

CURS nr. 2 – TEHNICI DE SIMULARE

1. Recapitulare: **noțiuni de statistică**
2. Metode generale de simulare a v.a.:
metoda inversă

Lect. dr. Bianca Mogoș

◀ ▶ ↻ 🔍

Scopul experimentului aleator

- ▶ *Experimentul aleator* se realizează pentru *colectarea de date* necesară pentru a obține informații privind un anumit fenomen de interes.
- ▶ Pe baza datelor se emit *concluzii* care, în general, ies din sfera experimentului particular.
- ▶ Cercetătorii *generalizează concluziile experimentului* pentru clasa tuturor experimentelor similare.
- ▶ Problema acestui demers este că *nu putem garanta corectitudinea concluziilor* obținute.
- ▶ Totuși, folosind tehnici statistice, putem măsura și administra *gradul de incertitudine* al rezultatelor.

◀ ▶ ↻ 🔍

Populație țintă

- ▶ *Populația țintă* este definită ca fiind întreaga colecție de obiecte sau indivizi despre care vrem să obținem anumite informații.
- ▶ Populația țintă trebuie *bine definită* indicând
 - ▶ ce constituie *membrii* acesteia (de ex, populația unei zone geografice, o anumită firmă care construiește componente hardware, etc.)
 - ▶ *caracteristicile* populației (de ex, starea de sănătate, numărul de defecțiuni, etc.).
- ▶ În majoritatea cazurilor este imposibil sau nerealist să observăm întreaga populație; cercetătorii măsoară numai o parte a populației țintă, denumită *eșantion*.
- ▶ Pentru a face inferențe privind întreaga populație este important ca *mulțimea eşantion* să fie *reprezentativă* relativ la întreaga populație.

◀ ▶ ↻ 🔍

Conținut

Partea I

- ▶ Scopul experimentului aleator. Statistică inferențială.
 - ▶ Populație țintă. Eșantion
 - ▶ Model probabilist. Selecție. Repartiția selecției
 - ▶ Model statistic. Statistică.
 - ▶ Convergența în probabilitate
 - ▶ Estimator. Estimație. Consistența unui estimator
- Partea a II - a*
- ▶ Numere aleatoare. Variabile uniforme
 - ▶ Metode generale de simulare a variabilelor aleatoare: Metoda inversă

◀ ▶ ↻ 🔍

Statistică inferențială

- ▶ *Statistica inferențială* este o colecție de *metode* care permit cercetătorilor să observe o submulțime a obiectelor de interes și folosind informația obținută pe baza acestor observații să facă afirmații sau inferențe privind întreaga populație.
- ▶ Câteva dintre aceste *metode* sunt:
 - ▶ estimarea parametrilor unei populații
 - ▶ verificarea ipotezelor statistice
 - ▶ estimarea densității de probabilitate.

◀ ▶ ↻ 🔍

Model probabilist

- ▶ Fie X o v.a. cu densitatea $f(x, \theta)$, $x \in \mathbb{R}$, $\theta \in \mathbb{R}$.

Definiție

Mulțimea densităților de repartiție $f(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}$, ce depind de parametrul θ se numește *model probabilist unidimensional*.

$$\{f(x, \theta) | x \in \mathbb{R}, \theta \in \Theta\}. \quad (1)$$

- ▶ Fie $X = (X_1, X_2, \dots, X_n)$ un vector aleator cu densitatea de repartiție

$$f(x_1, x_2, \dots, x_n; \theta), (x_1, x_2, \dots, x_n) \in \mathbb{R}^n, \theta \in \mathbb{R}^k.$$

Definiție

Mulțimea densităților de repartiție $f(x_1, x_2, \dots, x_n; \theta)$ cu parametrul $\theta \in \Theta \subset \mathbb{R}^k$ se numește *model probabilist multidimensional*.

$$\{f(x_1, x_2, \dots, x_n; \theta) | \theta \in \Theta\}. \quad (2)$$

◀ ▶ ↻ 🔍

Selecție. Repartiția selecției

Definiție

O *selecție* este o mulțime de v.a. X_1, X_2, \dots, X_n având aceeași densitate de repartiție $f(x, \theta)$.

