Středoškolská odborná činnost

Obor 01 - Matematika a statistika

Metoda Monte Carlo a její aplikace

Autorka:	Alena Harlenderová
Škola:	Slovanské gymnázium Olomouc, tř. Jiřího z Poděbrad 13, Olomouc, 771 10
Studijní obor:	všeobecné 8. ročník osmiletého studia
Konzultant:	RNDr. Karel Hron, Ph.D. Přírodovědecká fakulta UP Olomouc, Katedra matematické analýzy a aplikací matematiky

Olomouc, 2012

Čestné prohlášení

Prohlašují tímto, že jsem soutěžní práci vypracovala samostatně pod vedením pana RNDr. Karla Hrona, Ph.D. a použila jsem pouze podklady (literaturu, www stránky atd.) uvedené v seznamu. Nemám závažný důvod proti zpřístupňování této práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění.

V Olomouci dne Podpis:

Poděkování

Tímto bych chtěla moc poděkovat panu doktoru Hronovi, vedoucímu této práce, který mi věnoval mnoho času a trpělivosti, nejen při vypracovávání této práce SOČ. Ukázal mi mnohá zajímavá zákoutí matematiky, která by určitě stála za další prozkoumání, a jistě mi pomohou při dalším studiu. Největší dík mu však patří za to, že spolu s dalšími neustále utvrzuje mou víru v to, že na světě existují lidé, kteří vám dokážou pomoci, a to ještě s radostí.

Anotace

Spolu s překotným rozvojem výpočetní techniky se zdokonalují i stochastické numerické metody. Mezi ně patří také metoda nazývaná Monte Carlo. Tato metoda nám často efektivně pomáhá s problémy, jejichž řešení klasickými analytickými metodami je buď příliš složité nebo dokonce nemožné.

Práce se nejprve zabývá nastíněním podstaty metody Monte Carlo. Další část je věnována důležitým pojmům, zejména z teorie pravděpodobnosti. Uvedeny jsou základní definice a vztahy, které jsou nutné k pochopení problematiky. Následující kapitola se věnuje využití metody Monte Carlo při simulaci rozdělení pravděpodobnosti náhodné veličiny. Poznatky z této kapitoly byly následně aplikovány při simulaci rozdělení pravděpodobnosti testovací statistiky Anderson-Darlingova testu normality náhodného výběru. Experimentální část práce přitom ukazuje, za jakých podmínek je možné s dostatečnou přesností a přitom co nejrychleji odhadnout rozdělení Anderson-Darlingova testu, a to užít při přibližném stanovení p-hodnoty odpovídající dané realizaci této testovací statistiky.

Klíčová slova
Metoda Monte Carlo
rozdělení pravděpodobnosti náhodné veličiny
simulace
testování statistických hypotéz

Anotation

With rapid development of the computer technology, also stochastic numerical methods are being improved. One of them is called the Monte Carlo metod. This method helps effectively with problems, where the analytical solutions are either difficult or impossible.

At the first this text engages in the essence of the Monte Carlo method. The next section is devoted to the important concepts from the probability theory. Here basic definitions and properties that are necessary to understand the problems connected with the Monte Carlo method are introduced. The third section shows how the Monte Carlo method can be utilized for simulation of the probability distribution of a random variable. Finally, results of this section are applied to simulation of the probability distribution of the Anderson-Darling test of normality of a random sample. In particular, the goal is to find out how to estimate, with sufficient accuracy on one hand and numerical efficiency on the other one, the probability distribution of the Anderson-Darling test and to use it for estimation of p-value corresponding to a given realization of the test statistic.

Key words
Monte Carlo method
probability distribution of a random variable
simulation
statistical hypotheses testing

Obsah

Úvod

Ačkoli lidstvo disponuje množstvím důležitých poznatků, stále není možné nebo je příliš technicky obtížné zodpovědět některé důležité otázky aplikované matematiky pomocí klasických analytických metod. Proto se začaly používat simulační numerické metody, mezi něž patří i metoda Monte Carlo, které umožňují obdržet alespoň přibližné řešení. Pokud k těmto aspektům přičteme i to, že se výpočetní technika čím dál více zdokanaluje, začíná být jasné, proč se metoda Monte Carlo stala významným pomocníkem v oblasti nejen matematické a fyzikální, ale také např. v geografii a medicíně.

Prvním cílem této práce je popis samotné metody Monte Carlo a pojmů nutných k jejímu pochopení. Proto se zvídavý zájemce o tuto problematiku nejprve dozví, co to vlastně metoda Monte Carlo je, a následně jsou, často i na příkladech, vysvětleny základní pojmy z teorie pravděpodobnosti, které úzce souvisí s již zmiňovanou metodou. Mezi ně patří hlavně náhodný pokus, jev, Kolmogorova definice pravděpodobnosti a různé pravděpodobnostní modely, podmíněná pravděpodobnost, nezávislé jevy a náhodná veličina. Na ně navazuje rozdělení pravděpodobnosti, distribuční funkce a výpočet funkce k ní inverzní, někdy nazývané jako kvantilová funkce. Následně se práce zabývá generováním pseudonáhodných čísel a druhým vytyčeným cílem, využitím metody Monte Carlo při simulacích rozdělení pravděpodobnosti náhodných veličin. To je ukázáno také na příkladech s konkrétními rozděleními.

Nakonec práce pojednává o Anderson-Darlingově testu normality. Dosud se v praxi pracovalo především s hodnotami testovací statistiky spojenými s často užívanými hladinami testu. Tato práce si však klade za cíl zjistit, jak s dostatečnou přesností a co nejrychleji odhadnout rozdělení Anderson-Darlingova testu a to užít při přibližném stanovení p-hodnot odpovídajících libovolné realizaci této testovací statistiky.

A teď s chutí do toho.

1 Podstata metody Monte Carlo a její vznik

Metoda Monte Carlo je stochasticko-numerická metoda používaná především pro simulaci dějů souvisejících s náhodou, kde je velmi těžké nebo dokonce nemožné použít klasické analytické metody [6]. Když se zamyslíme nad tím, kde všude v reálném světě něco závisí na náhodě, zřejmě dojdeme k závěru, že metoda simulující tyto děje je velice užitečná. Z našeho pohledu je totiž ovlivněno náhodou vše to, co má výsledek, který nedokážeme s jistotou předpovědět. Ten přitom určují mnohé vlivy. Nejjednodušším způsobem řešení otázek tohoto typu se často stává simulace pomocí metody Monte Carlo. Proto tato metoda nalézá uplatnění v mnohých oborech jako je ekonomie, finance, pojišťovnictví, fyzika, medicína, technika, demografie a dokonce i samotná matematika.

Typ simulace, kdy umíme modelovat danou situaci způsobem podobným samotnému ději, se nazývá *analogový model*. Mezi analogové modely patří právě metoda Monte Carlo. Vymysleli ji polský vědec Stanislaw Marcin Ulam a jeho maďarský kolega John von Neumann. Tehdy byla mimo jiné využita při vývoji vodíkové bomby během druhé světové války v amerických laboratořích v Los Alamos. Jak uvádí [1], řešili tehdy, jaké procento ze sprškы neutronů projde daným materiálem (např. barelem vody). I když disponovali takovými znalostmi jako je pravděpodobnost srážky s atomem vodíku nebo kyslíku, pravděpodobnost pohlcení neutronu při srážkách a pravděpodobnost, že se po srážce neutron bude pohybovat daným způsobem, nebyli schopni problém vyřešit tehdejšími metodami. Pomohlo jim až vytvoření metody Monte Carlo. Jak je vidět z názvu, nechali se inspirovat městečkem v Monackém knížectví plném heren. Základní myšlenka metody Monte Carlo má totiž co do činění s ruletou. Jednotlivé části „života neutronu“ se dají představit na základě točení ruletou. Jestliže známe pravděpodobnost srážky s vodíkem, můžeme si zatočit odpovídajícím kolem rulety a zjistíme, zda daný nasimulovaný neutron narazil do atomu vodíku. Pokud víme, že narazil a pravděpodobnost jeho pohlcení je v tomto případě jedna setina, můžeme si zatočit ruletou se sto články, z nichž jeden reprezentuje pohlcení. Takto můžeme pokračovat i dále, dokud nezjistíme, zda daný neutron prošel barelem, nebo byl pohlcen.

Pokud bychom chtěli danou situaci nasimulovat pomocí kol rulety, trvalo by to příliš dlouho. Počítače nám mohou velice usnadnit práci, stejně tak, jako ji již usnadnily v mnoha dalších oborech. Stačí dát dohromady správný algoritmus, mít k dispozici pseudonáhodná čísla s daným rozdělením pravděpodobností, a můžeme spustit příslušný program. Výsledky analýzy pak zpracujeme vhodnými statistickými metodami. Z toho je vidět, že základy metody Monte Carlo tvoří především teorie pravděpodonosti a statis-

tika. Samotný název metoda Monte Carlo se datuje od roku 1949, kdy vyšel v americkém matematickém časopise stejnojmenný článek.

Kromě analogových modelů existují také *neanalogové*. Jsou to ty, při jejichž realizaci nepoužíváme model skutečného děje. Mezi ně patří např. výpočet určitého integrálu nebo obsahu ohraničené plochy.

Mějme čtverec s obsahem S_1 a v něm ohraničenou plochu s neznámým obsahem S_2, který chceme zjistit. Stačí nagenerovat dostatečné množství „náhodných“ bodů do tohoto čtverce a jsme schopni přibližně určit obsah ohraničeného útvaru. Počet nagenerovaných bodů označíme n a počet bodů na ohraničené ploše m, pak

$$\frac{S_2}{S_1} \sim \frac{m}{n} \Rightarrow S_2 \sim \frac{m \cdot S_1}{n}.$$

Chceme-li určit hodnotu čísla π, stačí za ohraničenou oblast považovat kruh.

