



Výpočtové a simulační metody štatistiky

Ivan Žežula

30. apríla 2008

Obsah

1	Druhy štatistických výpočtov	1
1.1	Úvod	1
1.2	Štandardné štatistické výpočty	3
1.2.1	Programové systémy pre numerickú matematiku	3
1.2.2	Tabulkové procesory	3
1.2.3	Malé štatistické systémy	3
1.2.4	Laické systémy	4
1.2.5	Veľké systémy	4
1.3	Neštandardné štatistické výpočty	4
1.4	Výpočty tabulkových hodnôt	5
1.5	Generovanie náhodných čísel a simulácie	5
1.6	Data mining	5
2	Niektoré praktické výpočtové metódy	7
2.1	Výpočet funkcií rozdelenia pravdepodobnosti	7
2.1.1	Normálne rozdelenie	7
2.1.2	Iné rozdelenia	9
2.2	Maticové výpočty	9
2.2.1	Inverzia symetrickej matice	10
2.2.2	Riešenie symetrickej sústavy lineárnych rovníc	10
2.2.3	Výpočet g-inverznej matice	11
3	Generovanie náhodných čísel	13
3.1	Rovnomerné rozdelenie	13
3.1.1	Lineárne rekurentné generátory	15
3.1.2	Seriálna korelácia	16
3.1.3	Bitové rekurentné generátory	18
3.1.4	Nelineárne generátory	21
3.2	Ostatné rozdelenia	21
3.2.1	Metóda inverznej transformácie	22
3.2.2	Zamietacia metóda	22
3.2.3	Podielová metóda	25
3.2.4	Kompozičná metóda	27
3.3	Špeciálne metódy	28

3.3.1	Normálne rozdelenie	28
3.3.2	Gama a beta rozdelenie	29
3.3.3	Rozdelenia súvisiace s normálnym	33
3.3.4	Náhodné permutácie	34
4	Využitie náhodných čísel	37
4.1	Simulácie	37
4.2	Približný výpočet integrálu	38
4.3	Metóda opakovaných výberov (bootstrap)	41

Kapitola 1

Druhy štatistických výpočtov

1.1 Úvod

Nástup počítačov silne ovplyvnil štatistiku a jej prístup k mnohým úlohám. Niektoré úlohy, ktoré boli predtým neriešiteľné, sa s použitím počítačov stali ľahkými (mnohé výpočty kritických hodnôt, p-hodnôt či odhadov), k iným sa zmenil prístup (postupný prechod od testovania pomocou kritických hodnôt na testovanie s využitím p-hodnôt) a objavili sa aj celkom nové, bez použitia počítačov nemysliteľné úlohy (bootstrap, data mining, simulácie, iteračné výpočty bez explicitných vzorcov).

Spočiatku sa k štatistickým výpočtom používali bežné programovacie jazyky (Fortran, Cobol, PL/I, Pascal...) a každý užívateľ si potrebné metódy programoval sám, príp. všeobecne použiteľné výsledky boli publikované v časopisoch a knihách (najmä kritické hodnoty testov). Neskôr sa objavili prvé špecializované štatistické programy (SPSS, BMDP, neskôr SAS) alebo knižnice podprogramov (Naglib, Linpack,...). Napokon sa objavili špeciálne simulačné jazyky (GPSS, Simscript, Simula, Covers,...). V neskoršom vývoji knižnice podprogramov takmer vymizli a boli nahradené špecializovanými štatistickými programovacími jazykmi (napr. Gauss, S+; často sa vyvinuli aj z veľkých štatistických balíkov, napr. SPSS, SAS), čím sa odstránila ich hlavná slabina, totiž vstup a výstup dát (väčšinou bol na nízkej úrovni, keďže každý si ho programoval sám).

Dnešná situácia - programové vybavenie môžeme deliť podľa niekoľkých hľadísk:

- podľa účelu:
 1. programy so širším zameraním, ktoré poskytujú (rôzne) silnú podporu štatistickým metódam (napr. tabuľkové procesory, Maple, Linpack)

2. špecializované štatistické programy (napr. SAS, S+)

- podľa veľkosti:

1. malé systémy (napr. Quickstat, Nlinreg)
2. laické systémy (napr. Statgraphics)
3. veľké systémy (napr. SAS, Statistica)

- podľa formy:

1. knižnice podprogramov (napr. Eispack, Linpack, Naglib)
2. uzavreté systémy (napr. Statgraphics)
3. štatistické alebo simulačné programovacie jazyky (napr. SPSS, SAS, GLIM, PC-GUHA, Resampling Stats, Simula)
4. otvorené systémy (napr. SAS-IML, S+, R, Gauss)

Voľba štatistického softwaru nie je vždy jednoduchá, lebo okrem charakteru úlohy závisí aj od schopností užívateľa, potrebnej miery všeobecnosti (či program bude použitý jednorazovo alebo opakovane, ktoré parametre pri opakovanom používaní sa budú meniť atď.) a v neposlednom rade aj od ceny programového systému. Do úvahy treba zobrať aj numerickú stránku výpočtov, keďže rôzne programy sú rôzne náchylné ku zlyhaniu vo výpočtovo náročných situáciách (napr. jazyk APL, tabuľkové procesory). Každý programový systém má pritom určitú optimálnu oblasť použitia.

Z hľadiska výpočtových prostriedkov môžeme rozlíšiť zhruba 5 okruhov štatistických úloh:

1. Štandardné štatistické spracovanie dát
2. Neštandardné štatistické spracovanie dát (nové metódy, ktoré ešte nie sú v komerčných programových systémoch)
3. Výpočty tabuľkových hodnôt (kritické hodnoty a p-hodnoty, obvykle numerická integrácia)
4. Generovanie náhodných čísel a simulácie (práca s empirickým rozdelením, napr. bootstrap)
5. Data mining (prieskumová analýza dát, automatické hromadné testovanie)

Týmito typmi úloh a zodpovedajúcim programovým vybavením sa budeme zaoberať v nasledujúcich kapitolách.

1.2 Štandardné štatistické výpočty

Ponuka programov pre oblasť štandardných štatistických metód je veľmi bohatá. Funkcie pre výpočet aspoň základných charakteristík výberového súboru ponúkajú mnohé programy zamerané na prácu s dátami, t.j. najmä rôzne databázové systémy a systémy zamerané na numerickú matematiku. Nie všetky sú však naozaj vhodné pre štatistické spracovanie dát (prácnosť, presnosť atď.). My sa zoznámime s charakteristikami niektorých z nich.

1.2.1 Programové systémy pre numerickú matematiku

Typickými zástupcami sú Mathematica a Maple. Poskytujú pomerne širokú podporu štatistických metód, ale iba v potenciálnej forme. Sú to v podstate programovacie jazyky - interprety, v ktorých si užívateľ príslušné metódy (odhady, testy) môže sám naprogramovať. Má pritom k dispozícii mnohé pokročilé funkcie (distribučné a kvantilové funkcie základných rozdelení, riešenie sústav rovníc, výpočet vlastných čísel matice, funkcie poradia atď.). Niektoré systémy tejto kategórie ponúkajú za príplatok aj balíky štatistických procedúr (napr. Gauss), iné podporujú výmenu programov medzi užívateľmi (napr. Maple, Mathematica). To tejto kategórie patria aj knižnice podprogramov typu Naglib. Spoločným nedostatkom všetkých programov tejto kategórie je slabá podpora editovania a práce s dátami.

1.2.2 Tabuľkové procesory

V súčasnosti zostali na trhu prakticky iba dva: MS Excel a Lotus 1-2-3. Funkčne sú takmer zhodné. Kvôli ľahkosti programovania sú v súčasnosti veľmi populárne pri výuke štatistiky a veľa neštatistikov ich používa aj na bežné spracovanie dát. Súčasťou programov sú aj moduly so základnými štatistickými procedúrami (t-testy, lineárna regresia...). Pre programovanie ponúkajú distribučné a kvantilové funkcie základných rozdelení, funkcie poradia, maticové operácie atď. Na trhu je aj viacero doplnkových (add-in) modulov s väčšou ponukou štatistických metód (napr. Analyze-It!, Xplore). Ich hlavným problémom je rýchlo rastúca prácnosť pri programovaní zložitejších metód a zlé numerické vlastnosti vstavaných štandardných funkcií (napr. zlé hodnoty kvantilov pre určité hodnoty parametrov). Silnou stránkou je naopak ľahká editácia a prenositeľnosť dát (takmer všetky väčšie systémy podporujú vstup dát z tabuľky Excel alebo Lotus).

1.2.3 Malé štatistické systémy

Existuje ich veľmi veľa. Často ide o jednoúčelové programy, venované len jednej štatistickej metóde (napr. Nlinreg). Typicky obsahujú niekoľko najpoužívanejších štatistických metód (napr. QuickStat) s minimálnymi možnosťami nastavenia či diagnostiky. Sú vhodné na rutinné spracovanie dát presne zodpovedajúcich príslušnému modelu. Ich najväčším problémom obvykle býva práca s

dátami (editovanie, výmena dát s inými programami, prenos dát medzi procedúrami). Pred začatím používania je tiež potrebné otestovať numerické vlastnosti programu.

1.2.4 Laické systémy

Typickým zástupcom je Statgraphics. Ide o systémy s pomerne širokou ponukou štandardných štatistických metód, ale s minimálnou možnosťou nastavovania. Typická je aj pomoc s interpretáciou výsledkov. Niektoré systémy ponúkajú tiež pomoc s určením správnej procedúry pre spracovanie dát (napr. Prophet). Sú teda určené pre ľudí, ktorí sa v štatistike príliš nevyznajú, nekladú príliš hlboké otázky, ale potrebujú štandardné spracovanie dát. Možnosti editovania bývajú veľmi dobré, numerické vlastnosti sú rôznej kvality. V prípade aj malej odchýlky dát od štandardného modelu sa môžu stať nepoužiteľné (napr. pri chýbajúcich dátach).

1.2.5 Veľké systémy

Typickými predstaviteľmi sú SAS a SPSS (ale patrí sem aj S+, GLIM...). Sú to vlastne programovacie jazyky zamerané na manipuláciu s dátami a ich štatistické spracovanie. Jednotlivé príkazy sú buď mená štatistických procedúr alebo prepínače slúžiace na ich detailné nastavenie. Ponúkajú veľké množstvo možností, obsahujú množstvo štatistických metód vrátane ich diagnostiky. Umožňujú hĺbkové analýzy aj v menej štandardných situáciách. Vyžadujú však poučeného užívateľa, ktorý má prehľad o štatistike. Numerické vlastnosti sú obvykle veľmi dobré. Spravidla vyžadujú dlhšie zaškolenie, ale odmenou je vynikajúci výkon. Sú vhodné i pre dávkové spracovanie dát, t.j. automatické spracovanie podľa vopred pripraveného programu.

1.3 Neštandardné štatistické výpočty

Množstvo štatistických metód z rôznych dôvodov nie je obsiahnutých v štandardných komerčných programových systémoch. Predovšetkým ide o metódy nové, ktoré ešte nevnikli do povedomia užívateľov, a teda po nich nie je dostatočný dopyt. Môže ísť aj o metódy výpočtovo a užívateľsky náročné, či príliš zriedkavo potrebné. Často ide o iteračné metódy, metódy založené na riešení sústav nelineárnych rovníc a pod. Užívateľ, ktorý ich potrebuje, stojí teda pred úlohou ich naprogramovať. Neodporúčame k tomuto účelu použiť numericky nestabilné prostredie (tabuľkové kalkulátory a pod.) Vzhľadom k dôležitosti editovania dát nie sú vhodné ani numericky vynikajúce programy typu Maple. Najvhodnejšie sú teda otvorené štatistické systémy ako je S+ , R, Gauss (s doplnkami) či SAS-IML. Tie zachovávajú všetky výhody štandardného prostredia pre štatistické analýzy (editovanie dát, štatistické funkcie, maticové funkcie...) a pritom dovoľujú programovanie vlastných procedúr.

1.4 Výpočty tabulkových hodnôt

Ide predovšetkým o výpočet kritických hodnôt štatistických testov, v menšej miere p-hodnôt a ďalších veličín. Často sa jedná o numerickú integráciu vo vysokodimenzionálnych priestoroch, teda o numericky veľmi náročné výpočty. Pri týchto výpočtoch obvykle odpadá potreba prostredia na editovanie dát, ale sú veľké nároky na presnosť výpočtov. Vhodné sú teda systémy zamerané na numerickú matematiku (Maple, Mathematica) alebo špecializované knižnice podprogramov (Linpac, Eispac, Naglib) spolu s príslušným programovacím jazykom. Samozrejme je potrebné mať prehľad o používaných numerických metódach.

1.5 Generovanie náhodných čísel a simulácie

Prakticky každý prekladač programovacieho jazyka má dnes vstavaný generátor náhodných čísel. Tie sú dnes samozrejmom súčasťou aj tabulkových procesorov i štatistických programových systémov (s výnimkou malých). Tieto generátory sú však rôznej kvality a je potrebné ich pred použitím otestovať, či majú dobré vlastnosti vzhľadom k účelu, na ktorý ich chceme použiť. To je veľmi dôležité najmä v prípade, že potrebný počet generovaných čísel je veľký (rádovo tisíce a viac). Nevhodný generátor náhodných čísel môže totiž silne skresliť výsledky simulácie (napr. ak sú čísla korelované). Pretože pri simuláciách obvykle nepotrebujeme dátové editory, ale dosť záleží na rýchlosti a presnosti generátora, najlepším prostredím sú knižnice podprogramov (NAGLIB a pod.), systémy pre numerickú matematiku (Maple a pod.), výkonné štatistické systémy (S+ a pod.) alebo špecializované simulačné jazyky (Simula, Covers). Podrobnejšie o simuláciách pojednáme v ďalších kapitolách.

1.6 Data mining

Úlohou data miningu je nájsť „zaujímavé“ skutočnosti v dátach, o ktorých nemáme žiadny apriórny názor. Niektoré jeho metódy sa už udomácnili v univerzálnych štatistických systémoch (napr. zhluková analýza), iné zatiaľ nie (napr. GUHA, projection pursuit). Data mining sa obvykle robí na veľmi rozsiahlych súboroch dát; typicky sú to databázy s desiatkami premenných a 100 000 a viac záznamami. Často preto vyžaduje spoluprácu viacerých programov: databázový program zbiera, triedi, kontroluje a prípadne transformuje dáta, štatistický program dáta zobrazuje (príp. to robí špecializovaný grafický program), robí prieskumovú analýzu a/alebo automatické testovanie a upozorňuje užívateľa na zaujímavé skutočnosti. Na niektoré metódy sú najlepšie špecializované programy (napr. PC-GUHA), na iné univerzálne systémy (najmä S+). Data mining vyžaduje veľkú rýchlosť a presnosť výpočtov a dobrú dynamickú grafiku. Veľké nároky kladie i na dôvtip výskumníka.

