Analiza datelor în R

Curs 4

Variabile aleatoare

- discrete
$\longrightarrow$ tablou de repartiţie

$$X \leftrightarrow \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix} \text{ cu } \sum_i p_i = 1$$

$$P(X \in I) = \sum_{x_i \in I} P(X = x_i) = \sum_{x_i \in I} p_i$$

- continue
$\longrightarrow$ densitate de repartiţie $f : \mathbb{R} \to [0, \infty)$ cu $\int\limits_{-\infty}^{\infty} f(t)dt = 1$.

$$P(X \in [a, b]) = \int\limits_a^b f(t)dt, \quad \forall a, b \in [-\infty, \infty]$$

- Funcţia de repartiţie a variabilei aleatoare X este

$F : \mathbb{R} \to [0, 1]$, $F(x) = P(X < x)$.

Distribuţia uniformă discretă

Fie $n \in \mathbb{N}^*$. Vom spune că variabila aleatoare X are distribuţie uniformă discretă pe $[1, n]$ (notaţie: $X \sim U(n)$) dacă

$$P(X = k) = \frac{1}{n},\ \forall k \in \{1, 2, \ldots, n\}$$

Dacă $X \sim U(n)$, atunci $M(X) = \dfrac{n+1}{2}$ şi $D^2(X) = \dfrac{n^2 - 1}{12}$.

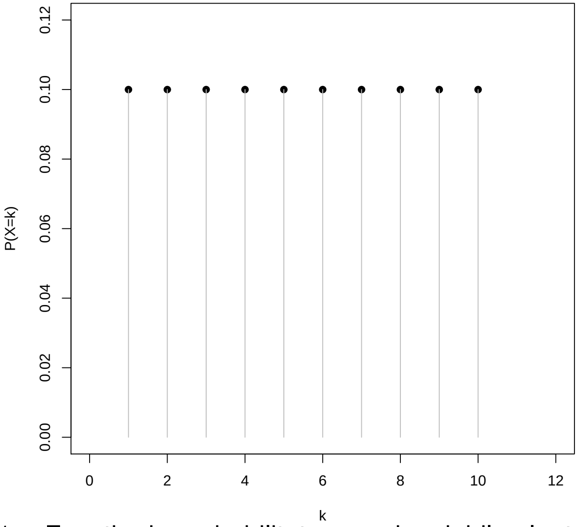

Figura 1 : Funcţia de probabilitate a unei variabile aleatoare $U(10)$

Distribuţia binomială

Fie $n \in \mathbb{N}^*$, $p \in (0,1)$. Vom spune că X are distribuţie binomială de parametri n şi p (notaţie: $X \sim B(n,p)$) dacă

$$P(X = k) = \begin{cases} C_n^k p^k (1-p)^{n-k}, & k \in \{0, 1, \ldots, n\} \\ 0, & \text{în rest} \end{cases}$$

Distribuţia binomială modelează numărul de realizări ale unui eveniment ("succese") în n repetări independente ale unui experiment, dacă la fiecare repetare probabilitatea de realizare a evenimentului este p.

Dacă $X \sim B(n,p)$, atunci $M(X) = np$ şi $D^2(X) = np(1-p)$.

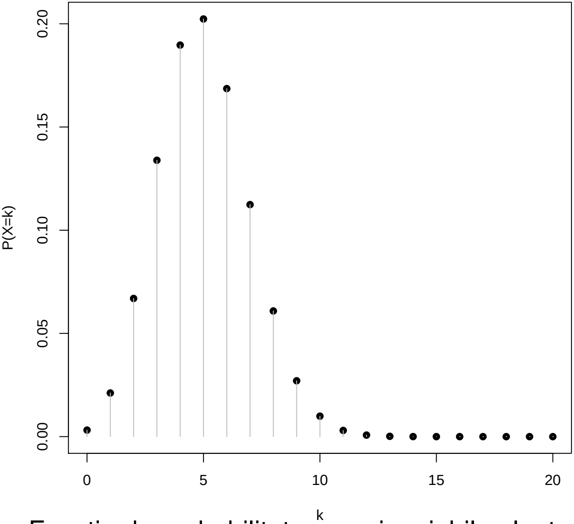

Figura 2 : Funcţia de probabilitate a unei variabile aleatoare $B(20, 0.25)$

Distribuţia geometrică

Fie $p \in (0,1)$. Vom spune că X are distribuţie geometrică de parametru p (notaţie: $X \sim Geom(p)$) dacă