Na závěr této kapitoly je třeba říci, že vzhledem k neustálému zdokonalování výpočetní techniky, stoupá aplikovatelnost metody Monte Carlo. Jak uvádí [1], přesnost samotného výpočtu pomocí výpočetní techniky je dána těmito faktory: *kvalitou generátoru pseudonáhodných čísel, výběrem racionálního algoritmu výpočtu a kontrolou přesnosti získaného výsledku*. Podrobněji se o nich zmíníme v dalších kapitolách.

2 Základy teorie pravděpodobnosti

Na střední škole jsme se naučili intuitivně vnímat mnohé základní pojmy teorie pravděpodobnosti. Většinou jsme se dověděli o tom, že existuje náhodný pokus a množina všech jeho možných výsledků. Také jsme se doslechli o jevech, relativní četnosti i o tom, že se dají za určitých podmínek pravděpodobnosti jednotlivých jevů sčítat a násobit. Mnozí z nás byli, aniž by o tom věděli, seznámeni s prvními dvěma axiomy Kolmogorovovy definice pravděpodobnosti. To vše nám bylo většinou vysvětleno na příkladech a pokud jsme se o pravděpodobnost nezajímali více, chápeme ji pouze intuitivně. Pro potřeby velké části lidí, kteří prošli střední školou, je to nejspíš dostačující. Pokud však chceme studovat některé části matematiky, je třeba pochopit základní pojmy teorie pravděpodobnosti poněkud zevrubněji. Mezi takové oblasti matematiky patří mimo jiné i metoda Monte Carlo. Proto se následující část práce bude věnovat podrobnějšímu vysvětlení a definování základních pojmů, které jsou k hlubšímu pochopení vlastností pravděpodobnosti potřeba [2, 3].

2.1 Pokusy

Definice 1 *Pokus* je jednorázové uskutečnění vymezeného souboru definičních podmínek.

Lze ho mnohonásobně opakovat. Opakováním souboru definičních podmínek rozumíme opět pokus. Podmínky, které nejsou definiční, se mohou měnit. Proto můžeme opakováním stejného pokusu dojít k jiným výsledkům.

Příklad 1 Házíme-li kostkou, vykonáváme pokus. Tento pokus můžeme zopakovat tím, že hodíme kostkou se stejným tvarem, rozložením hmoty a číslicemi na odpovídajích stranách (definiční podmínky). Specifická rotace, která kostku „donutila“ spadnout tak, jak spadla, není definiční podmínka. Při dalším pokusech se nemusí opakovat.

Deterministickým pokusem rozumíme takový pokus, jehož definiční podmínky jsou natolik „těsné“, že opakováním dostaneme vždy stejný výsledek.

Příklad 2 Pokud se nacházíme na Zemi a zároveň ne za polárním kruhem (definiční podmínky), slunce vyjde každé ráno (výsledek deterministického pokusu).

Náhodným pokusem nazýváme takový pokus, jehož realizací dostaneme právě jeden z více možných disjunktních (navzájem se vylučujících) výsledků. Jeden konkrétní náhodný pokus je prezentován v Příkladu 1, další je uveden níže.

Příklad 3 Při hodu mincí padne buď panna nebo orel. Definiční podmínky při tom neurčují, která z těchto dvou položek padne.

Nyní by bylo vhodné rozebrat, jakým způsobem se označují některé důležité pojmy, které budeme v teorii pravděpodobnosti dále potřebovat. Písmenem Ω se značí množina všech výsledků náhodného pokusu. Např. při hodu kostkou je to množina $\Omega = \{1; 2; 3; 4; 5; 6\}$. Pro tuto množinu platí, že $\Omega \neq \{\}$. Je zřejmé, že hod kostkou vždy skončí tím, že padne nějaké číslo. A stejně tak je to i s ostatními náhodnými pokusy. Jeden možný výsledek náležící Ω lze označit písmenem ω. Při hodu kostkou se ω může rovnat např. 6 nebo 5.

2.2 Náhodné jevy a pravděpodobnost

Definice 2 Každá množina $A \subseteq \Omega$ se nazývá jev. Jednoprvkové podmnožiny jsou *elementární jevy.*

Jev označuje nějakou událost, která při realizování pokusu buď nastane, nebo nenastane. Jevy se značí zpravidla velkými písmeny A, B, C,... Říkáme, že nastal jev A, jestliže byl realizací náhodného pokusu obdržen výsledek $\omega \in A$.

Prázdná množina je tzv. *nemožný jev*, protože při realizaci definičních podmínek nikdy nenastane. Je tomu tak, jelikož pokus má vždy nějaký výsledek. Opakem je *jistý jev* Ω, který nastane vždy.

Pro operace s jevy platí stejná pravidla jako pro operace s množinami.

Příklad 4 Při hodu kostkou se nikdy nestane, že by nepadlo žádné číslo (nemožný jev). Naopak se vždy stane, že padne jedno z čísel 1 až 6 (jistý jev). Elementárním jevem je třeba padnutí jedničky. Jevem může být např. i to, že padne sudé číslo. Následující definice je převzata z [2].

Definice 3 Nechť $\Omega \neq \{\}$ je libovolná množina. Neprázdný systém $\mathcal{A}$ podmnožin množiny Ω se nazývá *jevové pole*, jestliže platí:
a) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
b) $A_n \in \mathcal{A}$, $n = 1, 2, \ldots \Rightarrow \bigcup_1^\infty A_n \in \mathcal{A}$
Prvky $A \in \mathcal{A}$ se nazývají *náhodné jevy.*

Jevové pole značí určitou "rozumnou"množinu podmnožin Ω (pro konečnou a spočetnou množinu výsledků je tvořeno všemi jejími podmnožinami) a jeho zavedení je důležité pro následující definici pravděpodobnosti.

Základem dnešního pojetí pravděpodobnosti se stala definice ruského matematika Kolmogorova z roku 1933.

Definice 4 Nechť je dána neprázdná množina Ω a na ní jevové pole $\mathcal{A}$. *Pravděpodobností* nazveme každou reálnou funkci $P(\ )$ definovanou na $\mathcal{A}$, která vyhovuje následujícím axiomům:
a1 $P(\Omega) = 1$;
a2 $P(A) \geq 0, \forall A \in \mathcal{A}$;
a3 Pro libovolnou posloupnost $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ neslučitelných náhodných jevů (a tedy $A_i \cap A_j = \{\},\ i, j = 1, 2, \ldots,\ i \neq j$) platí

$$P\left(\bigcup_1^\infty A_n\right) = \sum_1^\infty P(A_n).$$

Uspořádaná trojice $(\Omega, \mathcal{A}, P)$ se označuje jako *pravděpodobnostní prostor.*

2.3 Různé pravděpodobnostní modely

Kolmogorova definice se vyznačuje jistou volností interpretace. Její axiomy netvoří „příliš přísný systém“. Proto je možné ji podle potřeby přizpůsobovat různým konkrétním případům. Existují tak i další, specifizující definice pravděpodobnosti. Ve školních lavicích jsme nejspíše slyšeli pouze o jedné této definici. Jmenuje se *klasická definice pravděpodobnosti.* Mezi další definice patří také geometrická a neklasická pravděpodobnost.

2.3.1 Klasická pravděpodobnost

Klasická definice pravděpodobnosti je založena na principu *stejné možnosti* nastoupení libovolného z elementárních jevů. Tím pádem se stává spíše logickým, vykonstruovaným systémem než nástrojem užitečným v praxi. Realita je často mnohem složitější než hod kostkou, která má všechny stěny stejné až na počet ok (tj. je pravidelná). Stačí uvažovat např. nekonečný počet výsledků pokusu nebo to, že všechny elementární jevy nenastanou se stejnou pravděpodobností, a tato definice selhává. Hodí se pouze v situacích, kdy se dá počet možných výsledků náhodného pokusu (a následně pravděpodobnost) určit s využitím kombinatorických úvah.

Popularizovaná definice (klasické) pravděpodobnosti, která se učí na středních školách:

$$P(A) = \frac{n(A)}{n},$$

$n(A)\ldots$ počet výsledků příznivých jevu A,
$n\ldots$ počet všech výsledků náhodného pokusu.
V souladu s Kolmogorovou definicí pravděpodobnosti se definuje klasická pravděpodobnost takto:

Definice 5 Nechť $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ je konečná množina, $\mathcal{A}$ je jevové pole, které obsahuje všechny podmnožiny množiny Ω. Funkce $P(\ )$ definovaná na $\mathcal{A}$ vztahy

$$P(\omega_j) = \frac{1}{n}, \qquad j = 1, 2, \ldots, n, \qquad P(A) = \sum_{j:\omega_j \in A} P(\{\omega_j\})$$

se nazývá *klasická pravděpodobnost.*

2.3.2 Geometrická pravděpodobnost

Geometrická pravděpodobnost se vyznačuje tím, že kvůli jejímu výpočtu neurčujeme nějaký počet, ale míru. Když chceme vyjádřit např. délku

úsečky, těžko tak učiníme pomocí počtu jejích bodů. Nemůžeme však existenci jednotlivých bodů zanedbat. Proto je pro korektní definici geometrické pravděpodobnosti třeba uvažovat tzv. Lebesgueovu míru. Její pomocí dokážeme objektu tvořenému nekonečným (nespočetným) počtem bodů přidělit číslo, které tento objekt charakterizuje. Nám však postačí zjednodušená definice převzatá z [2].

