Curs 3: Variabile aleatoare. Distribuţia de probabilitate. Funcţia de repartiţie. Media. Dispersia.

1.1 Variabile aleatoare: generalităţi

Teoria probabilităţilor studiază variabile aleatoare, concept pe care il vom prezenta în cele ce urmează. Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate asociat unui experiment. Dacă experimentul constă în observarea/ măsurarea valorilor unei variabile (diametrul unei piese, suma punctelor aruncării a două zaruri), valori care sunt afectate de întâmplare, atunci variabila respectivă se numeşte variabilă aleatoare.

Din punct de vedere matematic o variabilă aleatoare este o funcţie $X : \Omega \to \mathbb{R}$ pentru care are loc

$$\{\omega \in \Omega : X(\omega) \leq x\} =^{not} (X \leq x) \in \mathcal{K}, \quad \forall x \in \mathbb{R}. \tag{1}$$

Spre deosebire de funcţiile uzuale pentru care se precizează legea de corespondenţă (de exemplu, $f(x) = e^{-x}x^2$), în cazul variabilelor aleatoare prezintă interes nu legea de corespondenţă, ci probabilităţile evenimentelor ca variabila să ia valori intr-un interval de forma (a, b), $[a, b)$, $(a, b]$, $[a, b]$ $(-\infty, b)$, $(-\infty, b]$, (a, ∞), $[a, \infty)$. Asocierea probabilităţii evenimentelor relativ la valorile unei variabile aleatoare diferă în funcţie de tipul variabilei: discretă sau continuă. Variabilele aleatoare discrete iau un număr finit de valori sau o mulţime infinit numărabilă de valori, în timp ce variabilele aleatoare continue pot lua orice valori dintr-un interval real, marginit sau nu.

Exemplul 1. Pe internet informaţia se transmite sub formă de pachete, via routere (un router este un calculator specializat să transfere pachetele către un altul mai apropiat de destinatar). Un router stochează într-un buffer finit pachetele ce intră, înainte de a le retransmite mai departe. Numărul de pachete ce stau la coadă să fie transmise este o variabilă aleatoare discretă X. Evident că nu putem prescrie o expresie sau formulă care să indice valoarea precisă a numărului de pachete din buffer, la un moment dat. Dar este util pentru studiul performanţei reţelei să cunoaştem probabilitătile ca această variabilă să ia valoarea n, $n \in \{0, 1, 2, \ldots N\}$.

Într-un acelaşi experiment de monitorizare a intrării pachetelor în buffer, putem urmări nu numai numărul pachetelor, ci şi intervalele de timp dintre sosirea a două pachete.

Variabila aleatoare, Y, ce indică lungimea intervalelor dintre două sosiri nu mai ia valori discrete, pentru că o astfel de lungime poate fi, teoretic, orice număr real pozitiv. Din acest punct de vedere variabila Y, este o variabilă aleatoare continuă (a nu se confunda cu noţiunea de continuitate din analiza matematică).

1.2 Distribuţia de probabilitate a unei variabile aleatoare discrete

Presupunem că un experiment aleator poate avea n realizări, ce pot fi cuantificate numeric de valorile reale $x_1, x_2, \ldots, x_n$. Notăm cu X variabila ce ia aceste valori. Evenimentele elementare ale experimentului sunt

$$E_i = (X = x_i), i = \overline{1,n}. \tag{2}$$

$(X = x_i)$ se citeşte "evenimentul ca X să ia valoarea x_i". Notăm cu $p_i = P(X = x_i)$ probabilitatea evenimentului $(X = x_i)$, $i = \overline{1,n}$. Evenimentul sigur al experimentului este $\Omega = \cup_{i=1}^n (X = x_i)$. Cum evenimentele $(X = x_1)$, $(X = x_2), \ldots, (X = x_n)$ sunt mutual exclusive (producerea unuia exclude realizarea celorlalte), avem:

$$P(\Omega) = P(X = x_1) + P(X = x_2) + \cdots P(X = x_n). \tag{3}$$

Probabilitatea evenimentului sigur fiind 1, rezultă că:

$$p_1 + p_2 + \cdots p_n = 1. \tag{4}$$

Tabloul (matricea cu două linii) ce înregistrează pe prima linie valorile variabilei X şi pe a doua probabilităţile p_i asociate, tablou notat la fel, cu X:

