# Curs 10: Inegalitatea Markov. Inegalitatea Cebîșev. Covarianta. Coeficientul de corelație. Mixuri de probabilitate

Conf.dr. Maria Jivulescu

Departamentul de Matematică
UPT

# Cuprins

# Inegalitatea Markov.



Fie $X$ o v.a. astfel încât $X \geq 0$, adică $X$ ia valori nenegative. Dacă $X$ are medie finită, atunci, pentru $a > 0$, avem

$$P(X \geq a) \leq \frac{M(X)}{a}.$$

Demo:

- $X \geq 0 \to P(X \geq 0) = 1$, deci $P(X < 0) = 0 = \int_{-\infty}^{0} f(x)\,dx$.
- presupune că funcția $f$ este zero pe întreg intervalul $(-\infty, 0)$
-

$$M(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{0} xf(x)\,dx}_{=0} + \int_{0}^{a} xf(x)\,dx}_{\geq 0} + \int_{a}^{\infty} xf(x$$
$$\geq \int_{a}^{\infty} xf(x)\,dx.$$

- $M(X) \geq \int_a^\infty af(x)\,dx = a\int_a^\infty f(x)\,dx = a\int_{[a,\infty)} f(x)\,dx = aP(X \geq a)$

# Inegalitatea Markov



- Demonstrația este identică și pentru variabile aleatoare discrete, doar că integralele se înlocuiesc cu sume.
- se folosește în studiul algoritmilor probabiliști pentru a estima probabilitatea ca timpul de execuție al unui algoritm cu intrări de volum precizat să depășească $a$ unități de timp, când se cunoaște doar timpul mediu de execuție, $M(X)$.

# Inegalitatea Cebîșev



Fie $X$ o v.a. arbitrară de medie $M(X)$ și dispersie $\sigma^2(X)$ finite. Atunci:

$$P(|X - M(X)| \geq a) \leq \frac{\sigma^2(X)}{a^2}, \ a > 0.$$

Demo:

- Fie variabila aleatoare $Y = (X - M(X))^2 \geq 0$;
- Evident $M(Y) = \sigma^2(X)$;
- inegalitatea Markov pentru v.a. $Y$:

  $P(|X - M(X)| \geq a) = P((X - M(X)^2) \geq a^2) = P(Y \geq a^2) \overset{\text{Markov}}{\leq} \frac{M(Y)}{a^2}$.
- $M(Y) = \sigma^2(X)$, rezultă inegalitatea Cebîșev.

# O alta versiune a inegalității Cebîsev



Cazul $a = k\sigma(X)$, $k \in \mathbb{N}^*$:

$$P(|X - M(X)| \geq a) \leq \frac{\sigma^2(X)}{k^2\sigma^2(X)} = \frac{1}{k^2}.$$

Echivalent:

$$P(|X - M(X)| < k) = 1 - P(|X - M(X)| \geq k) \geq 1 - \frac{1}{k^2}.$$

- pentru $k = 2$ avem: $P(m - 2\sigma < X < m + 2\sigma) \geq 1 - \frac{1}{4} = 0.75$.
- pentru $k = 3$ avem $P(m - 3\sigma < X < m + 3\sigma) \geq 1 - \frac{1}{9} = \frac{8}{9} = 0.88$.
- pentru $k = 4$ avem
  $P(m - 4\sigma < X < m + 4\sigma) \geq 1 - \frac{1}{16} = \frac{15}{16} = 0.9375$.