Definice 6 Nechť $\Omega \subset \mathbb{R}^n, n = 1, 2, \ldots,$ je taková množina, jejíž míru $\mu(\Omega)$ umíme určit. Nechť $\mathcal{A}$ je třída všech podmnožin Ω se stejnou vlastností. Je-li $0 < \mu(\Omega) < \infty$, definujeme na $\mathcal{A}$ funkci P takto:

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}, \qquad A \in \mathcal{A}.$$

Lebesqueova míra v $\mathbb{R}^1$ reprezentuje délku, $\mathbb{R}^2$ obsah atd. Stejně jako klasická definice pravděpodobnosti, je i tato poměrně omezená, co se týče spektra problémů, které zahrnuje. Přesto je v praxi velmi užitečná, a to i co se týká metody Monte Carlo.

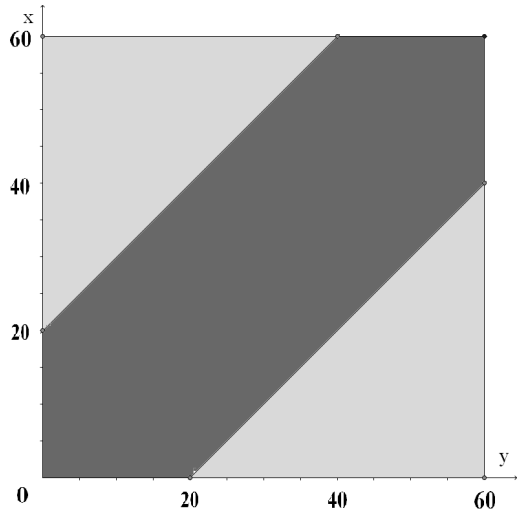
Příklad 5 (o setkání) James Bond a agent FBI chtějí během stejné hodiny vtrhnout do skladu s trhavinou a zneškodnit ji. Jeden o druhém však nevědí, přijdou tedy v určitý časový okamžik během dané hodiny a po nalezení a případné deaktivaci bomby, na což jim stačí 20 minut, sklad neprodleně opustí. S jako pravděpodobností se tam oba agenti potkají?

Řešení Při řešení této úlohy si můžeme pomoci jednoduchým obrázkem. Osa x (respektive y) znázorňuje dobu příchodu Jamese Bonda (agenta FBI) v minutách. Aby se agenti setkali, musí druhý z nich přijít do dvaceti minut po příchodu prvního. Na obrázku 1 je tato situace znázorněna tmavě šedou barvou. Případy, kdy se nesetkají, jsou reprezentovány světle šedou barvou. Ze zadání obržíme

$$\Omega = \{(x;\, y) : 0 \leq x \leq 60;\, 0 \leq y \leq 60\},$$
$$\mathcal{A} = \{(x;\, y) \in \Omega :\, |x - y| \leq 20\},$$
$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{60^2 - 40^2}{60^2} = \frac{5}{9}.$$

Pravděpodobnost, že se agenti setkají, je rovna $\frac{5}{9}$.

Další úlohu podle [11] formuloval roku 1777 francouzský matematik Georges Louis Leclerc de Buffon. Její výpočet se dá uskutečnit pomocí geometrické pravděpodobnosti, zároveň ovšem byla jednou z motivací k zavedení



Obrázek 1: Grafické znázornění úlohy o setkání.

metody Monte Carlo.

Příklad 6 (Buffonova úloha) Uvažujme rovinu, v níž jsou v pravidelných vzdálenostech d umístěny rovnoběžky. Na tuto rovinu náhodně vrháme jehlu jejíž délka l splňuje podmínku $l < d$. Jaká je pravděpodobnost, že tato jehla protne některou z rovnoběžek?

Řešení Označíme si vzdálenost středu jehly od nejbližší přímky písmenem x a úhel, který svírá jehla s nejbližší přímkou φ. Je zřejmé, že jehla protne nejbližší přímku, pokud platí

$$x \leq \frac{l}{2} \cdot \sin \varphi,$$

přitom x může nabývat hodnot od 0 do $\frac{d}{2}$, φ od 0 do π. To určuje množinu všech možných výsledků, viz obrázek 2. Pro

$$\Omega = \{(x; \varphi) : 0 \leq x \leq \frac{d}{2}; 0 \leq \varphi \leq \pi\},$$

$$A = \{(x; \varphi) \in \Omega : x \leq \frac{l}{2} \cdot \sin \varphi\}$$

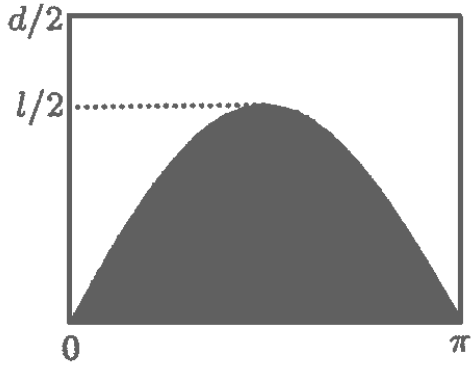
tedy dostaneme

$$P(A) = \frac{1}{\frac{d\pi}{2}} \int_0^\pi \frac{l}{2} \sin \varphi \, d\varphi = \frac{2l}{d\pi}.$$

Nyní je možné použít metodu Monte Carlo pro výpočet přibližné velikosti π. Stačí experiment s jehlou uskutečnit ručně, nebo ho nasimulovat pomocí počítače. Počet pokusů si označíme n, z toho počet pokusů, kdy jehla protla přímku, označíme m. Pro velká n platí

$$P(A) \sim \frac{m}{n} \Rightarrow \frac{m}{n} \sim \frac{2l}{d\pi} \Rightarrow \pi \sim \frac{2nl}{md},$$

kde l, d a n známe, m jsme určili pomocí experimentu nebo simulace. Zbývá tedy pouze dosadit do vzorce a známe přibližnou hodnotu π.



Obrázek 2: Grafické znázornění Buffonovy úlohy [4].

3 Podmíněná pravděpodobnost a nezávislé náhodné jevy

3.1 Podmíněná pravděpodobnost

Dosud jsme se v této práci setkali pouze s pravděpodobností nějakého jevu za uskutečnění definičních podmínek. Tato pravděpodobnost se někdy označuje jako *nepodmíněná pravděpodobnost*. Můžeme se však také setkat s případem, kdy víme více než pouze to, že nastaly definiční podmínky, tudíž se pokus zdařil. Víme, že nastal nějaký jev B. Naším úkolem je zkoumat, jak to ovlivňuje pravděpodobnost nastoupení daného jevu A. Zjišťujeme tedy, jakých hodnot nabývá pravděpodobnost jevu A za podmínky B. Hledaná pravděpodobnost se značí $P(A/B)$.

Definice 7 Nechť je dán náhodný jev $B \in \mathcal{A}$, pro který platí $P(B) > 0$. Funkce $P(\ /B)$ definovaná předpisem

$$P(A/B) \ = \frac{P(A \cap B)}{P(B)}, A \in \mathcal{A} \tag{1}$$

se nazývá podmíněná pravděpodobnost jevu A za podmínky B.

Příklad 7 Podle internetových stránek [8] v České republice v roce 2005 onemocnělo rakovinou plic přibližně 6000 lidí a výskyt rakoviny plic u kuřáků ve srovnání s nekuřáky je asi desetinásobný. Jak uvádí [9], v ČR kouří asi 26 % lidí. Jaká je pravděpodobnost, že český kuřák onemocněl v roce 2005 rakovinou plic, jestliže ČR obývá přibližně 10 milionů lidí?

Řešení Jev A bude reprezentovat výskyt rakoviny. Jako jev B si označíme kouření. Nyní spočítáme $P(A \cap B)$, pravděpodobnost, že je člověk nemocný a zároveň kuřák. Pravděpodobnost, že se jedná o člověka, který onemocněl, je vyjádřena podílem lidí, kteří onemocněli ku počtu lidí v ČR. Tuto pravděpodobnost vynásobíme poměrem vyjadřujícím, kolik z nemocných lidí byli kuřáci. $P(B)$ známe ze zadání, dohromady tedy dostaneme

$$P(A \cap B) = \frac{6000}{10^7} \cdot \frac{10}{11} = \frac{3}{5500}, \qquad P(B) = 26\% = \frac{13}{50}.$$

Potom stačí pravděpdobnost, že kuřák onemocněl, vydělit pravděpodobností, že byl obyvatel ČR kuřák.

$$P(A/B) = \frac{\frac{3}{5500}}{\frac{13}{50}} = \frac{15}{1430} \doteq 0{,}0105.$$

Pravděpodobnost, že v roce 2005 onemocněl určitý český kuřák rakovinou, je přibližně 1%.

3.2 Nezávislé náhodné jevy

Skutečnost, že nastoupení jevu A neovlivní pravděpodobnost, že nastal jev B, lze matematicky zapsat následujícím způsobem

$$P(B/A) = P(B). \tag{2}$$

Ze vztahu (1) vyplývá

$$P(A \cap B) = P(A) \cdot P(B/A) \tag{3}$$

a ze vztahů (2) a (3) pak dostaneme definici nezávislosti náhodných jevů A a B.

Definice 8 Řekneme, že dva jevy A a B jsou nezávislé, jestliže platí

$$P(A \cap B) = P(A) \cdot P(B).$$

S nezávislostí náhodných jevů se v teorii pravděpodobnosti a statistice setkáváme často, nevyhneme se jí ani v této práci.

