Universitatea din București
Facultatea de Matematică și Informatică

CURS nr. 7 – TEHNICI DE SIMULARE

Simularea unor variabile aleatoare discrete. Validarea generatorilor.

Lect. dr. Bianca Mogoș

Conținut

1.1 Repartiția Bernoulli(p), $p \in (0,1)$

- Fie un *eveniment observabil A* care are probabilitatea constantă $p = P(A) > 0$. Într-un experiment se poate produce A cu probabilitatea p sau evenimentul contrar A^C cu probabilitatea $q = 1 - p$. Un astfel de experiment se numește *probă Bernoulli*. Când se produce A spunem că s-a realizat un "succes", iar când A nu se produce spunem că avem un "eșec".
- Asociem unei probe Bernoulli variabila aleatoare Z astfel încât $Z = 1$ dacă se produce A și $Z = 0$ dacă se produce A^C. *Variabila Z are repartiția*

$$Z : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, E[Z] = p, V(Z) = pq = p(1-p). \quad (1)$$

Funcția de repartiție a lui Z este

$$F(x) = P(Z \le x) = \begin{cases} 0, & \text{dacă} \quad x < 0 \\ q, & \text{dacă} \quad 0 \le x < 1 \\ 1, & \text{dacă} \quad x \ge 1 \end{cases}. \quad (2)$$

1.2 Repartiția Binomială(n,p), $n \in \mathbb{N}, p \in (0,1)$ (1)

- Spunem că variabila discretă *$X \in \mathbb{N}$ este o variabilă binomială $Bin(n,p), n \in \mathbb{N}^+, 0 < p < 1$* dacă $X =$ numărul de succese în n probe Bernoulli independente, adică

$$X = \sum_{i=1}^{n} Z_i \quad (3)$$

unde Z_i sunt variabile identic și independent repartizate Bernoulli.
- *Repartiția variabilei binomiale X* este

$$X : \begin{pmatrix} 0 & 1 & 2 & \dots & x & \dots & n \\ q^n & npq^{n-1} & C_n^2 p^2 q^{n-2} & \dots & C_n^x p^x q^{n-x} & \dots & p^n \end{pmatrix}, \quad (4)$$

unde $q = 1 - p$.
- *Media și dispersia variabilei X* sunt date de formulele

$$E[X] = np \text{ și } Var(X) = npq. \quad (5)$$

1.2 Repartiția Binomială(n,p) – Algoritmi de simulare (2)

- *Simularea variabilei X* se poate realiza *direct*, prin numărarea de succese în n probe Bernoulli.
- Din *Teorema limită centrală* se deduce că pentru n suficient de mare $(n \to \infty)$ variabila

$$W_n = \frac{X - np}{\sqrt{npq}} \quad (6)$$

este repartizată $N(0,1)$.

1.3 Repartiția Geometrică(p), $p \in (0,1)$

- Variabila *X are repartiția $Geom(p), 0 < p < 1$* dacă $X =$ numărul de eșecuri până la apariția unui succes într-un șir oarecare de probe Bernoulli independente.
- *Repartiția variabilei $X \sim Geom(p)$* este

$$X : \begin{pmatrix} 0 & 1 & 2 & \dots & x & \dots \\ p & pq & pq^2 & \dots & P(X = x) = pq^x & \dots \end{pmatrix}, q = 1 - p. \quad (7)$$

- *Funcția de repartiție a variabilei X* se poate calcula cu formula

$$F(x) = P(X \le x) = \sum_{i=0}^{x} pq^i = 1 - q^{x+1}, x = 0, 1, 2, \dots \quad (8)$$

- *Media și dispersia variabilei X* sunt date de formulele

$$E[X] = \frac{q}{p} \text{ și } V(X) = \frac{q}{p^2}. \quad (9)$$

- *Interpretarea cu urnă:* $X =$ numărul de bile negre extrase cu întoarcere până când se obține o bilă albă.

1.4 Repartiția Pascal(k, p), $k \in \mathbb{N}, p \in (0,1)$

- Variabila *X are repartiția Pascal$(k,p), k \in \mathbb{N}^+, 0 < p < 1$* dacă $X =$ numărul de eșecuri până la apariția a k succese într-un șir oarecare de probe Bernoulli independente.
- *Repartiția variabilei $X \sim Pascal(k,p)$*, este

$$X : \begin{pmatrix} 0 & 1 & 2 & \dots & x & \dots \\ p^k & kp^kq & C_{k+1}^{k-1}p^kq^2 & \dots & C_{x+k-1}^{k-1}p^kq^x & \dots \end{pmatrix},$$

$$q = 1 - p, x = 0, 1, 2, \dots. \tag{10}$$

- *Media și dispersia variabilei X* sunt date de formulele

$$E[X] = \frac{kq}{p} \text{ și } V(X) = \frac{kq}{p^2}. \tag{11}$$

- *Interpretarea cu urnă:* X numărul de bile negre extrase cu întoarcere până când se obțin k bile albe.