- Deoarece selecția este o mulțime de variabile aleatoare asociate unui model probabilist, selecția trebuie să aibă o repartiție, pe care o vom numi *repartiția selecției*.

Definiție

Repartiția selecției X_1, X_2, \dots, X_n este definită ca fiind repartiția vectorului $X = (X_1, X_2, \dots, X_n)$, notată prin $f(x_1, x_2, \dots, x_n; \theta)$.

- Cea mai folosită formă de selecție este selecția aleatoare și este bazată pe ideea experimentului aleator.

Model statistic. Statistică

Definiție

Modelul probabilist

$\{f(x; \theta), \theta \in \Theta\}$

împreună cu selecția $X = (X_1, X_2, \dots, X_n)$ definesc *modelul statistic*.

Definiție

Statistica este o funcție $t_h : S \rightarrow \Theta \subset \mathbb{R}^k$ care nu conține niciun parametru necunoscut.

- ▶ Cele mai utilizate statistici sunt momentele de selecție:

Convergență în probabilitate

Definiție

Șirul de v.a. $(X_n)_n$ converge în probabilitate la v.a. X dacă

$$\lim_{n \rightarrow \infty} P(\{\omega \in \Omega, |X_n(\omega) - X(\omega)| < \epsilon\}) = 1. \quad (7)$$

Propoziție

Avem relativa

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j \xrightarrow[n \rightarrow \infty]{P} E[X] = \mu$$

Selecție aleatoare

Definiție

Spunem că X_1, X_2, \dots, X_n este o *selectie aleatoare* asupra v.a. X care are densitatea de repartiție $f(x; \theta)$ dacă X_1, X_2, \dots, X_n sunt v.a. independente și identic repartizate ca X .

X_1, X_2, \dots, X_n se numesc *variabile de selecție*.

- ▶ În cazul selecției aleatoare, densitatea de repartiție comună a variabilelor de selecție este
$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$$
- ▶ O selecție aleatoare poate fi construită prin repetarea unui experiment aleator de n ori.
- ▶ Un *rezultat al selecției aleatoare* se notează prin (x_1, x_2, \dots, x_n) și mulțimea tuturor rezultatelor definesc *spațiul observațiilor* $S \subseteq \mathbb{R}^n$.

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$$

Momente de selecție

- ▶ *Momentul de selecție de ordin 1* – media de selecție
- (3)
$$m'_r(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n x_j^r$$
- ▶ *Momentul de selecție de ordin 1 – media de selecție*
- (4)
$$m'_1(x_1, x_2, \dots, x_n) = \bar{X}_n = \frac{1}{n} \sum_{j=1}^n x_j$$
- ▶ *Momentul centrat de selecție de ordin r*
- (5)
$$m_r(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{X}_n)^r.$$
- ▶ *Momentul centrat de selecție de ordin 2 – dispersia de selecție*
- (6)
$$m_2(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{X}_n)^2.$$

$$m'_r(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n x_j^r \quad (3)$$

$$m'_1(x_1, x_2, \dots, x_n) = \overline{X}_n = \frac{1}{n} \sum_{j=1}^n x_j \quad (4)$$

$$m_r(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{X}_n)^r. \quad (5)$$

$$m_2(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{X}_n)^2. \quad (6)$$

Estimator. Estimatie.

Definiție

Se numește *estimator*, variabila aleatoare

$$t_n(X) : \Omega \rightarrow \Theta, \quad (8)$$

unde $X = (X_1, X_2, \dots, X_n)$, Ω este spațiul de selecție; $t_n(x)$, $x \in S$,
 S spațiul observațiilor se numeste *estimatie*.

Definiție

Un estimator $t_n = t_n(X)$ se numește *consistent* pentru θ dacă

$$\lim_{n \rightarrow \infty} P(|t_n - \theta| < \epsilon) = 1 \quad (9)$$

și notăm $t_n \xrightarrow{P} \theta$.

Consistența unui estimator

- Consistența unui estimator reprezintă o proprietate asimptotică a estimatorului.
- Un estimator bun pentru parametrul θ trebuie să aibă o repartiție cu o valoare centrală în vecinătatea lui θ .
- Definiția următoare cere ca estimatorul să aibă o valoare centrală în vecinătatea lui θ nu numai pentru valori mari ale lui n , ci pentru orice n .