Příklad 8 V jednom skladišti na mouku se rozšířily dva druhy škůdců, roztoč moučný a obaleč moučný. Ve skladu je celkem 21 000 kg mouky. Každý je zabalen zvlášť. Roztoč moučný napadl celkem 300 kg mouky zatímco obaleč moučný dokonce 2400 kg. Počet balení napadených oběma je roven 60. Mají škůci v tomto skladu raději balíčky napadené druhým druhem, vyhýbají se mu, nebo je jim to jedno?

Řešení Jako $P(A)$ si označíme pravděpodobnost napadení roztočem moučným, $P(B)$ reprezentuje pravděpodobnost poškození balení obalečem moučným. $P(A \cap B)$ vyjadřuje pravděpodobnost, že balíček napadli oba škůdci.

$$P(A) = \frac{300}{21000} = \frac{1}{70}, \quad P(B) = \frac{2400}{21000} = \frac{8}{70}, \quad P(A \cap B) = \frac{60}{21000} = \frac{2}{700}.$$

Mohou nastat 3 případy:

1. Jestliže $P(A) \cdot P(B) = P(A \cap B)$, pak je škůdcům jedno, jaký balíček napadají.
2. Pokud $P(A) \cdot P(B) < P(A \cap B)$, škůdci raději napadají stejné balíčky.
3. Když platí, že $P(A) \cdot P(B) > P(A \cap B)$, škůdci se sobě navzájem vyhýbají.

Dosazením konkrétních hodnot dostaneme

$$P(A) \cdot P(B) = \frac{1}{70} \cdot \frac{8}{70} = \frac{8}{4900} < \frac{2}{700} = \frac{14}{4900}.$$

Škůdci v tomto skladu raději napadají balíčky napadené druhým druhem.

4 Náhodná veličina

Snad každý výsledek náhodného pokusu se dá vyjádřit číslem. Nejvíce z vás si jistě vzpomene na hod kostkou. Výsledek tohoto náhodného pokusu označujeme čísly jedna až šest. Stejně tak můžeme číslem označit výšku člověka, počet aut prodaných v roce 2011, počet atomů smolince, které se rozpadly za poslední minutu, a další. Jednoduše řečeno je *náhodná veličina* reálná funkce, která výsledkům náhodného pokusu přiřazuje čísla. Náhodné veličiny označujeme velkými písmeny z konce abecedy, $X, Y, Z, \ldots$; jejich konkrétní realizace pak malými písmeny, $x, y, z, \ldots$. Následující zjednodušená definice náhodné veličiny je uvažovaná pro Ω s konečně a spočetně mnoha prvky (pro nespočetné Ω by vypadala analogicky, bylo by ovšem potřeba uvažovat další teoretické pojmy, které by neúměrně zvětšily rozsah práce).

Definice 9 Libovolná reálná funkce $X : \Omega \to \mathbb{R}$, která každému možnému výsledku $\omega \in \Omega$ přiřazuje reálné číslo $X(\omega)$, se nazývá *náhodná veličina* a číslo $x = X(\omega)$ je *číselná realizace veličiny* X příslušná možnému výsledku ω.

Příklad 9 Při hodu mincí si zavedeme náhodnou veličinu, která nabývá dvou hodnot. Padnutí panny jako výsledek hodu si označíme např. číslem 0 a orla číslem 1. Jestliže padne panna, hodnota náhodné veličiny $X(\omega_0) = 0$. Jestliže padne orel, $X(\omega_1) = 1$.

4.1 Rozdělení pravděpodobnosti

Rozdělení pravděpodobnosti náhodné veličiny je soubor pravidel, která každému výsledku přiřazují určitou pravděpodobnost, se kterou náhodná veličina nabude příslušné realizace. Konkrétní rozdělení je určeno všemi hodnotami, kterých může náhodná veličina nabývat, a odpovídajícími pravděpodobnostmi.

Příklad 10 (dle [5], str. 37) Charakterizujte rozdělení pravděpodobnosti náhodné veličiny označujících počet panen při současném hodu třemi mincemi, tj. že padnou 0 $1, 2,$ nebo 3 panny.

Řešení Orla si označíme číslem 1 a pannu číslem 0. Množina všech možných výsledků obsahuje všechny možné výsledky náhodného pokusu,

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}.$$

Výsledky jsou trojice uspořádaných nul a jedniček,

$$\omega_1 = \{1,1,1\},$$
$$\omega_2 = \{1,1,0\}, \omega_3 = \{1,0,1\}, \omega_4 = \{0,1,1\},$$
$$\omega_5 = \{1,0,0\}, \omega_6 = \{0,1,0\}, \omega_7 = \{0,0,1\},$$
$$\omega_8 = \{0,0,0\}.$$

Náhodná veličina X, která udává počet padlých panen, nabude následujících hodnot s těmito pravděpodobnostmi:

$$P(X=0) = P(\{\omega_1\}) = \tfrac{1}{8},$$
$$P(X=1) = P(\{\omega_2\} \cup \{\omega_3\} \cup \{\omega_4\}) = \tfrac{3}{8},$$
$$P(X=2) = P(\{\omega_5\} \cup \{\omega_6\} \cup \{\omega_7\}) = \tfrac{3}{8},$$
$$P(X=3) = P(\{\omega_8\}) = \tfrac{1}{8}.$$

Pravděpodobnost, že nepadne panna (náhodná veličina nabývá hodnoty 0) je $\frac{1}{8}$. Pravděpodobnost, že padne právě 1 panna (náhodná veličina nabývá hodnoty 1) je rovna hodnotě $\frac{3}{8}$. Dvě panny (náhodná veličina nabývá hodnoty 2) se na horní straně mince objeví také s pravděpodobností $\frac{3}{8}$. 3 panny (náhodná veličina nabývá hodnoty 3) padnou s pravděpodbností $\frac{1}{8}$.

V uvedeném příkladu jsme se setkali s tzv. *diskrétním rozdělením.* Jedná se o rozdělení pravděpodobnosti náhodné veličiny, která nabývá konečně nebo spočetně mnoha hodnot, tedy obecně hodnot z množiny $\{x_1, x_2, x_3, \ldots\}$. Pro pravděpodobnost diskrétní náhodné veličiny platí tyto vlastnosti převzaté z [1] :

$$p_j = P(X = x_j) > 0, \qquad \sum_j p_j = 1.$$

Mezi náhodné veličiny s tímto typem rozdělení patří také počet prodaných výrobků za měsíc, počet narozených dětí v rodině či počet atomů sloučených během reakce.

4.1.1 Spojité rozdělení

Co by se však stalo, kdybychom zkoumali rozdělení pravděpodobnosti výšky nebo váhy člověka? Hodnoty těchto náhodných veličin by nenabývaly pouze hodnot $\{2\,kg, 3\,kg, 4\,kg, \ldots\}$ v případě váhy a hodnot $\{50\,cm, 51\,cm, 52\,cm, \ldots\}$ v případě výšky. Výška dospělého člověka je totiž reálné číslo z určitého

„rozumného“ reálného intervalu, např. $\langle 150, 210\rangle$. Za takových podmínek se jedná o *rozdělení spojitého typu*.

Náhodná veličina se spojitým rozdělením nabývá hodnot z nějakého intervalu, přitom pravděpodobnost realizace náhodné veličiny $P(X = x)$ je rovna nule, nenulových pravděpodobností může nabýt až realizace z nějakého celého podintervalu. Stejně tak, jako se dá do grafu zaznačit rozdělení pravděpodobnosti diskrétní náhodné veličiny, tak to jde i u rozdělení pravděpodobnosti veličiny spojitého typu. Funkce $f(x)$, charakterizující rozdělení pravděpodobnosti tohoto typu, se nazývá *hustota pravděpodobnosti*,

$$\lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = f(x), \ \forall\, x \in \mathbb{R}.$$

Obdobně jako pravděpodobnosti $p_j \geq 0$ u diskrétní náhodné veličiny, také pro tuto funkci platí $f(x) \geq 0$. Rozdíl je však v tom, že $f(x)$ nemusí nabývat pouze hodnot z intervalu $\langle 0, 1\rangle$, ale $f(x) \in \langle 0, \infty)$.

Další vlastnost hustoty, která představuje analogii u diskrétního případu, lze zapsat jako

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Poznamenejme, že náhodnou veličinou s tímto typem rozdělení je také velikost chyby fyzikálního měření nebo doba (v časových jednotkách) jasného počasí během dne.

4.2 Distribuční funkce náhodné veličiny

Distribuční funkce slouží k popisu rozdělení pravděpodobnosti náhodné veličiny a říká nám, jaká je pravděpodobnost, že náhodná veličina X bude mít hodnotu, která je menší nebo rovna stanovené reálné hodnotě. Je definovaná pro náhodnou veličinu s diskrétním i spojitým rozdělením. Značí se $F(x)$, pokud v textu není více náhodných veličin, které by se mohly plést. V opačném případě píšeme $F_X(x)$.

Jestliže se jedná o náhodnou veličinu s diskrétním rozdělením, má její distribuční funkce „skoky“. Způsobuje to skutečnost, že $P(X = x_j) > 0$. „Schody“ se tedy nacházejí v bodech x_j, které představují realizace této veličiny s nenulovými pravděpodobnostmi.

Definice 10 Nechť X je náhodná veličina. Reálná funkce F_X definovaná na $\mathbb{R}$ předpisem

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R},$$

se nazývá *distribuční funkce náhodné veličiny* X.