$$P(X=k) = \begin{cases} (1-p)^{k-1}p, & k \in \mathbb{N}^* \\ 0, & \text{în rest} \end{cases}$$

Distribuţia geometrică modelează numărul de repetări independente ale unui experiment până la prima realizare a evenimentului de interes ("succes"), dacă la fiecare repetare probabilitatea de succes este p.

Dacă $X \sim Geom(p)$, atunci $M(X) = \dfrac{1}{p}$ şi $D^2(X) = \dfrac{1-p}{p^2}$.

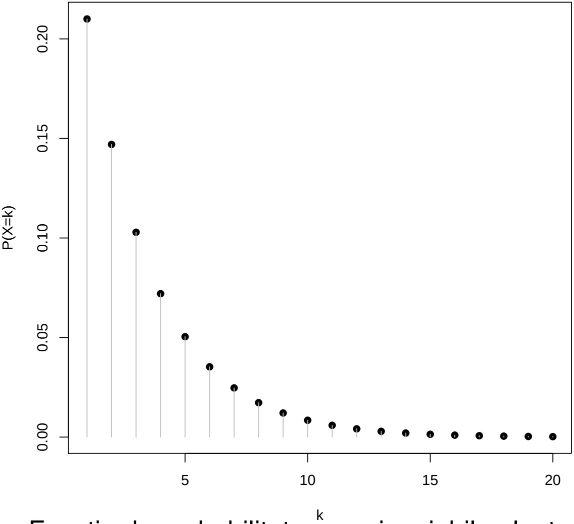

Figura 3 : Funcţia de probabilitate a unei variabile aleatoare *Geom*(0.3)

Distribuţia Poisson

Fie $\lambda > 0$. Variabila aleatoare X are distribuţie Poisson de parametru λ (notaţie: $X \sim Po(\lambda)$) dacă

$$P(X = k) = \begin{cases} \dfrac{\lambda^k}{k!} e^{-\lambda}, & k \in \mathbb{N} \\ 0, & \text{în rest} \end{cases}$$

Dacă $X \sim Po(\lambda)$, atunci $M(X) = \lambda$ şi $D^2(X) = \lambda$.

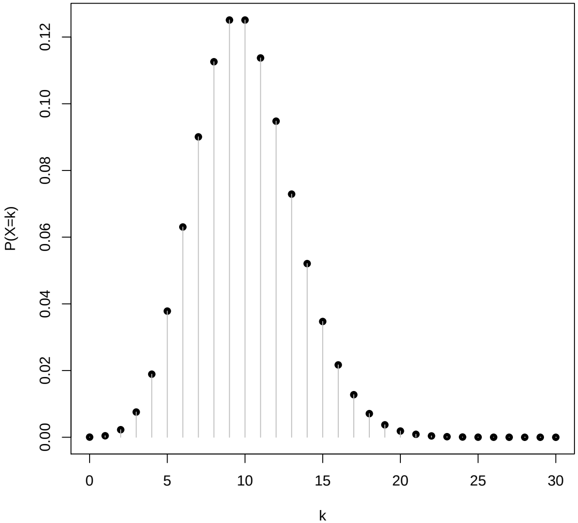

Figura 4 : Funcţia de probabilitate a unei variabile aleatoare $Po(10)$

Distribuţia uniformă

Fie $a, b \in \mathbb{R}$, $a < b$. Variabila aleatoare X are distribuţie uniformă pe $[a, b]$ (notaţie: $X \sim U[a, b]$) dacă densitatea de repartiţie a lui X este

$$f_X(t) = \begin{cases} \dfrac{1}{b-a}, & t \in [a, b] \\ 0, & t \notin [a, b] \end{cases}$$

Dacă $X \sim U[a, b]$, atunci $M(X) = \dfrac{a+b}{2}$ şi $D^2(X) = \dfrac{(b-a)^2}{12}$.

