Generování náhodných čísel
Zpracování stochastických dat

Ing. Martin Hrubý

28. listopadu 2000

# 1 Generování náhodných čísel

## 1.1 Definice

### 1.1.1 Distribuční funkce $F(x)$ náhodné veličiny

$$F(x) = P(X \leq x)$$

Distribuční funkce $F(x)$ náhodné veličiny udává pravděpodobnost, že hodnota náhodné veličiny podle tohoto rozložení je menší (nebo rovno) $x$.

- je neklesající funkce (jenom roste nebo je konstantní)
- pro $x \in < -\propto, \propto >$ platí $F(x) \in < 0, 1 >$
- pro $x \to -\propto$ platí $F(x) \to 0$
- pro $x \to \propto$ platí $F(x) \to 1$

### 1.1.2 Hustota pravděpodobnosti $f(x)$

- někdy taky nazývaná *frekvenční funkce*
- $f(x) = P(X = x)$
- integrací hustoty pravděpodobnosti dostáváme *distribuční funkci*

### 1.1.3 Charakteristiky náhodné veličiny

**Charakteristika polohy:** udává posunutí rozložení vzhledem k počátku

**Charakteristika variability:** udává kolísání hodnot

**Ch. šikmosti:** nesymetričnost rozložení

**Ch. špičatosti:** porovnání variability rozložení ve středu na krajích definičního oboru

Tyto charakteristiky bývají často vyjádřeny pomocí *momentů*. Rozlišujeme *obecné momenty* (ve vztahu k počátku) a *centrální momenty* (ve vztahu ke středu). Vztahy předvedu na *diskrétní náhodné veličině.*

$$m_k(X) = \sum_{x=-\propto}^{\propto} x^k \cdot f(x) \; k = 1, 2, ...$$

Obecný moment prvního řádu se nazývá *střed* nebo *střední hodnota* ($\mu$) a vyjadřuje charakteristiku polohy. Další charakteristiky vycházejí z *centrálního momentu k-tého řádu* (opět *číslo*):

$$M_k(X) = \sum_{x=-\propto}^{\propto} (x - E(X))^k \cdot f(x)$$

Centrální moment 2. řádu se nazývá *disperze* nebo taky *rozptyl,* charakterizuje variabilitu náhodné veličiny. Značí se $D(X)$. Další charakteristiky jsou pro šikmost a špičatost:

$$\sqrt{\beta_1} = \frac{M_3(X)}{(D(X))^3}$$

$$\beta_2 = \frac{M_4(X)}{(D(X))^4}$$

Vzorce momentů pro jednotlivá rozložení se odvozují z definičních vztahů rozložení. Toto může být někdy obtížné.

### 1.1.4 Bodové odhady

Bodové odhady vycházejí z konkrétního souboru dat. Jde o odhad momentů:

**střední hodnota:** nevychýlený odhad

$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$

**rozptyl:** v případě známé (analyticky odvozené) střední hodnoty $\mu$, nevychýlený odhad

$D(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

$D(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2$ je vychýlený odhad, jeho vychýlenost roste snižováním počtem prvků $n$

$D(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))^2$ je nevychýlený odhad

## 1.2 Poznatek o rozloženích

- každé rozložení je vyjádřeno distribuční funkcí s volitelnými parametry, s distribuční funkcí souvisí i funkce hustoty
- z definice rozložení lze odvodit čtveřici momentů rozložení, které charakterizují vlastnosti rozložení
- lze odvodit vzorce pro převod momentů na parametry rozložení a naopak (z bodových odhadů získat parametry rozložení)
- ze souboru dat (posloupnosti čísel) s neznámým rozložením lze vypočítat bodový odhad středu a rozptylu

## 1.3 Rovnoměrné rozložení

- Základní rozložení, používá se pro generování ostatních pravděpodobnostních rozložení
- $f(x) = \frac{1}{b-a}$
- $F(x) = \frac{x-a}{b-a}$

```
var
  ix : real;
function Random : real ;
begin
  ix := ix * 1220703125;
  if (ix < 0) then ix := ix + maxint + 1;
  Random := ix / maxint;
end;
```

Je to *kongruentní generátor* s modulem *maxint* a multiplikátorem 1220703125. Je založen na číselné chybě při přetečení čísla. Předpokládá se určitá počáteční hodnota $ix$. Dalšími transformacemi takto získaných čísel lze generovat náhodná čísla jiných rozložení.

