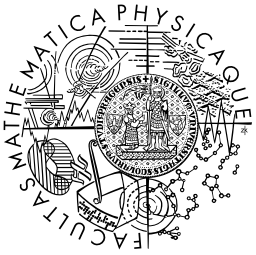


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Tomáš Rajtmajer

Zamítací metoda pro generování vzorků ze složitých rozdělení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jiří Dvořák, Ph.D.
Studijní program: Finanční matematika
Studijní obor: Finanční matematika

Praha 2024

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

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Rád bych poděkovat vedoucímu své bakalářské práce, RNDr. Jiřímu Dvořákovi, Ph.D. za cenné rady, pomoc a obětovaný čas.

Název práce: Zamítací metoda pro generování vzorků ze složitých rozdělení

Autor: Tomáš Rajtmajer

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jiří Dvořák, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Nejjednodušší a nejběžnější způsob generování vzorků z daného rozdělení je inverzní metoda. Tato metoda však používá inverzní funkci k distribuční funkci daného rozdělení. Proto ji nelze použít v případech, kdy tuto inverzní funkci nejsme schopni získat. V takovém případě lze použít zamítací metodu generování vzorků. Bakalářská práce se zabývá generováním vzorků ze složitých rozdělení pomocí zamítací metody. Cílem je představit tuto metodu a popsat její fungování. V praktické části použijeme tuto metodu k získání vzorků z rozdělení s hustotou danou tvarem hory Říp a pomocí testů normality se pokusíme se dokázat, že Říp není normální.

Klíčová slova: simulace, zamítací metoda, Monte Carlo metody, testování normality

Title: Rejection sampling

Author: Tomáš Rajtmajer

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jiří Dvořák, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The simplest and most common way to generate samples from a given distribution is the inverse transform sampling method. Since this method uses the inverse function of the distribution's cumulative distribution function, it cannot be used in cases where it is impossible to obtain this inverse function. In such cases, we can use the rejection sampling method. This thesis focuses on generating samples from complex distributions using the rejection sampling method. The goal is to introduce this method and describe how it works. In the practical part, we will apply this method to obtain samples from the distribution with the density defined by the shape of Říp Mountain. Using normality tests, we will attempt to demonstrate through that Říp does not have normal distribution.

Keywords: simulation, rejection sampling, Monte Carlo methods, normality testing

Obsah

Úvod

K získání vzorku z nějakého pravděpodobnostního rozdělení lze použít více způsobů. Nejjednodušší a zároveň i nejvíce využívaný způsob je inverzní metoda generování vzorků. Tato metoda je založena na dosazování do inverzní funkce k distribuční funkci cílového rozdělení. Jsou ale případy, kdy máme rozdělení dané hustotou se složitým předpisem, ze kterého je složité nebo nemožné získat inverzní funkci k distribuční funkci tohoto rozdělení. V takových případech lze použít zamítací metodu pro generování vzorků z rozdělení.

V první kapitole bakalářské práce si představíme inverzní metodu generování vzorků a dokážeme, že výstupem této metody je skutečně vzorek z cílového rozdělení.

V další kapitole se již zaměříme na samotnou zamítací metodu generování vzorků. Uvedeme a dokážeme si zde věty, ze kterých zamítací metoda vychází. Představíme si algoritmus zamítací metody a popíšeme jeho fungování. Pojednáme o volbě pomocné funkce a konstanty v algoritmu zamítací metody. Dále v práci uvedeme variaci zamítacího algoritmu, obálkovou metodu (anglicky the squeeze principle). Algoritmus zamítací metody je demonstrován na příkladu společně s různou volbou pomocné funkce.

Na závěr práce použijeme zamítací metodu k vygenerování vzorků z rozdělení s hustotou danou obrázkem hory Říp. Na získané vzorky aplikujeme testy normality a pokusíme se tak dokázat, že rozdělení s hustotou danou tvarem hory Říp není normální.

Mým přínosem v práci je rozšíření důkazů, ze kterých vychází inverzní a zamítací metoda. Vytvoření příkladu fungování zamítací metody a implementace zamítacího algoritmu v programu R. Poslední kapitola, ve které používáme zamítací metodu pro generování vzorků z rozdělení s hustotou danou tvarem hory Říp a následně testujeme normalitu těchto vzorků, je také mým přínosem.

1. Inverzní metoda

Inverzní metoda je popsána v knize Devroye (1986, str. 30) jako způsob generování realizace náhodné veličiny z rozdělení s libovolnou spojitou distribuční funkcí, a to za předpokladu, že její inverzní funkce F_X^{-1} je explicitně známá.

1.1 Princip inverzní metody

Věta 1. *(Devroye (1986, str. 28)) Nechť F_X je spojitá distribuční funkce na $\mathbb{R}$ s inverzní funkcí F_X^{-1} definovanou následovně:*

$$F_X^{-1}(u) = inf\{x : F(x) = u\},\ 0 < u < 1.$$

Pokud U je náhodná veličina z rovnoměrného rozdělení na intervalu [0,1]*, pak $F_X^{-1}(U)$ má distribuční funkci F_X. Navíc pokud X má distribuční funkci F_X, pak $F_X(X)$ má rovnoměrné rozdělení na intervalu* [0,1].

Důkaz. Nejdříve dokážeme první část tvrzení. Chceme najít distribuční funkci funkce F_X^{-1}, tedy chceme $Y \equiv F_{F_X^{-1}(U)}(x)$. Z definice distribuční funkce víme, že

$$Y = P(F_X^{-1}(U) \leq x)$$

pro všechna $x \in \mathbb{R}$. Protože F_X je spojitá a neklesající, lze aplikovat F_X na obě strany nerovnosti.

$$P(F_X^{-1}(U) \leq x) = P(F_X(F_X^{-1}(U)) \leq F(x)) = P(U \leq F_X(x)) = F_X(x)$$

pro všechna $x \in \mathbb{R}$. Poslední rovnost vychází z definice distribuční funkce, která říká, že $F_X(x) = P(X \leq x)$ pro všechna $x \in \mathbb{R}$. Rovnost platí, zvolíme-li za x libovolné číslo, tedy platí i pro $F_X(x)$.

Pro důkaz druhé části tvrzení chceme zjistit $F_{F_X}(u)$. Vycházíme opět z definice distribuční funkce, dále ze spojitosti distribuční funkce a z faktu, že $0 < u < 1$:

$$\begin{aligned} F_{F_X(X)}(u) &= P(F_X(X) \leq u) = P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(u)) = \\ &= P(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u \end{aligned}$$

□

Algoritmus inverzní metody pro generování realizace náhodné veličiny z rozdělení vychází z věty 1. Algoritmus funguje následujícím způsobem: Nejdříve vygenerujeme realizaci náhodné veličiny U z rovnoměrného rozdělení na intervalu [0,1]. Poté dosadíme U do F_X^{-1}, tedy $Y = F_X^{-1}(U)$, kde Y je náhodná veličina z rozdělení s distribuční funkcí F_X.

Inverzní metoda je přesná, pokud známe explicitní tvar F_X^{-1}. Co když ale přesný tvar této funkce neznáme? V takovém případě lze rovnici $F_X(X) = U$ řešit numericky. Nalezené řešení však nebude přesné a vygenerovaná realizace náhodné veličina nebude mít přesně požadované rozdělení. Při vhodné volbě numerické metody řešení inverzní úlohy však můžeme najít řešení s libovolnou požadovanou přesností.