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad \sum_{i=1}^{n} p_i = 1 \tag{5}$$

sau pentru o variabilă cu o mulţime numărabilă de valori $(x_i)_{i \in \mathbb{N}}$:

$$X = \begin{pmatrix} x_k \\ p_k \end{pmatrix}, \quad \sum_{k=0}^{\infty} p_k = 1 \tag{6}$$

defineşte *distribuţia de probabilitate* a variabilei aleatoare. Numele de distribuţie vine de la faptul că evidenţiază cum sunt distribuite probabilităţile p_k, cu $\sum_{k=1}^n p_k = 1$, valorilor posibile ale variabilei.

Riguros matematic, o variabilă aleatoare discretă asociată unui experiment aleator ce are evenimentul sigur Ω şi familia evenimentelor ce se pot produce în acest experiment este familia de evenimente, $\mathcal{K}$, se defineşte astfel:

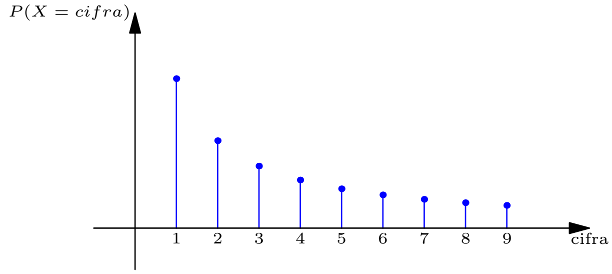

Fig. 1: Vizualizarea distribuţiei de probabilitate a unei v.a. discrete Benford. O astfel de vizualizare se numeşte vizualizare prin impulsuri.

Definiţia 1.2.1 *Fie* $(\Omega, \mathcal{K}, P)$ *un câmp de probabilitate asociat unui experiment aleator. O variabilă aleatoare discretă* X *asociată acestui experiment este o aplicaţie* $X : \Omega \to \mathbb{R}$*, ce ia valorile* $x_i \in \mathbb{R}$*,* $i \in I$*, cu* I *mulţime cel mult numărabilă, astfel încât mulţimea* $\{\omega \in \Omega \,|\, X(\omega) = x_i\}$ *este din* $\mathcal{K}$*, oricare ar fi* $i \in I$*, adică* $\{\omega \in \Omega \,|\, X(\omega) = x_i\}$ *este un posibil eveniment în experimentul considerat.*

Distribuţia de probabilitate a unei variabile aleatoare discrete, X, se vizualizează marcând pe o axă orizontală valorile x_i, şi segmente verticale între punctele de coordonate $(x_i, 0)$ şi (x_i, p_i). In fig.1 ilustrăm distribuţia de probabilitate a primei cifre, numită distribuţia Benford.

Distribuţia de probabilitate Benford a fost definită după ce s-a constatat experimental că multimile de înregistări numerice (date de pe piaţa financiară, date rezultate din simulări numerice pe calculator, în general date arbitrare şi nu coduri cum ar fi cod numeric personal, etc) au distribuţia de probabilitate a primei cifre (cifra din poziţia cea mai semnificativă; de exemplu prima cifră a numărului 4321.05 este 4) cu o structură neaşteptată. Şi anume, cifra ce se află în prima poziţie a unui număr, ce are frecvenţa cea mai mare este cifra 1, apoi descrescător 2,3,4,5,6,7,8,9. Fizicianul Benford a definit variabila aleatoare:

$$X = \begin{pmatrix} 1 & 2 & \cdots & k & \cdots & 8 & 9 \\ \log_{10}(2/1) & \log_{10}(3/2) & \cdots & \log_{10}((k+1)/k) & \cdots & \log_{10}(9/8) & \log_{10}(10/9) \end{pmatrix}$$

numită acum variabila aleatoare Benford. Probabilitatea ca un număr să aibă în prima poziţie cifra k este:

$$P(X = k) = p_k = \log_{10} \frac{k+1}{k}, \quad k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\},$$

Calculând logaritmii avem următorul tabel al probabilităţilor de apariţie a cifrelor în prima poziţie a numerelor:

cifra k	P(X=k)
1	0.30102999566398
2	0.17609125905568
3	0.12493873660830
4	0.09691001300806
5	0.07918124604760
6	0.06694678963061
7	0.05799194697769
8	0.05115252244738
9	0.04575749056068

