Curs 7: Distribuţia normală

1.1 Distribuţia Normala

Distribuţia normală este distribuţia de probabilitate cel mai mult folosită, ca model, în inteligenţa artificială, *machine learning*, *data mining*, analiza şi procesarea imaginilor, şi n statistică.

Definiţia 1.1.1 *O variabilă aleatoare continuă, X, ce are densitatea de probabilitate*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-m)^2}{2\sigma^2}, \ \forall\, x \in \mathbb{R}, \tag{1}$$

se numeşte variabilă aleatoare normal distribuită, de parametri m, σ, unde $m \in \mathbb{R}$, iar $\sigma > 0$ (Fig.1). Notăm cu $N(m, \sigma)$ clasa variabilelor aleatoare ce au această distribuţie.

Distribuţia normală standard: Pentru o variabilă aleatoare normal distribuită, de parametri $m = 0, \sigma = 1$, densitatea de probabilitate (1) se notează:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \ t \in \mathbb{R}. \tag{2}$$

Funcţia φ este pară, adică $\varphi(-t) = \varphi(t)$, $\forall\, t \in \mathbb{R}$ (Fig. 2) şi graficul său se numeşte clopotul lui Gauss.

Propoziţia 1.1.1 *O variabilă aleatoare X, ce are densitatea de probabilitate φ, are valoarea medie 0, iar dispersia $\sigma^2 = 1$.*

Demonstraţie: **Opţional.** Valoarea medie a variabilei X este: $M(X) = \int\limits_{-\infty}^{\infty} t\varphi(t)\, dt =$ $\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} t\, e^{-t^2/2}\, dt = \underbrace{\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{0} t\, e^{-t^2/2}\, dt}_{I_1} + \underbrace{\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} t\, e^{-t^2/2}\, dt}_{I_2}$. Pentru calculul integralei I_1 efectuăm schimbarea de variabilă $t = -y$. Astfel $I_1 = \frac{1}{\sqrt{2\pi}} \int\limits_{\infty}^{0} -ye^{-y^2/2}(-dy) = -I_2$ şi deci $M(X) = I_1 + I_2 = -I_2 + I_2 = 0$.

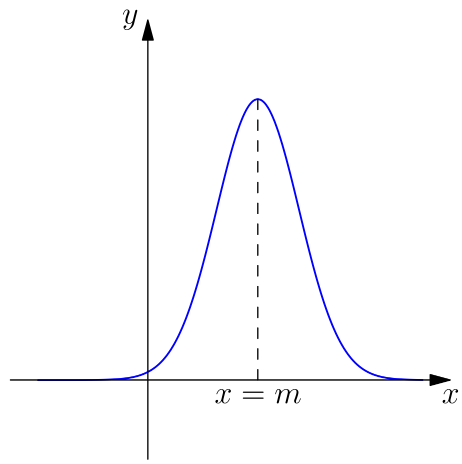

Fig.1: Graficul densităţii de probabilitate pentru $X \sim N(m = 2, \sigma = 0.75)$.

Dispersia variabilei X este $\sigma^2(X) = M(X^2) - (M(X))^2 = M(X^2) - 0 = M(X^2)$. Calculăm

$$M(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2}\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t(t\, e^{-t^2/2})\, dt. \tag{3}$$

Ultimă integrală se calculează prin părţi şi obţinem $M(X^2) = 1$. □

End Opţional.

• Notăm cu $N(0, 1)$ clasa variabilelor aleatoare ce au distribuţia normală de medie 0 şi dispersie 1, numită şi *distribuţia normală standard*.

• Există convenţia internaţională de a nota prin Z o variabilă aleatoare ce are distribuţia normală standard.

• Funcţia de repartiţie a unei astfel de variabile aleatoare se notează cu Φ:

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{4}$$

Propoziţia 1.1.2 *Funcţia de repartiţie, Φ, a unei variabile aleatoare $Z \sim N(0, 1)$ are proprietatea că:*

$$\Phi(-x) = 1 - \Phi(x),\ \forall x \in \mathbb{R} \tag{5}$$

Demonstraţie: **Opţional.** $\Phi(-x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{-x} e^{-t^2/2}\, dt$. Efectuând schimbarea de vari-

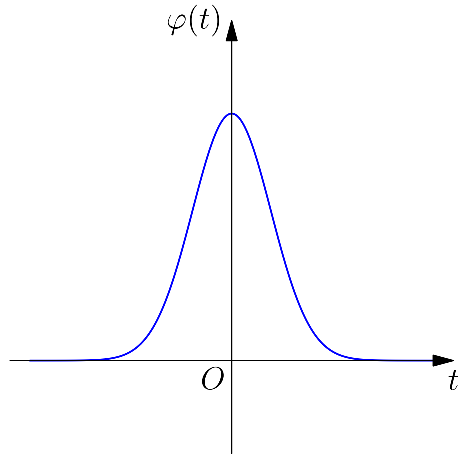

Fig.2: Densitatea de probabilitate a unei variabile aleatoare $Z \sim N(0,1)$.