2. Zamítací metoda

Jak je popsáno v knize Robert a Casella (2004, str. 47) Existuje spousta složitých rozdělení, pro které je obtížné či nemožné získat náhodnou veličinu pomocí inverzní metody. V některých případech dokonce není možné reprezentovat rozdělení v použitelné formě, jako je například transformací z jiného rozdělení nebo směsí rozdělení. V takových případech nelze simulovat náhodný výběr přímo. Jedna z možností, jak v takových případech postupovat je zamítací metoda.

Zamítací metoda využívá jednodušší rozdělení k získání náhodného výběru z cílového (složitého) rozdělení. Tuto metodu lze použít, pokud neznáme inverzní funkci k distribuční funkci cílového rozdělení. Nevyžaduje totiž její podrobnou znalost.

2.1 Potřebné vlastnosti hustoty

Následující věty ukazují vlastnosti hustoty, ze kterých vychází zamítací metoda.

Věta 2. *(Devroye (1986, str. 40)) Nechť X je náhodný vektor s hustotou f na $\mathbb{R}^d$, nechť U je náhodná veličina z rovnoměrného rozdělení na intervalu [0,1] nezávislá na X. Pak $(X,c\,Uf(X))$ má rovnoměrné rozdělení na $A = \{(x,u) : x \in \mathbb{R}^d, 0 \leq u \leq cf(x)\}$, kde $c > 0$ je konstanta. Naopak když (X,U) je náhodný vektor v $\mathbb{R}^{d+1}$ rovnoměrně rozdělený na A, pak X má hustotu f na $\mathbb{R}^d$.*

Důkaz. Pro důkaz první části nechť $S \subseteq A$ a nechť S_x je řez S u x , tedy $S_x = \{u : (x,u) \in S\}$. Označme $f_{(X,cUf(X))}(x,y) \equiv g(x,y)$ a $f_{cUf(X)|X}(y|x) \equiv g(y|x)$.

$$P((X,cUf(X)) \in S) = \int_S f_{(X,cUf(X))}(x,y)\ \mathrm{d}(x,y) = \int_S g(x,y)\ \mathrm{d}(x,y).$$

Z definice podmíněné hustoty a Tonelliho věty můžeme psát:

$$\begin{aligned}\int_S g(x,y)\ \mathrm{d}(x,y) &= \int_{\mathbb{R}^d} \left(\int_{S_x} f_{cUf(X)|X}(y|x)\ \mathrm{d}y \right) f(x)\ \mathrm{d}x = \\ &= \int_{\mathbb{R}^d} \left(\int_{S_x} g(y|x)\ \mathrm{d}y \right) f(x)\ \mathrm{d}x\end{aligned}$$

Víme, že $U \sim R[0,1]$. Z toho plyne, že $cUf(X)|X \sim R[0,cf(X)]$, protože $c > 0$ je konstanta a díky podmínění je $f(X)$ také konstanta. Hustotu $g(y|x)$ lze přepsat jako:

$$g(y|x) = \frac{1}{cf(x) - 0} \mathbb{1}_{[0,cf(x)]}(y).$$

Celkem tedy máme:

$$\begin{aligned}P((X,cUf(X)) \in S) = \int_S g(x,y)\ \mathrm{d}(x,y) &= \int_{\mathbb{R}^d} \left(\int_{S_x} g(y|x)\ \mathrm{d}y \right) f(x)\ \mathrm{d}x = \\ &= \int_{\mathbb{R}^d} \left(\int_{S_x} \frac{1}{cf(x)} \mathbb{1}_{[0,cf(x)]}(y)\ \mathrm{d}y \right) f(x)\ \mathrm{d}x = \frac{1}{c} \int_S \mathrm{d}y\ \mathrm{d}x.\end{aligned}$$

Poslední rovnost vychází z toho, že $S \subseteq A = \{(x,u) : x \in \mathbb{R}^d, 0 \leq u \leq cf(x)\}$, kde $c > 0$ je konstanta. S_x je řez S. Indikátor tak můžeme vynechat. Protože obsah A je c, dokázali jsme první část věty.

V druhé části chceme dokázat, že pro všechny borelovské množiny B z $\mathbb{R}^d$ platí, že

$$P(X \in B) = \int_B f(x)\ \mathrm{d}x.$$

Označme $B_1 \equiv \{(x,u) : x \in B, 0 \leq u \leq cf(x)\}$. Protože náhodný vektor (X,U) je rovnoměrně rozdělený na A, tak platí:

$$P(X \in B) = P((X,U) \in B_1) = \int_{B_1} f_{(X,U)}(x,u)\ \ \mathrm{d}(x,u) =$$
$$= \int_{B_1} \mathbb{1}_A(x,u)\frac{1}{|A|}\ \ \mathrm{d}(x,u) = \int_{B_1} \frac{1}{|A|}\ \ \mathrm{d}(x,u) = \frac{|B_1|}{|A|}.$$

Předposlední rovnost vychází z toho, že $B_1 \subset A$, $\mathbb{1}_A(x,u)$ tedy bude 1 pro $(x,u) \in B_1$. Dále pomocí Tonelliho věty a definici B_1 máme:

$$|A| = \int_A 1\ \mathrm{d}(x,y) = \int_{\mathbb{R}} \int_{A_x} 1\ \mathrm{d}y\ \mathrm{d}x = \int_{\mathbb{R}} cf(x)\ \mathrm{d}x = c\int_{\mathbb{R}} f(x)\ \mathrm{d}x = c$$
$$|B_1| = \int_B \int_{B_{1_x}} 1\ \mathrm{d}y\ \mathrm{d}x = \int_B cf(x)\ \mathrm{d}x.$$

Díky tomu můžeme psát:

$$P(X \in B) = \frac{|B_1|}{|A|} = \frac{1}{c}\int_B cf(x)\ \mathrm{d}x = \int_B f(x)\ \mathrm{d}x,$$

což jsme chtěli dokázat.

□

Věta 3. *(Devroye (1986, str. 41)) Nechť $X_1, X_2, \ldots$ je posloupnost nezávislých, stejně rozdělených náhodných vektorů nabývajících hodnot v $\mathbb{R}^d$. Dále nechť $A \subseteq \mathbb{R}^d$ je borelovská množina taková, že $P(X_1 \in A) = p > 0$. Nechť Y je první X_i, které nabývá hodnot v A. Potom Y má rozdělení určené předpisem:*

$$P(Y \in B) = \frac{P(X_1 \in A \cap B)}{p},$$

kde B je borelovská množinu R^d.

Speciálně, když X_1 je rovnoměrně rozdělená v A_0, kde $A_0 \supseteq A$, pak Y je rovnoměrně rozdělená v A.

Důkaz. $\{(Y = X_i), i \in \mathbb{N}\}$ je úplný systém jevů. Pro libovolnou borelovskou množinu B si můžeme všimnout, že

$$P(Y \in B) = \sum_{i=1}^{\infty} P(Y \in B, Y = X_i) = \sum_{i=1}^{\infty} P(X_1 \notin A, \ldots, X_{i-1} \notin A, X_i \in B \cap A) =$$
$$= \sum_{i=1}^{\infty}(1-p)^{i-1}P(X_1 \in A \cap B) = \frac{1}{1-(1-p)}P(X_1 \in A \cap B) =$$
$$= \frac{P(X_1 \in A \cap B)}{p},$$

což jsme chtěli dokázat.