Z definice vyplývá, že se v případě distribuční funkce spojité náhodné veličiny $F(x)$ se jedná o primitivní funkci k hustotě $f(x)$,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

Vlastnosti všech distribučních funkcí, převzaté z [1] a [2]:

1. $0 \leq F(x) \leq 1$, $\forall x \in \mathbb{R}$,
2. $F(x)$ je neklesající,
3. $\lim_{x\to\infty} F(x) = 1$,
4. $\lim_{x\to-\infty} F(x) = 0$,
5. pro libovolná reálná $a < b$ platí $P(a < X \leq b) = F(b) - F(a)$,
6. distribuční funkce je *zprava spojitá*.

Příklad 11 Maloobchodní řetězec potřebuje každý týden dodávku nealkoholického nápoje, který prodává spotřebitelům. Na základě dlouhodobého pozorování lze týdenní požadavek v tisících litrech vyjádřit pomocí spojité náhodné veličiny X s hustotou

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \text{jinak.} \end{cases}$$

Najděte distribuční funkci náhodné veličiny.

Řešení Je jasné, že pro $x \leq 1$ je $F(x) = 0$ a pro $x \geq 2$ je $F(x) = 1$. Otázkou je, jak funkce vypadá na intervalu $(1, 2)$. Abychom to zjistili, stačí funkci vyjadřující hustotu náhodné veličiny na intervalu $(1, 2)$ zintegrovat. Distribuční funkce je totiž funkce primitivní k funkci vyjadřující hustotu,

$$\int_1^x 2(t-1)\,dt = 2\left[\frac{t^2}{2} - t\right]_1^x = x^2 - 2x - (-1) = x^2 - 2x + 1.$$

Distribuční funkce je tedy tato:

$$F(x) = \begin{cases} 0, & x \leq 1, \\ x^2 - 2x + 1, & 1 < x < 2, \\ 1, & x \geq 2. \end{cases}$$

Stejně jako ostatní distribuční funkce, i tato splňuje podmínky 1 až 6.

5 Některá konkrétní rozdělení pravděpodobnosti spojitých náhodných veličin

Protože v našich dalších úvahách využijeme především náhodné veličiny se spojitým rozdělením, zmiňme se nyní o některých známých rozděleních (zejména) tohoto typu, která se v praxi zřejmě vyskytují nejčastěji.

5.1 Rovnoměrné rozdělení

Toto rozdělení patří mezi nejjednodušší. Každé hodnotě náhodné veličiny je přiřazena stejná pravděpodobnost. Náhodná veličina s tímto rozdělením může být diskrétní i spojitá.

Z toho, že rovnoměrné rozdělení přiřazuje každé hodnotě náhodné veličiny stejnou pravděpodobnost, se dá vyvodit, že se diskrétní rovnoměrné rozdělení náhodné veličiny X vyznačuje následujícím předpisem,

$$P(X = x_j) = \frac{1}{n},\ j = 1, 2 \ldots, n.$$

Pro potřeby této práce bude důležitá zejména spojitá forma rovnoměrného rozdělení. Uplatníme ji při generování pseudonáhodných čísel (viz kapitola 6). Rovnoměrné rozdělení spojité náhodné veličiny X na intervalu (a, b) má tuto hustotu:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b), \\ 0, & x \notin (a, b). \end{cases}$$

Distribuční funkce je dána následovně:

$$F(x) = \begin{cases} 0, & x \leq a, \\ \int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

Píšeme $X \sim Ro(a, b)$, kde a a b jsou parametry tohoto rozdělení.

5.2 Exponenciální rozdělení

Toto rozdělení se často využívá v případech, kdy nás zajímá rozdělení pravděpodobnosti délky časového intervalu, než nastane nějaká událost Jako příklad lze uvést dobu, za kterou se porouchá dané zařízení (uvažujeme–li poruchy z náhodných příčin a nikoli v důsledku mechanického opotřebení, příkladem jsou různá elektronická zařízení).

Exponenciální rozdělení náhodné veličiny X označujeme $X \sim Ex(\lambda)$. Hustota a distribuční funkce tohoto rozdělení vypadají následovně:

$$f(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0; \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

5.3 Normální rozdělení

Toto rozdělení je snad nejznámějším rozdělením pravděpodobnosti náhodných veličin spojitého typu a též nejčastěji se vyskytujícím rozdělením ve fyzikálních měřeních. Charakterizuje mnoho náhodných veličin, na které v aplikacích narazíme téměř na každém kroku. Mimo jiné mezi ně patří náhodné veličiny popisující velikost chyby měření, výšku člověka, vlnovou délku fotonu vyzářeného Sluncem nebo jinou hvězdou. Normální rozdělení se dělí na dva základní typy, *normální normované rozdělení* a *obecné normální rozdělení.*

Normální normované rozdělení náhoné veličiny X označujeme $X \sim N(0, 1)$. Jeho hustota se značí $\varphi(x)$ a distribuční funkce $\Phi(x)$. Toto rozdělení má náhodná veličina s hustotou

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$$

a distribuční funkcí

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt, x \in \mathbb{R}.$$

Hustota normovaného normálního rozdělení je sudá funkce, takže stačí tabelovat pouze její hodnoty pro $x \geq 0$.

Obecné normální rozdělení náhodné veličiny X značíme $X \sim N(\mu, \sigma^2)$. Přitom μ a σ jsou parametry (konstanty), pro které platí $\mu \in \mathbb{R}$ a $\sigma \in \mathbb{R}^+$. Obecné normální rozdělení má pak hustotu

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

a distribuční funkci

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, x \in \mathbb{R}.$$

Jak je psáno v [2], pro získání hodnot distribuční funkce rozdělení $N(\mu, \sigma^2)$ stačí tabelovat hodnoty distribuční funkce $\Phi(x)$ normovaného normálního rozdělení, protože platí

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), x \in \mathbb{R}.$$

Je zřejmé, že normální normované rodělení je pouze speciálním případem (obecného) normálního rozdělení. Pro uvedené konstanty zde platí $\mu = 0$ a $\sigma = 1$.

5.4 Trojúhelníkové rozdělení

Trojúhelníkové rozdělení se používá např. v ekonomických aplikacích. Má tři reálné parametry a, b a c, pro které platí $a < c < b$. Podle [10] je jeho hustota

$$f(x) = \begin{cases} 0, & x < a, \\ \frac{2(x-a)}{(b-a)(c-a)}, & a \leq x < c, \\ \frac{2}{b-a}, & x = c, \\ \frac{2(b-x)}{(b-a)(b-c)}, & c < x \leq b, \\ 0, & x > b \end{cases}$$

a distribuční funkce

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{(x-a)^2}{(b-a)(c-a)}, & a \leq x < c, \\ \frac{c-a}{b-a}, & x = c, \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)}, & c < x \leq b, \\ 1, & x > b. \end{cases}$$

V kapitole 7 budeme při transformaci náhodné veličiny X s rovnoměrným rozdělením na $(0, 1)$ na výběrové hodnoty náhodné veličiny Y s předepsaným rozdělením potřebovat výpočet inverzní funkce k distribuční funkci (tzv. kvantilové funkce), proto je zde zařazen následující příklad.

Příklad 12 Vypočítejte inverzní funkci k distribuční funkci trojúhelníkového rozdělení pravděpodobnosti na intervalu (a, b).

Řešení Výpočet inverzní funkce pro $x \in \langle a, c\rangle$: Vyměníme x a y ve vhodné části distribuční funkce:

$$x = \frac{(y-a)^2}{(b-a)(c-a)}.$$

Jmenovatelem můžeme násobit, protože $a \neq b \wedge a \neq c$. Po úpravě dostaneme

$$0 = y^2 - 2ay + a^2 - x(b-a)(c-a),$$

$$y_{1,2} = a \pm \sqrt{x(b-a)(c-a)}.$$

Za x dosadíme $\frac{c-a}{b-a}$ a vybereme + nebo − podle toho, kdy se výraz rovná c:

$$y_{1,2} = a \pm \sqrt{\frac{c-a}{b-a}(b-a)(c-a)} = a \pm (c-a).$$

Z toho je jasné, že vybereme +. Výsledek:

$$y = a + \sqrt{x(b-a)(c-a)}$$

Výpočet inverzní funkce pro $x \in \langle c, b\rangle$: Zaměníme x a y ve vhodné části distribuční funkce:

$$x = 1 - \frac{(b-y)^2}{(b-a)(b-c)}.$$

Jmenovatelem můžeme násobit, protože $b \neq a \wedge b \neq c$. Po úpravě:

$$0 = y^2 - 2by + b^2 + (x-1)(b-a)(b-c),$$

$$y_{1,2} = b \pm \sqrt{(1-x)(b-a)(b-c)}.$$

Za x dosadíme $\frac{c-a}{b-a}$ a vybereme + nebo − podle toho, kdy se výraz rovná c:

$$y_{1,2} = b \pm \sqrt{(1 - \frac{c-a}{b-a})(b-a)(b-c)} = b \pm (b-c).$$

Z toho je jasné, že vybereme −. Výsledek je roven

$$y = b - \sqrt{(1-x)(b-a)(b-c)}.$$

Inverzní funkce k distribuční funkci trojúhelníkového rozdělení na (a, b) je tedy rovna

$$F^{-1}(x) = \begin{cases} a + \sqrt{x(b-a)(c-a)}, & 0 \leq x < \frac{c-a}{b-a}, \\ c, & x = \frac{c-a}{b-a}, \\ b - \sqrt{(1-x)(b-a)(b-c)}, & \frac{c-a}{b-a} < x \leq 1. \end{cases}$$

6 Generování náhodných čísel

Abychom mohli simulovat rozdělení pravděpodobností náhodných veličin pomocí metody Monte Carlo, potřebujeme náhodná čísla s k desetinnými místy z intervalu $(0, 1)$.Tato čísla by po vygenerování měla mít rovnoměrné rozdělení s příslušnými parametry. Realizace rovnoměrného rozdělení se pak v případě potřeby dají transformovat na realizace předepsaného rozdělení (viz dále). Jak ale náhodná čísla získat?