Având o variabilă aleatoare discretă X, pornind de la evenimentele elementare $(X = x_i)$ ale experimentului aleator construim evenimente de tipul $(X < x)$, $(X \leq x)$, $(X > x)$, $X \geq x)$, $(a < X < b)$, $(a \leq X < b)$, $(a \leq X \leq b)$, $(a < X \leq b)$. Cum aceste evenimente se pot scrie ca reuniune de evenimentele mutual exclusive $(X = x_i)$, probabilitatea lor este suma probabilităţilor evenimentelor elementare componente.

Exemplul 2. Fie X variabila aleatoare având distribuţia de probabilitate:

$$X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0.35 & 0.4 & 0.15 & 0.05 & 0.03 & 0.02 \end{pmatrix} \tag{7}$$

Să se calculeze probabilitatea ca X să ia valoarea cel mult 3.

Rezolvare:

Evenimentul a cărui probabilitate se cere este: $(X \leq 3) = (X = 0) \cup (X = 1) \cup (X = 2) \cup (X = 3)$. Astfel $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$ $= 0.35 + 0.4 + 0.15 = 0.9 + 0.05$.

Pe lângă forma tablou a unei variabile aleatoare discrete X, se mai specifică distribuţia de probabilitate printr-o funcţie definită pe mulţimea valorilor pe care variabila le poate lua, $p_X : D \to [0, 1]$, $p_X(x) = P(X = x)$ (probabilitatea ca variabila aleatoare X să ia valoarea x).

Exemplul 3. Funcţia $p : \mathbb{R} \to \mathbb{R}$ definită prin

$$p(x) = \begin{cases} \frac{1}{10}(5 - x) & \text{dacă } x = 1, 2, 3, 4 \\ 0 & \text{în rest} \end{cases} \tag{8}$$

este distribuţia de probabilitate a unei variabile aleatoare discrete X, deoarece pentru orice $x \in \{1, 2, 3, 4\}$ avem $0 \leq \frac{1}{10}(5 - x) \leq 1$, funcţia este nenulă pe mulţimea $D = \{1, 2, 3, 4\}$ şi $\sum_{x=1}^{4}(5 - x)/10 = 1$.

Exemplul 4. Funcţia $p : \mathbb{N}^* \to [0,1]$ definită prin:

$$p(k) = (0.7)^{k-1}0.3, \quad \forall\, k \in \mathbb{N}^* = \{1, 2, \dots, n, \dots\}$$

este distribuţia de probabilitate a unei variabile aleatoare discrete deoarece,

$$\sum_{k=1}^{\infty}(0.7)^{k-1}0.3 = 1,$$

mai precis limita şirului sumelor parţiale ale seriei este:

$$\begin{array}{l}\lim_{n\to\infty} 0.3 + (0.7)0.3 + (0.7)^2 0.3 + \cdots + (0.7)^{n-1}0.3 = \\ \lim_{n\to\infty}(0.3)(1 + 0.7 + (0.7)^2 + \cdots + (0.7)^{n-1}) = \\ (0.3)\lim_{n\to\infty} \dfrac{1-(0.7)^n}{1-0.7} = 1.\end{array}$$

Un exemplu particular de distribuţie de probabilitate discretă este *distribuţia uniformă* pe o mulţime finită. O variabilă aleatoare X, discretă, ce ia valori în multimea D de n elemente şi are distribuţia de probabilitate, p_X, definită prin:

$$p_X(x) = \frac{1}{n}, \quad \forall\, x \in D, \tag{9}$$

se numeşte *variabilă aleatoare discretă, uniform distribuită* pe D. Numele este justificat de faptul ca valorile sunt echiprobabile. Dacă $D = \{x_1, x_2 \dots, x_n\}$, atunci variabila aleatoare

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ \dfrac{1}{n} & \dfrac{1}{n} & \dots & \dfrac{1}{n} \end{pmatrix}$$

are distribuţia uniformă discretă.