Pokud X_1 je rovnoměrně rozdělená v A_0, pak:

$$p = P(X_1 \in A) = \int_A \frac{\mathbb{1}_{A_0}(x,u)}{|A_0|} \, \mathrm{d}(x,u) = \frac{|A \cap A_0|}{|A_0|}.$$

Chceme dokázat, že

$$P(Y \in B) = \int_B \frac{\mathbb{1}_A(x,u)}{|A|} \, \mathrm{d}(x,u) = \frac{|B \cap A|}{|A|}.$$

Z již dokázané části máme:

$$P(Y \in B) = \frac{P(X_1 \in A \cap B)}{P(X_1 \in A)} = \frac{\frac{|A_0 \cap A \cap B|}{|A_0|}}{\frac{|A_0 \cap A|}{|A_0|}} = \frac{|A_0 \cap A \cap B|}{|A_0 \cap A|} = \frac{|A \cap B|}{|A|}.$$

Poslední rovnost vychází z toho, že $A_0 \supseteq A$.

□

2.2 Algoritmus zamítací metody

Mějme rozdělení s hustotou f, ze kterého chceme vygenerovat realizaci náhodné veličiny. K tomu použijeme pomocné rozdělení s hustotu g, ze kterého můžeme snadno generovat realizaci náhodné veličiny. Dále budeme potřebovat konstantu $c \geq 1$ takovou, aby pro všechna $x \in \mathbb{R}$ platila nerovnost $f(x) \leq cg(x)$.

Algoritmus zamítací metody funguje následujícím způsobem:

1. Vygenerujeme realizaci náhodné veličiny X z pomocného rozdělení s hustotou g.

2. Vygenerujeme realizaci náhodné veličiny U z rovnoměrného rozdělení na intervalu [0,1].

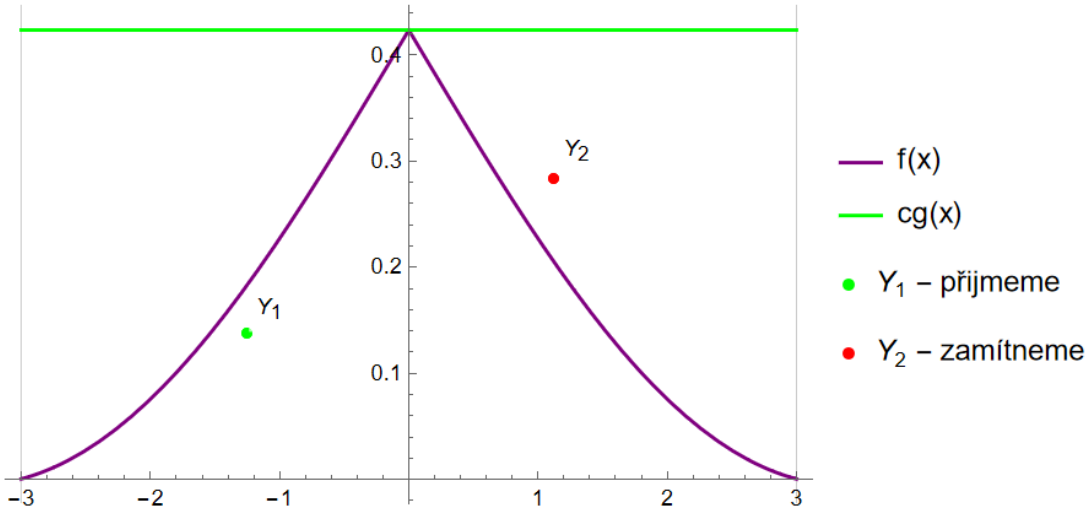
3. Porovnáme $Y \equiv cUg(X)$ s $f(X)$. Pokud je $Y > f(X)$, hodnotu X zamítneme a vracíme se k bodu 1. V opačném případě hodnotu přijmeme jako vzorek z cílového rozdělení.

Na obrázku 2.1 máme pro Y_1: $Y_1 = cU_1g(X_1) \leq f(X_1)$. Hodnotu X_1 přijmeme jako vzorek z cílového rozdělení. Pro Y_2 máme: $Y_2 = cU_2g(X_2) > f(X_2)$. Hodnotu X_2 zamítneme.

Jako pomocnou hustotu g většinou volíme hustotu rozdělení, ze kterého lze snadno generovat realizaci náhodné veličiny standardními dostupnými metodami. Například hustotu nějakého známého rozdělení. Díky tomu jsme schopni snadno vygenerovat realizaci náhodné veličiny. Při volbě pomocné funkce g se také musíme ujistit, že $\frac{g(X)}{f(X)}$ lze snadno spočítat.

Pokud chceme, aby byl algoritmus co nejefektivnější, je třeba konstantu c zvolit co nejmenší. Stále však musí platit nerovnost $f(X) \leq cg(X)$.

Formálně můžeme algoritmus zamítací metody zapsat takto (Devroye (1986, str. 42)):



Obrázek 2.1: Zamítnutí a přijmutí vzorku

Algoritmus 1 Zamítací metoda

repeat

Generuj dvě nezávislé veličiny: X (s hustotou g na $\mathbb{R}^d$) a U (rovnoměrně rozdělená na [0,1]).

Označ $T \leftarrow c\frac{g(X)}{f(X)}$

until $UT \leq 1$

return X

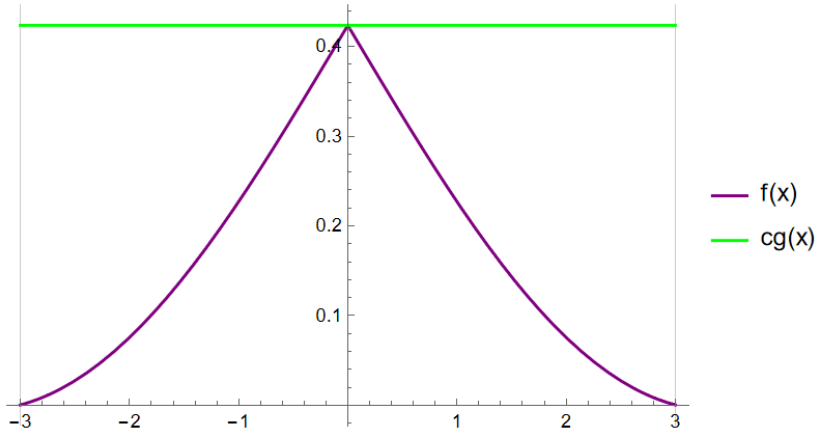
Z vět 2 a 3 vyplývá, že algoritmem získáme realizaci náhodné veličiny z rozdělení s hustotou f na $\mathbb{R}^d$. První část věty 2 říká, že pro X a U z algoritmu platí, že $(X,cUg(X))$ má pod křivkou cg rovnoměrné rozdělení v $\mathbb{R}^{d+1}$. Z věty 3 víme, že $(X,cUg(X))$ z tohoto algoritmu má rovnoměrné rozdělení pod křivkou f. Z druhé části věty 2 pak máme, že projekce $(X,cUg(X)) \to X$ do d-dimenzí má hustotu f.

