

Laborator 8

Lab 4

§3. Funcția de repartitie pt. v.a.

From where we stand, the rain seems random.

If we could stand somewhere else, we would see the order in it.

— T. Hillerman (1990) *Coyote Waits.*

- ?) În unele probleme concrete disponem numai de observații dintr-o anumită distribuție (necunoscută), fără a avea acces la distribuția exactă. Cum procedăm pt. studierea repartitiei datelor?
- ?) IDEE: folosin frevențe relative în locul probabilităților corespondătoare x_i , astfel, putem approxima funcția de repartitie (teoretică) cu funcția de repartitie EMPIRICĂ (`ecdf()` în R).

Definiție!

Fie x_1, \dots, x_n un eșantion de talie n dintr-o populație necunoscută \mathcal{F} . Se numește funcție de repartitie EMPIRICĂ asociată eșantionului x_1, x_2, \dots, x_n funcția definită astfel:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i),$$

unde $I_A(z) = \begin{cases} 1, & \text{dacă } z \in A \\ 0, & \text{dacă } z \notin A \end{cases}$ este funcția indicatoare a mulțimii A .

Remarcă (teoretic vs empiric):

teoretic, precis: $F(x) = P(X \leq x)$

practic, empiric: $F(x) \approx \hat{F}_m(x) = \frac{\text{nr. observații} \leq x}{\text{nr. total de obs.}} = \frac{\sum_{i=1}^m \text{obs. } \leq x}{m}$

§2. Generarea unei v.a. discrete

Problema: Dăm și generăm un eveniment de talie n dintr-o distribuție discretă cu multimea valorilor $\{x_1, \dots, x_N\}$ și având probabilități corespondătoare $\{p_1, \dots, p_N\}$.

Soluție:

$$X: \begin{pmatrix} x_1 & \dots & x_N \\ p_1 & \dots & p_N \end{pmatrix}$$

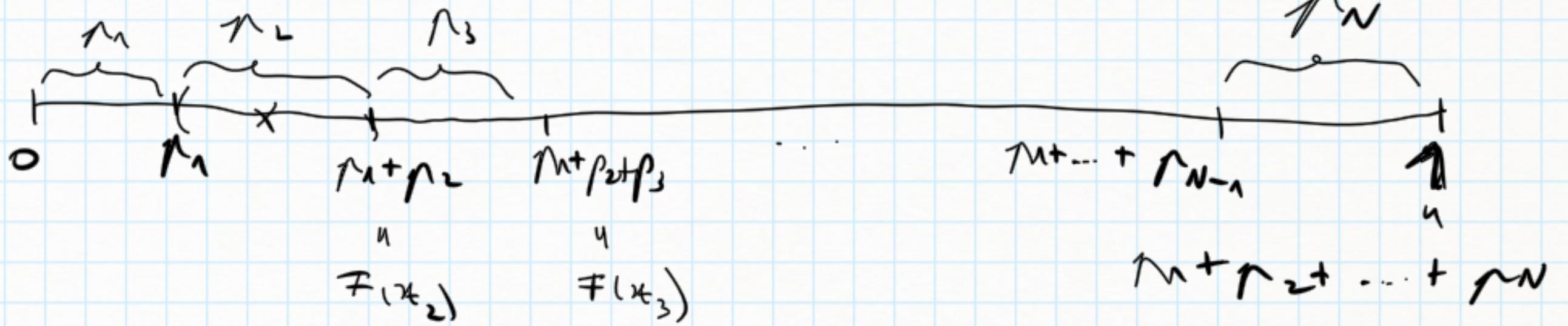
• Considerăm o v.a. auxiliară $U \sim \text{Unif}(0,1)$.

• Împărțim intervalul $(0,1)$ în N subintervale astfel: #cumsum (-)

→ primul subinterval are lungimea p_1

→ al doilea subinterval are lungimea p_2

→ al N -lea subinterval are lungimea p_N



• Observăm că:

$$P(X = x_i) = p_i, \quad \forall i = \overline{1, N}$$

$p_i = P(p_1 + \dots + p_{i-1} \leq U < p_1 + \dots + p_i)$, fiindcă U este uniformă pe $(0, 1)$

De aici rezultă că X ia valoarea x_i cu aceeași probabilitate ca și U să se găsească în intervalul $[p_1 + \dots + p_{i-1}, p_1 + \dots + p_i] = [F(x_{i-1}), F(x_i))$. Deci, practic, pentru a genera o observație x_n din distribuția cunoscută! a) Dacă X este suficient să generăm ună observație uniformă pe $(0, 1)$ și să verificăm în care interval de tipul $[F(x_{n-1}), F(x_n))$ se găsește.

? A se vedea R pt. implementare.

Exercițiu

Un site este accesat în fiecare minut cu probabilitate $1/4$. Care este probabilitatea ca site-ul să fie accesat de cel mult 5 ori în 10 minute?

Soluție:

Ne situăm în schema binomială.

IDE (reducere la un context clasic): E similar cu a da cu banul de 10 ori, cu $P(H) = \frac{1}{4}$.

$\underbrace{H}_{\text{n min}} \quad \underbrace{\text{non-H}}_{\text{n min}} \quad \dots \quad \underbrace{\text{non-H}}_{\text{n min}}$

Notăm cu $X = nr$ de accesări ale site-ului în 10 minute
Avem $X \sim Bin(n=10, p=1/4)$, din datele problemei.

Vrem:

$$\begin{aligned} P(X \leq 5) &= P(X=0) + P(X=1) + \dots + P(X=5) \\ &= \sum_{k=0}^5 C_{10}^k \cdot \left(\frac{1}{4}\right)^k \cdot \left(\frac{3}{4}\right)^{10-k} \\ &= \text{.plinom}(5, nre = 10, prob = 0.25) \\ &= 0.9802 \quad \blacksquare \end{aligned}$$

Temă: Alegeti 2 repartiții uscate studiate pînă acum și reprezentați pentru fiecare împărte, pe un același grafic, funcția de repartiție empirică și funcția de repartiție teoretică corespunzătoare. În script-ul corespondator laboratorului acesta am vizualizat ceva similar folosind pachetul "EnvStats". Încercati de data aceasta să nu folosiți pachete suplimentare, ci să utilizati explicit funcțiile prezentate azi.