1.5 Repartiția Hipergeometrică(N, p, n), $n, N \in \mathbb{N}, n < N, p \in (0,1)$ (1)

- Considerăm *experimentul cu urna*: se extrag n bile la întâmplare din urnă *fără întoarcere*.
- Notăm cu u evenimentul: s-a extras o bilă albă și cu v evenimentul: s-a extras o bilă neagră.
- *Probabilitățile de a extrage în prima extragere* o bilă albă, respectiv neagră sunt:

$$p = P(u) = A/N \text{ și } P(v) = B/N \tag{12}$$

unde A, B reprezintă numărul de bile albe, respectiv negre extrase din urnă, iar N numărul total de bile.

- *Probabilitățile de extragere* a unei bile albe sau negre *în a doua extragere* sunt condiționate de rezultatele primei extrageri:

$$\begin{aligned} P(u\,|u) &= \frac{A-1}{N-1}, P(u\,|v) = \frac{A}{N-1}, \\ P(v\,|u) &= \frac{B}{N-1}, P(v\,|v) = \frac{B-1}{N-1} \end{aligned} \tag{13}$$

1.5 Repartiția Hipergeometrică(N, p, n) (2)

- Se definește v.a. $X =$ numărul de bile albe extrase. Spunem că *$X \sim H(N, p, n)$*.
- Rezultă $A = round(Np)$, $B = N - A$.
- *Probabilitatea* ca în n extrageri succesive fără întoarcere, *să se extragă "a" bile albe* este:

$$P(X = a) = \frac{C_A^a C_B^{n-a}}{C_N^n}, 0 \le a \le n, n < N \tag{14}$$

- *Media și dispersia v.a. X* sunt date de

$$E[X] = np \text{ și } Var(X) = np(1-p)\frac{N-n}{N-1} \tag{15}$$

1.6 Repartiția Poisson(λ), $\lambda > 0$ (1)

- O variabilă aleatoare *X este repartizată Poisson(λ), $\lambda > 0$* dacă are funcția de probabilitate dată prin

$$P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}, \lambda > 0 \tag{16}$$

- *Media și dispersia variabilei aleatoare $X \sim Poisson(\lambda)$* sunt

$$E[X] = \lambda \text{ și } V(X) = \lambda. \tag{17}$$

- Repartiția Poisson poate fi utilizată în numeroase aplicații. Câteva situații în care o variabilă aleatoare discretă poate avea o distribuție Poisson sunt:
  - numărul erorilor de tipografie dintr-o pagină;
  - numărul concediilor dintr-o firmă în decursul unei luni
  - numărul defectelor de-a lungul unui fir.

1.6 Repartiția Poisson(λ) – Algoritm de simulare (2)

- *Repartiția Poisson* poate fi dedusă din *repartiția binomială*. Pentru valori mari ale lui n și mici ale lui p (deci valori moderate pentru np), numărul de succese apărute în n probe poate fi aproximat de variabila aleatoare Poisson cu parametrul $\lambda = np$.
- *Algoritmul de simulare a v.a. $X \sim Poisson(\lambda)$*:
  Pas 1: Se alege o probabilitate $p \approx 0$ (de exemplu, $p = 0.001$)
  Pas 2: Se determină $n = [\lambda/p]$
  Pas 3: Se simulează $Y \sim Bin(n, p)$
  Pas 4: Se returnează $X = Y$.

2 Validarea algoritmilor de simulare a unor variabile aleatoare (1)

- *Algoritm de simulare*: definirea unei v.a. X având o funcție de repartiție dată $F(x) = P(X \le x), \forall x \in \mathbb{R}$ și observarea variabilei X
- *Validarea algoritmilor de simulare* înseamnă:
  - verificarea corectitudinii formale a algoritmilor: se arată că v.a. X construită în algoritm are funcția de repartiție $F(x)$
  - analiza valorilor de selecție asupra v.a. X returnate de algoritm: pe baza unei mulțimi de selecție $X_1, X_2, \dots, X_n$, se verifică ipoteza statistică $H_0 : X \hookrightarrow F(x)$.