2.3 Volba majorizující funkce

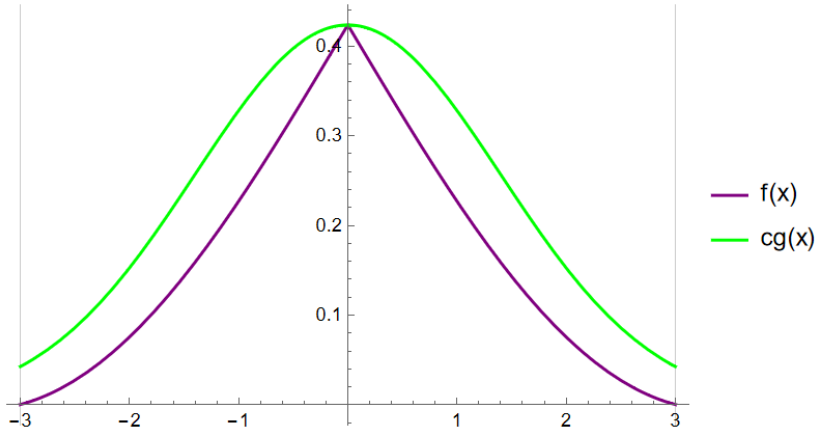
Efektivita algoritmu zamítací metody závisí na volbě majorizující funkce $cg(x)$. Při správné volbě majorizující funkce lze efektivitu algoritmu značně zvýšit.

Obrázek 2.2 ukazuje použití rovnoměrného rozdělení jako majorizující hustotu $g(x)$ (vynásobenou konstantou c). V takovém případě by byl algoritmus pro naši cílovou funkci $f(x)$ značně neefektivní a docházelo by k poměrně častému zamítání.

V našem případě je tedy mnohem vhodnější požít jako $g(x)$ hustotu normálního rozdělení (obrázek 2.3) v tomto případě nebude docházet k tak častému zamítání a algoritmus bude efektivnější.



Obrázek 2.2: Použití hustoty rovnoměrného rozdělení v majorizující funkci



Obrázek 2.3: Použití hustoty normálního rozdělení v majorizující funkci

2.4 Obálková metoda (the squeeze principle)

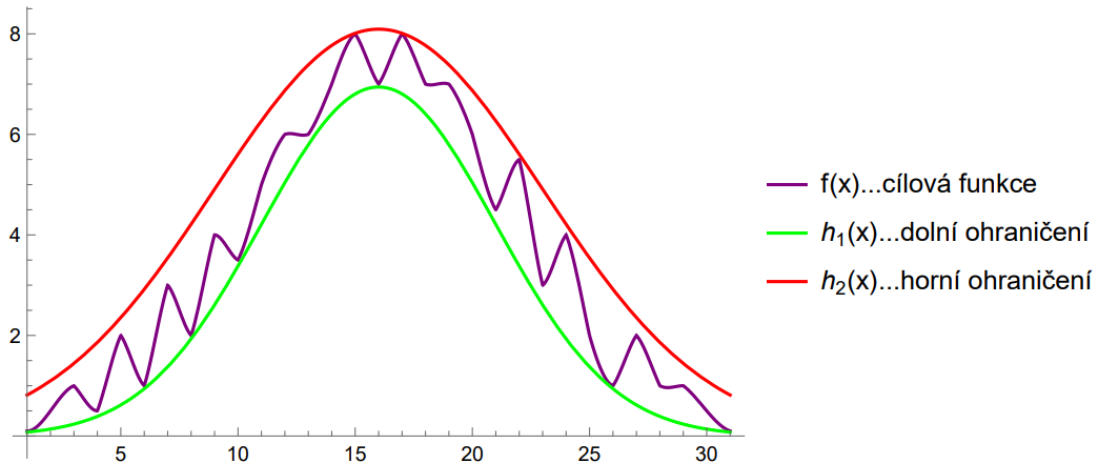
V základním algoritmu zamítací metody je potřeba neustále vyhodnocovat poměr $c\frac{g(X)}{f(X)}$. V některých případech má naše cílová funkce f složitý předpis. Abychom se tedy vyhnuli složitému vyhodnocování, můžeme použít obálkovou metodu (the squeeze principle).

Abychom se vyhnuli vyhodnocování funkce $f(X)$, používáme v algoritmu dvě pomocné funkce, $h_1(X)$ a $h_2(X)$. Těmito funkcemi si ohraničíme funkci $f(X)$.

Pokud $W \leq h_1(X)$, pak z předpokladu $h_1(X) \leq f(X)$ pro všechna x víme, že $W \leq f(X)$. Hodnotu tedy můžeme přijmout i bez vyhodnocování $f(X)$. Pokud $W \geq h_2(X)$, pak z předpokladu $h_2(X) \geq f(X)$ pro všechna x víme, že $W \geq f(X)$. Hodnotu tedy můžeme zamítnout i bez vyhodnocování $f(X)$. Náročnější porovnání W a $f(X)$ provádíme pouze v případě, kdy $h_1(X) \leq W \leq h_2(X)$.

Funkce $f(x)$ na obrázku 2.4 má složitý průběh a mohlo by být výpočetně náročné ji vyhodnocovat v jednotlivých bodech. Abychom nemuseli funkci $f(x)$ vyhodnocovat pokaždé, lze ji ohraničit funkcemi $h_1(x)$ a $h_2(x)$, které vznikly z hustot normálního rozdělení přenásobením konstantou. Jejich předpis tak lze snadno vyhodnotit.

Mějme funkce h_1 a h_2, které lze snadno vyčíslit a splňují nerovnost $h_1(x) \leq f(x) \leq h_2(x)$ pro všechna x. Formální zápis algoritmu je následující (Devroye



Obrázek 2.4: Obálková metoda

(1986), str. 54):

Algoritmus 2 Obálková metoda

repeat
Generuj realizaci náhodné veličiny U z rovnoměrného rozdělení na [0,1].
Generuj realizaci náhodné veličiny X z rozdělení s hustotou g.
Označ $W \leftarrow Ucg(X)$.
if $W \leq h_1(X)$ **then** *přijmi*
else
if $W \leq h_2(X)$ **then**
if $W \leq f(X)$ **then** *přijmi*
end if
end if
end if
until *přijmi*
return X

Obálkovou metodu lze použít k urychlení výpočtu. V současné době však není problém na počítačích provádět složité výpočty. Ve velkém množství případů je dostačující použít základní zamítací metodu. Obálková metoda měla své využití spíše v dřívějších dobách, kdy počítače nebyly tolik výkonné a bylo potřeba každý výpočet provádět co nejefektivněji.

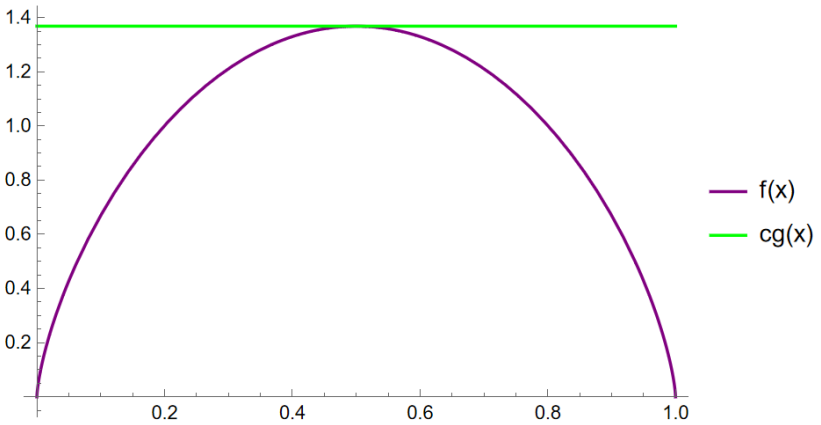
2.5 Příklad

Mějme beta rozdělení s parametry $\alpha = 1{,}7$ a $\beta = 1{,}7$, ze kterého budeme chtít získat náhodnou veličinu. Hustota tohoto rozdělení bude $f(x)$. Jako majorizující funkci $cg(x)$ zvolíme hustotu rovnoměrného rozdělení vynásobenou konstantou c tak, aby platila nerovnost $f(x) \leq cg(x)$.