V posledných rokoch sa objavujú špecializované programy na data mining

(Clementine, Statistica Data Miner), ktoré dosahujú zaujímavé výsledky. Ich nevýhodou však je, že sú to "čierne skrinky" užívateľ nevie, čo vlastne program robí. Laickému užívateľovi to síce obvykle nevadí, ale pokročilejším užívateľom to bráni v možnosti vidieť slabiny použitej metódy a prípadne neopakovať jej postupy pri ďalšom skúmaní dát.

Kapitola 2

Niektoré praktické výpočtové metódy

2.1 Výpočet funkcií rozdelenia pravdepodobnosti

Pre štatistické výpočty sú podstatné najmä hodnoty distribučných funkcií a hodnoty kvantilov. Metódy ich výpočtu sú väčšinou netriviálne, keďže matematická definícia obvykle nebýva použiteľná k priamemu výpočtu. My spomenieme iba niektoré z nich; úplný prehľad je nad naše časové a priestorové možnosti.

2.1.1 Normálne rozdelenie

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \Phi(x) = \int_{-\infty}^x \varphi(t) dt, \quad \Phi_0(x) = \int_0^x \varphi(t) dt.$$

Zrejme platí:

$$\begin{aligned}\varphi(-x) &= \varphi(x) \\ \Phi(-x) &= 1 - \Phi(x) \\ \Phi(x) &= \frac{1}{2} + \Phi_0(x) \\ \Phi_0(-x) &= -\Phi_0(x).\end{aligned}$$

Keďže $\Phi(x)$ nevieme vyjadriť v elementárnych funkciách, používajú sa rôzne aproximácie pomocou radov či reťazových zlomkov.

Pre malé x možno použiť vzorec

$$\Phi_0(x) = \varphi(x) \left(x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \frac{x^7}{3 \cdot 5 \cdot 7} + \dots \right) ;$$

pre väčšie hodnoty sa dá použiť semidivergentný rad

$$1 - \Phi(x) = \varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} + \frac{1 \cdot 3}{x^5} - \frac{1 \cdot 3 \cdot 5}{x^7} + \frac{1 \cdot 3 \cdot 5 \cdot 7}{x^9} - \dots \right), \quad x > 0.$$

Pre zvyšok r_n tohto radu platí $|r_n| < \varphi(x) \frac{(2n-1)!!}{x^{2n+1}}$, čo umožňuje vybrať vhodné n pre dané x , aby aproximácia bola dobrá.

Často sa tiež používajú aproximácie

$$\Phi_0(x) \doteq \frac{1}{2} \left[1 - \left(\sum_{i=0}^6 b_i x^i \right)^{-16} \right], \quad x > 0,$$

kde pre konštanty b_i platí

$$\begin{aligned} b_0 &= 1 \\ b_1 &= 0.049\ 867\ 347 \\ b_2 &= 0.021\ 141\ 006 \\ b_3 &= 0.003\ 277\ 626 \\ b_4 &= 0.000\ 038\ 004 \\ b_5 &= 0.000\ 048\ 891 \\ b_6 &= 0.000\ 005\ 383 \end{aligned}$$

(chyba je menšia ako $1.5 \cdot 10^{-7}$), resp.

$$\Phi(x) \doteq 1 - \varphi(x) \sum_{i=1}^5 a_i \frac{1}{(1 + a_0 x)^i}, \quad x \geq 0,$$

kde pre konštanty a_i platí

$$\begin{aligned} a_0 &= +0.231\ 641\ 9 \\ a_1 &= +0.319\ 381\ 5 \\ a_2 &= -0.356\ 563\ 8 \\ a_3 &= +1.781\ 478 \\ a_4 &= -1.821\ 256 \\ a_5 &= +1.330\ 274 \end{aligned}$$

(chyba je menšia ako $7 \cdot 10^{-7}$).

Pre výpočet kvantilovej funkcie sa často používa aproximácia

$$u_\alpha \doteq -w + \frac{\sum_{i=0}^2 a_i w^i}{\sum_{i=0}^2 b_i w^i}, \quad \alpha \in (0; 0.5), \quad w = \sqrt{-2 \ln \alpha},$$

kde pre konštanty a_i a b_i platí

$$a_0 = 2.515\ 517 \quad b_0 = 1$$

$$\begin{aligned} a_1 &= 0.802\ 853 & b_1 &= 1.432\ 788 \\ a_2 &= 0.010\ 328 & b_2 &= 0.189\ 269 \\ & & b_3 &= 0.001\ 308 \end{aligned}$$

(chyba je menšia ako $4.5 \cdot 10^{-4}$). Pre iné hodnoty α sa využije symetria kvantilovej funkcie.

Presnejšiu aproximáciu dostaneme pomocou Taylorovho rozvoja kvantilovej funkcie. Najprv vyrátame prvotnú aproximáciu z zo vzťahu

$$z = w - \frac{\sum_{i=0}^2 a_i w^i}{\sum_{i=0}^3 b_i w^i}, \quad w = \sqrt{-2 \ln(1 - \alpha)}, \quad \alpha \in (0.5; 1),$$

kde pre konštanty a_i a b_i platí

$$\begin{aligned} a_0 &= 1\ 673.72 & b_0 &= 659.935 \\ a_1 &= 494.877 & b_1 &= 908.401 \\ a_2 &= 7.473\ 95 & b_2 &= 117.9407 \\ & & b_3 &= 1 \end{aligned}$$

Potom Taylorov polynóm n -tého stupňa pre u_α má tvar

$$u_\alpha^{(n)} = z + \sum_{i=1}^n \frac{c_i(z)}{i!} \left[\frac{\alpha - \Phi(z)}{\varphi(z)} \right]^i,$$

kde $c_1(z) = 1$ a $c_{i+1}(z) = i \cdot z \cdot c_i(z) + \frac{d}{dz} c_i(z)$. Je teda $c_2(z) = z$, $c_3(z) = 2z^2 + 1$, $c_4(z) = 6z^3 + 7$ atď. (Chyba je menšia ako $1.2 \cdot 10^{-16}$ pre $n = 4$.)

2.1.2 Iné rozdelenia

Odkazujeme čitateľa na špecializovanú literatúru, najmä [13] a [1]. V blízkom čase bude nové vydanie [1] k dispozícii na <http://dlmf.nist.gov/>.

2.2 Maticové výpočty

Maticový počet sa v štatistike silne využíva, preto požiadavky na maticové výpočty sú veľké. Najčastejšími úlohami sú riešenie sústavy rovníc (špeciálnymi prípadmi sú výpočet inverznej či pseudoinverznej matice) a výpočet vlastných čísel a vektorov matice. Hoci takmer všetky štatistické programy tieto metódy majú implementované, predsa niekedy narazíme na problémy. Najčastejšie je to v situácii, že matica sústavy je takmer singulárna¹. Na zistenie „miery singularity“ matice sa používajú koeficienty kolinearity, napr.

$$CN_1(A) = \|A\| \cdot \|A^{-1}\|, \quad \text{kde } \|A\| = \max_j \sum_i |a_{ij}|,$$

¹V anglickej literatúre „ill-conditioned“=chorá, čo je v našej literatúre často otrocky (a zle) prekladané ako „zle podmienená“.

alebo

$$CN_2(A) = \sqrt{\frac{\lambda_{max}(AA')}{\lambda_{min}(AA')}}.$$

Čím je tento koeficient väčší, tým je matica A bližšie k singularite. Zrejme platí $CN_i(A) \geq 1$. Ak je matica takmer singulárna, rôzne programy - v závislosti od použitého algoritmu a programovacieho jazyka - môžu dať výrazne odlišné výsledky, niekedy aj nezmyselné. Takisto rozsah sústavy môže byť prekážkou úspešného výpočtu (geodetické výpočty vyžadujú niekedy riešenie sústav s rádovo 10^5 rovnicami). Najlepším riešením v podobnej situácii býva použitie špecializovaného podprogramu (Lapack, Linpack, Eispack...), ktorých algoritmy sú z numerického hľadiska na vysokej úrovni.

Ak sme nútení metódu si sami programovať, je potrebné využiť všetky apriórne informácie pre správnu voľbu algoritmu. Napr. bolo by chybou použiť všeobecnú metódu inverzie matice tam, kde vieme, že matica je pozitívne definitná. To je typické napr. pre regresný model.

V tejto časti uvedieme princípy troch najčastejšie potrebných metód maticového počtu v matematickej štatistike.

2.2.1 Inverzia symetrickej matice

Symetrickú maticu A obvykle rozkladáme na súčin trojuholníkových matíc:

$$A = S'S,$$

kde S je horná trojuholníková matica. Prvky matice S dostaneme riešením danej sústavy rovníc. Potom invertujeme trojuholníkovú maticu S a dostaneme

$$A^{-1} = S^{-1}(S')^{-1}.$$

Modifikáciou tohto postupu je Choleského metóda, pri ktorej A rozložíme na súčin

$$A = C'DC,$$

kde C je horná trojuholníková matica s diagonálnymi prvkami rovnými 1 a D je diagonálna matica. Invertovanie matíc C a D je ľahká úloha; potom

$$A^{-1} = C^{-1}D^{-1}(C')^{-1}.$$

Inverzia trojuholníkovej matice sa často ráta rekurentne z rovnice $CC^{-1} = I$.

2.2.2 Riešenie symetrickej sústavy lineárnych rovníc

Sústavu $Ax = b$ je možné riešiť pomocou inverznej matice, t.j. $x = A^{-1}b$. Choleského metóda však používa iný prístup. Po rozložení

$$A = C'DC$$

rieši dve ľahšie sústavy

$$C'y = b \quad \text{a} \quad Cx = D^{-1}y ,$$

ktoré sú ekvivalentné pôvodnej sústave (keďže $Ax = C'DCx = C'DD^{-1}y = C'y = b$).

Pri numerickom riešení je dôležité skontrolovať presnosť riešenia. Ak označíme $r = Ax_0 - b$ rezíduum (približného) riešenia sústavy x_0 , môžeme spresniť získané korene x_0 riešením sústavy $Ad = r$. Potom zrejme $x = x_0 - d$.

2.2.3 Výpočet g-inverznej matice

Najčastejšie je potrebné vypočítať Moore-Penroseovu g-inverziu danej matice $A_{m \times n}$. Ak označíme r hodnotu matice A , potom matica AA' má práve r kladných vlastných čísel (keďže je pozitívne semidefinitná) a jej spektrálny rozklad má tvar

$$AA' = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix} \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_r' \\ U_{m-r}' \end{pmatrix}$$

V matici U_r sú teda vlastné vektory zodpovedajúce kladným vlastným číslam, ktoré sú v diagonálnej matici Λ_r . MP-inverzia je potom

$$A^+ = A'U_r\Lambda_r^{-1}U_r'.$$

Dôkaz. Keďže $U = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix}$ je ortogonálna matica, platí

$$U_rU_r' + U_{m-r}U_{m-r}' = UU' = I_m = U'U = \begin{pmatrix} U_r'U_r & U_r'U_{m-r} \\ U_{m-r}'U_r & U_{m-r}'U_{m-r} \end{pmatrix}.$$

Špeciálne, platí

$$U_r'U_r = I_r, \quad U_r'U_{m-r} = 0_{r \times m-r}.$$

Zrejme tiež

$$AA' = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix} \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_r' \\ U_{m-r}' \end{pmatrix} = U_r\Lambda_rU_r'.$$

Je teda

$$AA'U_r = U_r\Lambda_r \quad \text{a} \quad AA'U_{m-r} = 0,$$

z čoho ďalej

$$U_{m-r}'A = 0 \quad \text{a} \quad A = (U_rU_r' + U_{m-r}U_{m-r}')A = U_rU_r'A.$$

Potom platí:

matica $AA^+ = AA'U_r\Lambda_r^{-1}U_r' = U_r\Lambda_r\Lambda_r^{-1}U_r' = U_rU_r'$ je symetrická;

matica $A^+A = A'U_r\Lambda_r^{-1}U_r'A$ je zrejme tiež symetrická.

Využívajúc vzťah pre AA^+ ďalej dostávame

$$AA^+A = U_rU_r'A = A$$

a

$$A^+AA^+ = A'U_r\Lambda_r^{-1}U_r'U_rU_r' = A'U_r\Lambda_r^{-1}I_rU_r' = A^+.$$

Matica A^+ má teda všetky požadované vlastnosti. ■

Takisto sme mohli využiť spektrálny rozklad matice $A'A$ a definovať

$$A^+ = U_r\Lambda_r^{-1}U_r'A'$$

(pozor: tu je iné U_r). Výhodné je vybrať si na rozkladanie z dvojice $A'A$ a AA' tú, ktorá má menšiu dimenziu.

Podobná je metóda založená na SVD: ak $A = S\Lambda_r^{1/2}T'$, $S'S = T'T = I_r$, potom $A^+ = T\Lambda_r^{-1/2}S'$.

Všeobecnú g-inverznú maticu môžeme z MP-inverzie dostať napr. pomocou vzťahu

$$A^- = A^+ + Q - A^+AQA^+,$$

kde Q je ľubovoľná matica typu $n \times m$.

Problematikou výpočtu vlastných vektorov a vlastných čísel sa nebudeme zaoberať.

Problém určenia hodnosti matice:

Podstata problému je v tom, že po transformácii matice na hornú trojuholníkovú býva ťažké rozhodnúť, či diagonálne prvky, ktoré sú blízke k strojovej presnosti, sú naozaj nenulové alebo nenulové číslo vzniklo iba v dôsledku zaokrúhľovacích chýb. Toto sa ľahko rieši u idempotentných matíc, kde hodnosť je rovná stope. Vyrátame teda stopu a tú zaokrúhlime na najbližšie prirodzené číslo. U všeobecnej matice sa obvykle odporúča určiť hodnosť ako počet singularných čísel (odmocnín z vlastných čísel matice AA'), ktoré sú väčšie ako $\sqrt{\lambda_{max}(AA')} \cdot \text{rozmer}(A) \cdot \varepsilon$, kde ε je strojová presnosť.

