Generarea unor variabile particulare

Curs 8

Validarea generatorilor

Se verifică faptul că algoritmul de generare produce valori de selecţie ale variabilei aleatoare care se doreşte a fi generată.

Acest lucru se poate face prin verificarea ipotezei statistice de concordanţă:

$$H : X \sim F(x)$$

pe baza unei selecţii $X_1, X_2, ..., X_n$ cu n suficient de mare, produse de algoritmul de generare.

Repartiţia empirică (sau de selecţie) ar trebui să se “asemene” cu repartiţia teoretică $F(x) \Rightarrow$ este necesară construcţia histogramei rezultate din valorile de selecţie $X_1, X_2, ..., X_n$.

Construcţia unei histograme

Pentru a construi o histogramă atunci când se cunosc valorile de selecţie $X_1, X_2, ..., X_n$ se aplică următorul algoritm:

Algoritm Histograma1

Intrare: $X_1, X_2, ..., X_n$;

P1: Se determină $m = \min\{X_1, X_2, ..., X_n\}$ şi $M = \max\{X_1, X_2, ..., X_n\}$, adică intervalul de variaţie $[m, M]$;

P2: Se alege k (se recomandă $15 \leq k \leq 40$) reprezentând numărul de intervale ale histogramei;

P3: Se împarte intervalul $[m, M]$ în k intervale $I_i = [a_{i-1}, a_i)$ de lungimi egale, cu $1 \leq i \leq k$, $a_0 = m$, $a_k = M$;

P4: Se determină f_i = nr. valorilor de selecţie din intervalul I_i, $1 \leq i \leq k$, adică frecvenţele absolute.

P5: Pentru $i = 1, 2, ..., k$ se determină frecvenţele relative $r_i = \frac{f_i}{n}$

Ieşire: Frecvenţele relative $r_1, r_2, ..., r_k$.

Prin urmare repartiţia empirică obţinută este:

$$\begin{pmatrix} I_1, & I_2, & ... & I_k \\ r_1 & r_2 & ... & r_k \end{pmatrix} \quad (1)$$

Folosind (10), o histogramă se poate reprezenta grafic astfel: se consideră în plan un sistem de axe $x0y$, pe axa Ox se marchează intervalele I_i, $1 \le i \le k$ şi se construiesc dreptunghiuri care au ca bază intervalele I_i şi ca înălţime frecvenţele relative r_i, $1 \le i \le k$.

Implementarea algoritmului Histograma1 creează anumite probleme pentru că determinarea valorilor m şi M şi apoi construirea intervalelor I_i şi calcularea frecvenţelor f_i implică memorarea selecţiei sau repetarea simulării selecţiei folosind aceeaşi sămânţă a generatorului, acest lucru însemnând consum de memorie sau de timp de calcul.

Idei pentru construcţia eficientă a histogramei:

- Generăm mai întâi $n_1 \ll n$ valori de selecţie $X_1, X_2, ..., X_{n_1}$ pe care le memorăm;
- Determinăm $a_1 = \min\{X_1, X_2, ..., X_{n_1}\}$, $a_{k-1} = \max\{X_1, X_2, ..., X_{n_1}\}$, $a_0 = a_1$, $a_k = a_{k-1}$;
- Se consideră
$$h = \frac{a_{k-1} - a_1}{k-2}$$
şi $a_i = a_{i-1} + h$, $i = 2, 3, ..., k-1$.
- Se iau $f_1 = 0$, $f_k = 0$ şi se determină $f_2, f_3, ..., f_{k-1}$ numărul de valori X_i cu $1 \leq i \leq n_1$ din intervalele $I_2, I_3 ..., I_{k-1}$.
- Se simulează celelalte $n - n_1$ valori de selecţie. Fie X o astfel de valoare, atunci:
    - Dacă $X < a_1$ atunci $a_0 = \min\{a_0, X\}$, $f_1 = f_1 + 1$;
    - Dacă $X > a_{k-1}$ atunci $a_k = \max\{a_k, X\}$, $f_k = f_k + 1$;