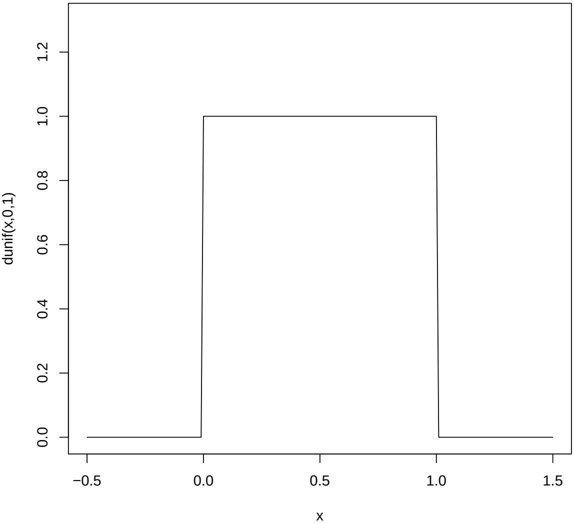

Figura 5 : Densitatea de repartiţie a unei variabile aleatoare $U[0, 1]$

Distribuţia normală

Fie $\mu \in \mathbb{R}, \sigma > 0$. O variabilă aleatoare X având densitatea de repartiţie

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (t \in \mathbb{R})$$

se numeşte variabilă aleatoare normală de parametri μ, σ^2. (Notaţie: $X \sim N(\mu, \sigma^2)$)

- Dacă $X \sim N(\mu, \sigma^2)$, atunci $M(X) = \mu$ şi $D^2(X) = \sigma^2$.
- Dacă $\mu = 0$ şi $\sigma^2 = 1$, vom spune că X are distribuţie normală standard.
- $X \sim N(\mu, \sigma^2) \Leftrightarrow \dfrac{X - \mu}{\sigma} \sim N(0, 1)$.

Distribuţia normală

- Densitatea de repartiţie este simetrică în raport cu media μ.
- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$

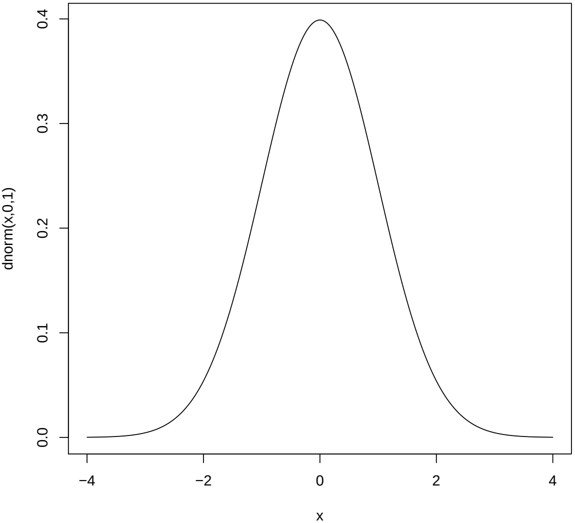

Figura 6 : Densitatea de repartiţie a unei variabile aleatoare $N(0, 1)$

Distribuţia χ^2

Fie $n \in \mathbb{N}^*$. Spunem că o variabilă aleatoare X are distribuţie χ^2 cu n grade de libertate (notaţie: $X \sim \chi^2(n)$) dacă densitatea sa de repartiţie este

$$f_X(t) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} \quad (t \geq 0).$$

- $X \sim N(0,1) \Rightarrow X^2 \sim \chi^2(1)$.
- $X_1, X_2, \ldots, X_n \sim \chi^2(1)$ independente $\Rightarrow \sum\limits_{i=1}^{n} X_i \sim \chi^2(n)$.