Algoritmus zamítací metody lze v programu R implementovat následujícím způsobem:

```
a=1.7
b=1.7
```

```
c=dbeta(.5,a,b)
u=2
t=1
while(u*t>1) {     #ekvivalentní příkazu DOKUD u*t<=1
x=runif(1,0,1)     #realizace náhodné veličiny X
u=runif(1,0,1)     #realizace náhodné veličiny U
t=c*(dunif(x,0,1)/dbeta(x,a,b))
}
x
```



Obrázek 2.5: Příklad - beta rozdělení

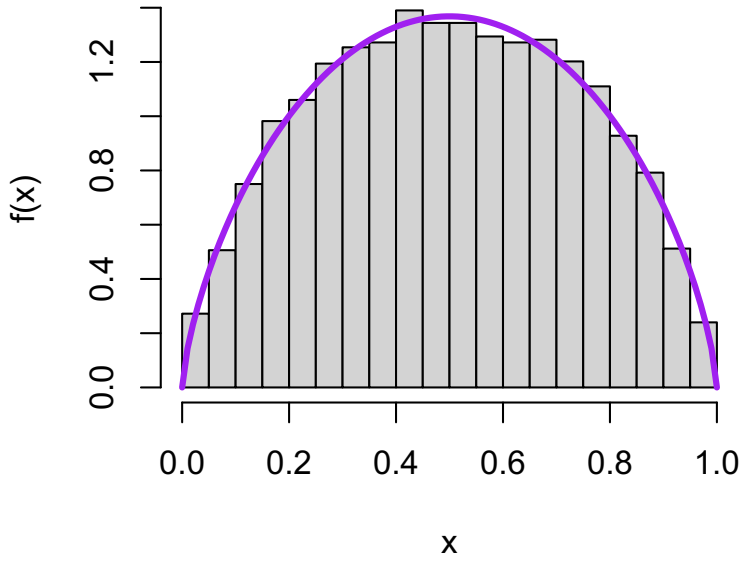
S námi zvolenými parametry je hustota beta rozdělení symetrická kolem bodu 0,5 a unimodální (obrázek 2.5). Za konstantu c lze tedy zvolit hodnotu hustoty v bodě 0,5.

Na obrázku 2.6 můžeme vidět histogram hodnot X vygenerovaných algoritmem zamítací metody po 10000 opakování. Tvar histogramu přibližně odpovídá grafu hustoty $f(x)$ našeho cílového rozdělení.

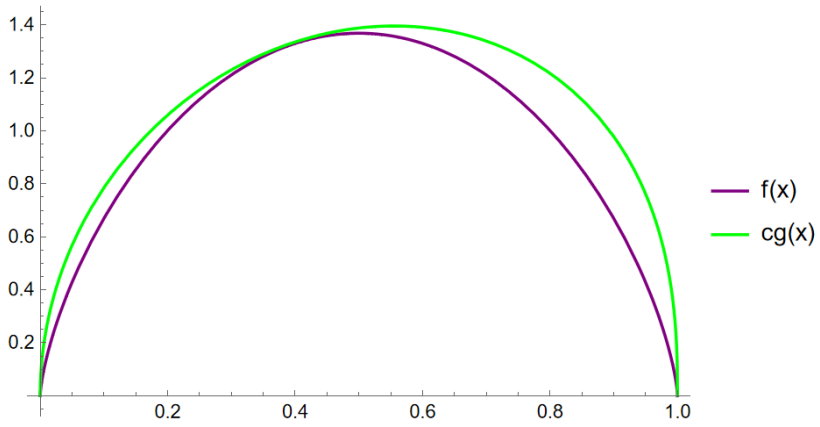
Při použití rovnoměrného rozdělení v majoritní funkci získáme efektivitu algoritmu přibližně 73%. Tedy přibližně v 27% případů budeme zamítat. Pro zvýšení efektivity algoritmu bychom mohli požít jinou funkci $g(x)$.

Zvolíme-li jako majorizující funkci $cg(x)$ hustotu beta rozdělení s parametry $\alpha = 1{,}5$ a $\beta = 1{,}3$, které vynásobíme konstantou c, aby platila podmínka $f(x) \leq cg(x)$, pak můžeme dosáhnout efektivity algoritmu přibližně 90%. Tedy budeme zamítat pouze v přibližně 10% případů.

Nutno poznamenat, že příklad je pouze ilustrativní. V praxi nelze předpokládat, že pokud bychom neuměli simulovat náhodnou veličinu z beta rozdělení s danou sadou parametrů, pak nelze předpokládat, že bychom uměli simulovat náhodnou veličinu s jinou sadou parametrů.



Obrázek 2.6: histogram vygenerovaných hodnot s hustotou beta rozdělení



Obrázek 2.7: Hustota beta funkce v majorizující funkci

3. Má Říp normální rozdělení?

Tvar hustoty normálního rozdělení je často přirovnáván k tvaru Řípu. Má však Říp skutečně normální rozdělení? Tuto hypotézu se můžeme pokusit ověřit pomocí zamítací metody. Díky zamítací metodě můžeme získat vzorky z rozdělení s hustotu danou tvarem Řípu a pomocí testů normality otestovat nulovou hypotézu, že tvar Řípu odpovídá tvaru hustoty normálního rozdělení. Tento postup ověření normality Řípu není jediný. Nemusí se ani jednat o nejvhodnější postup. Naším cílem cílem v této kapitole je ilustrace zamítací metody.

3.1 Testy normality

Nejdříve si definujeme několik testů normality, které použijeme k ověřování nulové hypotézy.

3.1.1 Shapirův-Wilkův test

Zdroj testu: Shapiro a Wilk (1965, str. 591)
Nulová hypotéza: Náhodný výběr X_1,...,X_n pochází z normálního rozdělení.
Alternativa: Náhodný výběr X_1,...,X_n nepochází z normálního rozdělení.
Testová statistika:

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Kde $X_{(i)}$ je i-tá pořadová statistika, $\overline{X} = \frac{(X_1+...+X_n)}{n}$, $(a_1,...,a_n) = \frac{m^T V^{-1}}{C}$, $C = ||V^{-1}m|| = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$. $m = (m_1,...,m_n)^T$ je vektor středních hodnot pořadové statistiky vzorku o velikosti n ze standardního normálního rozdělení, $V = (v_{ij})$ je kovarianční matice těchto hodnot o rozměrech. Tedy nechť $Y_{(i)}$ je i-tá pořadová statistika vzorku o velikosti n z normálního rozdělení s nulovou střední hodnotou a rozptylem rovný jedné, pak $m_i = E(Y_i)$, $v_{ij} = cov(Y_i,Y_j)$.
Asymptotické rozdělení: Testová statistika W má poměrně složité rozdělení. Kritické hodnoty byly určené pomocí Monte Carlo simulací a jsou tabelovány. Další podrobnosti lze najít v knize Shapiro a Wilk (1965, str. 592).

V programu R použijeme příkaz *shapiro.test*.