6.1 Základní způsoby generování náhodných čísel

V minulosti se používaly různé nástroje generování náhodných čísel. Tím nejprimitivnějším se ovšem stala ruleta. Stačí zatočit k-krát ruletou s deseti částmi reprezentujícími číslice 0 až 9. Jednotlivé číslice postupně zapisujeme za desetinnou čárku. Tím jsme získali jedno vyhovující číslo. Také můžeme házet desetistěnnou kostkou nebo tahat čísla 0 až 9 z urny. Dále bychom postupovali jako v případě rulety. Tento způsob generování náhodných čísel je ovšem neefektivní a v praxi nepoužitelný. Trvalo by totiž celou věčnost, než bychom vygenerovali 10^4 až 10^5 potřebných náhodných čísel v případě jednodužších simulací, 10^6 až 10^9 v případě náročnějších simulací.

V minulosti se také používaly *fyzikální* neboli *hardwarové metody* generování náhodných čísel. Jedná se o využití náhodného fyzikálního děje. K tomu se hodil např. rozpad radioktivní látky. Měřila se doba mezi jednotlivými rozpady atomů. Mezi další fyzikální generátory patří také snímání šumu elektronky.

Dnes se používají *výpočetní* neboli *softwarové generátory pseudonáhodných čísel*. Čísla již nejsou opravdu náhodná, protože jsou vygenerována pomocí algoritmu. U vhodných algoritmů se jejich vlastnosti velice podobají vlastnostem náhodných čísel, proto dochází ke stírání pojmů *náhodné* a *pseudonáhodné číslo*.

Tyto generátory fungují nejčastěji na stejném základním principu. Máme číslo x_0. Další čísla získáváme pomocí rekurentního vztahu $x_{n+1} = f(x_n)$. Dojde ke vzniku posloupnosti pseudonáhodných čísel. Musíme si však dávat pozor, protože takové posloupnosti bývají periodické. Po K opakováních získáme číslo, které se rovná číslu x_k, kde $k < K$. Množina čísel $x_0, x_1, x_2, \ldots, x_K$ se nazývá *úsek aperiodičnosti*. Hodnota K je pak *délka úseku aperiodičnosti*, $P = K - k$ se nazývá *délka periody*. K výpočtu se obvykle nedoporučuje použít více než K čísel. V některých případech, kdy je málo pravděpodobné, že se číslo použije k simulování stejného děje, se může použít i více čísel.

Dnes pro potřeby metody Monte Carlo nemusíme znát systém generování čísel. Zpravidla tyto algoritmy neprogramujeme sami, ale jsou součástí matematických a speciálně statistických softwarů. Mezi ně se řadí např. statistický software R (`www.r-project.org`), ve kterém jsou také provedeny simulace v této práci.

Pro zajímavost se v závěru této kapitoly zmíníme o jedné metodě generování pseudonáhodných čísel. Konkrétně, historicky nejstarší softwarovou metodou generování pseudonáhodných čísel se stala *metoda středu kvadrátu*. Byla navržena Johnem von Neumannem v roce 1951.

Mějme $2k$-místné číslo x_0. Toto číslo umocníme na druhou a vznikne $4k$-

místné číslo. Z něj vyjmeme prostředních $2k$ číslic a vzniklé číslo považujeme za x_1. Postup opakujeme. Daná poslupnost však rychle degeneruje a rozdělení takto vygenerovaných čísel se liší od rovnoměrného.

7 Modelování hodnot náhodné veličiny

Z hlediska metody Monte Carlo je velice užitečné umět transformovat hodnoty náhodné veličiny X s rovnoměrným rozdělením na $\langle 0, 1\rangle$ na hodnoty náhodné veličiny Y se zadaným rozdělením pravděpodobnosti.

7.1 Diskrétní náhodná veličina

Zde uvedený postup pro získání výběrových hodnot veličiny Y je převzat z [5]. Pokud má Y pravděpodobnostní funkci $P(Y = y_k) = p_k$, kde $k = 1, 2, \ldots, n$, platí

$$P(Y = y_m) = p_m = P\left(\sum_{k=1}^{m-1} p_k < X \leq \sum_{k=1}^{m} p_k\right),\ m = 1, \ldots, n.$$

Postup hledání výběrových hodnot náhodné veličiny Y vychází z této skutečnosti. Pro každé x_j z posloupnosti realizací $\{x_j\}$ rovnoměrného rozdělení určíme takové m, aby platilo

$$\sum_{k=1}^{m-1} p_k < x_j \leq \sum_{k=1}^{m} p_k.$$

Potom říkáme, že v $j-$tém pokuse nastal jev $A = (Y = y_m)$. Tímto způsobem získáme posloupnost výběrových hodnot veličiny Y.

Jinak řečeno, pravděpodobnosti p_m si můžeme zobrazit pomocí úsečky o délce 1, jejíž krajní body si označíme P_0 a P_n. Na ní zvýrazníme bod, jehož vzdálenost od P_0 je rovna velikosti p_1. Označíme ho P_1. Dále si označíme P_2 bod, jehož vzdálenost od P_1 je rovna velikosti p_2, atd. Je zřejmé, že vzdálenost P_m od P_0 je rovna $\sum_{k=1}^{m} p_k$. Na stejnou úsečku si také můžeme vyznačit jednotlivá x_j z posloupnosti $\{x_j\}$. Náleží-li hodnota x_j úsečce ohraničené body P_{m-1} a P_m říkáme, že v $j-$tém pokuse nastal jev $A = (Y = y_m)$. Metoda vychází z toho, že čím je větší úsečka ohraničená body P_{m-1} a P_m, tedy i pravděpodobnost p_m, tím spíše hodnota x_j padne na zmíněnou úsečku.

7.2 Spojitá náhodná veličina

Při generování hodnot náhodné veličiny s rozdělením pravděpodobnosti spojitého typu se užívá následující věta z [5].

Věta 1 Nechť náhodná veličina X má rovnoměrné rozdělení na $(0,1)$ a nechť G^{-1} je inverzní funkce k nějaké rostoucí spojité distribuční funkci G. Potom náhodná veličina $Y = G^{-1}(X)$ má distribuční funkci G.

Důkaz Tvrzení věty dostaneme užitím definice distribuční funkce a vlastností funkce G,

$$P(Y \leq y_m) = P(G^{-1}(X) \leq y) = P(X \leq G(y)) = G(y).$$

□

Máme-li tedy k dispozici posloupnost $\{x_j\}$ hodnot náhodné veličiny X s rovnoměrným rozdělením na $(0,1)$, můžeme pomocí vztahů

$$y_j = G^{-1}(x_j), \qquad G(y_j) = x_j$$

získat hodnoty náhodné veličiny Y se spojitou a rostoucí distribuční funkcí $G(y)$.

Příklad 13 Nechť má Y trojúhelníkové rozdělení s distribuční funkcí

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{(x-a)^2}{(b-a)(c-a)}, & a \leq x < c, \\ \frac{c-a}{b-a}, & x = c, \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)}, & c < x \leq b, \\ 1, & x > b. \end{cases}$$

Transformujte hodnoty náhodné veličiny $X \sim Ro(0,1)$ na hodnoty náhodné veličiny Y se zadanou distribuční funkcí.

Řešení Z příkladu 12 v páté kapitole o některých konkrétních rozděleních víme, že inverzní funkce k zadané distribuční funkci je

$$F^{-1}(x) = \begin{cases} a + \sqrt{x(b-a)(c-a)}, & 0 \leq x < \frac{c-a}{b-a}, \\ c, & x = \frac{c-a}{b-a}, \\ b - \sqrt{(1-x)(b-a)(b-c)}, & \frac{c-a}{b-a} < x \leq 1. \end{cases}$$

Z věty 1 vyplývá, že posloupnost hodnot veličiny Y je dána vztahem

$$y_j = \begin{cases} a + \sqrt{x_j(b-a)(c-a)}, & 0 \leq x_j < \frac{c-a}{b-a}, \\ c, & x_j = \frac{c-a}{b-a}, \\ b - \sqrt{(1-x_j)(b-a)(b-c)}, & \frac{c-a}{b-a} < x_j \leq 1. \end{cases}$$

Příklad 14 Transformujte hodnoty náhodné veličiny $X \sim Ro(0,1)$ na hodnoty náhodné veličiny Y s hustotou

$$g(y) = \begin{cases} 0, & y \leq 0, \\ \lambda e^{-\lambda y}, & y > 0. \end{cases}$$

Řešení Nejprve vypočítáme odpovídající distribuční funkci ke stanovené hustotě (viz kapitola 5.2),

$$\int_0^y \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^y = -e^{-\lambda y} + e^{-0\lambda} = 1 - e^{-\lambda y},$$

tedy

$$F(y) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-\lambda y}, & y > 0. \end{cases}$$

Z toho víme, že $1 - e^{-\lambda y_j} = x_j$. Nalezení inverzní funkce k této distribuční funkci odpovídá vyjádření y_j, které hledáme. Ze vztahu

$$1 - x_j = e^{-\lambda y_j} \Rightarrow \ln(1 - x_j) = -\lambda y_j$$

dojdeme k výsledku

$$y_j = -\frac{1}{\lambda} \ln(1 - x_j).$$

8 Testování normálního rozdělení

Již zmiňované normální rozdělení (viz kapitola 5.3) patří mezi klíčová rozdělení matematické statistiky. Proto se mnohdy setkáváme s otázkou, zda má nějaká zkoumaná náhodná veličina X normální rozdělení. K zodpovězení samozřejmě potřebujeme dostatečný počet jejích realizací $x_1, x_2, \ldots, x_n$ (realizace tzv. *náhodného výběru*), se kterými můžeme dále pracovat.