abilă $t = -u$, avem:

$$\begin{aligned}\Phi(-x) &= -\frac{1}{\sqrt{2\pi}}\int_{\infty}^{x} e^{-u^2/2}\,du = \frac{1}{\sqrt{2\pi}}\int_{x}^{\infty} e^{-u^2/2}\,du \\ &= \frac{1}{\sqrt{2\pi}}\left(\int_{x}^{\infty} e^{-u^2/2}\,du + \int_{-\infty}^{x} e^{-u^2/2}\,du - \int_{-\infty}^{x} e^{-u^2/2}\,du\right) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2}\,du - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\,du = 1 - \Phi(x). \qquad (6)\end{aligned}$$

□

End Opţional.

Valoarea funcţiei Φ într-un punct x se poate calcula doar prin metode aproximative, deoarece nu se poate determina analitic nici o primitivă a funcţiei $e^{-t^2/2}$. În general, cărţile de Teoria Probabilităţilor conţin tabele de valori ale lui Φ în argumente pozitive. Pentru x negativ valoarea se exprimă în prelabil prin $\Phi(x) = 1 - \Phi(-x)$. $-x$ fiind pozitiv, $\Phi(-x)$ se caută în tabel.

α–cvantila distribuţiei normale standard.

Pentru o variabilă aleatoare, de distribuţie de probabilitate arbitrară, ce are funcţia de repartiţie, $F : \mathbb{R} \to (0,1)$, strict crescătoare, inversa $F^{-1} : (0,1) \to \mathbb{R}$ se numeşte funcţia cvantilă. α–cvantila, $\alpha \in (0,1)$, unei astfel de distribuţii de probabilitate este unicul număr real $x_\alpha = F^{-1}(\alpha)$. Aplicând funcţia F aceste egalităţi obţinem $F(x_\alpha) = \alpha$, sau echivalent $P(X \leq x_\alpha) = \alpha$. α–cvantila unei variabile aleatoare având distribuţia normală standard, $Z \sim N(0,1)$, se notează z_α.

Propoziţia 1.1.3 *Între cvantilele $z_{1-\alpha}$ şi z_α ale unei variabile aleatoare $Z \sim N(0,1)$, există relaţia:*

$$z_{1-\alpha} = -z_\alpha, \forall\, \alpha \in (0,1). \qquad (7)$$

Demonstraţie: **Opţional.** Fie Φ funcţia de repartiţie a variabilei aleatoare $Z \sim N(0,1)$ şi $z_{1-\alpha}$, z_α cvantilele sale, adică

$$\Phi(z_{1-\alpha}) = 1-\alpha \tag{8}$$
$$\Phi(z_\alpha) = \alpha. \tag{9}$$

Prin urmare putem scrie: $\Phi(z_{1-\alpha}) = 1-\Phi(z_\alpha)$. Ţinând seama de (5), avem $1-\Phi(z_\alpha) = \Phi(-z_\alpha)$. Astfel $\Phi(z_{1-\alpha}) = \Phi(-z_\alpha)$ şi cum Φ este strict crescătoare, deci injectivă, rezultă că $z_{1-\alpha} = -z_\alpha$, $\forall \alpha \in (0,1)$. □

End Opţional.

α–cvantilele distribuţiei normale standard se folosesc în statistică. Valorile α–cvantilelor se găsesc fie în tabele, fie apelând funcţiile adecvate, din pachetele software ce le pun la dispoziţie. α–cvantilele distribuţiei normale standard, vor intervini în partea de Statistică a cursului.

Studiul variabilelor aleatoare ce au distribuţie normală arbitrară, $N(m,\sigma)$, se realizează, stabilind o relaţie a acestora cu variabilele aleatoare normal distribuite, standard. Şi anume:

Fie variabila aleatoare $X \sim N(m,\sigma)$, având densitatea de probabilitate, f_X, definită în (1). Observăm că această densitate se poate exprima în funcţie de densitatea φ, a distribuţiei normale standard, $N(0,1)$, astfel:

$$f_X(x) = \frac{1}{\sigma}\,\varphi_Z\left(\frac{x-m}{\sigma}\right) \tag{10}$$

Propoziţia 1.1.4 *Funcţia de repartiţie a variabilei $X \sim N(m,\sigma^2)$, F_X, se exprimă în funcţie de repartiţia normală standard prin:*

$$F_X(x) = \Phi\left(\frac{x-m}{\sigma}\right) \tag{11}$$

Demonstraţie: Într-adevăr, $F_X(x) = P(X \leq x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sigma}\varphi\left(\frac{t-m}{\sigma}\right) dt$. Efectuând schimbarea de variabilă, $y = \dfrac{t-m}{\sigma}$, obţinem:

$$F_X(x) = \int_{-\infty}^{(x-m)/\sigma} \varphi(y)\,dy = \Phi\left(\frac{x-m}{\sigma}\right)$$