Kapitola 3

Generovanie náhodných čísel

3.1 Rovnomerné rozdelenie

Skutočne náhodné čísla môže dať iba fyzikálny generátor. Pri počítačovom generovaní hovoríme o *pseudonáhodných číslach*: ide o deterministickú postupnosť čísel, ktorá však má mnohé vlastnosti náhodnej postupnosti.

Definícia 3.1.1 *Konečnú postupnosť dekadických čísel, ktorú je možné považovať za náhodný výber z $R(0, \dots, 9)$, budeme nazývať náhodné číslice (t.j. $P(X = i) = 0.1$, $i = 0, 1, \dots, 9$). Konečnú postupnosť čísel z $(0; 1)$, ktorú je možné považovať za náhodný výber z $R(0; 1)$, budeme nazývať náhodné čísla.*

Veta 3.1.2 *Nech X_1, X_2, \dots je postupnosť nezávislých, rovnako rozdelených náhodných veličín (NV) s $R(0, \dots, 9)$. Potom NV $Y = \sum_{i=1}^{\infty} X_i 10^{-i}$ má rozdelenie $R(0; 1)$. Naopak, ak Y má rozdelenie $R(0; 1)$, kde $Y = \sum_{i=1}^{\infty} X_i 10^{-i}$, potom X_1, X_2, \dots sú nezávislé NV s $R(0, \dots, 9)$.*

Dôkaz. Vezmime $y \in (0; 1)$, $y = \sum_{i=1}^{\infty} x_i 10^{-i}$. Potom

$$\{Y < y\} = \bigcup_{i=1}^{\infty} \left[\bigcap_{k=1}^{i-1} \{X_k = x_k\} \cap \{X_i < x_i\} \right],$$

z čoho vzhľadom k disjunktnosti uvažovaných javov a nezávislosti veličín X_i vyplýva

$$\begin{aligned} P[Y < y] &= \sum_{i=1}^{\infty} P \left[\bigcap_{k=1}^{i-1} \{X_k = x_k\} \cap \{X_i < x_i\} \right] = \\ &= \sum_{i=1}^{\infty} \left[\prod_{k=1}^{i-1} P(X_k = x_k) \right] P(X_i < x_i) = \sum_{i=1}^{\infty} 10^{-i+1} x_i 10^{-1} = \\ &= \sum_{i=1}^{\infty} x_i 10^{-i} = y. \end{aligned}$$

Je teda $F_Y(y) = y$, čiže Y má rovnomerné rozdelenie na $(0; 1)$.

Naopak, vezmeme dekadickú číslicu x_1 . Potom pre $Y \sim R(0; 1)$ platí

$$P[X_1 = x_1] = P[10^{-1}x_1 \leq Y < 10^{-1}(x_1 + 1)] = 10^{-1}(x_1 + 1) - 10^{-1}x_1 = 10^{-1}.$$

Veličina X_1 má teda rozdelenie $R(0, \dots, 9)$. Ak X_1, \dots, X_{n-1} sú nezávislé s rozdelením $R(0, \dots, 9)$, a x_1, \dots, x_n sú dekadické číslice, potom

$$P[X_1 = x_1, \dots, X_n = x_n] = P \left[\sum_{i=1}^n x_i 10^{-i} \leq Y < \sum_{i=1}^n x_i 10^{-i} + 10^{-n} \right] = 10^{-n}.$$

Podľa indukčného predpokladu platí $P[X_1 = x_1, \dots, X_{n-1} = x_{n-1}] = 10^{-n+1}$, z čoho $P[X_n = x_n] = 10^{-1}$. Teda aj X_n je nezávislá NV s rozdelením $R(0, \dots, 9)$. ■

Generovanie náhodných číslic a náhodných čísel je teda ekvivalentné. Generovanie pomocou číslic je však veľmi pomalé a v praxi sa takmer nepoužíva. Ak máme k dispozícii náhodný výber z $R(0; 1)$, môžeme pomocou neho generovať náhodné čísla aj z iných rozdelení. O konkrétnych metódach pojednáme neskôr.

Základným problémom je teda generovanie náhodných čísel z $R(0; 1)$. Dnes sa používajú tri typy generátorov:

1. Lineárne rekurentné (kongruenčné) generátory.

Pri nich sa generuje najprv postupnosť celých čísel z intervalu $\langle 0; m \rangle$ pomocou rekurentného vzťahu

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \dots + a_k X_{n-k} + b \pmod{m}, \quad n = k, k+1, \dots$$

kde a_1, a_2, \dots, a_k, b a m sú nezáporné celé čísla a X_0, X_1, \dots, X_{k-1} sú počiatočné hodnoty. Potom definujeme

$$U_i = \frac{X_i}{m};$$

je teda $U_i \in \langle 0; 1 \rangle$. Toto je postupnosť pseudonáhodných čísel. Štatistické vlastnosti závisia na konštantách a_1, \dots, a_k, b, m a zvolených počiatočných hodnotách.

2. Bitové rekurentné generátory.

Pri nich sa generuje postupnosť bitov rekurentným vzťahom

$$B_n = a_1 B_{n-1} + a_2 B_{n-2} + \dots + a_k B_{n-k} \pmod{2}, \quad n = k, k+1, \dots$$

kde $a_n \in \{0, 1\} \quad \forall n = 1, \dots, k-1$, $a_k = 1$, $B_n \in \{0, 1\} \quad \forall n$. Potom definujeme m -bitové pseudonáhodné číslo

$$X_i = \sum_{j=1}^m 2^{m-j+1} B_{ri+j},$$

kde $r \geq m$. Prirodzene, $U_i = X_i 2^{-m-1}$.

Do tejto kategórie patria aj (v praxi veľmi používané) posuvné registrové generátory.

3. Nelineárne kongruenčné generátory

Pri nich sa využíva rekurentný vzťah

$$X_n = f(X_{n-1}) \pmod{m}, \quad n = 1, 2, \dots$$

kde f je nejaká nelineárna celočíselná funkcia.

Príkladom takejto funkcie je $f(z) = az^{-1} + b$, $z \in \{1, 2, \dots, m-1\}$, kde $a, b \in \mathbb{N}$, $m \geq 5$ je prvočíslo a z^{-1} je prirodzené číslo menšie ako m , pre ktoré platí $zz^{-1} \equiv 1 \pmod{m}$. Číslo z^{-1} sa nazýva multiplikatívna inverzia z modulo m ; z teórie grúp vyplýva jeho existencia i jednoznačnosť.

3.1.1 Lineárne rekurentné generátory

Najčastejšie sa používajú jednoduché generátory typu

$$X_n = aX_{n-1} + b \pmod{m}$$

(zmiešané kongruenčné generátory). Keďže X_i môže nadobúdať iba hodnoty $0, 1, \dots, m-1$, musí sa postupnosť po konečnom počte krokov, maximálne m , začať opakovať. Počet krokov, po ktorých sa postupnosť začne opakovať, sa nazýva *perióda generátora*. Ak je perióda rovná m , hovoríme o *plnej perióde*.

Tvrdenie 3.1.3 *Nech b a m sú nesúdeliteľné. Nech $a \equiv 1 \pmod{p}$ pre každý prvočiniteľ p prvočíselného rozkladu m . Nech $a \equiv 1 \pmod{4}$, ak m je násobok 4. Potom má generátor $X_n = aX_{n-1} + b \pmod{m}$ plnú periódu m pre každé X_0 .*

Niekedy sa používajú iba multiplikatívne generátory:

Tvrdenie 3.1.4 *Nech a, m a X_0, m sú nesúdeliteľné čísla. Nech prvočíselný rozklad m je $m = 2^\alpha \prod_{i=1}^r p_i^{\beta_i}$, kde p_1, \dots, p_r sú rôzne nepárne prvočísla. Položme*

$$\lambda(p^\beta) = (p-1)p^{\beta-1}, \quad p \text{ nepárne}$$

$$\lambda(2^\alpha) = \begin{cases} 1 & \alpha = 0, 1 \\ 2 & \alpha = 2 \\ 2^{\alpha-2} & \alpha > 2 \end{cases}$$

Nech $a^n \not\equiv 1 \pmod{p_i^{\beta_i}}$ pre $0 < n < \lambda(p_i^{\beta_i})$, $i = 1, \dots, r$

$$a \equiv \begin{cases} 1 \pmod{2} & \alpha = 1 \\ 3 \pmod{4} & \alpha = 2 \\ 3 \text{ alebo } 5 \pmod{8} & \alpha > 2 \end{cases}$$

Potom je perióda generátora $X_n = aX_{n-1} \pmod{m}$ maximálna a je rovná $h(m) = NSN(\lambda(2^\alpha), \lambda(p_1^{\beta_1}), \dots, \lambda(p_r^{\beta_r}))$.

Multiplikatívne generátory teda nemôžu mať plnú periódu.

Aplikáciu predchádzajúcich všeobecných viet na dôležité konkrétne prípady dostaneme:

Dôsledok 3.1.5 Nech $a \equiv 1 \pmod{4}$ a $b \equiv 1 \pmod{2}$. Potom má generátor $X_n = aX_{n-1} + b \pmod{2^\alpha}$, $\alpha \geq 1$, plnú periódu pre každé X_0 .

Nech $a \equiv 3$ alebo $5 \pmod{8}$. Potom má generátor $X_n = aX_{n-1} \pmod{2^\alpha}$, $\alpha \geq 3$, maximálnu periódu $2^{\alpha-2}$ pre každé nepárne X_0 .

Dôsledok 3.1.6 Nech $a \equiv 1 \pmod{20}$ a $b \equiv 1, 3, 7$ alebo $9 \pmod{10}$. Potom má generátor $X_n = aX_{n-1} + b \pmod{10^\beta}$, $\beta \geq 2$, plnú periódu pre každé X_0 .

Nech $a \equiv 3, 13, 27, 37, 53, 67, 77, 83, 117, 123, 133, 147, 163, 173, 187$ alebo $197 \pmod{200}$ a nech $X_0 \equiv 1, 3, 7$ alebo $9 \pmod{10}$. Potom má generátor $X_n = aX_{n-1} \pmod{10^\beta}$, $\beta > 3$, maximálnu periódu $5 \cdot 10^{\beta-2}$.

Vidíme, že pre kvalitný generátor je dobré mať hodnoty m veľmi veľké.

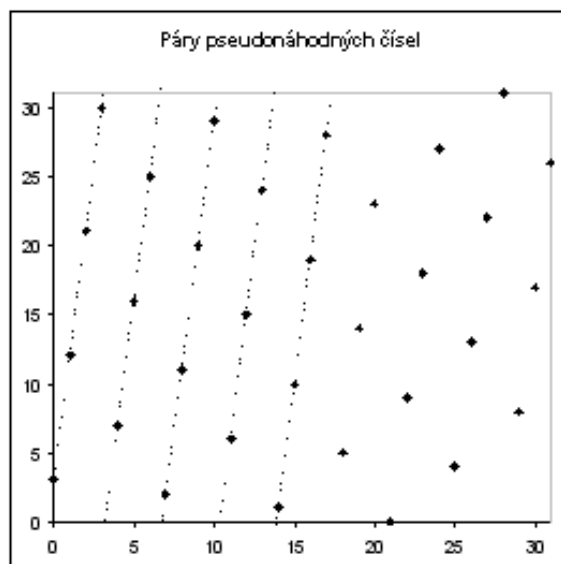
3.1.2 Seriálna korelácia

Keďže hodnoty pseudonáhodných čísel vznikajú rekurentne, sú nutne korelované. Hovoríme o *seriálnej korelácii*. Ak máme generátor $X_n = aX_{n-1} + b \pmod{m}$ s plnou periódou, potom pre korelačný koeficient dvoch susedných čísel platí

$$\rho_{X_{n-1}, X_n} \doteq \frac{1}{a}.$$

Je teda žiaduce mať aj hodnotu a dost veľkú.

Keby sme mali náhodné veličiny X, Y nezávislé s rozdelením $R(0, \dots, m-1)$, usporiadaná dvojica (X, Y) by mala rovnomerné rozdelenie na m^2 možných hodnotách (i, j) , $i = 1, \dots, m$, $j = 1, \dots, m$. Pri kongruenčnom generátore však páry (X_{n-1}, X_n) nadobúdajú iba m rôznych hodnôt (ak má plnú periódu), čiže $m(m-1)$ párov nemôže byť vygenerovaných. Graf takýchto dvojíc vytvára rovnoobežné „priamky“ v štvorci $\langle 0; m-1 \rangle \times \langle 0; m-1 \rangle$.



Niekedy potrebujeme aj seriálne korelácie vyšších rádoov. Vtedy definujeme

$$\rho_k = \text{corr}(X_{n-k}, X_n), \quad k = 1, 2, \dots$$

Závislosť po sebe idúcich pseudonáhodných čísel však býva najväčším problémom. Preto niekedy používame *kombinované generátory*, ktoré v sebe kombinujú niekoľko rôznych generátorov s cieľom oslabiť závislosť po sebe idúcich čísel. Používajú sa dve základné metódy kombinovania generátorov: miešanie a sumácia. Napr.:

- Nech $\{X_n\} \sim R(0; 1)$ a $\{Y_n\} \sim R(0, \dots, k-1)$. Urobíme k -miestnu tabuľku a vložíme do nej počiatočné hodnoty X_0, X_1, \dots, X_{k-1} . Postupnosť $\{Z_n\}$ vytvárame takto:
 - 1) generujeme Y_0 , položíme $Z_0 = X_{Y_0}$, potom generujeme X_k a dáme ho do bunky Y_0 .
 -
 - m) generujeme Y_m , položíme $Z_m = X_{Y_m}$, potom generujeme X_{k+m} a dáme ho do bunky Y_m .
- Nech $\{X_n^{(i)}\} \sim R(0; 1)$, $i = 0, 1, \dots, k-1$ sú rôzne generátory a $\{Y_n\} \sim R(0, \dots, k-1)$. V každom kroku generujeme najprv Y_i , a potom za Z_i vezmeme číslo vygenerované generátorom č. Y_i .
- Nech $\{X_n^{(1)}\}, \{X_n^{(2)}\} \sim R(0; 1)$ sú dva rôzne generátory. Potom definujeme $Z_n = X_n^{(1)} + X_n^{(2)} \pmod{1}$.

Alternatíva: ak $\{Y_n^{(1)}\} \sim R(0, \dots, m_1 - 1)$ a $\{Y_n^{(2)}\} \sim R(0, \dots, m_2 - 1)$, potom definujeme $Z_n = \frac{Y_n^{(1)}}{m_1} + \frac{Y_n^{(2)}}{m_2} \pmod{1}$ resp. $Z_n = \frac{Y_n^{(1)}+1}{m_1} + \frac{Y_n^{(2)}}{m_2} \pmod{1}$ (v druhom prípade dostávame čísla z $(0; 1)$ a nie z $\langle 0; 1 \rangle$).