Definiție

Estimatorul t_n se numește nedepășat pentru θ dacă

$$E[t_n] = \theta. \quad (10)$$

Generarea numerelor uniforme în Matlab (1)

- ▶ În Matlab există funcția **rand** pentru generarea variabilelor aleatoare uniforme.
- ▶ **rand(n)**, unde n este un număr natural, returnează o matrice de dimensiune $n \times n$ având ca elemente numere aleatoare uniform distribuite între 0 și 1.
- ▶ **rand(m, n)**, unde m, n sunt numere naturale, returnează o matrice de dimensiune $m \times n$ având ca elemente numere aleatoare uniform distribuite între 0 și 1.

Generarea numerelor uniforme în Matlab (3)

Pentru a genera numere aleatoare uniform distribuite pe un interval (a,b) , și scriem $X \sim \mathcal{U}(a,b)$, pornind de la un număr generat uniform pe intervalul $(0,1)$ se poate folosi transformarea

$$X = (b - a) \cdot U + a, \quad (11)$$

unde $U \sim \mathcal{U}(0, 1)$

Numere aleatoare

- Majoritatea metodelor de generare a v.a. se bazează pe generarea unor numere aleatoare uniform distribuite pe intervalul $(0,1)$.
 - Datorită calculatorului avem posibilitatea de a genera foarte ușor numere aleatoare uniforme.
 - Totuși, trebuie cunoscut faptul că numerele generate de calculator sunt pseudo-aleatoare, deoarece acestea sunt generate cu un algoritm determinist.
- ### Generarea numerelor uniforme în Matlab (2)
- O secvență de numere aleatoare generată în Matlab depinde de "sămânța" sau starea generatorului. Starea este resetată la valoarea implicită în momentul pornirii Matlab-ului, astfel aceeași secvență de variabile aleatoare este generată la o nouă pornire a Matlab-ului. Acesta poate fi un avantaj în situațiile în care analistul are nevoie să reproducă rezultatele unei simulări pentru a verifica anumite concluzii.
 - Folosind sintaxa **rand('state',0)** Matlab-ul resetează generatorul la starea inițială.
 - Se folosește sintaxa **rand('state',j)** pentru a seta generatorul la starea j .
 - Pentru a obține vectorul de stări se apelează **S = rand('state')**, S va reprezenta vectorul conținând cele 35 de stări posibile.

Generarea numerelor uniforme în Matlab (4)

Exemplu care ilustrează utilizarea funcției **rand**.

```
% Generam un vector de numere aleatoare pe intervalul (0,1):
```

```
x = rand(1,1000);
```

% Histograma eșantionului generat în x.

[N,X] = hist(x,15);

% x: mulțimea eșantion

% 15 reprezintă numărul de dreptunghiuri ale histogramei

% N: vector conținând numărul de elemente din fiecare dintre dreptunghiuri.

% X: vector conținând centrele dreptunghiurilor

% Folosirea funcției bar pentru reprezentarea grafică a histogramei.

$$\text{bar}(X, N, 1, w')$$

title('Histogram')

xlabel('X')

ylabel('Frecventa absoluta')

Histograma rezultată este prezentată în Figura 1.

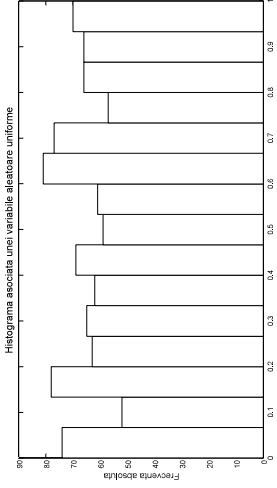


Figura: 1 – Histograma asociată unui eșantion de numere aleatoare uniform distribuite

Funcții de o variabilă aleatoare (2): Exemlu 1

Propoziție

Dacă $X \sim N(m, \sigma^2)$ atunci $Y = e^X$ are densitatea de repartiție:

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y - m)^2}{2\sigma^2}}, y > 0 \quad (13)$$

Demonstrație

$$\begin{aligned} F_Y(y) &= P(\{\omega \in \Omega \mid Y(\omega) \leq y\}) = P(e^X \leq y) = P(X \leq \ln y) = \\ &= F_X(\ln y) = \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\ln y} e^{-\frac{(t-m)^2}{2\sigma^2}} dt. \end{aligned}$$

Definiție

Variabila aleatoare Y având densitatea de probabilitate (13) se numește lognormală.