---

Inverzní transformace na libovolný interval:
$U = \frac{x-a}{b-a} \Rightarrow \quad U(b-a) = x-a \Rightarrow \quad x = U \cdot (b-a) + a$

---

**Momenty:**

$E(X) = \frac{a+b}{b}$

$D(X) = \frac{(b-a)^2}{12}$

## 1.4 Transformace a generování

Účelem transformace je převést náhodnou veličinu z jednoho pravděpodobnostního rozložení na jiné a tím generovat náhodná čísla. Nejčastěji se používá transformace z rovnoměrného rozložení, které je výpočetně nejjednodušší. Proto je základem transformace rychlý a především kvalitní generátor pseudonáhodných čísel rovnoměrného rozložení.

Transformačních technik je několik a její výběr závisí na typu cílového rozložení a požadavků na přesnost výpočtu:

- metoda inverzní transformace
- metoda vylučovací
- metoda kompoziční
- jiná metoda založená na znalosti rozložení, charakteristik

### 1.4.1 Metoda inverzní transformace

Tato metoda vychází z distribuční funkce $F$ cílového rozložení, k této odvodíme její inverzní podobu (funkci). Dosazením čísel $R \in< 0,1 >$ získáme náhodná čísla z cílového rozložení. V některých případech nelze odvodit inverzní funkci, pak volíme nějakou aproximaci.

#### Příklad s exponenciálním rozložením

$$F(x) = 1 - exp(-\frac{x-A}{b})$$

Po převodu:

$$U = 1 - exp(-\frac{x-A}{b})$$

$$X = A - b * lnU$$

### 1.4.2 Metoda vylučovací

U této metody se vychází z funkce hustoty cílového rozložení a opět z náhodné veličiny $R$.

Požadovaná funkce hustoty je: $f : (x_1, x_2) \to (0, M)$, kde $(x_1, x_2)$ znamená ohraničený definční obor funkce a $(0, M)$ je ohraničený obor hodnot funkce. Náhodnou veličinu $\xi$s funkcí hustoty $f$ budeme generovat takto:

1. vygenerujeme číslo $x$ s rozložením $R(x_1, x_2)$
2. vygenerujeme číslo $y$ s rozložením $R(0, M)$
3. je-li $y < f(x)$ , pak je $x$ náhodné číslo veličiny $\xi$, jinak postup opakujeme

Efektivnost metody je závislá na poměru plochy funkce hustoty $f$ a obdélníka $(x_1, 0, x_2, M)$

### 1.4.3 Metoda kompoziční

Kompoziční metoda se používá v případě, že definice rozložení je příliš složitá. Funkce hustoty rozložení se rozdělí na sadu jednodušších funkcí. Důvodem pro tento postup může být i usnadnění použití vylučovací metody.

$f(x) = \sum_{i=1}^{n} f_i$ rozklad funkce hustoty na podfunkce

## 1.5 Ukázky rozložení, použití

### 1.5.1 Exponenciální rozložení

Exponenciální rozložení se používá v modelování systémů hromadné obsluhy, jako rozložení dob obsluhy, časových intervalů mezi příchody požadavků a intervalů mezi poruchami. Je to nesymetrické rozložení.

### 1.5.2 Normální rozložení

Normální rozložení (též Gaussovo) je nejdůležitější rozložení spojité náhodné veličiny. Je použitelné všude, kde na změny náhodné veličiny má současný vliv větší množství nezávislých faktorů, které se uplatňují přibližně stejnou měrou.

$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

**generování:** nelze použít metodu inverzní transformace, protože nelze odvodit inverzní funkci k distribuční funkci (pouze její aproximaci). Vhodná je metoda vylučovací. Zpravidla se volí interval $(-3\sigma, 3\sigma)$. Metoda může být značně zdlouhavá, neboť poměr ploch je asi 6:1

**metoda sčítání náh. čísel:** speciální forma transformace z rozložení $R(0,1)$

$$X = \sigma\left(\sum_{i=1}^{12} U_i - 6\right) + \mu$$

Počet iterací (12) se volí s ohledem na efektivnost výpočtu, s rostoucím počtem iterací kvalita generátoru samozřejmě roste.

