Curs 4

V.a. discrete

4.1 Caracteristici numerice ale variabilelor aleatoare discrete

Definiţia 4.1.1 *Se spune că o variabilă aleatoare discretă X, având matricea de repartiţie de forma*

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix} \tag{4.1}$$

admite medie şi dispersie dacă seriile $\sum\limits_{n \geq 1} x_i p_i$ şi $\sum\limits_{n \geq 1} x_i^2 p_i$ sunt convergente. În acest caz se definesc ***media*** *lui X*

$$M[X] = \sum_{n \geq 1} x_i p_i = m \text{ (suma seriei)},$$

momentul de ordin doi

$$M\left[X^2\right] = \sum_{n \geq 1} x_i^2 p_i$$

şi ***dispersia***

$$D[X] = M\left[X^2\right] - [M[X]]^2 = \sum_{n \geq 1} x_i^2 p_i - \left(\sum_{n \geq 1} x_i p_i \right)^2.$$

Dacă X are un număr finit de valori, atunci evident X are medie şi dispersie şi sumele anterioare sunt finite.

Definiţia 4.1.2 *Se spune că o variabilă aleatoare discretă X, având matricea de repartiţie de forma (4.1)* ***admite moment de ordin*** *k dacă seria $\sum\limits_{n \geq 1} x_i^k p_i$ este convergentă. În acest caz se definesc* ***momentul de ordin*** *k al lui X*

$$M\left[X^k\right] = \sum_{n \geq 1} x_i^k p_i$$

şi ***momentul centrat de ordin*** *k al v.a. X*

$$M\left[(X - M[X])^k\right] = \sum_{n \geq 1} (x_i - M[X])^k p_i.$$

Definiţia 4.1.3 *Se numeşte* ***abarere media pătratică*** *$\sigma = \sqrt{D[X]}$.*

Exerciţiul 1 Demonstraţi că momentul centrat de ordin doi este chiar dispersia, adică $D\left[X\right]=\sum\limits_{n\geq 1}\left(x_i-M\left[X\right]\right)^2 p_i$.

Definiţia 4.1.4 *Variabila aleatoare $X-M[X]$ se numeşte* ***abaterea de la medie*** *a variabilei aleatoare X.*

Propoziţia 4.1.5 *Valoarea medie a abaterii oricărei variabile aleatoare este nulă.*

Exerciţiul 2 Variabilele aleatoare discrete simple X_1 şi X_2, având repartiţiile:

$$X_1, X_2 : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix},$$

au valoarea medie

$$M[X_i] = 1p + 0q = p, \quad i = 1, 2,$$

iar

$$X_1 + X_2 : \begin{pmatrix} 0 & 1 & 2 \\ q^2 & 2pq & p^2 \end{pmatrix},$$
$$X_1X_2 : \begin{pmatrix} 0 & 1 \\ 2pq + q^2 & p^2 \end{pmatrix}$$

au media

$$M[X_1 + X_2] = 2pq + 2p^2 = 2p, M[X_1X_2] = p^2.$$

Calculăm şi dispersiile:

$$M[X_i^2] = p, D[X_i] = p^2 - p = pq,$$
$$M[(X_1 + X_2)^2] = 1^2 2pq + 2^2 p^2, D[X_1 + X_2] = 2pq + 4p^2 - 4p^2 = 2pq,$$
$$D[X_1X_2] = M[(X_1X_2)^2] - (M[X_1X_2])^2 = p^2 - p^4 = p^2q^2.$$

Observaţia 4.1.6 *Denumirea de medie este îndreptăţită dacă ţinem seama de sensul ei practic. Presupunem că am repetat de N ori o experienţă care ne-a condus la variabila aleatoare X. Dacă valoarea x_1 este luată de n_1 ori, valoarea x_2 de n_2 ori,..., valoarea x_m de n_m ori, unde $n_1 + n_2 + \ldots + n_m = N$, atunci suma valorilor luate de variabila aleatoare X este $n_1x_1 + n_2x_2 + \ldots + n_mx_m$, iar media aritmetică a valorilor luate de X va fi*

$$\frac{n_1x_1 + n_2x_2 + \ldots + n_mx_m}{N} = \frac{n_1}{N}x_1 + \frac{n_2}{N}x_2 + \ldots + \frac{n_m}{N}x_m.$$