3.1.2 Jarqueův a Beryho test

Zdroj testu: Jarque a Bera (1987, str. 165)
Nulová hypotéza: Náhodný výběr X_1,...,X_n pochází z normálního rozdělení.
Alternativa: Náhodný výběr X_1,...,X_n nepochází z normálního rozdělení.
Testová statistika:

$$JB = n\left(\frac{(b_1)^2}{6} + \frac{(b_2-3)^2}{24}\right)$$

Kde n je počet pozorování, b_1 je výběrový koeficient šikmosti, b_2 je výběrový koeficient špičatosti, tedy:

$$b_1 = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^{\frac{3}{2}}}$$
$$b_1 = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^2}$$

Asymptotické rozdělení: $JB \sim \chi^2_2$ (testová statistika má asymptoticky chí kvadrát rozdělení o 2 stupních volnosti)

Jarqueův a Beryho test je založený na šikmosti a špičatosti a na histogramu 3.6 lze vidět, že špičatost vygenerovaných hodnot z rozdělení daných horou Říp bude nejspíš jiná než špičatost normálního rozdělení. Proto by tento test mohl fungovat dobře. Na druhou stranu, test funguje lépe s větším počtem dat.

V programu R použijeme příkaz *jarque.Bery.test* z balíčku *tseries*.

3.2 Generování vzorků z Řípu

Obrázek 3.1: Hora Říp

Pro získání vzorku z rozdělení s hustotou danou tvarem Řípu použijeme program R.

Vezměme si fotku Řípu (obrázek 3.1, zdroj: Turistická mapa.cz), kterou vhodným způsobem ořízneme tak, aby na obrázku byl jenom samotný kopec (obrázek 3.2).

Obrázek 3.2: Oříznutý obrázek

Díky stromům, které se na Řípu nachází, není tvar vrchní části kopce hladký. Samotná hora Říp, ze které chceme vygenerovat vzorek, ale hladká je. Proto před

použitím zamítací metody tedy ještě tvar vyhladíme. Stromy se nachází na celém povrchu Řípu. Po vyhlazení by měl tedy získaný tvar odpovídat tvaru kopce, který se nachází pod nimi. Většinu nerovností můžeme zahladit tím, že snížíme rozlišení obrázku tak, aby většina nerovností splynula (obrázek 3.3).

Obrázek 3.3: Obrázek po snížení rozlišení

Aby se s fotkou lépe pracovalo, použijeme segmentaci barev k tomu, abychom z barevného obrázku vytvořili černobílý obrázek, na kterém je pozadí bílé a kopec černý. K tomu použijeme příkaz *image_convert* z balíčku *magick*. Implementace v programu R vypadá následovně:

```
library(magick)
obrázek=image_read("C:/Users/rajtm/Desktop/říp.jpg")
čb=obrázek %>%
  image_convert(colorspace = "Gray") %>%
  image_threshold(type = "black", threshold = "70%") %>%
  image_threshold(type = "white", threshold = "30%")
```

Obrázek 3.4: Oříznutý obrázek po barevné segmentaci

Nyní máme černobílý obrázek (3.4), kde bílá část je pozadí a černá část je kopec. Díky převedení obrázku na černobílý se na obrázku projevilo ještě pár nerovností, které zůstaly i po snížení rozlišení obrázku. Tyto zbylé nerovnosti můžeme zarovnat ručně v programu na úpravu obrázků.

Obrázek 3.5: Obrázek po ruční úpravě

Po vyhlazení (3.5) je obrázek již připravený na použití zamítací metody. Pomocí příkazu *image_raster* z balíčku *magick* převedeme obrázek na dataset udávající barvu jednotlivých pixelů.

```
čb=image_read("C:/Users/rajtm/Desktop/čbb.jpg")
                          #načtení upraveného obrázku
b=image_raster(čb)        #vytvoření tabulky barev každého pixelu
```

Dataset obsahuje tři sloupečky. V prvním se nachází poloha pixelu na ose x, v druhém poloha pixelu na ose y a ve třetím barva tohoto pixelu. V našem případě je to buď #ffffffff(bílá), nebo #000000ff (černá).

Nyní použijeme algoritmus 1. Jako pomocnou hustotu $g(x)$ si zvolíme hustotu rovnoměrného rozdělení na intervalu od 0 do 300 (šířka obrázku) vynásobenou příslušnou konstantou $c = 46$ (výška obrázku). Implementace v programu R bude vypadat následovně:

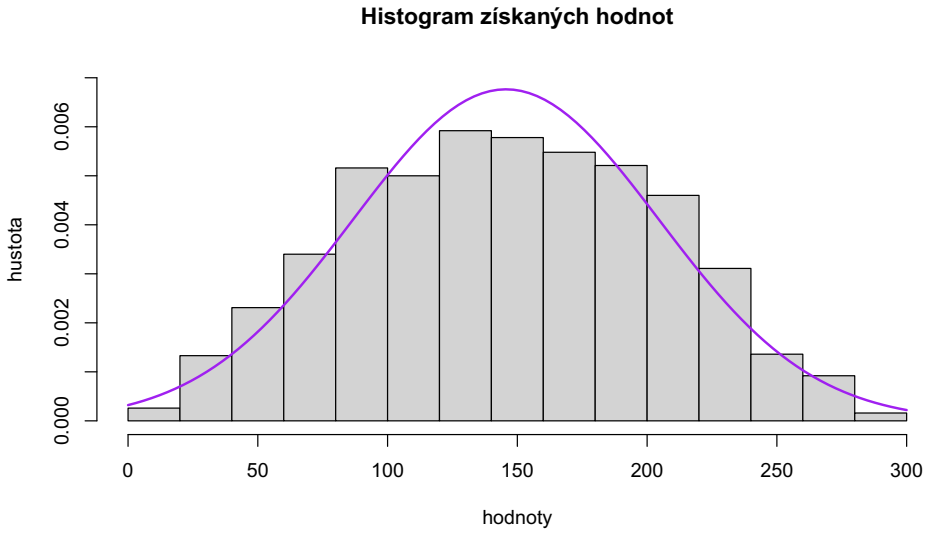
```
barva=1
while(barva!="#000000ff"){  #opakujeme, dokud nedostaneme černý pixel
x=runif(1,0,300)
u=runif(1,0,1)
t=ceiling((u*46))           #obrázek má na výšku 46 pixelů
xs=ceiling(x)               #hodnoty zaokrouhlíme na celá čísla
barva=b[b$x==xs&b$y==t,3]   #barva pixelu na vygenerované pozici
}
x
```

V algoritmu 1 jsme ověřovali kritérium $Uc\frac{g(X)}{f(X)} \leq 1$. V našem případě je funkce $f(x)$ tvar Řípu, pro který máme daný předpis obrázkem. Pro porovnání tak můžeme využít toho, že na námi upraveném obrázku je barva kopce černá a barva okolí bílá. V našem algoritmu tedy vygenerujeme realizace nezávislých náhodných veličin X a U jako v algoritmu 1. Poté pro $Ucg(X)$ zjistíme, zda-li je barva pixelu na daném místě bílá, nebo černá. V případě, že je černá, vzorek přijmeme. V opačném případě vzorek zamítneme.

Na obrázku 3.6 je vidět histogram hodnot získaných po 5000 opakování algoritmu společně s hustotou normálního rozdělení se střední hodnotou danou výběrovým průměrem z hodnot vygenerovaných algoritmem a výběrovým rozptylem z hodnot vygenerovaných algoritmem. Lze vidět, že histogram vygenerovaných hodnot má sice tvar podobný hustotě normálního rozdělení, nicméně se liší ve špičatosti, která je u normálního rozdělení vyšší. Špičatost normálního rozdělení je 3, ale výběrová špičatost vygenerovaných dat je 2,314091. Výběrová šikmost vygenerovaných dat je $-0{,}000570527$, což je blízko nulové šikmosti normálního rozdělení. Lze tedy očekávat nízké p-hodnoty u testů normality aplikovaných na vygenerovaná data.