### 1.5.3 Pearsonovo rozložení $\chi^2$

Vychází z $m$ hodnot generovaných z $N(0,1)$.

$$\chi^2 = \sum_{i=1}^{m} Z_i^2$$

Pearsonovo rozložení se používá na testování statistických hypotéz, nejvíce při testu dobré shody. Počet iterací by neměl převýšit číslo 50, protože pak hodnota výstupu konverguje k číslu *jedna*.

### 1.5.4 Studentovo rozložení

Náhodná hodnota $t$ se Studentovým rozložením je dána transformačním vztahem

$$t = \frac{z}{\sqrt{\frac{\chi^2}{m}}}$$

kde $z$ je podle $N(0,1)$ a $\chi^2$ má $m$ stupňů volnosti. Toto rozložení se opět používá k testování statistických hypotéz a někdy i k dokonalejší aproximaci stupních dat místo normálního rozložení.

### 1.5.5 Weibullovo rozložení

Weibullovo rozložení je mírně atypické, podobá se normálnímu, je ale nesymetrické. Pro generování je použita metoda inverzní transformace:

$$W = (\frac{-ln(1-U)}{c})^{\frac{1}{P}} + \mu$$

Toto rozložení je velmi flexibilní, pro $p = 1$ přechází v exponenciální se středem $1/c$, pro $p = 2$ zase v Rayleighovo rozložení.

Využití je hlavně při modelování v teorii životnosti a spolehlivosti (intenzity poruch, kde neplatí exponenciální rozložení).

## 1.6 Testy rozložení

### 1.6.1 Test dobré shody (chí-kvadrát)

Test se používá k posouzení, jestli generátor náhodné veličiny odpovídá teoretickým vztahům. Lze též použít ke zjištění, jak určitý soubor dat odpovídá určitému rozložení.

**Algoritmus:**

1. Vygenerování souboru hodnot náhodné veličiny s daným rozložením ($n$ čísel)
2. Získání srovnávacího souboru se správnými parametry a tvarem
3. Srovnání obou souborů výpočtem hodnoty $\chi^2$
4. Zhodnocení výsledků testu podle předem zvolené hodnoty hladiny významnosti

**Bod 1**

Vygenerujeme velký soubor hodnot (řádově tisíce) náhodné veličiny. Vypočteme *histogram* souboru - graf četnosti. Histogram má $k$ položek, rozděluje obor hodnot rozložení na $k$ stejných dlouhých intervalů. *k-tá* složka histogramu říká, kolik čísel ze souboru padá do *k-tého* intervalu.

**Bod 2**

Kontrolní soubor získáme nejlépe z distribuční funkce rozložení. Opět vytváříme histogram:

$$n_j = n * (F(h_{j+1}) - F(h_j))$$

**Bod 3**

Srovnání obou histogramů se provádí podle algoritmu

$$\chi^2 = \sum_{j=1, n_j \geq 5}^{k} \frac{(N_j - n_j)^2}{n_j}$$

$N_j$ je kontrolovaný histogram, $n_j$ je teoretický histogram. Provádí se pouze iterace, kde $n_j$ je větší než pět (podmínka správnosti testu).

**Bod 4 - vyhodnocení**

Podle získaného čísla $\chi^2$ se určí koeficient volnosti $v = m - 1 - r$, kde $m$ je počet použitých vzorků histogramu a $r$ je počet volitelných parametrů daného rozložení. Hodnota hladiny významnosti $\alpha$ je číslo z generátoru *Pearsonova rozložení* s parametrem $v$. Často se používá k hodnocení koeficient zvaný *spolehlivost*:

$$S = 1 - \alpha$$

# 2 Zpracování stochastických dat

## 2.1 Stochastická data

Stochastická data jsou soubor čísel vyjadřujících určitý časový úsek stochastického (náhodného jevu). Příkladem může být obslužná linka s příchozími požadavky, kde interval mezi příchody je náhodný. To znamená, že příchody požadavků jsou náhodný (stochastický) jev. Budeme nějakou dobu měřit intervaly a dostaneme stochastický soubor dat.