Príklady niektorých konkrétnych kongruenčných generátorov:

- $X_n = 65\,539X_{n-1} \pmod{2^{31}}$, X_0 nepárne
- $X_n = 8\,192X_{n-1} \pmod{67\,099\,547}$
- $X_n = 24\,298X_{n-1} + 9\,991 \pmod{199\,017}$
- $X_n = 5^{2q+1}X_{n-1} \pmod{2^\alpha}$
- $X_n = 7^{4q+1}X_{n-1} \pmod{10^\alpha}$
- $X_{n+1} = X_{n-1} + X_{n-2} \pmod{3\,137}$

Cvičenie: pokúste sa nájsť periódu týchto generátorov!

3.1.3 Bitové rekurentné generátory

Myšlienka generovať radšej binárne ako desiatkové čísla vznikla z pozorovania, že rekurentné vzťahy medzi binárnymi číslami sa na počítačoch dajú hardwarovo implementovať pomocou posuvného registra so spätnou väzbou. To výrazne urýchlí proces generovania.

Definícia 3.1.7 *Nech $f(x)$ je polynóm nad telesom \mathcal{F} taký, že $f(0) \neq 0$. Ak jeho vedúci koeficient je 1, nazývame ho monický polynóm. Potom najmenšie prirodzené číslo n také, že f delí $x^n - 1$, sa nazýva rád polynómu f a označuje $\text{ord}(f)$.*

Definícia 3.1.8 *Nech m je prvočíslo. Polynóm f stupňa r nad Galoisovým telesom rádu m sa nazýva primitívny polynóm, ak je monický a $\text{ord}(f) = m^r - 1$.*

Tvrdenie 3.1.9 *Nutná a postačujúca podmienka k tomu, aby lineárny kongruenčný generátor*

$$X_n = a_1X_{n-1} + a_2X_{n-2} + \dots + a_kX_{n-k} \pmod{m}, \quad n = k, k+1, \dots$$

mal plnú periódu $m^k - 1$, je, že pre jeho charakteristický polynóm

$$f(x) = x^k - a_1x^{k-1} - a_2x^{k-2} - \dots - a_k$$

musí platiť $a_k \neq 0 \pmod{m}$ a $f \in \mathfrak{F}_m[x]$ musí byť primitívny polynóm nad Galoisovým telesom \mathfrak{F}_m .

Ak toto uplatníme pre špeciálny prípad Galoisovho telesa \mathfrak{F}_2 , dostávame rekurentný generátor pseudonáhodných bitov B_n tvaru

$$B_n = a_1 B_{n-1} + a_2 B_{n-2} + \dots + a_k B_{n-k} \pmod{2}, \quad n = k, k+1, \dots,$$

kde koeficienty a_i sú takisto bitové veličiny ($a_k = 1$). Ak je jeho charakteristický polynóm

$$f(x) = 1 - a_1 x - \dots - a_k x^k$$

je primitívny nad \mathfrak{F}_2 , má generátor plnú periódu $2^k - 1$. Nájdenie primitívneho polynómu v binárnom prípade je pomerne ľahké; boli publikované rôzne ich tabuľky (napr. <http://www.commsys.isy.liu.se/en/staff/mikael/polynomials/>). Ak má generátor plnú periódu, každá konečná binárna postupnosť dĺžky $n < k$ sa v ňom opakuje 2^{k-n} -krát (postupnosť n núl $2^{k-n} - 1$ -krát). Pseudonáhodné číslo z intervalu $(0; 1)$ sa potom definuje vzťahom

$$X_i = \sum_{j=1}^m 2^{-j} B_{ir-j},$$

kde $m < k$, $m \leq r$ a r je číslo nesúdeliteľné s $2^k - 1$. Pritom index $ir - j$ sa berie modulo $2^k - 1$. Takéto generátory sa dajú hardwarovo realizovať pomocou posuvného registra so spätnou väzbou; preto sa im hovorí tiež posuvné registrové generátory (FSR-generátory). Z praktických dôvodov sa používajú najmä primitívne trinómy tvaru

$$f(x) = 1 + x^s + x^k,$$

$s < k$, ktoré zodpovedajú generátorom tvaru

$$B_n = B_{n-s} + B_{n-k} \pmod{2}$$

(indexy sú brané modulo $2^k - 1$). Ak $\mathbf{x}_i = B_{ir-1} B_{ir-2} \dots B_{ir-m}$ je binárny tvar čísla X_i , potom pre generované čísla platí

$$\mathbf{x}_i = \mathbf{x}_{i-s} \oplus \mathbf{x}_{i-k},$$

kde \oplus označuje (bitovú) vylučujúcu disjunkciu (xor). Pre ich výpočet teda nie je potrebné násobenie. Navyše môžeme pracovať priamo so slovami v registroch.

Pri práci priamo v registroch sa opäť vynorila otázka či sa všetky bity generovaných čísel správajú rovnako. To nás vedie k nasledujúcej definícii:

Definícia 3.1.10 *Postupnosť pseudonáhodných m -bitových celých čísel \mathbf{x}_i s periódou p sa nazýva t -rovnomerná s v -bitovou presnosťou, ak platí nasledujúce tvrdenie: Nech $\text{trunc}_v(\mathbf{x})$ označuje binárne číslo tvorené vedúcimi v bitmi čísla \mathbf{x} a uvažujme p tv -bitových vektorov*

$$(\text{trunc}_v(\mathbf{x}_i), \text{trunc}_v(\mathbf{x}_{i+1}), \dots, \text{trunc}_v(\mathbf{x}_{i+t-1})), \quad 0 \leq i < p.$$

Potom každá z 2^{tv} možných kombinácií bitov sa v rámci jednej periódy vyskytuje rovnaký počet krát, s výnimkou kombinácie tvorenej samými nulami, ktorá má početnosť o 1 menšiu. Pre každé $v = 1, 2, \dots, m$ bude $t(v)$ označovať maximálne číslo také, že uvažovaná postupnosť je $t(v)$ -rovnomerná s v -bitovou presnosťou.

Hlavnou nevýhodou týchto generátorov je potreba pracovať s rozšírenou presnosťou, ak chceme zvýšiť dĺžku periódy generátora. Tento nedostatok odstraňujú zovšeobecnené posuvné registrové generátory (GFSR-generátory), kde sa pseudonáhodné číslo z intervalu $(0; 1)$ definuje vzťahom

$$X'_i = \sum_{j=0}^{m-1} 2^{-j-1} B_{i-jp} .$$

Binárny tvar tohto čísla je zrejme $\mathbf{x}'_i = B_i B_{i-p} \dots B_{i-p(m-1)}$, kde p sa nazýva *oneskorenie*. Z pôvodnej postupnosti bitov teda vyberáme tie, ktoré sú od seba vzdialené práve p krokov, takže každá postupnosť $\mathbf{x}'_i, \mathbf{x}'_{i+1}, \dots, \mathbf{x}'_{i+p-1}$ nemá spoločné bity $\forall p \leq 2^k/m$. Ak použijeme primitívny trinóm, dostaneme vzťahy

$$B_{i-jp} = B_{i-jp-s} + B_{i-jp-k} \pmod{2}$$

a

$$\mathbf{x}'_i = \mathbf{x}'_{i-s} \oplus \mathbf{x}'_{i-k} .$$

Pre každú hodnotu oneskorenia p však dostaneme iný cyklus $2^k - 1$ čísel a takýto generátor sa dá ľahko implementovať aj pre k výrazne väčšie ako je dĺžka slova m počítača. Napr. na 32-bitových počítačoch sa používa generátor s primitívnym trinómom $f(x) = 1 + x^{32} + x^{521}$. Nevýhodou je pomerne zložitá inicializácia týchto generátorov, keďže počiatočná postupnosť bitov musí mať dĺžku $k + p(m-1)$ a navyše voľba počiatočnej postupnosti $B_0, \dots, B_{k+p(m-1)-1}$ dosť ovplyvňuje rôzne aspekty náhodnosti generovaných čísel. Ďalšie technické detaily je možné nájsť napr. v [5].

Testovaním sa neskôr zistilo, že primitívne trinómy nemajú príliš dobré vlastnosti z hľadiska náhodnosti, je teda vhodnejšie používať primitívne polynómy s väčším počtom členov. Túto nevýhodu GFSR-generátorov, spolu s problematickou inicializáciou, odstraňujú pokrivené GFSR-generátory (twisted – TGFSR-generátory). Sú založené na vzťahu

$$\mathbf{x}_i = \mathbf{x}_{i-s} \oplus \mathbf{x}_{i-k} A ,$$

kde A je nejaká bitová matica typu $m \times m$ a čísla \mathbf{x}_j berieme ako riadkové vektory. Pri vhodnej voľbe k, m a A dosahuje takýto generátor maximálnu periódu $2^{mk} - 1$.

Tvrdenie 3.1.11 *Nech $s < k$ sú prirodzené čísla, A je $m \times m$ matica a \mathbf{x} riadkový m -vektor nad \mathbb{F}_2 a $\varphi_A(t)$ je charakteristický polynóm matice A . Perióda generátora*

$$\mathbf{x}_n = \mathbf{x}_{n-s} + \mathbf{x}_{n-k} A \pmod{2} , \quad n = k, k+1, \dots$$

s nenulovou počiatočnou postupnosťou $(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{k-1})$ je $2^{mk}-1$ práve vtedy, keď $\varphi_A(t^k + t^s)$ je primitívny polynóm stupňa mk nad \mathbb{F}_2 . V tom prípade jednotlivé bity postupnosti $\mathbf{x}_0, \mathbf{x}_1, \dots$ vytvárajú binárnu rekurentnú postupnosť s maximálnou periódou.

Keďže charakteristický polynóm TGFSR-generátora má obvykle veľa nenulových členov, má veľmi dobré vlastnosti z hľadiska náhodnosti i rovnomernosti generovaných čísel. Podrobnosti možno nájsť v článkoch [10], [11], a [12].

3.1.4 Nelineárne generátory

Pseudonáhodné číslo z Galoisovho telesa $\mathfrak{F}_m = \{0, 1, \dots, m-1\}$ je generované vzťahom

$$X_n = f(X_{n-1}) \pmod{m}, \quad n = 1, 2, \dots$$

Pseudonáhodné číslo z intervalu $\langle 0; 1 \rangle$ je potom

$$U_i = \frac{X_i}{m}.$$

Dva najpoužívanejšie druhy sú kvadratické generátory a invertujúce generátory.

Tvrdenie 3.1.12 *Nech p je prvočíslo, $k \geq 2$ je prirodzené číslo a čísla $a, b, c \in \{0, 1, \dots, p^k - 1\}$, $a \neq 0$.*

Ak (i) $a = 0 \pmod{p}$, $b = 1 \pmod{p}$, $c \neq 0 \pmod{p}$

(ii) $a = b - 1 \pmod{4}$ ak $p = 2$ a

(iii) $a = 0 \pmod{9}$, $ac = 6 \pmod{9}$ ak $p = 3$,

potom generátor

$$X_n = aX_{n-1}^2 + bX_{n-1} + c \pmod{p^k}, \quad n = 1, 2, \dots$$

má maximálnu periódu p^k pre každé $X_0 \in \{0, 1, \dots, p^k - 1\}$.

Tvrdenie 3.1.13 *Nech $m \geq 5$ je prvočíslo a nech pre každé $x \in \mathfrak{F}_m$ je x^{-1} jeho multiplikatívna inverzia v \mathfrak{F}_m (definujeme $0^{-1} = 0$). Ďalej nech $a, b \in \mathfrak{F}_m$ sú také, že $ab \neq 0$ a $f(x) = x^2 - bx - a$ je primitívny polynóm nad \mathfrak{F}_m . Potom generátor*

$$X_n = aX_{n-1}^{-1} + b \pmod{m}, \quad n = 1, 2, \dots$$

má plnú periódu m pre každé $X_0 \in \mathfrak{F}_m$.

Z praktického hľadiska majú nelineárne generátory určité prednosti i nedostatky vzhľadom k lineárnym, preto sa nedá dať jednoznačné odporúčanie, ktoré používať. Ďalšie informácie viď [5].

3.2 Ostatné rozdelenia

Existujú niekoľko všeobecných metód pre generovanie náhodných čísel z ľubovoľného rozdelenia a veľa špeciálnych určených na generovanie jedného konkrétneho rozdelenia (resp. rodiny rozdelení). My najprv spomenieme tri všeobecné metódy:

3.2.1 Metóda inverznej transformácie

Veta 3.2.1 *Nech $F(x)$ je distribučná funkcia, $F^{-1}(y) = \sup_{x \in (0;1)} \{x; F(x) \leq y\}$*

a $Y \sim R(0;1)$. Potom náhodná veličina $X = F^{-1}(Y)$ má rozdelenie s distribučnou funkciou $F(x)$.

Dôkaz. Najprv dokážeme, že $F(x) \leq y \Leftrightarrow x \leq F^{-1}(y)$.

“ \Rightarrow ” $F(x) \leq y \Rightarrow x \in \{z; F(z) \leq y\} \Rightarrow x \leq F^{-1}(y)$

“ \Leftarrow ” $x \leq F^{-1}(y) \Rightarrow \forall \varepsilon > 0 \exists x_\varepsilon : F(x_\varepsilon) \leq y \wedge x_\varepsilon + \varepsilon > F^{-1}(y) \Rightarrow x - \varepsilon \leq F^{-1}(y) - \varepsilon < x_\varepsilon \Rightarrow F(x - \varepsilon) < F(x_\varepsilon) \leq y \stackrel{\varepsilon \rightarrow 0}{\Rightarrow} F(x) \leq y$

Čiže $P(X \geq x) = P(F^{-1}(Y) \geq x) = P(Y \geq F(x))$, čo vzhľadom k skutočnosti, že $Y \sim R(0;1)$, je rovné $1 - F(x)$. ■

Príklad: Distribučná funkcia Weibullovoho rozdelenia $\mathcal{W}(\lambda, c)$ je

$$F(x) = 1 - e^{-(\lambda x)^c}, x > 0, \lambda > 0.$$

Lahko sa ukáže, že inverzná funkcia je

$$F^{-1}(y) = \frac{1}{\lambda} [-\ln(1-y)]^{\frac{1}{c}}.$$

Ak teda máme postupnosť $Y_0, Y_1, \dots \sim R(0;1)$, potom

$$X_n = \frac{1}{\lambda} [-\ln(1-Y_n)]^{\frac{1}{c}} \sim \mathcal{W}(\lambda, c).$$

Pri mnohých dôležitých rozdeleniach však nevieme kvantilovú funkciu explicitne vyjadriť resp. efektívne vyrátať. To je hlavný nedostatok tejto metódy.