3.3 Testování normality

Na vygenerovaná data použijeme testy normality, které jsme si definovali na začátku kapitoly. Konkrétně tedy zkusíme použít Shapirův-Wilkův test a Jarqueův a Beryho test. Zvolíme si hladinu významnosti 0,05. Z histogramu na obrázku 3.6 lze očekávat zamítnutí nulové hypotézy na této hladině významnosti. Testy provedeme pro různý počet vzorků.



Obrázek 3.6: Histogram vygenerovaných hodnot

Na obrázku 3.7 vidíme histogramy vygenerovaných hodnot pro různý počet vzorků společně s hustotou normálního rozdělení se střední hodnotou danou výběrovým průměrem z vygenerovaných hodnot a rozptylem daným výběrovým rozptylem z vygenerovaných hodnot.

	Počet vzorků				
Typ testu	50	100	500	1000	5000
S-W	0,6755	0,4762	0,0005308	$1{,}465 \cdot 10^{-06}$	$2{,}2 \cdot 10^{-16}$
J a B	0,5454	0,4171	0,003942	$3{,}922 \cdot 10^{-05}$	$2{,}2 \cdot 10^{-16}$

Tabulka 3.1: P-hodnoty testů normality

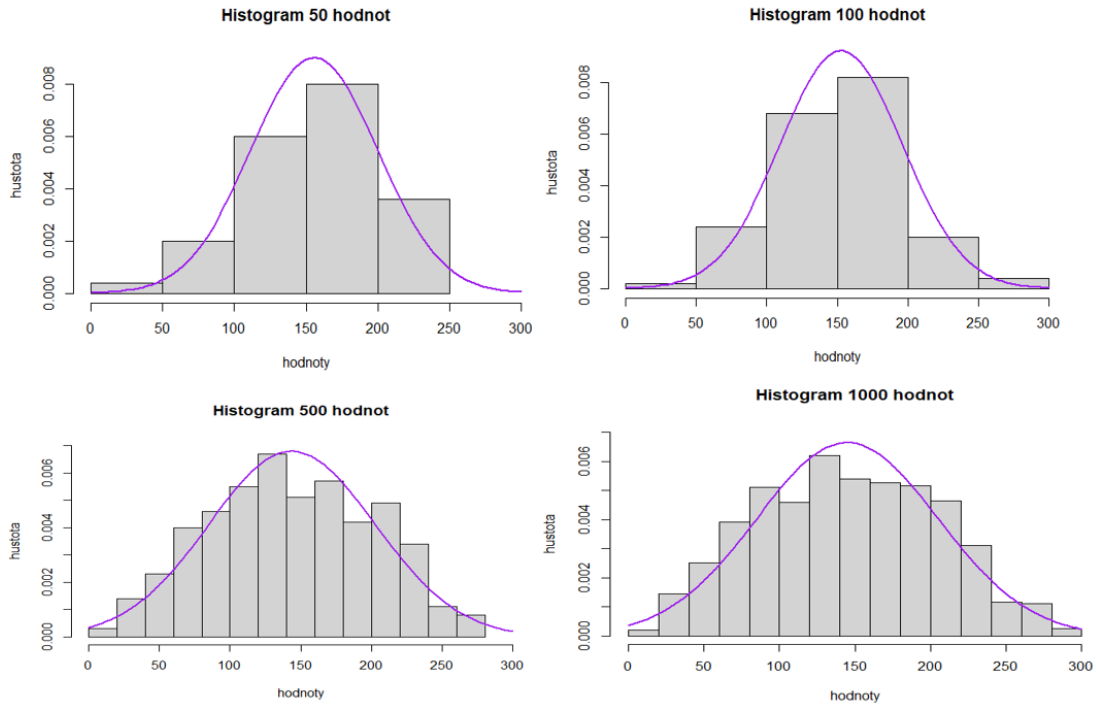
V tabulce 3.1 lze vidět p-hodnoty námi zvolených testů při různém rozsahu vzorků. Shapirův-Wilkův a Jarqueův a Beryho test zamítnou nulovou hypotézu o tom, že náhodný výběr pochází z normálního rozdělení na hladině 0,05 už při 500 vzorcích. Pro menší počet vzorků je p-hodnota větší. To může být způsobeno malým počtem vzorků. Na obrázku 3.7 můžeme vidět, že pro 50 a 100 vzorků je tvar histogramu celkem podobný hustotě normálního rozdělení. U většího počtu vzorků se tvar histogramu od normálního rozdělení odlišuje více.

V obou případech při dostatečném počtu vzorků zamítáme na hladině 0,05 nulovou hypotézu o tom, že rozdělení s hustotou danou tvarem Řípu je normální.

3.4 Rozdělení p-hodnoty

Protože p-hodnoty v tabulce vznikly z náhodně vygenerovaných dat, jsou také náhodné. Pokusíme se nahlédnout na rozdělení p-hodnoty u obou použitých testů normality, a to konkrétně pro 100 vzorků.

Zamítací metodu zopakujeme 100krát, abychom získali 100 vzorků. Z nich spočítáme p-hodnotu obou testů. Tento proces zopakujeme 1000krát. Tím získáme vzorek 1000 p-hodnot pro oba testy normality.



Obrázek 3.7: Histogramy 50, 100, 500, 1000 vzorků

Celá implementace v programu R bude vypadat následovně:

```
psha=c()
pjarque=c()
o=0
set.seed(127)
while(o<1000){                  #1000 p-hodnot
hodnoty=c()
z=0
barva=1
while(z<100){                   #získání 100 vzorků
  barva=1
  while(barva!="#000000ff"){    #původní zamítací algoritmus
  x=runif(1,0,300)
  u=runif(1,0,1)
  t=ceiling((u*46))
  xs=ceiling(x)
  barva=b[b$x==xs&b$y==t,3]
}
z=z+1
hodnoty=append(hodnoty,x)
}
psha=append(psha,shapiro.test(hodnoty)$p.value)   #uložení p-hodnot
pjarque=append(pjarque,jarque.Bery.test(hodnoty)$p.value)
o=o+1
}
```

	Výběrový průměr	Výběrový rozptyl	V původních datech
S-W	0,2570	0,0517	0,4762
J a B	0,3236	0,0301	0,4171

Tabulka 3.2: Naměřené p-hodnoty v porovnání s původními

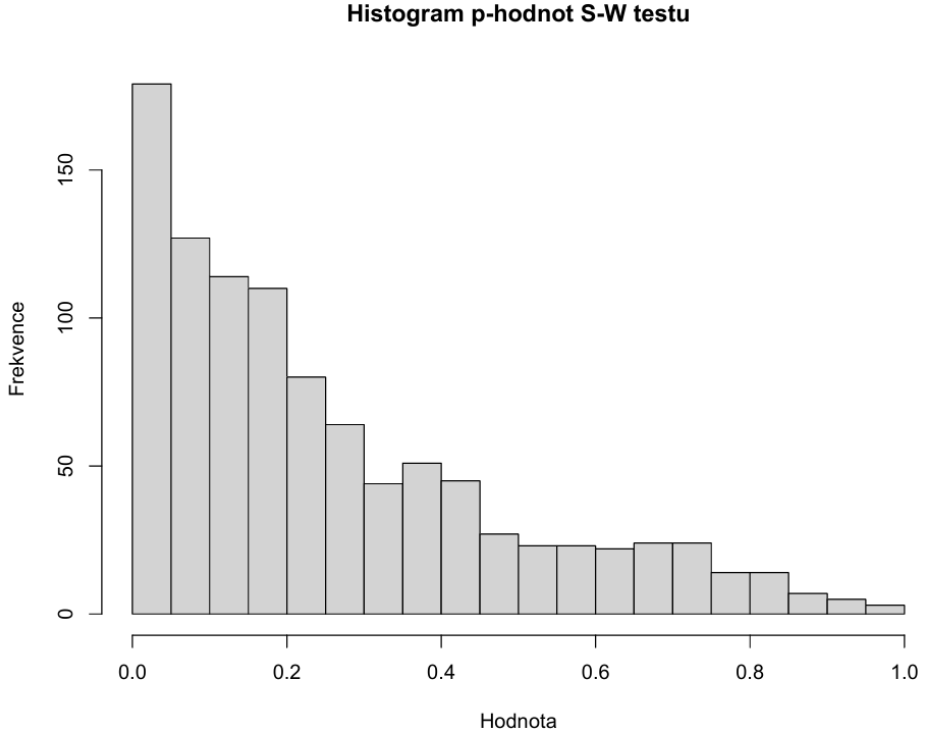
V tabulce 3.2 jsou výběrové průměry a výběrové rozptyly p-hodnot Shapirova-Wilkova testu a Jarqueova a Beryho testu získané opakováním algoritmu v porovnání s p-hodnotami z tabulky 3.1.