Podívejme se na testování statistických hypotéz rovnou v kontextu testování normality náhodného výběru, potažmo veličiny X. Hypotéze, která říká, že má veličina X normální rozdělení, říkáme *nulová hypotéza* a značíme ji H_0. Naopak H_A značíme tzv. *alternativní hypotézu*, která tvrdí, že náhodná veličina X nemá normální rozdělení. Rozhodnutí o H_0, kterou zamítáme nebo ji nelze zamítnout, provádíme pomocí testovací statistiky T. Hodnotu T získáme pomocí vztahu

$$T = h(X_1, X_2, \ldots, X_n),$$

kde $X_1, X_2, \ldots, X_n$ jsou navzájem nezávislé náhodné veličiny se stejným rozdělením jako má náhodná veličina X (náhodný výběr o rozsahu n) - veličina T je tak vlastně funkcí daného náhodného výběru. Za $X_1, X_2, \ldots, X_n$ dosazujeme hodnoty realizací náhodné veličiny X, $x_1, x_2, \ldots, x_n$. Tímto postupem získáme konkrétní realizaci náhodné veličiny T, kterou značíme t.

Zavádíme také číslo α, které nazýváme *hladina testu*. Rovná se námi zvolené hodnotě pravděpodobnosti, že platnou H_0 pomocí realizace statistiky T (nesprávně) zamítneme ve prospěch alternativní hypotézy H_A. Pro představu lze uvést nejčastější používané hodnoty hladiny testu α, a to $0,05; 0,1$ a $0,01$. Po výpočtu hodnoty testovací statistiky, t, stanovíme velikost α_0 neboli p-hodnoty. Odpovídá pravděpodobnosti, že bychom při testování normality náhodné veličiny dostali hodnotu testovací statistiky, která ještě více odporuje nulové hypotéze. Hladina testu α je tedy stabilní p-hodnota, se kterou p-hodnoty získané při testování porovnáváme. Jestliže platí, že $\alpha_0 < \alpha$, H_0 *zamítáme*. Jestliže $\alpha_0 \geq \alpha$, H_0 *nelze zamítnout*.

8.1 Anderson-Darlingův test normality

V dalším budeme uvažovat jeden konkrétní test normality, který se v praxi patří k nejčastěji užívaným. Údaje o Anderson-Darlingovu testu jsou převzaty z [7]. Pro potřeby testu nejprve z realizace náhodného výěru odhadneme parametry μ a σ^2 normálního rozdělení (testování totiž provádíme za předpokladu platnosti nulové hypotézy). Odhad parametru μ odpovídá aritmetickému průměru $x_1, x_2, \ldots, x_n$,

$$\mu \sim \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Odhad parametru σ získáme pomocí směrodatné odchylky s,

$$\sigma \sim s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Pomocí počítače a distribuční funkce normálního normovaného rozdělení dále získáme hodnoty

$$\Phi\left(\frac{x_i - \overline{x}}{s}\right) = z_i.$$

Ty dosadíme do testovací statistiky Anderson-Darlingova testu normality,

$$Q_\alpha = \left(\frac{25}{n^2} - \frac{4}{n} - 1\right)\left(\frac{1}{n} \sum_{r=1}^{n} (2r-1)[\ln z_i + \ln(1 - z_{n+1-r}) + n]\right).$$

Konkrétní realizaci Q_a, tedy q_a, porovnáme s tabulkovými hodnotami, abychom zjistili přibližnou velikost p-hodnoty α_0 a mohli dospět k rozhodnutí o nulové hypotéze. Porovnáním α_0 s námi určenou hladinou testu α zjistíme, zda nulovou hypotézu H_0 zamítáme či ji nelze zamítnout.

Například hladině testu 0,1 odpovídá realizace náhodné veličiny Q_a hodnotě $q_{a;0,1} = 0,656$, pro $\alpha = 0,05$ je určena hodnota $q_{a;0,05} = 0,787$ a pro $\alpha = 0,01$ je dáno $q_{a;0,01} = 1,092$.

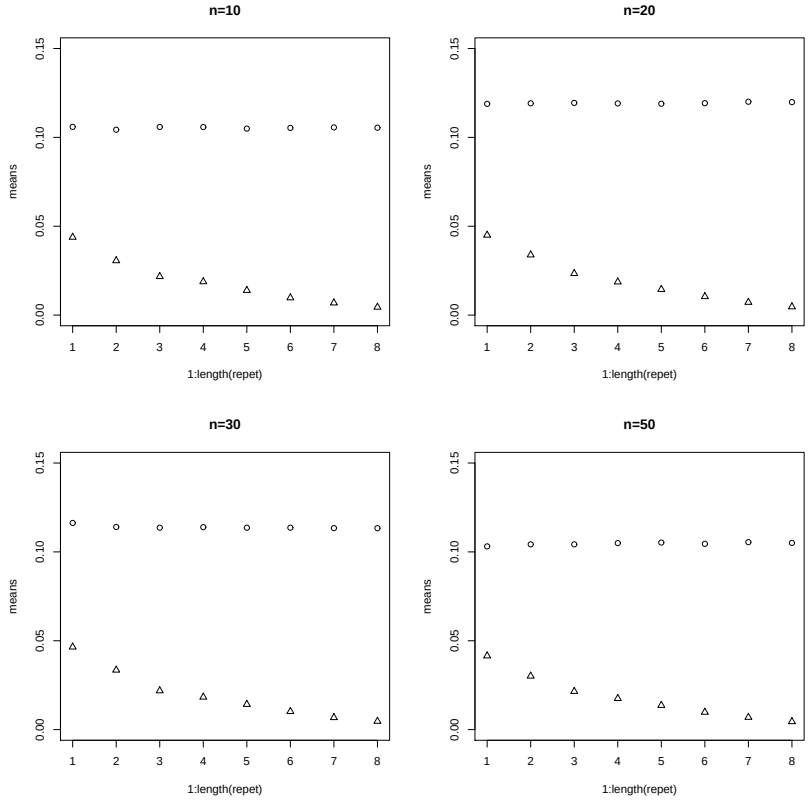
Při použití Anderson-Darlingova testu se obvykle používá porovnání s těmito hodnotami. Je–li realizace testovací statistiky větší než daná kritická hodnota, nulovou hypotézu na příslušné hladině zamítáme. Nevýhodou se však stává to, že nemáme příliš velkou představu o již zmíněné p-hodnotě. K tomu, abychom tuto hodnotu znali, bychom však potřebovali distribuční funkci náhodné veličiny Q_a, která se ovšem v praxi pro nemožnost explicitního vyjádření neužívá. Pro její přibližné určení můžeme užít metodu Monte Carlo. Stačí nagenerovat dostatečné množství R náhodných výběrů o daném rozsahu n z normálního rozdělení (díky konstrukci Anderson-Darlingova testu nezáleží na hodnotách parametrů tohoto rozdělení, můžeme tedy bez újmy na obecnosti generovat z normálního normovaného rozdělení). Poté na tyto výběry aplikujeme Anderson-Darlingův test a dostaneme tak R realizací testovací statistiky Q_a. Díky takto vytvořeným hodnotám náhodné veličiny Q_a jsme schopni určit přibližné rozdělení Q_a. Čím více výběrů nagenerujeme, tím více se dostáváme ke skutečnému rozdělení Q_a. To se v naší situaci projeví tak, že budeme schopni stále lépe odhadnout skutečnou p-hodnotu pro realizaci testovací statistiky v našem konkrétním případě. Tento odhad je přitom roven relativní četnosti hodnot Q_a větších než realizace testovaného náhodného výběru.

Jak ale určit ono dostatečné množství R nagenerovaných výběrů o rozsahu n, abychom obdrželi odhad hustoty Q_a s velkou přesností, která se následně projeví na vyhovující přesnosti odhadu p-hodnot? Nechceme totiž současně generování výběrů a následným výpočtům hodnot testovací statistiky věnovat příliš mnoho času, což je bohužel daň jakékoli rozsáhlejší simulace.