Metoda inversă (1): Teorema lui Hincin

Propoziție

Fie X o variabilă aleatoare (v.a.) cu funcția de repartiție inversabilă $F(x)$. Atunci variabila aleatoare $Y = F(X)$ este repartizată uniform pe $[0, 1]$.

Teoremă (Teorema lui Hincin)

Fie $U \sim \mathcal{U}([0, 1])$ și $F(x)$ o funcție de repartiție inversabilă, de tip continuu. Atunci variabila aleatoare

$$X = F^{-1}(U) \quad (14)$$

este o variabilă aleatoare continuă cu funcția de repartiție $F(x)$.

Demonstrație: Funcția de repartiție a v.a. X este

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x))$$

deoarece F este strict crescătoare.

Cum $U \sim \mathcal{U}(0, 1)$ rezultă $F_U(u) = u$ pentru $u \in (0, 1)$. Obținem astfel $P(X \leq x) = F(x)$.

- ▶ Fie X o v.a. $X : \Omega \rightarrow \mathbb{R}$ care ia valori în $D \subset \mathbb{R}$ și $\phi : D \rightarrow \mathbb{R}$, ϕ continuă.
- ▶ Atunci v.a. $Y = \phi(X) : \Omega \rightarrow \phi(D) \subset \mathbb{R}$ are repartiția dată de $P(\{\omega \in \Omega \mid Y(\omega) \in A\}) = P(\{\omega \in \Omega \mid X(\omega) \in \phi^{-1}(A)\})$ (12) pentru orice $A \subset \phi(D)$ și $\phi^{-1}(A) = \{z \mid \phi(z) \in A\}$.

Funcții de o variabilă aleatoare (3): Exemlu 2

Propoziție

Dacă $X \sim U(0, 1)$ atunci variabila aleatoare

$$Y = \left(-\frac{1}{a} \ln(1 - X) \right)^{1/b}$$

cu $a, b > 0$ are densitatea de probabilitate

$$f(y) = aby^{b-1} e^{-ay^b}, \quad 0 < y < \infty.$$

Definiție

O variabilă aleatoare cu densitatea de probabilitate (0.15) se numește Weibull cu parametrii a și b și este notată $W(a, b)$.

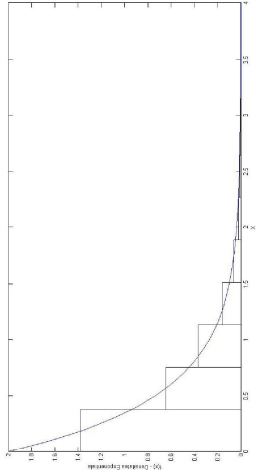
Metoda inversă (2): Descrierea metodei

- ▶ este introdusă ca o consecință directă a teoremei lui Hincin
- ▶ se aplică în cazul în care funcția de repartiție se poate inversa ușor
- ▶ dacă am putea produce valorile de selecție u_1, u_2, \dots, u_n asupra v.a. $U \sim \mathcal{U}(0, 1)$ și am cunoaște funcția de repartiție F a variabilei X atunci am putea produce valorile de selecție x_1, x_2, \dots, x_n asupra lui X cu formula $x_i = F^{-1}(u_i), 1 \leq i \leq n$

Metoda inversă (3): Algoritm pentru simularea unor variabile aleatoare continue

Intrare	$F(x)$: funcția de repartiție a variabilei X pe care ne propunem să o simulăm
Pas 1	Se generează o valoare de selecție u uniformă pe $[0,1]$
Pas 2	Se determină expresia inversei funcției de repartiție $F^{-1}(u)$
Pas 3	Se obține valoarea de selecție dorită $x = F^{-1}(u)$
Ieșire	Valoarea de selecție, x , a v.a. X