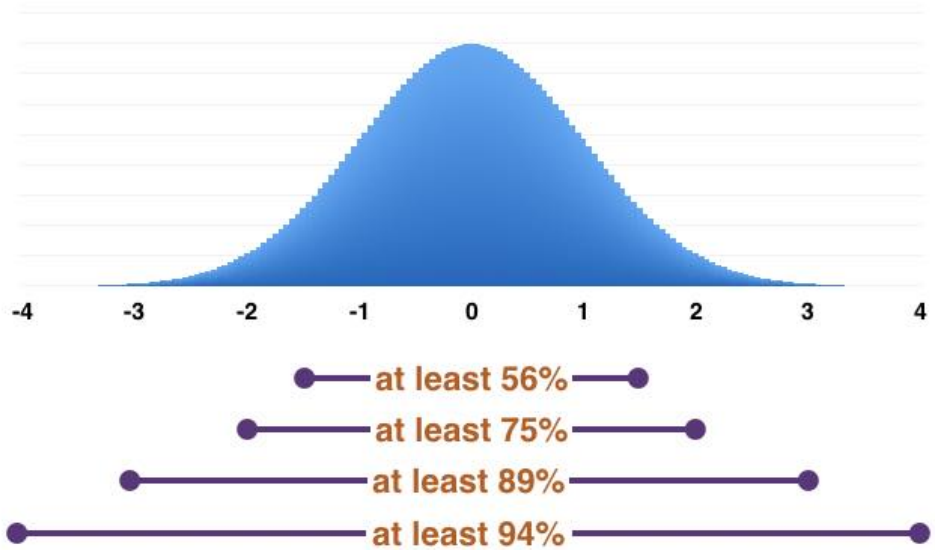
# Inegalitatea Cebîșev

**Concluzie**: probabilitatea ca variabila aleatoare $X$ să ia valori in intervale centrate în valoarea sa medie și de lungime $4\sigma$, $6\sigma$, $8\sigma$ este mai mare decât $0.75, 0.88$, respectiv $0.9375$.

## Chebyshev's Inequality

*The proportion of observations within k standard deviations is at least 1-1/k²*

| No. Std Devs | 1 - 1/k² |
|---|---|
| 1.5 | 0.56 |
| 2.0 | 0.75 |
| 3.0 | 0.89 |
| 4.0 | **0.94** |



# Covariația și coeficientul de corelație



**Intrebare:** Cum caracterizăm intensitatea legăturii dintre două variabile ce nu sunt independente?

Măsuri de studiu al dependenței:

- covariația
- coeficientul lor de corelație.

# Covarianța a două variabile aleatoare



## Definiție

**Covarianța** variabilelor aleatoare $X$ și $Y$, ce au mediile $m_X = M(X)$, $m_Y = M(Y)$ finite, este definită prin

$$cov(X, Y) = M((X - m_X)(Y - m_Y)).$$

**Observație:**

- Covarianța=generalizare la două variabile a dispersiei

$$\text{cov}(X, X) = M((X - m_X)(X - m_X)) = M((X - m_X)^2) = \sigma^2(X).$$

- formula mai simplă de calcul

$$\text{cov}(X, Y) = M(XY) - M(X)\, M(Y)$$

# Variabile aleatoare necorelate



**Variabile aleatoare necorelate:** $cov(X, Y) = 0$ .

**Definiție**

**Coeficientul de corelație** a două variabile aleatoare $X$ și $Y$, de abateri standard nenule, este un număr real, notat cu $\rho(X, Y)$, definit prin

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

unde $\sigma_X, \sigma_Y -$ abaterile standard ale variabilelor aleatoare $X$, respectiv $Y$.

- daca $Z_1 = \frac{X - m_X}{\sigma_X}$, $\quad Z_2 = \frac{Y - m_Y}{\sigma_Y}$, atunci

$$\rho(X, Y) = cov(Z_1, Z_2)$$

- $\rho(X, Y) \in [-1, 1]$.

## Proprietate

Dacă între variabilele aleatoare $X$ și $Y$ există o relație liniară de forma

$$Y = aX + b,\ a, b \in,\ a \neq 0,$$

atunci

$$\rho(X, Y) = \begin{cases} -1, & \text{dacă } a < 0, \\ 1, & \text{dacă } a > 0. \end{cases}$$

Reciproc, dacă $|\rho(X, Y)| = 1$, atunci între ele există o relație liniară,

$$Y = aX + b, \quad a \neq 0$$

În concluzie:

- $\rho(X, Y) = 0$, atunci $X$ și $Y$ sunt **necorelate**;
- $\rho(X, Y)$ este apropiat de zero, atunci $X$ și $Y$ sunt **slab corelate** (intensitatea legăturii dintre ele este redusă);
- $\rho(X, Y) = 1$, atunci $Y = aX + b, \quad a > 0$, $X$ și $Y$ sunt **pozitiv corelate**;
- $\rho(X, Y) = -1$, atunci $Y = aX + b, \quad a < 0$, $X$ și $Y$ sunt **negativ corelate**;
- $|\rho(X, Y)|$ are o valoare apropiată de 1, **relația dintre variabilele aleatoare este "aproape liniară"**, adică valorile $(x, y)$ ale vectorului aleator $(X, Y)$ sunt ușor dispersate în jurul unei drepte de ecuație $y = ax + b$.