Pentru a evidenţia modul în care variază probabilităţile evenimentelor de forma $(X \leq x)$, în funcţie de x se asociază unei variabile aleatoare o funcţie, denumită funcţia de repartiţie (cumulative distribution function):

Definiţia 1.2.2 *Fie X o variabilă aleatoare discretă. Funcţia $F_X : \mathbb{R} \to [0,1]$, definită prin $F_X(x) = P(X \leq x)$ (probabilitatea ca variabila aleatoare X să ia valori mai mici sau egale cu x), se numeşte funcţia de repartiţie a variabilei X.*

Dacă X ia valorile $x_1, x_2, \dots, x_n$ şi $P(X = x_i) = p_i$, atunci valoarea funcţiei de repartiţie în $x \in \mathbb{R}$ se calculează ca suma probabilităţilor p_i corespunzătoare valorilor $x_i \leq x$:

$$F_X(x) = \sum_{i | x_i \leq x} p_i$$

Graficul funcţiei de repartiţie este o funcţie în scară (Fig.2). Într-adevăr pentru o variabilă aleatoare

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad \sum_{i=1}^{n} p_i = 1 \tag{10}$$

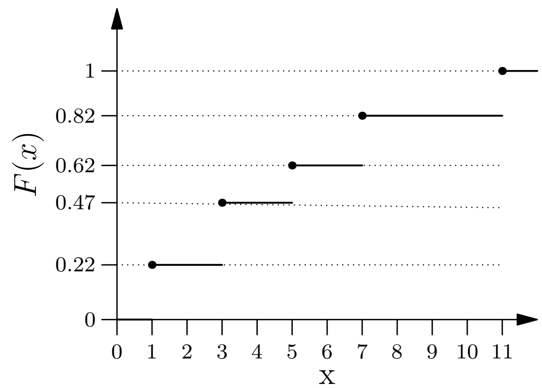

Fig. 2: Graficul functiei de repartitie a variabilei aleatoare X

ale cărei valori sunt date în ordine crescătoare $x_1 < x_2 < \cdots < x_n$ funcţia de repartiţie este:

$$F_X(x) = \begin{cases} 0 & \text{pentru } x < x_1 \\ p_1 & \text{pentru } x_1 \leq x < x_2 \\ p_1 + p_2 & \text{pentru } x_2 \leq x < x_3 \\ \vdots & \\ p_1 + p_2 + \cdots + p_k & \text{pentru } x_k \leq x < x_{k+1} \\ \vdots & \\ 1 & x \geq x_n \end{cases} \tag{11}$$

Atunci când nu există pericol de confuzie, notăm funcţia de repartiţie a variabilei X, fără indicele X.

De exemplu, dacă $X = \begin{pmatrix} 1 & 3 & 5 & 7 & 11 \\ 0.22 & 0.25 & 0.15 & 0.20 & 0.18 \end{pmatrix}$, atunci

$$F_X(x) = \begin{cases} 0 & \text{pentru } x < 1 \\ 0.22 & \text{pentru } 1 \leq x < 3 \\ 0.47 & \text{pentru } 3 \leq x < 5 \\ 0.62 & \text{pentru } 5 \leq x < 7 \\ 0.82 & \text{pentru } 7 \leq x < 11 \\ 1 & x \geq 11 \end{cases} \tag{12}$$

Graficul funcţiei F_X este dat de (Fig.2).

Se observă că funcţia de repartiţie a variabilei aleatoare X este crescătoare, aducă dacă $x_1 < x_2$, atunci $F_X(x_1) \leq F_X(x_2)$. În plus, are loc

$$\lim_{x\to-\infty} F_X(x) = 0 \text{ şi } \lim_{x\to\infty} F_X(x) = 1$$

Funcţia de repartiţie fiind crescătoare implică faptul că mulţimea punctelor ei de discontinuitate este cel mult numărabilă (finită sau numărabilă) şi orice punct de discontinuitate este de speţa întâi.

Funcţia de repartiţie este continuă la dreapta în orice punct $x_0 \in \mathbb{R}$, adică $\lim\limits_{x \to x_0, x > x_0} F_X(x) = F_X(x_0)$. Are loc relaţia

$$P(X = x_0) = F_X(x_0) - \lim_{x \to x_0, x < x_0} F_X(x) \tag{13}$$

Reciproc, se poate demonstra că dacă o funcţie $F : \mathbb{R} \to \mathbb{R}$ verifică proprietăţile

- F este crescătoare
- $\lim\limits_{x \to -\infty} F_X(x) = 0$ şi $\lim\limits_{x \to \infty} F_X(x) = 1$
- F este continuă la dreapta în orice punct,

atunci există o variabilă aleataore X, definită pe un anumit spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ având pe F ca funcţie de repartiţie, adică $F_X = F$.