□

Propoziţia 1.1.5 *Dacă variabila aleatoare X are distribuţia normală, $X \sim N(m,\sigma)$, atunci variabila standardizată asociată, $Z = (X-m)/\sigma$, are distribuţia $N(0,1)$ şi reciproc, dacă $Z \sim N(0,1)$, atunci oricare ar fi $m \in \mathbb{R}$ şi $\sigma > 0$, variabila aleatoare $X = \sigma Z + m \sim N(m,\sigma^2)$.*

Demonstraţie: Vom arăta că funcţia de repartiţie a variabilei Z este Φ. $P(Z \leq x) = P\left(\dfrac{X-m}{\sigma} \leq x\right) = P(X \leq m + \sigma\, x) = F_X(m + \sigma\, x) \overset{(11)}{=} \Phi(x)$.

Reciproca este evidentă. □

Ca şi în cazul variabilelor aleatoare ce au distribuţia normală standard, parametrii m şi σ ai unei variabile aleatoare, $X \sim N(m, \sigma)$, au semnificaţie probabilistă, şi anume m este media variabilei aleatoare, $m = M(X)$, iar σ^2 este dispersia sa.

Pentru a demonstra aceste egalităţi asociem variabilei X, variabila standardizată, $Z = \dfrac{X-m}{\sigma} \sim N(0,1)$. Astfel $X = m + \sigma Z$ şi $M(X) = m + \sigma M(Z) = m$. Pe de altă parte, $D^2(X) = 0 + \sigma^2 D^2(Z) = \sigma^2$ (am folosit proprietatea $D^2(aZ) = a^2 D^2(Z)$).

Aşa cum se observă din definiţia densităţii de probabilitate f, a unei variabile aleatoare $X \sim N(m, \sigma)$, graficul funcţiei f este simetric faţă de dreapta $x = m$. Cu cât σ este mai mic, cu atât probabilitatea ca valorile variabilei aleatoare să fie concentrate în jurul mediei m este mai mare (Fig.3).

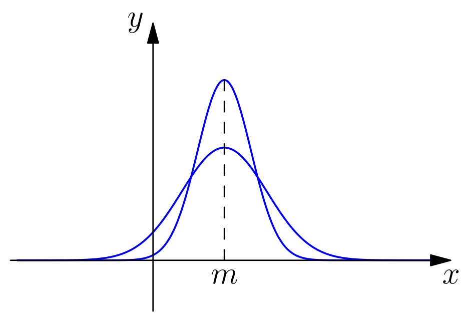

Fig.3: Graficele a două densităţi normale, $N(m, \sigma)$, ce au aceeaşi medie $m = 2$, iar abaterile standard sunt respectiv $\sigma = 0.75$, $\sigma = 1.2$.

Calculul probabilităţilor normale.

Fie v.a $X \sim N(m, \sigma)$. Funcţia sa de repartiţie fiind continuă, avem:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \\ = F_X(b) - F_X(a) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

Exemplul 1. Fie $X \sim N(1, 0.4^2)$. Să se calculeze:
a) $P(X > 0.8)$;
b) $P(0.75 < X \leq 1.3)$;
c) Să se determine $x \in \mathbb{R}$ astfel încât $P(X \leq x) = 0.95$

Rezolvare: a) $(X > 0.8) = \complement(X \leq 0.8)$. Astfel, $P(X > 0.8) = 1 - F_X(0.8)$. Dar $F_X(x) = \Phi((x-1)/0.4)$ şi deci $F_X(0.8) = \Phi(-0.2/0.4) = \Phi(-0.5)$. Ţinând seama că

$\Phi(-z) = 1 - \Phi(z)$, rezultă că $\Phi(-0.5) = 1 - \Phi(0.5)$ şi prin urmare $P(X > 0.8) = 1 - (1 - \Phi(0.5)) = \Phi(0.5)$.
b) $P(0.75 < X \leq 1.3) = F_X(1.3) - F_X(0.75) = \Phi((1.3 - 1)/0.4) - \Phi((0.75 - 1)/0.4) = \Phi(0.75) - \Phi(-0.625) = \Phi(0.75) - 1 + \Phi(0.625)$.

c) $P(X \leq x) = F_X(x) = \Phi((x - 1)/4) = 0.95$. Deci $(x - 1)/0.4 = \Phi^{-1}(0.95) = z_{0.95}$. Din tabele se află că 0.95–cvantila distribuţiei normale standard este $z_{0.95} = 1.64$ şi deci $x = 1 + 0.4 \cdot 1.64$.