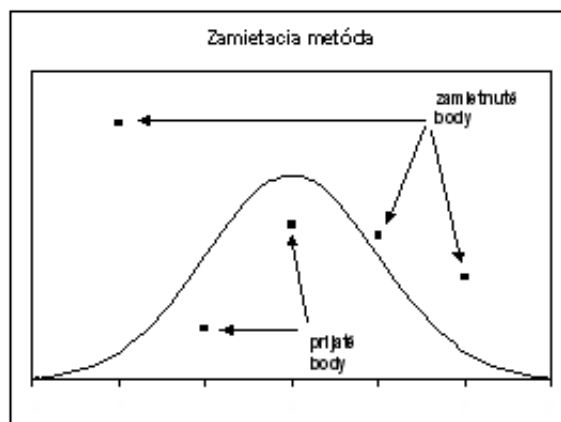
3.2.2 Zamietacia metóda

Princíp tejto metódy spočíva v tom, že generujeme body s rovnomerným rozdelením v nejakom obdĺžniku obsahujúcom graf žiadanej hustoty pravdepodobnosti a tie, ktoré nepadnú pod graf hustoty, zamietneme. X-ová súradnica zvyšných bodov má želané rozdelenie.

Veta 3.2.2 *Nech pre hustotu pravdepodobnosti $f(x)$ platí, že $f(x) = 0$ pre $x \notin \langle a; b \rangle$ a $f(x) \leq c$ pre $x \in \langle a; b \rangle$. Nech $Y \sim R(a; b)$ a $Z \sim R(0; c)$ sú nezávislé NV. Potom NV $X = Y$ za podmienky, že $f(Y) \geq Z$ má rozdelenie s hustotou $f(x)$.*

Dôkaz. Je

$$P(X < x) = P(Y < x | f(Y) \geq Z) = \frac{P(Y < x, f(Y) \geq Z)}{P(f(Y) \geq Z)}$$



Pre pravdepodobnosť v menovateli platí

$$\begin{aligned}
 P(f(Y) \geq Z) &= \iint_{\substack{f(y) \geq z \\ y \in (a;b), z \in (0;c)}} \frac{1}{c} \cdot \frac{1}{b-a} dz dy = \frac{1}{c(b-a)} \int_a^b \int_0^{f(y)} dz dy = \\
 &= \frac{1}{c(b-a)} \int_a^b f(y) dy = \frac{1}{c(b-a)}.
 \end{aligned}$$

Podobne pre čitateľa platí

$$\begin{aligned}
 P(Y < x, f(Y) \geq Z) &= \frac{1}{c(b-a)} \int_a^x \int_0^{f(y)} dz dy = \frac{1}{c(b-a)} \int_a^x f(y) dy = \\
 &= \frac{F(x)}{c(b-a)},
 \end{aligned}$$

kde $F(x)$ je distribučná funkcia prislúchajúca k hustote $f(x)$.

Z toho zrejme

$$P(X < x) = \frac{F(x)}{c(b-a)} \cdot c(b-a) = F(x).$$

■

Algoritmus teda vyzerá takto:

1. Generujeme nezávislé veličiny $U_1, U_2 \sim R(0;1)$.
2. Definujeme $Y = a + U_1(b-a)$, $Z = cU_2$.

3. Ak $f(Y) \geq Z$, potom položíme $X = Y$; ináč dvojicu U_1, U_2 zamietneme.

Hlavnými problémami tejto metódy sú predpoklad o konečnosti nosiča a obvykle nízka efektívnosť (navyše náhodne kolísajúca). Problémom môže byť aj predpoklad ohraničenosti hustoty.

Efektívnosť generátora vo všeobecnosti definujeme ako prevrátenú hodnotu strednej doby potrebnej na vygenerovanie jedného čísla. V prípade zamietacej metódy definujeme ešte iný druh efektívnosti (s predchádzajúcim zrejme spojený) vzťahom

$$E = \frac{\text{počet nezamietnutých dvojíc}}{\text{počet všetkých dvojíc}},$$

čo je náhodné číslo so strednou hodnotou

$$E_0 = \frac{\text{plocha pod hustotou}}{\text{plocha obdĺžnika}} = \frac{1}{c(b-a)} = P(\text{dvojicu nezamietneme}).$$

Napriek tomu však zamietacia metóda môže byť niekedy rýchlejšia (a teda efektívnejšia vo všeobecnom zmysle) ako metóda inverznej transformácie, ak je príslušná transformácia výpočtovo dostatočne zložitá.

Efektivita zamietacej metódy sa dá zlepšiť, ak vieme ľahko generovať náhodné čísla s hustotou, ktorá dominuje $f(x)$ na celom intervale, kde je nenulová. Nasledujúca veta dáva teoretický základ zovšeobecnenej zamietacej metódy:

Veta 3.2.3 *Nech $NV Y$ má rozdelenie s hustotou $g(x)$ a nech pre hustotu $f(x)$ existuje taká konštanta c , že $f(x) \leq c \cdot g(x) \forall x$. Nech podmienené rozdelenie $NV Z$ pri danom Y je $R(0; c \cdot g(Y))$. Potom $NV X = Y$ za podmienky, že $Z \leq f(Y)$ má rozdelenie s hustotou $f(x)$.*

Dôkaz. Platí

$$P(X < x) = P(Y < x | f(Y) \geq Z) = \frac{P(Y < x, f(Y) \geq Z)}{P(f(Y) \geq Z)}.$$

Zrejme pre podmienenú hustotu Z za podmienky Y platí $h_{Z|Y}(z|y) = \frac{1}{c \cdot g(y)}$, $z \in \langle 0; c \cdot g(y) \rangle$. Potom $h_{Z,Y}(z, y) = h_{Z|Y}(z|y) \cdot g(y) = \frac{1}{c}$ na $G = \{(y, z); 0 \leq z \leq c \cdot g(y)\}$. Vzhľadom k skutočnosti, že $z \leq f(y) \leq c \cdot g(y)$, je teda

$$P(f(Y) \geq Z) = \iint_{\substack{f(y) \geq z \\ (y, z) \in G}} \frac{1}{c} dz dy = \frac{1}{c} \int_{-\infty}^{+\infty} \int_0^{f(y)} dz dy = \frac{1}{c} \int_{-\infty}^{+\infty} f(y) dy = \frac{1}{c};$$

podobne v čitateli platí

$$\begin{aligned} P(Y < x, f(Y) \geq Z) &= \iint_{\substack{f(y) \geq z, y < x \\ (y,z) \in G}} \frac{1}{c} dz dy = \frac{1}{c} \int_{-\infty}^x \int_0^{f(y)} dz dy = \\ &= \frac{1}{c} \int_{-\infty}^x f(y) dy = \frac{F(x)}{c}, \end{aligned}$$

kde $F(x)$ je distribučná funkcia prislúchajúca k hustote $f(x)$. Z toho zrejme

$$P(X < x) = \frac{F(x)}{c} \cdot c = F(x).$$

■

Táto modifikácia nahradzuje predpoklad konečnosti nosiča predpokladom existencie majorantného rozdelenia (až na násobok), ktoré vieme efektívne generovať¹. Algoritmus je teda nasledujúci:

1. Generujeme náhodné číslo Y z rozdelenia s hustotou $g(x)$.
2. Generujeme náhodné číslo Z z $R(0; c \cdot g(Y))$.
3. Ak $f(Y) \geq Z$, potom položíme $X = Y$; ináč dvojicu Y, Z zamietneme.

Základná zamietacia metóda sa niekedy používa aj pre rozdelenia s nekonečným nosičom tak, že sa obmedzíme na konečný interval, ktorého pravdepodobnosť je takmer 1.

3.2.3 Podielová metóda

Myšlienkovy vychádza zo zamietacej metódy; v mnohých situáciách je však výpočtovo menej náročná. Často sa používa pre generovanie diskretných rozdelení.

Veta 3.2.4 *Nech $G = \left\{ (y; z) : 0 \leq y \leq \sqrt{f\left(a + b\frac{z}{y}\right)} \right\}$, kde $f(x)$ je hustota pravdepodobnosti, $a \in \mathbb{R}$, $b > 0$. Nech (Y, Z) má rovnomerné rozdelenie na ohraničenej množine $H \supset G$. Položíme $W = a + b\frac{Z}{Y}$; potom NV $X = W$ za podmienky, že $Y^2 \leq f(W)$, má rozdelenie s hustotou $f(x)$.*

Dôkaz. Združené podmienené rozdelenie (Y, Z) na G je rovnomerné s hustotou

$$g(y, z) = \frac{1}{|G|}, \quad (y, z) \in G.$$

¹V základnej metóde je ním vlastne $R(0; 1)$.

Všetky ďalšie hustoty budú teda podmienené, za podmienky $(Y, Z) \in G$. Označme $V = Y^2$; potom inverzné zobrazenie $k : (Y, Z) \rightarrow (V, W)$ je zrejme

$$\tau : \begin{pmatrix} V \\ W \end{pmatrix} \rightarrow \begin{pmatrix} \sqrt{V} \\ \frac{W-a}{b}\sqrt{V} \end{pmatrix}.$$

Jeho Jakobián je

$$D_\tau = \begin{vmatrix} \frac{1}{2\sqrt{V}} & 0 \\ \frac{W-a}{2b\sqrt{V}} & \frac{\sqrt{V}}{b} \end{vmatrix} = \frac{1}{2b}.$$

Združená hustota (V, W) potom je

$$h(v, w) = \frac{1}{2b|G|}, \quad 0 \leq v \leq f(w), \quad w \in \mathbb{R}$$

a hustota rozdelenia W je

$$\int_0^{f(w)} \frac{1}{2b|G|} dv = \frac{f(w)}{2b|G|} = f(w), \quad w \in \mathbb{R}$$

keďže

$$1 = \int_{-\infty}^{+\infty} \int_0^{f(w)} h(v, w) dv dw = \int_{-\infty}^{+\infty} \frac{f(w)}{2b|G|} dw = \frac{1}{2b|G|}$$

■

Maximálne hodnoty súradníc dosahuje uvažovaný útvar zrejme na svojej hranici $y = \sqrt{f\left(a + b\frac{z}{y}\right)}$ (pre $y \geq 0$). Položme $t = a + b\frac{z}{y}$. Zrejme musí platiť $y \in \langle 0; \sup \sqrt{f(t)} \rangle = \langle 0; y^* \rangle$. Keďže $z = \frac{t-a}{b}y = \frac{t-a}{b}\sqrt{f(t)}$, musí byť $z \in \langle \inf \frac{t-a}{b}\sqrt{f(t)}; \sup \frac{t-a}{b}\sqrt{f(t)} \rangle = \langle z_*; z^* \rangle$. Väčšinou sa preto za množinu H berie obdĺžnik $\langle 0; y^* \rangle \times \langle z_*; z^* \rangle$. Aby tento obdĺžnik bol ohraničený, musí byť funkcia $t^2 f(t)$ integrovateľná. Algoritmus potom vyzerá nasledovne:

1. Generujeme nezávislé náhodné čísla $Y \sim R(0; y^*)$ a $Z \sim R(z_*; z^*)$.
2. Definujeme $W = a + b\frac{Z}{Y}$.
3. Ak $Y^2 \leq f(W)$, položíme $X = W$, ináč dvojicu (Y, Z) zamietneme.

Stredná hodnota efektívnosti tohto generátora je zrejme

$$P(W \text{ prijmem}) = \frac{|G|}{|H|} = \frac{1}{2by^*(z^* - z_*)}.$$

Najčastejšie sa používajú hodnoty $a = 0$ a $b = 1$. Podielová metóda dáva obvykle rýchlejšie generátory ako zamietacia metóda; pri základnej variante

vďaka obvykle vyššej efektívnosti a pri zovšeobecnenej variante vďaka tomu, že nepotrebuje generovať iné rozdelenie ako rovnomerné. Efektívnosť metódy sa dá zvýšiť zmenou tvaru pokrývajúcej množiny H ; treba ale starostlivo uvážiť, či čas stratený generovaním rovnomerného rozdelenia na zložitejšej množine neprevýši zisk zo zvýšenia efektívnosti.

3.2.4 Kompozičná metóda

V zásade je určená pre generovanie náhodných čísel zo zmesi rozdelení; táto trieda rozdelení je však dosť široká.

Veta 3.2.5 *Nech $H(y)$ je distribučná funkcia a $\{g_y(x)\}$ systém hustôt pravdepodobnosti závislých od parametra y . Nech $NV Y$ má rozdelenie s distribučnou funkciou $H(y)$ a nech podmienené rozdelenie $NV X$ pri danom Y má hustotu g_Y . Potom $NV X$ má rozdelenie s hustotou $f(x) = \int_{-\infty}^{+\infty} g_y(x) dH(y)$.*

Dôkaz. Zrejme

$$\begin{aligned} P(X < x) &= \int_{-\infty}^{+\infty} P(X < x | Y = y) dH(y) = \int_{-\infty}^{+\infty} \int_{-\infty}^x g_y(z) dz dH(y) = \\ &= \int_{-\infty}^x \int_{-\infty}^{+\infty} g_y(z) dH(y) dz = \int_{-\infty}^x f(z) dz \end{aligned}$$

■

Príklad: Nech

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} a_n x^n \quad \text{pre } x \in \langle 0; 1 \rangle, \text{ kde } a_n \geq 0 \forall n \\ &= 0 \quad \text{ináč} \end{aligned}$$

Keďže

$$1 = \int_0^1 f(x) dx = \sum_{n=0}^{\infty} \frac{a_n}{n+1}$$

je $\left\{ \frac{a_n}{n+1} \right\}$ rozdelenie pravdepodobnosti. Je teda

$$f(x) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} \cdot (n+1) x^n,$$

kde $g_n(x) = (n+1)x^n$ je hustota pravdepodobnosti na intervale $\langle 0; 1 \rangle$. Z toho vyplýva algoritmus generovania čísel s rozdelením $f(x)$:

1. Generujeme náhodné číslo N s rozdelením $\left\{\frac{a_n}{n+1}\right\}$.
2. Generujeme náhodné číslo X s rozdelením s hustotou $g_N(x)$.