# Matricea de covarianță a unui vector aleator



## Definiție

**Matricea de covarianță** a vectorului aleator $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ este matricea notată cu $\Sigma$, ale cărei elemente sunt $\sigma_{ij} = \text{cov}(X_i, X_j)$, $i, j \in \{1, \ldots, n\}$.

**Observații:**

- $\sigma_{ii} = cov(X_i, X_i) = \sigma^2(X_i)$
- $\Sigma$ este simetrică și semipozitiv definită
- $\Sigma = M(\mathbf{Y}\mathbf{Y}^T)$, unde $\mathbf{Y} = \mathbf{X} - \mathbf{m} = (X_1 - m_1, X_2 - m_2, \ldots, X_n - m_n)^T$ iar $M(\mathbf{Y}\mathbf{Y}^\mathbf{T})$ notează matricea mediilor elementelor matricii $\mathbf{Y}\mathbf{Y}^\mathbf{T}$.

# Mixturi de probabilitate

Fie $p_1, p_2, \ldots, p_n \in (0,1)$ astfel încât $p_1 + p_2 + \cdots + p_n = 1$.

- Dacă $F_1, F_2, \ldots, F_n$ sunt funcțiile de repartiție ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci funcția

$$F = p_1 F_1 + p_2 F_2 + \cdots + p_n F_n$$

este o funcție de repartiție, numită **repartiție compusă**.

- Dacă $f_1, f_2, \ldots, f_n$ sunt densitățile de probabilitate ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci

$$f = p_1 f_1 + p_2 f_2 + \cdots + p_n f_n$$

este o densitate de probabilitate, numită **densitate compusă**.

**Definiție**

O variabilă aleatoare $X$ ce are densitatea de probabilitate compusă $f$ sau funcția de repartiție compusă $F$, se numește **mixtură de distribuții de probabilitate** sau, mai simplu, **mixtură de probabilitate**.

- Dacă $X$ este o v.a. cu densitatea $f = p_1 f_1 + p_2 f_2 + \ldots + p_n f_n \Longleftrightarrow X$ are densitatea $f_1$ cu probab. $p_1$, ..., $X$ are densitatea $f_n$ cu probab. $p_n$.
- Reprezentând fiecare densitate $f_k$ prin indicele său $k$, asociem unei densități compuse o variabilă aleatoare discretă:

$$H = \begin{pmatrix} 1 & 2 & \ldots & k & \ldots & n \\ p_1 & p_2 & \ldots & p_k & \ldots & p_n \end{pmatrix}.$$

# Variabile hiperexponențiale



**Definiție**

O v.a. $X$ a cărei densitate de probabilitate

$$f = p_1 f_1 + p_2 f_2 + \cdots + p_n f_n, \text{unde } f_i(x) = \begin{cases} \frac{1}{\theta_i} e^{-x/\theta_i}, & \text{dac'a } x \geq 0, \\ 0, & \text{dacă } x < 0, \end{cases}$$

este compusa a $n$ densități ale distribuției exponențiale de parametrii $\theta_1, \theta_2, \ldots, \theta_n$ se numește **variabilă aleatoare hiperexponențială**.

Variabilele aleat. hiperexponențiale modelează durata serviciului procesorului. Se folosesc în simularea rețelelor de cozi.

# Mixura Poisson

## Exemplu:

Facebook monitorizează atitudinea unui user față de postările pe wall-uri și îi asociază un număr de reacții ce sunt modelate de o mixtură Poisson.

- reacția $R_1$ (cu prob. $p_1$) = linkuri la articole din *Times New Roman* cu rata $\lambda_1$/oră
- reacția $R_2$ (cu prob. $p_1$) = like-uri la pozele amicilor, cu rata $\lambda_2$/oră

Astfel, numărul de reacții/manifestări ale userului pe oră este o variabilă aleatoare $X$ ce are ca distribuție de probabilitate mixtura Poisson:

$$P_X(k) := P(X = k) = p_1 e^{-\lambda_1} \frac{\lambda_1^k}{k!} + p_2 e^{-\lambda_2} \frac{\lambda_2^k}{k!}.$$

Mixtura de distribuții de probabilitate **nu** înseamnă că variabila $X$ este de forma $X = p_1 X_1 + p_2 X_2$, cu $X_i \sim Poiss(\lambda_i)$, $i = 1, 2$, ci că $X$ are distribuția Poisson de rată $\lambda_1$ cu probab. $p_1$, respectiv $X$ are distribuția Poisson de rată $\lambda_2$ cu probab. $p_2$.