Důvodem pro takový sběr dat je zařazení linky a procesu příchodu požadavků do modelu. Chceme analyticky vyjádřit pravděpodobností rozložení jevu příchodů.

## 2.2 První filtrace - ulítlé hodnoty

Při každém měření vznikají chyby (obsluha, měřící zařízení, záznam dat). Naměřený soubor dat pak může obsahovat *ulítlé hodnoty*, které nevhodně ovlivňují další zpracování.

- některé hodnoty nejsou fyzikálně možné (délka něčeho je záporná)
- některé jsou již na první pohled podezřelé
- zbytek může odstranit statistický test

### 2.2.1 Dixonův test

Dixonův test je statistický neparametrický test, který používá pro zhodnocení extrémích odchylek testovacích kritérií:

$$Q_{max} = (V_{max} - V_{max-1})/(V_{max} - V_{min})$$

$$Q_{min} = (V_{min+1} - V_{min})/(V_{max} - V_{min})$$

kde $V_{max}, V_{max-1}$ je největší a druhý největší prvek posloupnosti, $V_{min}, V_{min+1}$ je nejmenší a druhý nejmenší prvek posloupnosti. Účelem Dixonova testu je podle těchto extrémních prvků vyloučit nevhodné prvky posloupnosti, což jsou obvykle minimální a maximální hodnota. Kritérium pro vyloučení největšího (resp. nejmenšího) prvku je:

$$Q_{max} > Q(n, \alpha)$$

$$Q_{min} > Q(n, \alpha)$$

$n$ je počet prvků v posloupnsti a $\alpha$ je hladina významnosti. $Q$ je speciální tabulka s koeficienty. Po vyloučení prvku se test opakuje.

## 2.3 Postup aproximace

**Obecný postup:**

1. Výběr vhodných rozdělení
2. Odhad parametrů rozdělení
3. Testování spolehlivosti, ověření

**Bod 1 - výběr vhodných rozložení**

Pro řešení na počítačích je vhodná metoda navržená Pearsonem a Johnsonem. Tato metoda vychází z toho, že jednotlivým rozložením jsou jednoznačně přiřazeny hodnoty charakteristik $\beta_1$ a $\beta_2$, nebo vztahy mezi těmito hodnotami.

Postupuje se tak, že se provede bodový odhad charakteristiky šikmosti $B_1$ a špičatosti $B_2$. Ty se pak porovnají se známými hodnotami $\beta_1$ a $\beta_2$. Rozložení s parametry beta podobnými těm odhadnutým se zahrnou do dalšího zpracování. Rozdíl mezi $B_i$ a $\beta_1$ by neměl přesáhnout toleranční koeficient:
$Toler_{B_1} = \sqrt{\frac{6}{N}}$ $\qquad Toler_{B_2} = \sqrt{\frac{24}{N}}$
Tyto koeficienty jsou odvozeny z normálního rozložení a pro jiná rozložení přesnost roste s rostoucím $N$.

**Bod 2 - odhad parametrů rozložení**

Na to existuje několik metod, nejvhodnější se zdá být *momentová metoda.* Ta vychází z bodových odhadů momentů vstupního souboru a transformačních vzorců momenty na parametry.

```
void ExponentialParameters(MomentsArrayType Moments,
                        ParamArrayType Params)
{

        Params[1] = sqrtl(Moments[1]);
        Params[0] = Moments[0] - sqrtl(Moments[1]);
}
void ExponentialMoments(MomentsArrayType Moments,
                         ParamArrayType Params)
{

        Moments[0] = Params[0] + Params[1];
        Moments[1] = sqr(Params[1]);
        Moments[2] = 2.0;
        Moments[3] = 9.0;
}
```

## 2.4 Výsledky

Strojově vypočtená aproximace nemusí být vždy přesná, dokonce někdy je vybrané rozložení na první pohled úplně nevhodné (na normální se napasuje exponenciální). Závisí hodně na počtu vzorků v souboru.