Pretože $G_N(x) = x^{N+1}$, $x \in \langle 0; 1 \rangle$, je $G_N^{-1}(y) = y^{\frac{1}{N+1}}$, $y \in \langle 0; 1 \rangle$.
Podľa metódy inverznej transformácie teda platí

$$Y \sim R(0; 1) \Rightarrow Y^{\frac{1}{N+1}} \sim g_N$$

3.3 Špeciálne metódy

V tejto podkapitole spomenieme niektoré špeciálne metódy pre generovanie náhodných čísel z rozdelení dôležitých pre praktické aplikácie. Nerobíme si pritom nárok na úplnosť.

3.3.1 Normálne rozdelenie

Veta 3.3.1 *Nech U_1, U_2 sú nezávislé NV s rozdelením $R(0; 1)$. Potom*

$$X_1 = \sqrt{-2 \ln U_1} \sin(2\pi U_2) \quad a \quad X_2 = \sqrt{-2 \ln U_1} \cos(2\pi U_2)$$

sú nezávislé NV s rozdelením $N(0; 1)$.

Dôkaz. Máme zobrazenie

$$t : \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \rightarrow \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad z \quad (0; 1)^2 \quad \text{na} \quad \mathbb{R}^2;$$

inverzné zobrazenie je zrejme

$$\tau : \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \rightarrow \begin{pmatrix} e^{-\frac{1}{2}(X_1^2+X_2^2)} \\ \frac{1}{2\pi} \arctan\left(\frac{X_1}{X_2}\right) \end{pmatrix}.$$

Jeho Jakobián je

$$\begin{aligned} D_\tau &= \begin{vmatrix} -X_1 e^{-\frac{1}{2}(X_1^2+X_2^2)} & -X_2 e^{-\frac{1}{2}(X_1^2+X_2^2)} \\ \frac{1}{2\pi} \cdot \frac{X_2}{X_1^2+X_2^2} & \frac{1}{2\pi} \cdot \frac{-X_1}{X_1^2+X_2^2} \end{vmatrix} = \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(X_1^2+X_2^2)} \left(\frac{X_1^2}{X_1^2+X_2^2} + \frac{X_2^2}{X_1^2+X_2^2} \right) = \frac{1}{2\pi} e^{-\frac{1}{2}(X_1^2+X_2^2)} \end{aligned}$$

Združená hustota (X_1, X_2) potom je

$$g(x_1, x_2) = f(\tau(x_1, x_2)) |D_\tau(x_1, x_2)| = 1 \cdot \frac{1}{2\pi} e^{-\frac{1}{2}(X_1^2+X_2^2)}, \quad (x_1, x_2) \in \mathbb{R}^2.$$

Zrejme platí $g(x_1, x_2) = \varphi(x_1) \cdot \varphi(x_2)$, kde $\varphi(x)$ je hustota $N(0; 1)$, z čoho vyplýva tvrdenie. ■

Ak použijeme túto vetu na generovanie normálnych náhodných čísel s využitím jediného generátora $R(0; 1)$, bude medzi po sebe idúcimi číslami dosť silná závislosť (všetky generované body (X_i, X_{i+1}) budú ležať na špirále). Je preto lepšie používať dva nezávislé $R(0; 1)$ generátory, napr.

$$U'_{n+1} = 5U'_n \pmod{2^{35}} \quad \text{a} \quad U''_{n+1} = 131U''_n \pmod{2^{35}}.$$

3.3.2 Gama a beta rozdelenie

Veta 3.3.2 *Nech $\alpha > 0$, $p \in \mathbb{N}$ a U_1, \dots, U_p sú nezávislé NV s rozdelením $R(0; 1)$. Potom NV*

$$X = -\frac{1}{\alpha} \sum_{i=1}^p \ln(U_i)$$

má rozdelenie $\Gamma(\alpha, p)$.

Dôkaz. Definujme $Z_i = -\frac{1}{\alpha} \ln(U_i)$. Keďže $U_i = e^{-\alpha Z_i}$, hustota Z_i je

$$g(z) = \alpha e^{-\alpha z}, \quad z > 0,$$

čiže $Z_i \sim \text{Exp}(\alpha) = \Gamma(\alpha, 1)$ a Z_i sú nezávislé. Stačí teda dokázať, že súčet nezávislých Γ NV má opäť Γ rozdelenie s príslušnými parametrami. Nech $Y_1 \sim \Gamma(\alpha, p)$ a $Y_2 \sim \Gamma(\alpha, q)$ sú nezávislé NV. Potom združená hustota (Y_1, Y_2) je

$$h(x, y) = \frac{1}{\Gamma(p)} \alpha^p x^{p-1} e^{-\alpha x} \cdot \frac{1}{\Gamma(q)} \alpha^q y^{q-1} e^{-\alpha y}, \quad x, y > 0$$

a podľa vety o konvolúcii pre hustotu súčtu $Y_1 + Y_2$ platí

$$\begin{aligned} f(x) &= \int_0^\infty h(u, x-u) du = \int_0^x \frac{\alpha^{p+q}}{\Gamma(p)\Gamma(q)} u^{p-1} e^{-\alpha u} (x-u)^{q-1} e^{-\alpha(x-u)} du = \\ &= \int_0^x \frac{\alpha^{p+q}}{\Gamma(p)\Gamma(q)} e^{-\alpha x} u^{p-1} (x-u)^{q-1} du = \left\{ \begin{array}{l} \text{substitúcia: } \frac{u}{x} = z \\ u = xz, \quad du = x dz \end{array} \right\} \\ &= \frac{\alpha^{p+q}}{\Gamma(p)\Gamma(q)} e^{-\alpha x} \int_0^1 x^{p+q-2} z^{p-1} (1-z)^{q-1} x dz = \\ &= \frac{\alpha^{p+q}}{\Gamma(p)\Gamma(q)} x^{p+q-1} e^{-\alpha x} B(p, q) = \frac{\alpha^{p+q}}{\Gamma(p+q)} x^{p+q-1} e^{-\alpha x}, \end{aligned}$$

čo je hustota $\Gamma(\alpha, p+q)$. Indukciou teda dostaneme $\sum_{i=1}^p \Gamma(\alpha, 1) = \Gamma(\alpha, p)$. ■

Keďže výpočet logaritmu je časovo náročný, v praxi sa ráta podľa vzorca

$$X = -\frac{1}{\alpha} \ln \left(\prod_{i=1}^p U_i \right).$$

Veta 3.3.3 *Nech $\alpha > 0$, $p \in (0; 1)$ a $Y \sim B(p, 1-p)$ a $Z \sim \Gamma(1, 1)$ sú nezávislé NV. Potom*

$$X = \frac{1}{\alpha}YZ$$

má rozdelenie $\Gamma(\alpha, p)$.

Dôkaz. Máme zobrazenie

$$t : \begin{pmatrix} Y \\ Z \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{\alpha}YZ \\ Z \end{pmatrix} \quad z \in (0; 1) \times (0; \infty) \text{ na } (0; \infty)^2;$$

inverzné zobrazenie je zrejme

$$\tau : \begin{pmatrix} X \\ Z \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \frac{X}{Z} \\ Z \end{pmatrix} \quad \text{a jeho Jakobián je } D_\tau = \begin{vmatrix} \alpha \frac{1}{Z} & -\alpha \frac{X}{Z^2} \\ 0 & 1 \end{vmatrix} = \frac{\alpha}{Z}.$$

Keďže združená hustota (Y, Z) je

$$h(y, z) = \frac{1}{B(p, 1-p)} y^{p-1} (1-y)^{-p} e^{-z}, \quad y \in (0; 1), \quad z > 0,$$

pre združenú hustotu (X, Z) platí

$$g(x, z) = h\left(\alpha \frac{x}{z}, z\right) |D_\tau(x, z)| = \frac{1}{B(p, 1-p)} \left(\frac{\alpha x}{z}\right)^{p-1} \left(1 - \alpha \frac{x}{z}\right)^{-p} e^{-z} \cdot \frac{\alpha}{z},$$

kde $x > 0$, $z > 0$, $\alpha x < z$. Potom X má hustotu

$$\begin{aligned} f(x) &= \int_0^\infty g(x, z) dz = \int_{\alpha x}^\infty \frac{1}{B(p, 1-p)} \left(\frac{\alpha x}{z}\right)^{p-1} \left(1 - \alpha \frac{x}{z}\right)^{-p} \frac{\alpha}{z} e^{-z} dz = \\ &= \int_{\alpha x}^\infty \frac{\alpha^p x^{p-1}}{B(p, 1-p)} (z - \alpha x)^{-p} e^{-z} dz = \left\{ \begin{array}{l} \text{substitúcia:} \\ z = u + \alpha x, \quad dz = du \end{array} \right\} = \\ &= \int_0^\infty \frac{\alpha^p x^{p-1}}{B(p, 1-p)} u^{-p} e^{-u-\alpha x} du = \frac{\alpha^p x^{p-1}}{B(p, 1-p)} e^{-\alpha x} \Gamma(1-p) = \\ &= \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, \quad x > 0, \end{aligned}$$

čo je hustota $\Gamma(\alpha, p)$. ■

Ak teda máme $p \in \mathbb{R}^+$, môžeme ho rozložiť na celú a zlomkovú časť, $p = [p] + \{p\}$, vygenerovať nezávislé NV s rozdelením $\Gamma(\alpha, [p])$ a $\Gamma(\alpha, \{p\})$ a využiť skutočnosť, že $\Gamma(\alpha, p) = \Gamma(\alpha, [p]) + \Gamma(\alpha, \{p\})$. K úplnosti algoritmu nám však ešte chýba generátor B -rozdelenia.

Veta 3.3.4 *Nech Y, Z sú nezávislé $R(0; 1)$ NV, $p, q > 0$. Potom NV*

$$X = \frac{Y^{\frac{1}{p}}}{Y^{\frac{1}{p}} + Z^{\frac{1}{q}}}$$

za podmienky, že $Y^{\frac{1}{p}} + Z^{\frac{1}{q}} < 1$ má rozdelenie $B(p, q)$.

Dôkaz. Máme zobrazenie

$$t : \begin{pmatrix} Y \\ Z \end{pmatrix} \rightarrow \begin{pmatrix} X \\ T \end{pmatrix} \quad \text{kde } T = Y^{\frac{1}{p}} + Z^{\frac{1}{q}}, \quad z \ (0; 1)^2 \text{ na } (0; 1) \times (0; 2);$$

k nemu inverzné zobrazenie je

$$\tau : \begin{pmatrix} X \\ T \end{pmatrix} \rightarrow \begin{pmatrix} T^p X^p \\ T^q (1 - X)^q \end{pmatrix}.$$

Je teda

$$\begin{aligned} D_\tau &= \begin{vmatrix} pX^{p-1}T^p & pX^pT^{p-1} \\ -q(1-X)^{q-1}T^q & q(1-X)^qT^{q-1} \end{vmatrix} = \\ &= pqX^{p-1}(1-X)^{q-1}T^{p+q-1}(1-X+X) = \\ &= pqX^{p-1}(1-X)^{q-1}T^{p+q-1} \end{aligned}$$

a združená hustota (X, T) je

$$g(x, t) = 1 \cdot pqx^{p-1}(1-x)^{q-1}t^{p+q-1}, \quad x \in (0; 1), \quad t \in (0; 2).$$

Potom

$$\begin{aligned} P(T < 1) &= \int_0^1 \int_0^1 pqx^{p-1}(1-x)^{q-1}t^{p+q-1} dx dt = \\ &= \int_0^1 x^{p-1}(1-x)^{q-1} dx \int_0^1 pqt^{p+q-1} dt = B(p, q) \frac{pq}{p+q}, \end{aligned}$$

a

$$\begin{aligned} P(X < x, T < 1) &= \int_0^x \int_0^1 pqx^{p-1}(1-u)^{q-1}t^{p+q-1} du dt = \\ &= \int_0^x u^{p-1}(1-u)^{q-1} du \int_0^1 pqt^{p+q-1} dt = \\ &= \frac{pq}{p+q} \int_0^x u^{p-1}(1-u)^{q-1} du. \end{aligned}$$

Z toho vyplýva, že

$$\begin{aligned} P(X < x | T < 1) &= \frac{P(X < x, T < 1)}{P(T < 1)} = \frac{1}{B(p, q)} \int_0^x u^{p-1}(1-u)^{q-1} du = \\ &= IB(x; p, q), \end{aligned}$$

čo je distribučná funkcia rozdelenia $B(p, q)$. ■

Nevýhodou tejto metódy je, že pre veľké hodnoty p, q je pravdepodobnosť splnenia podmienky $T < 1$ veľmi malá. To znamená, že jej efektívnosť s rastúcimi hodnotami p, q rýchlo klesá. Pre generovanie doplnku Γ -rozdelenia, kde $p = 1 - q \in (0; 1)$, je však vyhovujúca.

Pokiaľ sú $p, q \in \mathbb{N}$, môžeme naopak generovať B -rozdelenie pomocou Γ -rozdelenia.

Veta 3.3.5 *Nech $p, q \in \mathbb{N}$, a nech $V \sim \Gamma(1, p)$ a $W \sim \Gamma(1, q)$ sú nezávislé NV. Potom*

$$X = \frac{V}{V + W}$$

má rozdelenie $B(p, q)$.

Dôkaz. Máme zobrazenie

$$t : \begin{pmatrix} V \\ W \end{pmatrix} \rightarrow \begin{pmatrix} X \\ T \end{pmatrix} \quad \text{kde } T = V + W, \text{ z } (0; \infty)^2 \text{ na } (0; 1) \times (0; \infty);$$

k nemu inverzné zobrazenie je

$$\tau : \begin{pmatrix} X \\ T \end{pmatrix} \rightarrow \begin{pmatrix} XT \\ (1 - X)T \end{pmatrix}.$$

Pre jeho Jakobián platí

$$D_\tau = \begin{vmatrix} T & X \\ -T & (1 - X) \end{vmatrix} = T(1 - X) + TX = T;$$

združená hustota (X, T) teda je

$$\begin{aligned} g(x, t) &= h(xt, (1 - x)t) |D_\tau(x, t)| = \\ &= \frac{1}{\Gamma(p)} (xt)^{p-1} e^{-xt} \cdot \frac{1}{\Gamma(q)} ((1 - x)t)^{q-1} e^{-(1-x)t} \cdot t = \\ &= \frac{1}{\Gamma(p)\Gamma(q)} x^{p-1} (1 - x)^{q-1} t^{p+q-1} e^{-t}, \quad x \in (0; 1), t > 0. \end{aligned}$$

Hustota X potom je

$$\begin{aligned} f(x) &= \int_0^\infty \frac{1}{\Gamma(p)\Gamma(q)} x^{p-1} (1 - x)^{q-1} t^{p+q-1} e^{-t} dt = \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1} (1 - x)^{q-1}, \quad x \in (0; 1), \end{aligned}$$

čo je hustota rozdelenia $B(p, q)$. ■

Pre generovanie gama a beta rozdelenia sa často používajú aj špeciálne upravené kombinácie všeobecných metód (metódy inverznej transformácie, zamietacej metódy, kompozičnej metódy a podielovej metódy).