Dar $\frac{n_1}{N} = p_1$ reprezintă raportul dintre numărul cazurilor în care s-a luat valoarea x_1 şi numărul total de experimentări, adică p_1. La fel $\frac{n_2}{N} = p_2, \ldots, \frac{n_m}{N} = p_m$. Media lui X ne arată la ce valoare putem să ne aşteptăm pentru media aritmetică a unui mare număr de valori ale lui X, obţinute în urma repetării experienţei date.

Proprietăţi ale mediei

1. Valoarea medie a unei constante este egală cu constanta,

$$X : \begin{pmatrix} c \\ 1 \end{pmatrix}, \quad M[X] = c.$$

2. Valoarea medie a unei variabile aleatoare care admite medie este cuprinsă între cea mai mică şi cea mai mare dintre valorile posibile ale variabilei aleatoare. Fie $a = \inf_i x_i$ şi $A = \sup_i x_i$. Atunci

$$\left.\begin{array}{l} M[X] = \sum\limits_{i\geq 1} x_i p_i \geq a \sum\limits_{i\geq 1} p_i = a, \\ M[X] = \sum\limits_{i\geq 1} x_i p_i \leq A \sum\limits_{i\geq 1} p_i = A, \end{array}\right. \Leftrightarrow a \leq M[X] \leq A.$$

3. Dacă X, Y sunt variabile aleatoare discrete care admit medie atunci

$$M[X+Y] = M[X] + M[Y]$$

Într-adevăr,

$$\begin{aligned} M[X+Y] &= \sum_{i\geq 1}\sum_{j\geq 1} (x_i + y_j)\, p_{ij} = \sum_{i\geq 1} x_i \sum_{j\geq 1} p_{ij} + \sum_{j\geq 1} y_j \sum_{i\geq 1} p_{ij} \\ &= \sum_{i\geq 1} x_i p_i + \sum_{j\geq 1} y_j q_j = M[X] + M[Y]. \end{aligned}$$

4. Dacă X, Y sunt variabile aleatoare discrete care admit medie sunt independente atunci media produsului este egală cu produsul mediilor variabilelor considerate.

$$M[XY] = M[X]M[Y].$$

   Fie
$$M[XY] = \sum_{i\geq 1}\sum_{j\geq 1} x_i y_j p_{ij}.$$

   Dacă variabilele sunt independente, se produc simplificări importante deoarece putem scrie $p_{ij} = p_i q_j$ şi atunci

$$M[XY] = \sum_{i\geq 1}\sum_{j\geq 1} x_i y_j p_i q_j = (\sum_{i\geq 1} x_i p_i)(\sum_{j\geq 1} y_j q_j) = M[X]M[Y].$$

5. Dacă X este o variabilă aleatoare discretă care admite medie şi $a \in \mathbb{R}$ o constantă, atunci au loc relaţiile:

$$M[a+X] = a + M[X], \quad M[aX] = aM[X].$$

Proprietăţi ale dispersiei

1. Dispersia unei constante este nulă deoarece, ţinând seama de proprietăţile mediei,

$$D[c] = M[c^2] - (M[c])^2 = 0.$$

2. $D[aX] = a^2 D[X]$ deoarece

$$D[aX] = \sum_{i\geq 1} p_i (a x_i - am)^2 = \sum_{i\geq 1} p_i a^2 (x_i - m)^2 = a^2 \sum_{i\geq 1} p_i (x_i - m)^2 = a^2 D[X].$$

3. Dacă variabilele aleatoare $X_1, X_2, \ldots, X_k$ sunt independente, atunci dispersia sumei este egală cu suma dispersiilor.

