

# Generarea unor variabile particulare

## Curs 8

December 9, 2020

# Validarea generatorilor

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

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

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

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

Repartiția empirică (sau de selecție) ar trebui să se “asemene” cu repartiția teoretică  $F(x) \Rightarrow$  este necesară construcția histogramei rezultate din valorile de selecție  $X_1, X_2, \dots, X_n$ .

## Construcția unei histograme

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

## Algoritm Histograma1

**Intrare:**  $X_1, X_2, \dots, X_n$ ;

P1: Se determină  $m = \min\{X_1, X_2, \dots, X_n\}$  și

$M = \max\{X_1, X_2, \dots, X_n\}$ , adică intervalul de variație  $[m, M]$ ;

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  $f_i = \text{nr. valorilor de selecție din intervalul } I_i$ ,  $1 \leq i \leq k$ , adică frecvențele absolute.

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

**Ieșire:** Frecvențele relative  $r_1, r_2, \dots, r_k$ .

Prin urmare repartiția empirică obținută este:

$$\begin{pmatrix} l_1 & l_2 & \dots & l_k \\ r_1 & r_2 & \dots & r_k \end{pmatrix} \quad (1)$$

Folosind (10), o histogramă se poate reprezenta grafic astfel: se consideră în plan un sistem de axe  $xOy$ , pe axa  $Ox$  se marchează intervalele  $l_i$ ,  $1 \leq i \leq k$  și se construiesc dreptunghiuri care au ca bază intervalele  $l_i$  și ca înălțime frecvențele relative  $r_i$ ,  $1 \leq i \leq k$ .

Implementarea algoritmului Histograma1 creează anumite probleme pentru că determinarea valorilor  $m$  și  $M$  și apoi construirea intervalelor  $l_i$  și calcularea frecvențelor  $f_i$  implică memorarea selecției sau repetarea simulării selecției folosind aceeași sămânță a generatorului, acest lucru însemnând consum de memorie sau de timp de calcul.

Idei pentru construcția eficientă a histogramei:

- ▶ Generăm mai întâi  $n_1 \ll n$  valori de selecție  $X_1, X_2, \dots, X_{n_1}$  pe care le memorăm;
- ▶ Determinăm  $a_1 = \min\{X_1, X_2, \dots, X_{n_1}\}$ ,  
 $a_{k-1} = \max\{X_1, X_2, \dots, X_{n_1}\}$ ,  $a_0 = a_1$ ,  $a_k = a_{k-1}$ ;

- ▶ Se consideră

$$h = \frac{a_{k-1} - a_1}{k - 2}$$

și  $a_i = a_{i-1} + h$ ,  $i = 2, 3, \dots, k - 1$ .

- ▶ Se iau  $f_1 = 0$ ,  $f_k = 0$  și se determină  $f_2, f_3, \dots, f_{k-1}$  numărul de valori  $X_i$  cu  $1 \leq i \leq n_1$  din intervalele  $I_2, I_3, \dots, I_{k-1}$ .
- ▶ Se simulează celelalte  $n - n_1$  valori de selecție. 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

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

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

## Algoritm Histograma2

**Intrare:**  $n, n_1, k$ ;

P1: Pentru  $i = 1, 2, \dots, k$   $f[j] := 0$ ;

P2: Pentru  $i = 1, 2, \dots, n_1$  se generează  $X$ ,  $x[i] := X$ ;

P3: Se determină  $a[1] := \min\{x[1], x[2], \dots, x[n_1]\}$ ,

$a_{k-1} = \max\{x[1], x[2], \dots, x[n_1]\}$ ,  $h = (a[k-1] - a[1]) / (k - 2)$ ;

P4: Pentru  $i = 1, 2, \dots, n_1$ ,  $j := \text{trunc}((x[i] - a[1]) / h) + 2$ ,  
 $f[j] := f[j] + 1$ ;

P5:  $a[0] := a[1]$ ,  $a[k] := a[k-1]$ ;

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

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

**leșire:** Frecvențele  $f[1], f[2], \dots, f[k]$ .

## Validarea generatorilor

### 1. Validarea cu histogramă și test de concordanță

Construcția histogramei  $\Rightarrow$  obținerea frecvențelor absolute  $f_1, f_2, \dots, f_k$  pentru fiecare interval de frecvență  $[a_0, a_1), [a_1, a_2), \dots, [a_{k-a}, a_k)$ . Pentru a valida generatorul se poate aplica testul de concordanță  $\chi^2$  pentru verificarea ipotezei:

$$H : X \sim F(x). \quad (2)$$

- Se calculează probabilitățile:

$$p_1 = F(a_1)$$



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

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

- Se calculează

$$\sum_{j=1}^k \frac{(f_j - np_j)^2}{np_j} \quad (3)$$

unde  $n$  reprezintă numărul de valori generate. S-a demonstrat că valoarea obținută în (2) ar trebui să fie valoarea unei variabile aleatoare  $\chi^2$  cu  $k-1$  grade de libertate.

- Fiind dat riscul de genul I,  $\alpha$  (o probabilitate apropiată de 0), se determină din tabele valoarea  $\chi_{k-1,\alpha}^2$  numită  $\alpha$ -cuantilă superioară, astfel încât dacă  $X$  este o variabilă aleatoare  $\chi^2$  cu  $k-1$  grade de libertate, atunci este verificată condiția

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

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

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

## 2. Validarea cu medie și dispersie

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

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

Cu ajutorul valorilor  $X_1, X_2, \dots, X_n$  obținute din algoritmul de simulare se calculează media empirică de selecție și dispersia empirică de selecție:

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

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