3.3.3 Rozdelenia súvisiace s normálnym

Pre rozdelenie χ^2 si stačí uviesť, že

$$\chi_n^2 = \Gamma\left(\frac{1}{2}, \frac{n}{2}\right) \quad \text{alebo} \quad \chi_n^2 = \sum_1^n (N(0; 1))^2.$$

Prvá možnosť je výrazne lepšia, keďže v druhom prípade veľmi záleží na nezávislosti členov súčtu.

Podobne

$$t_n = \frac{N(0; 1)}{\sqrt{\chi_n^2}} \sqrt{n}.$$

Táto štatistika je robustná na porušenie normality, ale **extrémne** citlivá na porušenie nezávislosti čitateľa a menovateľa. Treba preto používať na ich generovanie rôzne generátory. Častejšie sa však na generovanie t-rozdelenia používa kombinácia zovšeobecnenej zamietacej a podielovej metódy. Dominujúca hustota pre zamietaciu metódu je hustota $t_3(x) = \frac{2}{\pi\sqrt{3}} \left(1 + \frac{x^2}{3}\right)^{-2}$, pre ktorú je potrebný koeficient $c = \frac{8}{9} \sqrt{\frac{3\pi}{2e}}$, aby dominovala ľubovoľnú

$$t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Rozdelenie t_3 sa pritom generuje podielovou metódou s parametrami $a = 0$, $b = \frac{\sqrt{3}}{2}$.

Pre F -rozdelenie platí

$$F_{n,m} = \frac{\chi_n^2}{\chi_m^2} \cdot \frac{m}{n}.$$

Opäť treba dať veľký pozor na nezávislosť čitateľa a menovateľa, t.j. použiť pre ne rôzne generátory. Jednoduchšie je však generovanie pomocou beta-rozdelenia:

Veta 3.3.6 *Nech $Y \sim B\left(\frac{m}{2}, \frac{n}{2}\right)$. Potom NV*

$$X = \frac{nY}{m(1-Y)}$$

má rozdelenie $F(m, n)$.

Dôkaz. Hustota rozdelenia $B\left(\frac{m}{2}, \frac{n}{2}\right)$ je

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} y^{\frac{m}{2}-1} (1-y)^{\frac{n}{2}-1}, \quad y \in (0; 1).$$

Inverzná transformácia $k: t: Y \rightarrow X \in (0; +\infty)$ je

$$\tau: X \rightarrow Y = \frac{mX}{mX + n}.$$

Jej derivácia je

$$D_\tau = \frac{mn}{(mX + n)^2};$$

hustota NV X je teda

$$\begin{aligned} f(x) &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{mx}{mx+n}\right)^{\frac{m}{2}-1} \left(1 - \frac{mx}{mx+n}\right)^{\frac{n}{2}-1} \frac{mn}{(mx+n)^2} = \\ &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (mx+n)^{-\frac{m+n}{2}} = \\ &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, \end{aligned}$$

čo je hustota rozdelenia $F(m, n)$. ■

Jednoducho vieme generovať aj log-normálne rozdelenie pomocou transformácie

$$LN(\mu, \sigma^2) = \exp(N(\mu, \sigma^2)).$$

3.3.4 Náhodné permutácie

Niekedy potrebujeme generovať náhodné permutácie. Keďže priestor všetkých permutácií rádu n je pomerne zložitý a rozsiahly, nie je to celkom triviálna úloha. Pre nenáročnejšie aplikácie často postačí nasledujúca približná metóda, ktorá je výpočtovo veľmi jednoduchá:

1. Vložíme $1 \rightarrow P(1), \dots, n \rightarrow P(n)$
2. V cykle pre $i = 1, \dots, n$ vygenerujeme náhodné číslo $X \sim R(1, \dots, n)$ a zameníme obsahy $P(i)$ a $P(X)$.

Presná metóda je založená na nasledujúcej vete:

Veta 3.3.7 (Marshall Hall) *Nech $b = (b_0, b_1, \dots, b_{n-1})$ je permutácia čísel $0, 1, \dots, n-1$. Nech a_k označuje počet tých b_i , ktoré v postupnosti b_0, b_1, \dots, b_{n-1} nasledujú za k a sú menšie ako k , $k = 1, \dots, n-1$. Potom*

$$f(b) = a_1 1! + a_2 2! + \dots + a_{n-1} (n-1)!$$

je vzájomne jednoznačné zobrazenie množiny všetkých permutácií na množinu celých čísel $0, 1, \dots, n! - 1$.

Dôkaz. Zrejme minimum dosiahne funkcia $f(b)$ pre $b = (0, 1, \dots, n-1)$, kde všetky $a_k = 0$; je teda $\min_b f(b) = 0$. Maximum táto funkcia dosiahne pre $b = (n-1, n-2, \dots, 0)$, kde platí $a_k = k \forall k$, čiže

$$\begin{aligned} \max_b f(b) &= 1 \cdot 1! + 2 \cdot 2! + \dots + (n-1) \cdot (n-1)! = \\ &= n! - (n-1)! + (n-1)! - (n-2)! + \dots + 2! - 1! = n! - 1. \end{aligned}$$

Funkcia f teda zobrazuje množinu všetkých permutácií rádu n (ktorá má $n!$ prvkov) do množiny celých čísel medzi 0 a $n! - 1$. Predpokladajme, že táto funkcia nie je prostá, t.j. že existujú permutácie $b_1 \neq b_2$ také, že $f(b_1) = f(b_2)$. Z toho vyplýva, že musí existovať $n_0 \in \{0, \dots, n! - 1\}$ také, že $f(b) = n_0 \forall b$.

Z Euklidovho algoritmu pre delenie vyplýva, že $\exists! a_{n-1}, r_{n-1}$ také, že

$$n_0 = a_{n-1} (n-1)! + r_{n-1},$$

kde $0 \leq r_{n-1} < (n-1)!$. Navyiac z nerovnosti $n_0 \leq n! - 1$ vyplýva $a_{n-1} \leq n-1$. Podobne pre $r_{n-1} \exists! a_{n-2}, r_{n-2}$ také, že

$$r_{n-1} = a_{n-2} (n-2)! + r_{n-2},$$

kde $0 \leq r_{n-2} < (n-2)!$, $a_{n-2} \leq n-2$, atď., až pre $r_2 \exists! a_1$ také, že

$$r_2 = a_1 1! + 0,$$

kde $a_1 \leq 1$. Získali sme teda čísla a_{n-1}, \dots, a_1 , $a_j \leq j \forall j$ (uvedomme si, že čísla a_j musia mať túto vlastnosť vzhľadom k spôsobu ich definície) také, že

$$n_0 = \sum_{j=1}^{n-1} a_j j!.$$

Navyiac platí

$$\forall k \sum_{j=1}^{k-1} a_j j! < k!,$$

lebo výrok zrejme platí pre $k = 2$ a

$$\sum_{j=1}^k a_j j! = \sum_{j=1}^{k-1} a_j j! + a_k k! < k! + k \cdot k! = (k+1)!.$$

Predpokladajme, že existujú dva rôzne rozklady $\{a_j\}$, $\{c_j\}$ čísla n_0 s týmito vlastnosťami. Nech j_0 je najväčší index taký, že $a_{j_0} \neq c_{j_0}$ (bez ujmy na všeobecnosti nech $a_{j_0} < c_{j_0}$). Potom musí byť

$$\sum_{j=1}^{j_0} a_j j! = \sum_{j=1}^{j_0} c_j j!,$$

lebo vyššie členy sú rovnaké. Z toho ale vyplýva, že

$$\underbrace{\sum_{j=1}^{j_0-1} (a_j - c_j) j!}_{< j_0!} = \underbrace{(c_{j_0} - a_{j_0}) j_0!}_{\neq 0, \geq j_0!},$$

čo je spor. Pre dané n_0 teda existuje jediná $(n-1)$ -tica čísel a_{n-1}, \dots, a_1 , $a_j \leq j \forall j$ také, že

$$n_0 = \sum_{j=1}^{n-1} a_j j!.$$

Ukážeme, že k nej existuje permutácia b taká, že čísla $\{a_j\}$ majú vyššie uvedený význam. Túto permutácia budeme konštruovať postupne od menších čísel k väčším:

$$a_1 = \begin{cases} 1 & \dots 0, 1 \text{ sú v poradí } 1, 0 \\ 0 & \dots 0, 1 \text{ sú v poradí } 0, 1 \end{cases}$$

(teda 1 dáme na prvú alebo druhú pozíciu podľa hodnoty a_1 , vo všeobecnosti na $(2 - a_1)$ -tú pozíciu),

$$a_2 = \begin{cases} 2 & \dots 0, 1, 2 \text{ sú v poradí } 2, 1, 0 \text{ alebo } 2, 0, 1 \\ 1 & \dots 0, 1, 2 \text{ sú v poradí } 1, 2, 0 \text{ alebo } 0, 2, 1 \\ 0 & \dots 0, 1, 2 \text{ sú v poradí } 1, 0, 2 \text{ alebo } 0, 1, 2 \end{cases}$$

(teda 2 dáme na prvú, druhú alebo tretiu pozíciu podľa hodnoty a_2 , vo všeobecnosti na $(3 - a_2)$ -tú pozíciu), atď. V k -tom kroku teda zaradíme číslo k na $(k + 1 - a_k)$ -tú pozíciu v postupnosti; je zřejmé, že takto vznikne permutácia čísel $0, 1, \dots, n - 1$, v ktorej budú mať čísla a_k horeuvedený význam. ■

Algoritmus generovania je teda nasledujúci:

1. Generujeme $N \sim R(0, 1, \dots, n! - 1)$.
2. Nájdeme koeficienty a_j rozkladu $N = \sum_{j=1}^{n-1} a_j j!$ s danými vlastnosťami.
3. Na základe čísel $\{a_j\}$ vytvoríme permutáciu b spôsobom uvedeným v dôkaze.

Ďalšie informácie je možné nájsť napr. v [9].

Kapitola 4

Využitie náhodných čísel

4.1 Simulácie

Náhodné čísla sa v štatistike využívajú veľmi často. Typická je najmä situácia, keď máme nejakú štatistiku, ktorej rozdelenie je neznáme, ale my ho aspoň čiastočne potrebujeme poznať. Vtedy môžeme opakovane generovať náhodné výbery z príslušného základného rozdelenia, vypočítať danú štatistiku a z jej mnohých realizácií vypočítať empirické rozdelenie. Pritom niekedy nás môžu zaujímať iba niektoré charakteristiky (napr. rozptyl), niekedy celé rozdelenie.

Definícia 4.1.1 *Nech X_1, \dots, X_n je náhodný výber z rozdelenia s distribučnou funkciou $F(x)$. Ak $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ je tomu zodpovedajúci usporiadaný náhodný výber, potom definujeme empirickú distribučnú funkciu vzťahom*

$$F_n(x) = \begin{cases} 0 & \text{ak } x \leq X_{(1)} \\ \frac{k}{n} & \text{ak } X_{(r)} < x \leq X_{(k+1)} \text{ a } X_{(r-1)} < X_{(r)} = \dots = X_{(k)} < X_{(k+1)} \\ 1 & \text{ak } X_{(n)} < x \end{cases}$$

Funkcia $F_n(x)$ teda udáva podiel pozorovaní vo výbere, ktoré sú menšie ako x . Pre pevné x je $F_n(x)$ NV nadobúdajúca hodnoty $0, \frac{1}{n}, \dots, 1$. Keďže pravdepodobnosť, že jednotlivé pozorovanie bude menšie ako x je $F(x)$, platí

$$P\left(F_n(x) = \frac{k}{n}\right) = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}.$$

Z toho vyplýva, že

$$E F_n(x) = \sum_{k=0}^n \frac{k}{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k} = \frac{1}{n} n F(x) = F(x).$$

Podľa silného zákona veľkých čísel platí

$$F_n(x) \xrightarrow[n \rightarrow \infty]{} F(x) \text{ s.i.}$$

Platí dokonca ešte silnejšie tvrdenie:

Veta 4.1.2 (Glivenko) Postupnosť $F_n(x)$ konverguje k $F(x)$ pri $n \rightarrow \infty$ rovnomerne na \mathbb{R} skoro iste, t.j.

$$P\left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1.$$

Takéto výpočty sú potrebné najmä v nasledujúcich situáciách:

- Potrebujeme mať predstavu o rozdelení pravdepodobnosti danej štatistiky (odhadu), ktoré napriek znalosti základného rozdelenia nevieme odvodiť analyticky.
- Skúmame správanie sa štatistiky pri porušení predpokladov modelu. (O-dolnosť voči porušeniu predpokladov sa nazýva **robustnosť**.) Môže to byť napr. iné základné rozdelenie, narušenie predpokladu nezávislosti pozorovaní a pod.
- Chceme zistiť rýchlosť konvergencie rozdelenia danej štatistiky k známemu asymptotickému rozdeleniu, resp. určiť približný rozsah výberu umožňujúci aplikáciu asymptotických metód s rozumnou presnosťou.
- Potrebujeme zistiť približný priebeh silofunkcie daného testu atď.

Všetky tieto situácie majú spoločné to, že poznáme parametre simulovaného rozdelenia.

4.2 Približný výpočet integrálu

Veľa úloh štatistických simulácií sa dá formulovať ako výpočet hodnoty nejakého (eventuálne viacrozmerného) integrálu - je to napr. výpočet pravdepodobnosti, strednej hodnoty, rozptylu a pod. My sa budeme zaoberať len najjednoduchším prípadom, výpočtom integrálu

$$I = \int_0^1 f(x) dx$$

pre danú funkciu $f(x) \geq 0 \forall x \in (0; 1)$.