Metoda inversă (5): Histograma asociată variabilei aleatoare exponențiale



Metoda inversă (7): Algoritm pentru simularea unor variabile aleatoare discrete

- ▶ Regula de generare a unei valori de selecție asupra v.a. X :
$$X = x_i \quad \text{dacă } F(x_{i-1}) < u \leq F(x_i) \text{ și } x_0 < x_1. \quad (17)$$
- ▶ Algoritmul pentru simularea v.a. X :

Intrare	Repartiția variabilei X $P(X = x_i) = p_i, \sum_{i=1}^m p_i = 1, x_1 < x_2 < \dots < x_m.$
Pas 1	Se generează o valoare de selecție u uniformă pe $[0,1]$
Pas 2	Dacă $u \leq p_1$ atunci $x = x_1$ Altfel dacă $u \leq p_1 + p_2$ atunci $x = x_2$ Altfel dacă $u \leq p_1 + p_2 + p_3$ atunci $x = x_3$...
Ieșire	Altfel dacă $u \leq p_1 + p_2 + \dots + p_m$ atunci $x = x_m$ Valoarea de selecție, x , a v.a. X

Metoda inversă (4): Variabile aleatoare continue care pot fi simulate folosind metoda inversă

Repartiția	Densitatea f	Inversa F^{-1}
$Exp(\lambda), \lambda > 0$	$f(x) = \lambda e^{-\lambda x}, x > 0$	$x = -\frac{1}{\lambda} \ln(u)$
$Weib(0, 1, \nu), \nu > 0$	$f(x) = \nu x^{\nu-1} e^{-x^\nu}$	$x = (-\ln(u))^{1/\nu}$
<i>Cauchy</i>	$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, x \in \mathbb{R}$	$x = \tan \pi(u - 1/2)$
<i>Arcsin</i>	$f(x) = \frac{1}{\pi} \frac{1}{\sqrt{1-x^2}}, x \in [-1, 1]$	$x = \sin \pi(u - 1/2)$

Metoda inversă (6): Simularea unei variabile aleatoare discrete

- ▶ Fie v.a. discretă X definită prin repartiția

$$X : \left(\begin{array}{cccc} x_1 & x_2 & \dots & x_m \\ p_1 & p_2 & \dots & p_m \end{array} \right), \sum_{i=1}^m p_i = 1, x_1 < x_2 < \dots < x_m. \quad (15)$$

- ▶ Funcția de repartiție a v.a. X este dată de

$$F(x) = \begin{cases} 0 & \text{dacă } x < x_1 \\ p_1 & \text{dacă } x_1 \leq x < x_2 \\ p_1 + p_2 & \text{dacă } x_2 \leq x < x_3 \\ \dots & \dots \\ p_1 + p_2 + \dots + p_k & \text{dacă } x_k \leq x < x_{k+1} \\ \dots & \dots \\ 1 & \text{dacă } x \geq x_m \end{cases} \quad (16)$$

Metoda inversă (8): Exemplu simularea unei v.a. discrete

- ▶ Vrem să generăm o v.a. discretă X cu repartiția
$$X : \left(\begin{array}{ccc} 0 & 1 & 2 \\ 0.3 & 0.2 & 0.5 \end{array} \right) \quad (18)$$
- ▶ Funcția de repartiție este dată de
$$F(x) = \begin{cases} 0 & \text{dacă } x < 0 \\ 0.3 & \text{dacă } 0 \leq x < 1 \\ 0.5 & \text{dacă } 1 \leq x < 2 \\ 1 & \text{dacă } x \geq 2 \end{cases} \quad (19)$$
- ▶ Se generează valori de selecție asupra v.a. X conform regulilor
$$X = \begin{cases} 0 & U \leq 0.3 \\ 1 & 0.3 < U \leq 0.5 \\ 2 & 0.5 < U \leq 1 \end{cases} \quad (20)$$
- ▶ Dacă v.a. $u = 0.78$ atunci obținem valoarea de selecție $x = 2$.

Bibliografie I

- M. Craiu (1998), *Statistică matematică: teorie și probleme*, Editura Matrix Rom, București
- 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