Brno.

[25] Royston, P. 1993. *A pocket-calculator algorithm for the Shapiro-Francia W test for non-normality*: an application to medicine. Statistics in Medicine 12: 181–184.

[26] Thode, H. C., Jr.: *Testing for normality*, Marcel Dekker, New York 2002