2.1 Histograma – o validare empirică a algoritmului de simulare (1)

- Se verifică intuitiv dacă *repartiția empirică (de selecție)* este asemănătoare cu cea *teoretică*
- *Histograma* asociată mulțimii de valori de selecție $x_1, \ldots, x_n$ asupra variabilei aleatoare X având funcția de repartiție $F(x)$ și densitatea $f(x)$:
  - se determină $m = \min\{x_1, x_2, \ldots, x_n\}$ și $M = \max\{x_1, x_2, \ldots, x_n\}$
  - se alege k numărul de intervale/dreptunghiuri ale histogramei
  - se împarte intervalul $[m, M]$ în k intervale egale $I_i = (a_{i-1}, a_i]$, $2 \leq i \leq k$, $I_1 = [a_0, a_1]$, $a_0 = m, a_k = M$
  - se determină frecvențele relative $r_i = \frac{f_i}{n}$, unde f_i : numărul de valori de selecție ce cad în intervalul I_i, $1 \leq i \leq k$
  - se reprezintă grafic: se iau pe abscisă intervalele I_i și se construiesc dreptunghiurile având ca bază aceste intervale și ca înălțimi $h_i = r_i$.
- *Înălțimile h_i ale dreptunghiurilor se scalează a.î. $h_i \approx f(x), x \in I_i$. Definim $h_i = \frac{f_i}{nl} \approx f(x), l = a_i - a_{i-1}$.*

2.1 Histograma – o validare empirică a algoritmului de simulare (2)

- Din *Teorema lui Bernoulli – forma slabă* rezultă că pentru orice $\epsilon > 0$ avem

$$\lim_{n\to\infty} P\left(\left|\frac{f_i}{n} - p_i\right| \leq \epsilon\right) = 1, \tag{18}$$

unde $p_i = P(X \in I_i)$, $I_i = a_{i-1}, a_i$
- Numărul k de dreptunghiuri ale histogramei se consideră a.î. să se minimizeze *"media pătratelor erorilor (MSE)"* estimatorului $\hat{f}(x)$ (definit în orice punct x) al densității de repartiție $f(x)$, definită prin

$$MSE(\hat{f}(x)) = E\left[\left(\hat{f}(x) - f(x)\right)^2\right]. \tag{19}$$

Astfel, avem *Regula Sturges*

$$k = [1 + \log_2 n]. \tag{20}$$

2.2 Test bazat pe momentele de selecție

- Se determină *momentele teoretice* ale v.a. X:

$$\mu = E[X] \text{ și } \sigma = Var(X) \tag{21}$$

- Pe baza mulțimii de valori de selecție $\{x_1, x_2, \ldots, x_n\}$ se calculează *momentele de selecție*:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ și } s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 \tag{22}$$

- Ca o consecință a *Legii numerelor mari*, putem considera că generatorul este bun dacă pentru n suficient de mare ($n > 1000$)

$$\overline{X} \approx \mu \text{ și } s^2 \approx \sigma^2 \tag{23}$$

2.3 Testul X^2 (1)

- Considerăm *testul de concordanță X^2* pentru verificarea ipotezei *$H_0 : X \hookrightarrow F(x)$*.
- Definim *variabila aleatoare X^2*:

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{24}$$

unde $p_1 = F(a_1), p_i = F(a_i) - F(a_{i-1}), 2 \leq i \leq k-1, p_k = 1 - F(a_{k-1})$.
- *Observații*:
  - $f_i : \Omega \to \{0, 1, \ldots, n\}$ este o v.a. $Bin(n, p_i)$
  - Mulțimea tuturor valorilor posibile ale lui X^2 se obține făcând ca $f_1, f_2, \ldots, f_k$ să parcurgă toți întregii nenegativi a.î. $\sum_{i=1}^{k} f_i = n$
  - Pentru $n \to \infty$, X^2 este repartizată χ^2_{k-1} (hi pătrat cu $k-1$ grade de libertate).

2.3 Testul X^2 (2)

- Fie *α eroarea de tip I*: probabilitatea de a respinge ipoteza nulă H_0 când este adevărată. Valorile clasice pentru α sunt 0.01, 0.05, 0.1.
- Se determină *α – cuantila superioară*, notată $\chi^2_{k-1,\alpha}$ astfel încât

$$P\left(X^2 \leq \chi^2_{k-1,\alpha}\right) = 1 - \alpha \tag{25}$$

- Ipoteza *H_0 se acceptă* dacă în urma experimentului aleator s-a obținut evenimentul $\omega \in \Omega$ a.î.:

$$X^2(\omega) \leq \chi^2_{k-1,\alpha}, \tag{26}$$

în caz contrar se respinge.

Bibliografie I

- W. L. Martinez, A. R. Martinez (2002), *Computational Statistics Handbook with MATLAB*, Chapman & Hall/CRC, Boca Raton London New York Washington, D.C.
- I. Văduva (2004), *Modele de simulare: note de curs*, Editura Universității din București, București