$$D[X_1 + X_2 + \ldots + X_k] = D[X_1] + D[X_2] + \ldots + D[X_k].$$

Demonstrăm pentru două variabile aleatoare. Într-adevăr, fie $Y = X_1 + X_2$ atunci

$$\begin{aligned} D[Y] &= M[Y^2] - (M[Y])^2 = M[(X_1 + X_2)^2] - (M[X_1 + X_2])^2 \\ &= M[X_1^2 + X_2^2 + 2X_1X_2] - (M[X_1] + M[X_2])^2 \\ &= M[X_1^2] + M[X_2^2] + 2M[X_1X_2] - (M[X_1])^2 - 2M[X_1]M[X_2] - (M[X_2])^2 \\ &= M[X_1^2] - (M[X_1])^2 + M[X_2^2] - (M[X_2])^2 = D[X_1] + D[X_2] \end{aligned}$$

4. Două variabile aleatoare care diferă printr-o constantă au dispersiile egale.

Mai general, combinând Proprietăţile 2 şi 3, avem

$$D[\sum_{i=1}^{k} a_i X_i] = \sum_{i=1}^{k} a_i^2 D[X_i].$$

În particular,

$$D[X - Y] = D[X] + D[Y].$$

Arătăm că dispersia unei variabile aleatoare X este, într-un anume sens, cea mai bună valoare care caracterizează împrăştierea valorilor $x_1, x_2, \ldots, x_n$ sau, cu alte cuvinte, $M[X]$ este punctul cel mai potrivit faţă de care trebuie să măsurăm devierile acestor valori. Acest lucru rezultă din următoarea propoziţie.

Propoziţia 4.1.7 *Dacă X este o variabilă aleatoare care ia valorile $x_1, x_2, \ldots, x_n, ..$ cu probabilităţile $p_1, p_2, \ldots, p_n, ...$, iar m este media, atunci*

$$D[X] = \sum_{i \geq 1} p_i (x_i - m)^2 \leq \sum_{i \geq 1} p_i (x_i - x)^2, \quad \forall x \in \mathbb{R}.$$

Demonstraţie. Putem scrie $x_i - x = (x_i - m) + (m - x), \ \ i = \overline{1, n}$ şi

$$\begin{aligned} & p_i (x_i - x)^2 = \sum_{i \geq 1} p_i [(x_i - m) + (m - x)]^2 \\ &= \sum_{i \geq 1} p_i (x_i - m)^2 + 2(m - x) \sum_{i \geq 1} p_i (x_i - m) + (m - x)^2 \\ &= \sum_{i \geq 1} p_i (x_i - m)^2 + 2(m - x)(m - m) + (m - x)^2 = D[X] + (m - x)^2 \leq D[X]. \end{aligned}$$ □

Definiţia 4.1.8 *Fie X şi Y două variabile aleatoare care admit medie şi dispersie.* ***Covarianţa*** *(sau* ***corelaţia****) lui X şi Y este numărul real*

$$\operatorname{cov}(X, Y) = M\left[(X - M[X])(Y - M[Y])\right].$$

În cazul în care $\operatorname{cov}(X, Y) = 0$ *se spune că X şi Y sunt* ***necorelate****.*

Teorema 4.1.9 *Fie X şi Y două variabile aleatoare care admit medie. Atunci:*
a) $\operatorname{cov}(X, Y) = M[XY] - M[X]M[Y]$.
b) dacă X şi Y sunt independente, atunci X şi Y sunt necorelate. Reciproca nu este adevărată.

Demonstraţie. a) $\operatorname{cov}(X,Y) = M\left[XY - YM\left[X\right] - XM\left[Y\right] + M\left[X\right]M\left[Y\right]\right]$. Notând

$$M\left[X\right] = m_1, M\left[Y\right] = m_2,$$

atunci

$$\begin{aligned}\operatorname{cov}(X,Y) &= M\left[XY - Ym_1 - Xm_2 + m_1m_2\right] \\ &= M\left[XY\right] - m_1M\left[X\right] - m_2M\left[Y\right] + m_1m_2 = M\left[XY\right] - M\left[X\right]M\left[Y\right].\end{aligned}$$

b) Dacă X şi Y sunt independente, atunci $M[XY] = M[X]M[Y] \Rightarrow \operatorname{cov}(X,Y) = 0$. □