Průměrná p-hodnota vzorků by u obou testů stále nezamítala nulovou hypotézu na hladině 0,05. Průměrná p-hodnota u Shapirova-Wilkova testu je nižší, její rozptyl je ale o trochu vyšší. V obou případech nám z původních dat v tabulce 3.1 vyšla vyšší p-hodnota, než jakou bychom očekávali z těchto odhadů.

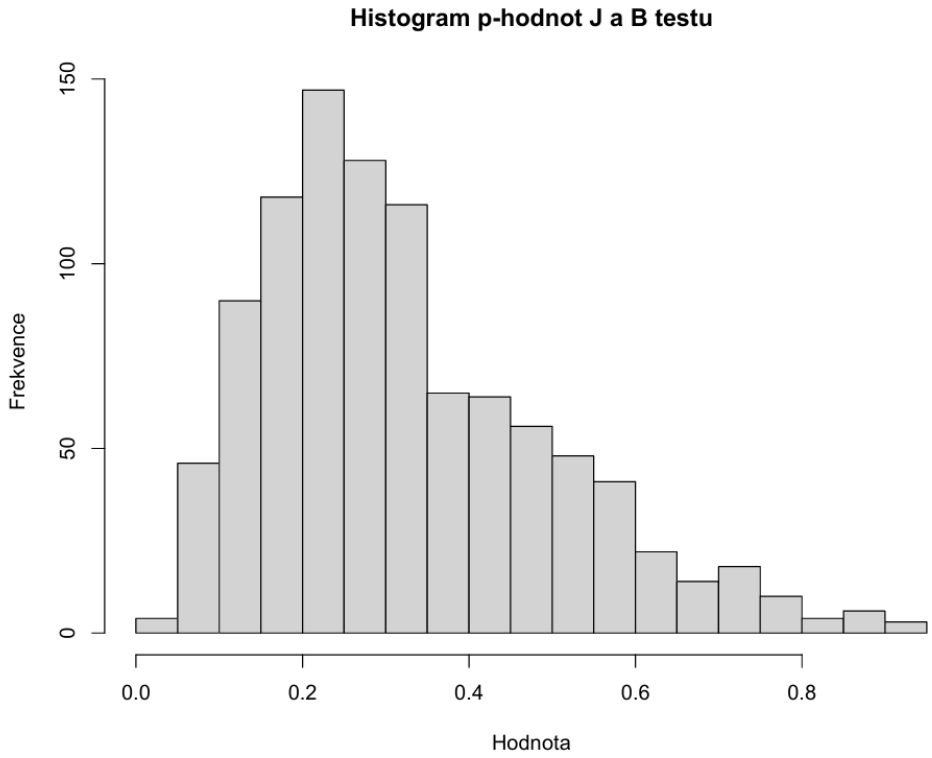
Nyní se můžeme podívat na histogramy získaných p-hodnot. Na obrázku 3.8 vidíme histogram p-hodnot Shapirova-Wilkova testu. Největší koncentrace p-hodnot je v intervalu od 0 do 0,05, tedy v intervalu, ve kterém bychom na hladině 0,05 zamítali nulovou hypotézu.

Na obrázku 3.9 vidíme histogram p-hodnot Jarqueova a Beryho testu. Největší koncentrace p-hodnot je v intervalu od 0,2 do 0,25, tedy v integrálu, ve kterém bychom nezamítali nulovou hypotézu.

Pro větší počet vzorků nám vyšly malé p-hodnoty (tabulka 3.1), pro které jsme nulovou hypotézu o tom, že vzorky získané pomocí zamítací metody nepochází z normálního rozdělení, zamítali na hladině významnosti 0,05. Výběrový průměr vygenerovaných p-hodnot je u Shapirova-Wilkova testu nižší (tabulka 3.2). V našem případě se tak Shapirův-Wilkův test zdá být přesnější. K tomuto závěru nás také vedou histogramy p-hodnot obou testů.



Obrázek 3.8: Histogram p-hodnot Shapirova-Wilkova testu



Obrázek 3.9: Histogram p-hodnot Jarqueova a Beryho testu

Závěr

V bakalářské práci jsme uvedli inverzní metodu generování vzorků z rozdělení a dokázali, že jejím výstupem je skutečně vzorek z požadovaného cílového rozdělení. Představili zamítací metodu generování vzorků z rozdělení, dokázali potřebné vlastnosti hustoty, ze kterých tato zamítací metoda vychází a popsali fungování zamítací metody. Zamysleli jsme se nad tím, jak volba pomocné funkce a konstanty ovlivní efektivitu této metody. Podívali jsme se na variaci zamítací metody, obálkovou metodu.

Ve třetí kapitole jsme použili zamítací metodu k vygenerování vzorků z rozdělení s hustotou danou tvarem hory Říp. Na vygenerované data o různém rozsahu jsme aplikovali Shapirův-Wilkův a Jarqueův a Beryho testy normality. Zjistili jsme, že na hladině významnosti 0,05 nám oba testy zamítají nulovou hypotézu o normalitě dat při 500 a více vzorcích s tím, že čím více vzorků, tím menší byla p-hodnota obou testů.

Protože p-hodnota vychází z vygenerovaných dat, které jsou náhodné, pak p-hodnota je také náhodná. Proto jsme se na závěr podívali na to, jak vypadá rozdělení p-hodnoty obou našich použitých testů normality, a to konkrétně pro vzorky o rozsahu 100. Spočítali jsme výběrové průměry a výběrové rozptyly pro p-hodnoty obou testů a podívali se na histogramy těchto p-hodnot. Zjisili jsme, že p-hodnoty, které jsme získali z původních dat, byly u obou testů větší, než jaké bychom předpovídali na základě rozdělení p-hodnot v histogramu a jejich výběrovému průměru a rozptylu. Usoudili jsme, že v našem případě je Shapirův-Wilkův test normality přesnější než Jarque a Beryho test normality.

Seznam použité literatury

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA.

Jarque, C. M. a Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, **55**(2), 163–172. ISSN 03067734, 17515823. URL `http://www.jstor.org/stable/1403192`.

Robert, C. a Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

Shapiro, S. S. a Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611. ISSN 00063444. URL `http://www.jstor.org/stable/2333709`.

Turistická mapa.cz (2021). Hora Říp. URL `https://turistickamapa.cz/data_fotos/hora-rip-2021_26_6-114040.jpg`. [Online, navštíveno 14.4.2024].