- Dacă $X \in [a_1, a_{k-1}]$ atunci

$$j := \left[\frac{X - a_1}{h}\right] + 2, \quad f_j := f_j + 1$$

Observăm că această construcţie produce $k-2$ intervale egale şi două intervale I_1 şi I_k de lungimi oarecare. Algoritmul rezultat astfel este:

Algoritm Histograma2
Intrare: n, n_1, k;
P1: Pentru $i = 1, 2, ..., k$ $f[j] := 0$;
P2: Pentru $i = 1, 2, ..., n_1$ se generează X, $x[i] := X$;
P3: Se determină $a[1] := \min\{x[1], x[2], ..., x[n_1]\}$, $a_{k-1} = \max\{x[1], x[2], ..., x[n_1]\}$, $h = (a[k-1] - a[1])/(k-2)$;
P4: Pentru $i = 1, 2, ..., n_1$, $j := trunc((x[i] - a[1])/h) + 2$, $f[j] := f[j] + 1$;
P5: $a[0] := a[1]$, $a[k] := a[k-1]$;

P6: Pentru $i = 1, 2, ..., n - n_1$ execută

- Se generează X;
- Dacă $a[1] \leq X \leq a[k-1]$ atunci $j = trunc\left(\frac{X - a[1]}{h}\right) + 2$, $f[j] := f[j] + 1$;
- Dacă $X < a[1]$ atunci $a[0] := \min\{a[0], X\}$, $f[1] := f[1] + 1$;
- Dacă $X > a[k-1]$ atunci $a[k] := \max\{a[k], X\}$, $f[k] := f[k] + 1$;

Ieşire: Frecvenţele $f[1], f[2], ..., f[k]$.

Validarea generatorilor

1. Validarea cu histogramă şi test de concordanţă

Construcţia histogramei (algoritm în Cursul 8) $\Rightarrow$ obţinerea frecvenţelor absolute $f_1, f_2, ..., f_k$ pentru fiecare interval de frecvenţă $[a_0, a_1)$, $[a_1, a_2)$,...,$[a_{k-a}, a_k)$. Pentru a valida generatorul se poate aplica testul de concordanţă χ^2 pentru verificarea ipotezei:

$$H : X \sim F(x). \tag{2}$$

- Se calculează probabilităţile:

$$p_1 = F(a_1)$$

$$p_i = F(a_i) - F(a_{i-1}), \quad 2 \leq i \leq k-2$$

$$p_k = 1 - F(a_{k-1})$$

- Se calculează
$$\sum_{j=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{3}$$
unde n reprezintă numărul de valori generate. S-a demonstrat că valoarea obţinută în (2) ar trebui să fie valoarea unei variabile aleatoare χ^2 cu $k-1$ grade de libertate.
- Fiind dat riscul de genul I, α (o probabilitate apropiată de 0), se determină din tabele valoarea $\chi^2_{k-1,\alpha}$ numită α-cuantilă superioară, astfel încât dacă X este o variabilă aleatoare χ^2 cu $k-1$ grade de libertate, atunci este verificată condiţia

$$P(X > \chi^2_{k-1,\alpha}) = \alpha$$

adică probabilitatea de a avea $X > \chi^2_{k-1,\alpha}$ este foarte mică.

- Ipoteza H se acceptă dacă valoarea dată de (2) este mai mică decât $\chi^2_{k-1,\alpha}$ (înseamnă că valoarea calculată poate fi valoarea unei variabile χ^2 cu $k-1$). Altfel, ipoteza H este respinsă.

2. Validarea cu medie şi dispersie

Presupunem că pentru variabila aleatoare X, ale cărei valori au fost generate printr-un algoritm de generare, sunt cunoscute media teoretică şi dispersia teoretică:

$$E[X] = m \quad Var[X] = \sigma^2$$

Cu ajutorul valorilor X_1, X_2,...,X_n obţinute din algoritmul de simulare se calculează media empirică de selecţie şi dispersia empirică de selecţie:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad s^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2.$$

Generatorul este considerat bun dacă pentru n destul de mare $(n > 1000)$ valorile mediei $\bar{X}$ şi ale dispersiei s^2 sunt apropiate de m şi σ^2 ca o consecinţă a *legii numerelor mari*.