8.2 Odhad p-hodnot v Anderson-Darlingově testu

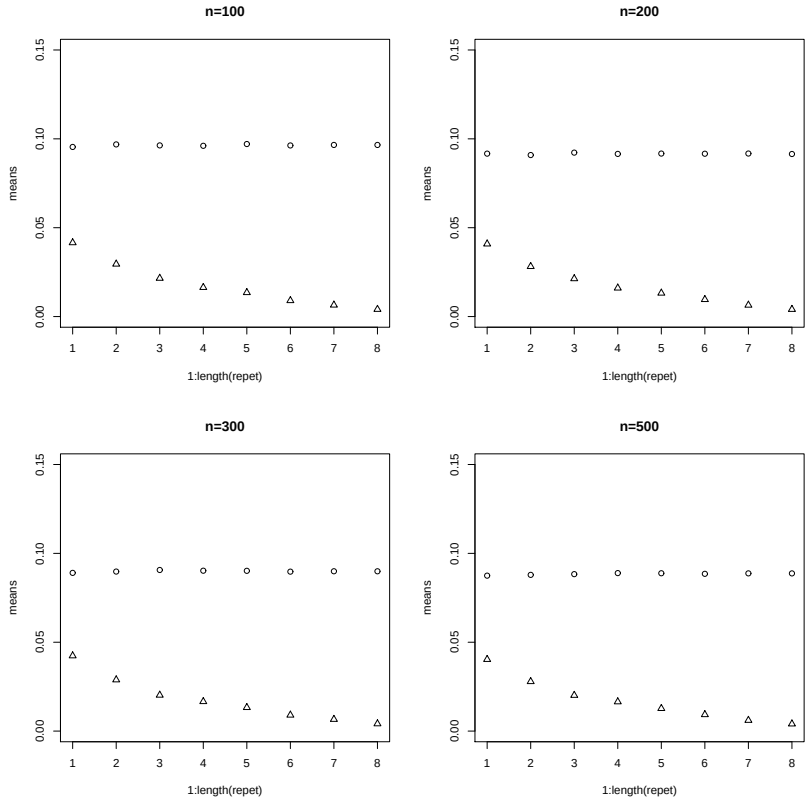
K účelu z konce předchozí kapitoly byla zapotřebí funkce `adtest` z knihovny `robCompositions` ve statistickém softwaru R. Byla přepracovaná tak, aby program v softwaru R po spuštění nageneroval R výběrů z normálního normovaného rozdělení o rozsahu n, pro každý výběr se spočítala hodnota Anderson-Darlingova testu a výsledky se porovnaly s hodnotou $q_{a;0,1} = 0,656$ (viz přiložené soubory). Dále program určí poměr počtu hodnot x_i pro $i = 1, 2, \ldots, n$ v každém výběru, které jsou větší než $0,656$ ku počtu R. Ideálně

by měl tento poměr odpovídat číslu $0,01$. Čím větší je R, tím více by se poměr měl tomuto číslu přibližovat. Uvedený postup se opakuje celkem N-krát, aby byly dosažené výsledky simulace pokud možno co nejstabilnější. Z N obdržených odhadů p-hodnot vypočítá program průměr a také určí odpovídající směrodatnou odchylka. Hodnota průměru a směrodatné odchylky se pro každý ze zvolených rozsahů výběrů zaznačí do grafu, viz obrázky 3-5. Pomocí nich si každý může udělat představu o tom, jaké R stačí pro daná n. Již na první pohled vidíme, že se p-hodnoty pro různé hodnoty R stále pohybují kolem $0,01$, což je správně, hodnoty směrodatných odchylek ovšem klesají. Čím menší je přitom směrodatná odchylka, můžeme si být více jisti, že při konkrétní realizaci Anderson-Darlingova testu se zvoleným počtem R simulací dostaneme přesnější odhad p-hodnoty.



Obrázek 3: Výsledky simulace p-hodnot (jejich průměry jsou označeny $\circ$ a směrodatné odchylky $\triangle$) při různých volbách R pro $n = 10, 20, 30, 50$.

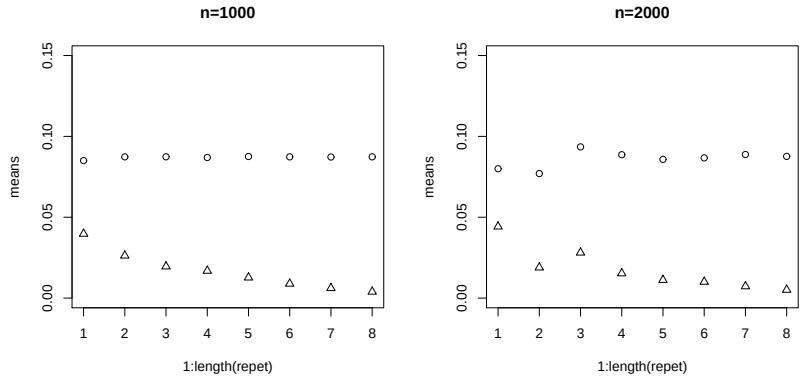
Ke generování p-hodnot testu bylo užito pro každou z voleb rozsahu výběru $n = 10, 20, 30, 50, 100, 200, 300, 500, 1000, 2000$ a stabilních hod-



Obrázek 4: Výsledky simulace p-hodnot (jejich průměry jsou označeny $\circ$ a směrodatné odchylky $\triangle$) při různých volbách R pro $n = 100,\ 200,\ 300,\ 500$.

not parametrů $N = 1000$ a $R = 50, 100, 200, 300, 500, 1000, 2000, 5000$ (těmto v grafech odpovídá kódové označení $1, \ldots, 8$). Každé volbě R v grafu odpovídá průměrnou hodnota odhadnuté p-hodnoty a odpovídající směrodatná odchylka. Jak je vidět, rozumné velikosti odchylek se objevují, když je velikost R v řádu tisíců, kdy se již setrvalý pokles směrodatných odchylek víceméně stabilizuje, a to bez ohledu na daný rozsah výběru n. K určování p-hodnot lze proto jako dolní hranici doporučit alespoň $R = 1000$. V případě velkých hodnot n a R (v řádu tisíců) už však výpočty trvají několik hodin. Ovšem, na druhou stranu, testujeme-li pro $n = 10$, trvá výpočet pouze několik minut i pro $R = 5000$, proto není důvod při nižších hodnotách n užívat hodnotu R menší než 5000.

Na grafech si můžete povšimnout toho, že průměry p-hodnot, které by se měly blížit k 0,1, se u $n = 10$ k hodnotě 0,1 skutečně blíží. Poté začínají pro vyšší n stoupat a potom zase klesat. U $n = 100$ se průměry opět velmi



Obrázek 5: Výsledky simulace p-hodnot (jejich průměry jsou označeny $\circ$ a směrodatné odchylky $\triangle$) při různých volbách R pro $n = 1000, 2000$.

blíží k $0,1$ a pro vyšší n klesají. Tento nežádoucí jev je zřejmě způsoben povahou generátoru náhodných čísel v softwaru R. Pro velmi přesná určení p-hodnot u některých n se proto zřejmě nehodí. Odchýlení od 0,1 však není, se zahrnutím informace o směrodatné odchylce, při vyšších hodnotách R až tak dramatické, takže většině uživatelů jistě bohatě postačí k relevantnímu rozhodnutí o nulové hypotéze.

Závěr

Na závěr této práce zhodnotíme splnění tří základních cílů, které jsem si před její tvorbou vytyčila.

Prvním cílem bylo popsat metodu Monte Carlo a základní pojmy nutné k jejímu pochopení. Věřím, že nejen díky slovnímu vysvětlení, ale také pomocí uvedených příkladů, se toto povedlo srozumitelnou formou. Nejlépe však tento aspekt jistě posoudí sám čtenář. Tato část práce se sice z časového hlediska stala tou nejnáročnější, ale jinak nebyla až tak obtížná jako část následující. Kromě toho mě velice bavilo pomalu pronikat do tajů teorie pravděpodobnosti, následně si kvalitu svého pochopení problematiky ověřovat na vysvětlení v této práci a nakonec tvořit i příklady.

Další cíle byly ukázat využití metody Monte Carlo při simulacích rozdělení pravděpodobností náhodných veličin a při stanovení optimálního postupu pro odhady p-hodnot při testování normality náhodného výběru pomocí Anderson-Darlingova testu. Tyto dva cíle jsem splnila na konci práce. Bylo třeba nagenerovat obrázky, z nich vyčíst potřebné údaje, zhodnotit je a nakonec si myšlenky utřídit tak, abych byla schopna tyto postupy co nejlépe popsat. Tato část práce se pro mě stala bez pochyby tou celkově nejtěžší.

Domnívám se, že výsledky ze závěru práce mohou být v praxi užitečné a snadno přenositelné i na další užívané testy normality. Konkrétním výsledkem je přitom například skutečnost, že pro přesnost určení p-hodnot je zřejmě podstatný pouze počet provedených simulací R, nikoli ovšem rozsah n uvažovaného výběru. Pokud tedy někomu při testování normality nebude stačit pouhé porovnávání kritické hodnoty z tabulky s realizací testovací statistiky odpovídající dané hladině testu, stačí využít výše uvedené poznatky. Práce se dá také dále rozvíjet, hlavně při studiu dalších testovacích statistik, jejichž rozdělení není vyjádřeno explicitně.

Doufám, že tato práce dobře poslouží zájemcům o tuto problematiku a pomůže jim nejen při jejím pochopení, ale také dalším rozvíjení některých myšlenek v ní uvedených.

Reference

[1] FABIAN, F., KLUIBER, Z., *Metoda Monte Carlo*. Praha: Prospektorum, 1998.

[2] KUNDEROVÁ, P., *Základy pravděpodobnosti a matematické statistiky*. Olomouc: VUP, 2004.

[3] BUDÍKOVÁ, M., *Statistika* (učební text). Brno: Masarykova univerzita, 2004.

[4] ZVÁRA, K., ŠTĚPÁN, J., *Pravděpodobnost a matematická statistika*. Praha: MATFYZPRESS, 2006.

[5] KUNDEROVÁ, P., *Metody Monte Carlo*. Olomouc: RUP, 1982.

[6] VIRIUS, M., *Aplikace matematické statistiky - Metoda Monte Carlo*. Praha: ČVUT, 1998.

[7] PAWLOWSKY-GLAHN, V., EGOZCUE, J.J., TOLOSANA-DELGADO, R., *Lecture notes on compositional data analysis*. [Cit. 5.4.2012]. Dostupné z URL: http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf

[8] http://www.angis.cz/angis_revue/ar_clanek.php?CID=127 [Cit. 13.10.2011]

[9] http://www.zdrav.cz/modules.php?op=modload&name=News &file=article&sid=8460 [Cit. 13.10.2011]

[10] http://www.en.wikipedia.org/wiki/Triangular_ distribution [Cit. 10.1.2012]

[11] http://cs.wikipedia.org/wiki/Buffonova_ jehla [Cit. 2.10.2011]