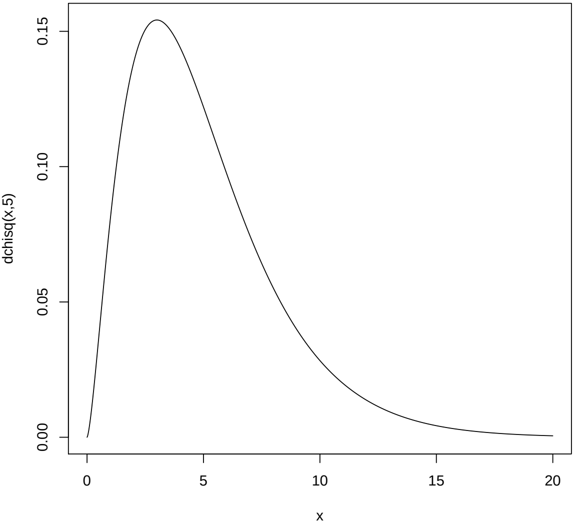

Figura 7 : Densitatea de repartiţie a unei variabile aleatoare $\chi^2(5)$

Distribuţia Student (t)

Fie $n \in \mathbb{N}^*$. Spunem că o variabilă aleatoare X are distribuţie t cu n grade de libertate (notaţie: $X \sim t(n)$) dacă densitatea sa de repartiţie este

$$f_X(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad (t \in \mathbb{R}).$$

- Dacă $X \sim t(n)$, atunci $M(X) = 0$ şi $D^2(X) = \dfrac{n}{n-2}$ $(n > 2)$.
- Densitatea de repartiţie este simetrică în raport cu media.
- $X \sim N(0,1)$, $Y \sim \chi^2(n)$ independente $\Rightarrow \dfrac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$

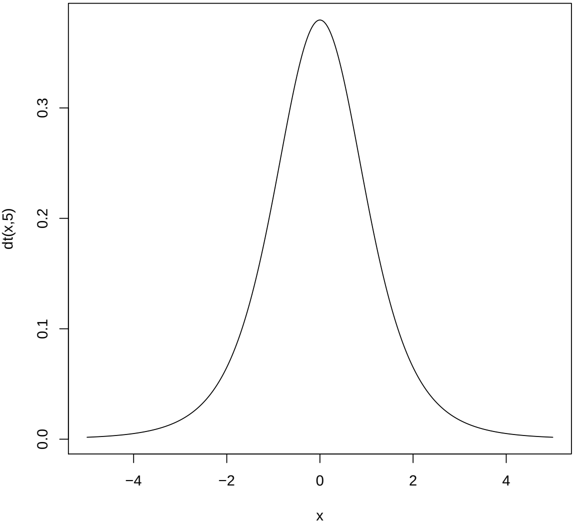

Figura 8 : Densitatea de repartiţie a unei variabile aleatoare $t(5)$

Distribuţii de probabilitate în R

- `d+numeDistributie` $\rightarrow$ funcţie de probabilitate (cazul discret), respectiv densitate de repartiţie (cazul continuu)
- `p+numeDistributie` $\rightarrow$ funcţie de repartiţie
- `q+numeDistributie` $\rightarrow$ quantile
- `r+numeDistributie` $\rightarrow$ generare de numere aleatoare care urmează distribuţia dată

unde numeDistributie= binom, geom, pois, unif, norm, chisq, t etc.

Dat $p \in (0, 1)$, se numeşte *p-quantila* variabilei aleatoare X (sau a distribuţiei de probabilitate corespunzătoare) numărul real x cu proprietatea că $P(X < x) = p$.