Teorema 4.1.10 *Fie X şi Y două variabile aleatoare care admit medie şi dispersie. Atunci: $D[X+Y] = D[X] + D[Y] + 2\operatorname{cov}(X,Y)$.*

Demonstraţie. Avem

$$\begin{aligned}D[X+Y] &= M\left[\left(X+Y-M\left[X+Y\right]\right)^2\right] = M\left[\left(\left(X-M\left[X\right]\right)+\left(Y-M\left[Y\right]\right)\right)^2\right] = \\ &= M\left[\left(X-M\left[X\right]\right)^2 + \left(Y-M\left[Y\right]\right)^2 + 2\left(X-M\left[X\right]\right)\left(Y-M\left[Y\right]\right)\right] \\ &= M\left[\left(X-M\left[X\right]\right)^2\right] + M\left[\left(Y-M\left[Y\right]\right)^2\right] + 2M\left[\left(X-M\left[X\right]\right)\left(Y-M\left[Y\right]\right)\right] \\ &= D[X] + D[Y] + 2\operatorname{cov}(X,Y).\end{aligned}$$

□

Teorema 4.1.11 (Inegalitatea lui Cebâşev) *Fie X o variabilă aleatoare care admite medie şi dispersie finite. Atunci, oricare ar fi $\varepsilon > 0$, are loc inegalitatea*

$$P(\{\mid X - m \mid < \varepsilon\}) \geq 1 - \frac{D[X]}{\varepsilon^2}, \quad m = M[X]. \tag{4.2}$$

Demonstraţie. O vom face în cazul continuu. □

O formă echivalentă a teoremei este: dacă media distribuţiei X este m şi abaterea medie pătratică este $\sigma = \sqrt{D[X]}$, probabilitatea ca valorile variabilei aleatoare să fie mai departate de medie cu cel puţin $k\sigma$ este cel mult $\frac{1}{k^2}$.

$$P(\{|X - m| \geq k\sigma\}) < \frac{1}{k^2}. \tag{4.3}$$

Dacă luăm $\epsilon = k\sigma$ această inegalitatea (4.2) se obţine $P(\{|X-m| < k\sigma\}) \geq 1 - \frac{1}{k^2}$ sau $P(\{m - k\sigma < X < m + k\sigma\}) \geq 1 - \frac{1}{k^2}$. Dar evenimentele $\{|X-m| < k\sigma\}$ şi $\{|X-m| \geq k\sigma\}$ sunt contrare, rezultă (4.3).

De exemplu, pentru $k = 3$ obţinem $P(\{m - 3\sigma < X < m + 3\sigma\}) \geq 1 - \frac{1}{9} = \frac{8}{9} = 0.88889$. Deci, cu o probabilitate cuprinsă între 0.88889 şi 1, orice variabilă ia valori cuprinse în intervalul $(m - 3\sigma, m + 3\sigma)$.

Exerciţiul 3 Numărul clienţilor care vizitează un showroom sâmbăta dimineaţă este o variabilă aleatoare cu media $m = 18$ şi $\sigma = 2,5$. Cu ce probabilitate putem presupune că numărul vizitatorilor va fi cuprins între 8 şi 28 de clienţi.

Deoarece $8 < X < 28 \Rightarrow -10 < X - 18 < 10 \Rightarrow \epsilon = 10$ şi folosind relaţia (4.2) obţinem: $P(\{\mid X - 18 \mid < 10\}) \geq 0.9375$,deci probabilitatea ca numărul vizitatorilor să fie cuprins între 8 şi 28 de clienţi este de cel puţin 0.9375.

Avantajul inegalităţii lui Cebâşev este că se poate aplica oricărei distribuţii pentru care media şi abaterea medie pătratică există.

Dezavantajul este că această inegaliate dă o margine inferioară a probabilităţii ca valorile variabilei aleatoare să fie cuprinse într-un interval de forma $(m - \epsilon, m + \epsilon)$.