Všeobecná metóda Ak $U_1, \dots, U_n \sim R(0; 1)$ sú nezávislé NV, potom $f(U_1), \dots, f(U_n)$ sú tiež nezávislé NV a platí $E f(U_i) = I$. Sú to teda nestranné odhady hodnoty I . Ich aritmetický priemer je potom tiež nestranným odhadom I :

$$\theta_1 = \frac{1}{n} \sum_{i=1}^n f(U_i).$$

Zamietacia metóda Nech $f(x) \leq c \forall x \in (0; 1)$. Generujeme náhodné body $(U_1, V_1), \dots, (U_n, V_n)$ s rovnomerným rozdelením na $(0; 1) \times (0; c)$, t.j. U_i tvoria náhodný výber z $R(0; 1)$ a V_j nezávislý náhodný výber z $R(0; c)$. Definujeme funkciu

$$g(x, y) = \begin{cases} 0 & \text{ak } f(x) < y \\ 1 & \text{ak } f(x) \geq y \end{cases} .$$

Keďže zrejme $f(x) = \int_0^1 g(x, y) dy$, je

$$I = \int_0^1 \int_0^1 g(x, y) dy dx .$$

Aplikovaním všeobecnej metódy dostaneme, že nestranným odhadom I je

$$\theta_2 = c \frac{1}{n} \sum_{i=1}^n g(U_i, V_i) = c \frac{n_1}{n} ,$$

kde n_1 je počet vygenerovaných bodov ležiacich pod grafom funkcie $f(x)$.

Metóda výberu podľa dôležitosti Nech $h(x)$ je hustota pravdepodobnosti taká, že $h(x) = 0 \iff f(x) = 0$. Potom platí

$$I = \int_0^1 \frac{f(x)}{h(x)} h(x) dx = \int_0^1 \frac{f(x)}{h(x)} dH(x) ,$$

kde $H(x)$ je zodpovedajúca distribučná funkcia. Ak máme náhodný výber X_1, \dots, X_n z rozdelenia s distribučnou funkciou $H(x)$, potom nestranným odhadom integrálu I je

$$\theta_3 = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{h(X_i)} .$$

Táto metóda má za cieľ vyberať viac bodov v oblastiach, ktoré sú nejakým spôsobom dôležité (napr. tam, kde sa hodnoty f prudko menia), a tým znížiť rozptyl odhadu.

Metóda stratifikovaného výberu Je to vlastne špeciálny prípad predchádzajúcej metódy. Celý interval $(0; 1)$ je rozdelený na k podintervalov bodmi $0 = a_0 < a_1 < \dots < a_k = 1$. Na i -tom podintervale potom generujeme pevný počet n_i čísel s rovnomerným rozdelením (na tomto podintervale). Ak teda $\sum_{i=1}^k n_i = n$ a U_1, \dots, U_n je náhodný výber z $R(0; 1)$, potom definujeme $U_{ij} = U_{n_1 + \dots + n_{i-1} + j}$ a nestranný odhad I zrejme je

$$\theta_4 = \sum_{i=1}^k \frac{1}{n_i} \sum_{j=1}^{n_i} (a_i - a_{i-1}) f(a_{i-1} + (a_i - a_{i-1}) U_{ij}) .$$

Metóda riadiacích premenných Nech $g(x)$ je funkcia, ktorá dobre aproximuje funkciu $f(x)$ taká, že integrál

$$J = \int_0^1 g(x) dx$$

vieme vypočítať analyticky. Zrejme

$$I = \int_0^1 g(x) dx + \int_0^1 f(x) - g(x) dx.$$

Nestranným odhadom I potom je

$$\theta_5 = J + \frac{1}{n} \sum_{i=1}^n (f(U_i) - g(U_i)),$$

kde U_1, \dots, U_n je náhodný výber z $R(0; 1)$.

Metóda antitetických premenných Nech $g(x)$ je taká funkcia, že

$$\int_0^1 g(x) dx = I.$$

Potom aj

$$\int_0^1 \frac{f(x) + g(x)}{2} dx = I.$$

Ak teda $U_1, \dots, U_n \sim R(0; 1)$ sú nezávislé NV, potom nestranný odhad I je

$$\theta_6 = \frac{1}{2n} \sum_{i=1}^n (f(U_i) + g(U_i)).$$

Vhodnou voľbou funkcie $g(x)$ môžeme výrazne znížiť rozptyl odhadu, najmä vplyv prudkých zmien hodnôt f . Špeciálne, ak funkcia $f(x)$ je monotónna, potom funkcia $g(x) = f(1-x)$ spĺňa naše požiadavky a platí

$$\theta_6 = \frac{1}{2n} \sum_{i=1}^n (f(U_i) + f(1-U_i)).$$

V takom prípade hovoríme o jednoduchšej symetrizácii odhadu. Zrejme je možné robiť aj symetrizáciu na podintervaloch.

V prípade jednorozmerných integrálov sú klasické numerické metódy obvykle účinnejšie ako metódy Monte Carlo. Ich sila je hlavne vo výpočte viacrozmerných integrálov, hoci zovšeobecnenie jednotlivých metód do viacrozmerného priestoru nie je vždy triviálne. Podrobnejšie o týchto metódach viď napr. [3].

4.3 Metóda opakovaných výberov (bootstrap)

V rôznych aplikačných úlohách sa stávalo, že na základe náhodného výberu bola vypočítaná štatistika (napr. nejaký odhad), o rozdelení ktorej nebolo z principiálnych dôvodov nič známe (napr. nebolo známe základné rozdelenie, alebo trieda možných základných rozdelení bola veľmi široká). Napriek tomu bolo potrebné čosi sa dozvedieť o jej rozdelení (napr. jej rozptyl).

Prvým pokusom riešiť túto situáciu bola **vynechávajúca metóda (jack-knife method)**:

Ak X_1, \dots, X_n bol pôvodný výber a $S_n(X_1, \dots, X_n)$ vypočítaná štatistika, potom séria odhadov

$$S_{n-1}(X_2, \dots, X_n), S_{n-1}(X_1, X_3, \dots, X_n), \dots, S_{n-1}(X_1, \dots, X_{n-1})$$

založených na systematicky vytvorených podmnožinách pôvodného výberu (vždy s vynechaním jedného pozorovania) umožnila skonštruovať empirické rozdelenie štatistiky S (ale fakticky S_{n-1} , nie S_n). Túto metódu navrhli Quenouille a Tukey.

Neskôr Bradley Efron (viď [4]) tento princíp zovšeobecnil a nazval ho **svojpomocnou metódou (bootstrap method)**. My budeme používať názov **metóda opakovaných výberov (MOV)**:

Nech X_1, \dots, X_n je náhodný výber z neznámeho rozdelenia F . Označme $X = (X_1, \dots, X_n)$ a $x = (x_1, \dots, x_n)$ jeho pozorovanú realizáciu. Chceme odhadnúť rozdelenie NV $R = R(X, F)$ na základe pozorovaného vektora x .

Obvykle sa používajú štatistiky dvoch druhov:

$$R(X, F) = t(X) - \theta(F)$$

resp.

$$R(X, F) = \frac{t(X) - \text{odhad vychýlenia } t(X) - \theta(F)}{\sqrt{\text{odhad rozptylu } t(X)}},$$

kdže $\theta(F)$ je parameter, ktorý nás zaujíma a $t(X)$ jeho odhad.

Algoritmus riešenia problému MOV je nasledujúci:

1. Skonštruujeme výberové rozdelenie $\hat{F} = F_n(x)$.
2. Pri pevnom \hat{F} urobíme náhodný výber rozsahu n z \hat{F} ; označíme ho X^* resp. x^* . Tento výber budeme nazývať MOV-výberom alebo svojpomocným výberom.
3. Rozdelenie $R(X, F)$ aproximujeme MOV-rozdelením (svojpomocným rozdelením) NV $R^* = R(X^*, \hat{F})$.

Všimnime si, že MOV-výber nie je permutáciou x_1, \dots, x_n , pretože ide o výber s vracaním. Rozdelenie R^* , ktoré by sme mali teoreticky vedieť vždy vypočítať, je rovné skutočnému rozdeleniu R , ak $\hat{F} = F$. Vzhľadom ku konzistencii \hat{F} je to teda dobrý odhad. MOV-rozdelenie v praxi môžeme skonštruovať tromi spôsobmi:

- a) Priamym teoretickým výpočtom. To sa dá iba vo veľmi jednoduchých situáciách.
- b) Pomocou Taylorovho rozvoja vypočítať približnú strednú hodnotu a rozptyl MOV-rozdelenia. Dá sa ukázať, že je to špeciálny prípad vynechávacej metódy.
- c) Monte Carlo aproximáciou. Robíme teda opakované náhodné výbery X^* rozsahu n z \hat{F} a z ich realizácií x_1^*, \dots, x_n^* zostrojíme empirické rozdelenie $G_N(R^*)$.

Najmä posledná metóda by nebola mysliteľná bez použitia počítačov a generátorov (pseudo)náhodných čísel.

MOV má už dnes veľa rôznych aplikácií a modifikácií. Prirodzené je napr. rozšírenie na dvojjvýberový problém:

Nech $X = (X_1, \dots, X_m)$ a $Y = (Y_1, \dots, Y_n)$ sú nezávislé náhodné výbery z rozdelení F a G . Nech x a y sú pozorované hodnoty X a Y . Výberové rozdelenie náhodnej veličiny $R((X, Y), (F, G))$ potom aproximujeme MOV-rozdelením $R^*((X^*, Y^*), (\hat{F}, \hat{G}))$, kde $\hat{F} = F_m(x)$ a $\hat{G} = G_n(y)$ sú empirické distribučné funkcie založené na vektoroch x, y a X^*, Y^* sú MOV-výbery z \hat{F} a \hat{G} . Toto rozdelenie sa najčastejšie aproximuje Monte Carlo simuláciou.

Veľký význam má MOV aj u všeobecných regresných modelov:

Nech

$$Y_i = g_i(\beta) + \varepsilon_i, \quad i = 1, \dots, n,$$

kde g_i sú známe funkcie neznámeho vektora parametrov β . Nech $\varepsilon_i \sim F$ sú nezávislé NV centrovane v nule, t.j.

$$E_F \varepsilon = 0 \quad \text{alebo} \quad M_{e_F} \varepsilon = 0.$$

Rozdelenie F je pritom neznáme. Ak sme pozorovali $Y = y$, odhadneme nejakou štandardnou metódou neznáme parametre β , napr. MNŠ:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - g_i(\beta))^2.$$

Zaujímá nás výberové rozdelenie $\hat{\beta}$.

Definujme $\hat{F} = F_n(\hat{\varepsilon})$ ako empirickú distribučnú funkciu vektora rezíduí $\hat{\varepsilon} = (y_i - g_i(\hat{\beta}))_{i=1}^n$. (Uvedomme si, že ak jednou zo zložiek vektora β je parameter polohy funkcií g , potom \hat{F} má strednú hodnotu 0. Ak by to tak nebolo a predpoklad $E_F \varepsilon = 0$ by bol potrebný, aj tak by sa posunutím \hat{F} dala dosiahnuť nulová stredná hodnota.) MOV-výber, pri danom $(\hat{\beta}, \hat{F})$, je potom

$$Y_i^* = g_i(\hat{\beta}) + \varepsilon_i^*, \quad i = 1, \dots, n,$$

kde ε^* je náhodný výber z \widehat{F} . Každá realizácia MOV-výberu nám dá novú hodnotu $\widehat{\beta}^*$, ktorú získame rovnakým postupom ako $\widehat{\beta}$, čiže napr.

$$\widehat{\beta}^* = \arg \min_{\beta} \sum_{i=1}^n (y_i^* - g_i(\beta))^2.$$

Nezávislé realizácie $\widehat{\beta}_1^*, \dots, \widehat{\beta}_N^*$ potom môžeme použiť k zostrojeniu empirického MOV-rozdelenia $\widehat{\beta}^*$.

V klasickom regresnom modeli je

$$g_i(\beta) = x_i' \beta,$$

kde x_i je známy vektor (obvykle s prvým prvkom rovným 1) z \mathbb{R}^k . Matica plánu je teda

$$X = \begin{pmatrix} x_1' \\ \dots \\ x_n' \end{pmatrix}$$

a MNŠ-odhad β je

$$\widehat{\beta} = (X'X)^{-1} X'Y.$$

Jeho variančná matica je

$$\sigma_F^2 (X'X)^{-1}.$$

Veličiny ε_i^* z MOV-výberu majú strednú hodnotu 0 a rozptyl

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i^* - g_i(\widehat{\beta}) \right)^2.$$

Z toho vyplýva, že pre MOV-odhad $\widehat{\beta}^* = (X'X)^{-1} X'Y^*$ platí

$$E_* \widehat{\beta}^* = \widehat{\beta}, \quad \text{var}_* \widehat{\beta}^* = \widehat{\sigma}^2 (X'X)^{-1}.$$

Zhoda s klasickou teóriou je teda dobrá.

Literatúra

- [1] Abramowitz, M., Stegun, I. (1972): Handbook of mathematical functions, Dover Publications, New York.
- [2] Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1994): Multivariate Analysemethoden, Springer, Berlin.
- [3] Deák, I. (1990): Random number generators and simulation, Akadémiai kiadó, Budapest.
- [4] Efron, B. (1979): Bootstrap methods: Another look at the jackknife, *Annals of Statistics* 7, pp. 1–26.
- [5] Fishman, G. S. (1996): Monte Carlo. Concepts, Algorithms, and Applications, Springer, New York.
- [6] Hájek, P., Havránek, T., Chytil, M. K. (1983). Metoda GUHA. Automatická tvorba hypotéz. ACADEMIA, Praha.
- [7] Hájek P., Sochorová A., Zvárová J. (1995): GUHA for personal computers. - *Computational Statistics & Data Analysis* 19, pp. 149-153.
- [8] Harmancová D. (1994): PC-GUHA - brief manual.
- [9] Hurt, J. (1982): Simulační metody, skriptá MFF UK, Praha.
- [10] Matsumoto, M., Kurita, Y. (1992): Twisted GFSR Generators, *ACM Transactions on Modeling and Computer Simulation* 2, No. 3, pp. 179–194.
- [11] Matsumoto, M., Kurita, Y. (1994): Twisted GFSR Generators II, *ACM Transactions on Modeling and Computer Simulation* 4, No. 3, pp. 254–266.
- [12] Matsumoto, M., Nishimura, T. (1998): Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator, *ACM Transactions on Modeling and Computer Simulation* 8, No. 1, pp. 3–30.
- [13] Olehla, M., Věchet, V., Olehla, J. (1982): Řešení úloh matematické statistiky ve Fortranu, Nadas, Praha.