1.3 Funcţii de o variabilă aleatoare

Fie X o variabilă aleatoare discretă de valori x_i şi probabilităţi corespunzătoare p_i, $i \in I$ şi D_X mulţimea valorilor sale. Dacă $\varphi : E \supset D_X \to \mathbb{R}$ este o funcţie reală continuă sau cu discontinuităţi de prima spetă, atunci $Y = \varphi \circ X$ este de asemenea o variabilă aleatoare discretă, având valorile $y_i = \varphi(x_i)$, $i \in I$.

Dacă X are distribuţia de probabilitate:

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

să determinăm distribuţia de probabilitate a variabilei aleatoare $Y = \varphi(X)$:

$$P(Y = y_i) = P(\varphi(X) = y_i) = P(X \in \varphi^{-1}(\{y_i\}), \tag{14}$$

unde $\varphi^{-1}(\{y_i\}) = \{x \in D_X \mid \varphi(x) = y_i\}$ este preimaginea mulţimii $\{y_i\}$, adică mulţimea elementelor din D_X ce sunt aplicate în y_i.

• dacă φ este injecţie, atunci preimaginea lui y_i este un singur element, x_i, şi deci:

$$P(Y = y_i) = P(\varphi(X) = y_i) = P(X \in \varphi^{-1}(\{y_i\}) = P(X \in \{x_i\}) = P(X = x_i) = p_i,$$

adică distribuţia de probabilitate a variabilei imagine este:

$$Y = \begin{pmatrix} \varphi(x_1) & \varphi(x_2) & \dots & \varphi(x_n) \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

• dacă φ nu este injecţie, atunci în tabloul de mai sus pot apărea mai multe valori egale, de exemplu $\varphi(x_3) = \varphi(x_7) = \varphi(x_8) = y$. În acest caz se rescrie tabelul de definiţie al lui Y, luând pe prima linie fiecare valoare posibilă, y, o singură dată, iar probabilitatea ca $\varphi(X)$ să ia valoarea y este $P(\varphi(X) = y) = P(X \in \varphi^{-1}(\{y\})) = \sum_i P(X = x_i)$, unde x_i sunt toate valorile variabilei X aplicate de φ în y.

Exemplul 5. Se dă variabila aleatoare:

$$X = \begin{pmatrix} -3 & -1 & 0 & 1 & 3 & 5 \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} & \frac{1}{3} & \frac{1}{12} & \frac{1}{4} \end{pmatrix} \tag{15}$$

Considerând funcţia $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(x) = x^2$, variabila aleatoare $X^2 = \varphi(X)$ ia valorile $\{(-3)^2, (-1)^2, 0, 1, 9, 25\} = \{0, 1, 9, 25\}$. Fie $y \in \{0, 1, 9, 25\}$. $P(X^2 = y) = P(X \in \varphi^{-1}(y)) = P((X = -\sqrt{y}) \cup (X = \sqrt{y}))= P(X = -\sqrt{y}) + P(X = \sqrt{y})$.

Astfel avem:
$P(X^2 = 0) = P(X = 0) = 1/12$
$P(X^2 = 1) = P(X = -1) + P(X = 1) = 1/6 + 1/3 = 1/2$
$P(X^2 = 9) = P(X = -3) + P(X = 3) = 1/12 + 1/12 = 1/6$.
$P(X^2 = 25) = P(X = -5) + P(X = 5) = 0 + 1/4 = 1/4$.
Deci distribuţia de probabilitate a variabilei X^2 este:

$$X^2 = \begin{pmatrix} 0 & 1 & 9 & 25 \\ \frac{1}{12} & \frac{1}{2} & \frac{1}{6} & \frac{1}{4} \end{pmatrix}. \tag{16}$$

1.4 Media şi dispersia unei variabile aleatoare discrete

Unei variabile aleatoare i se asociază câteva valori care dau informaţii sumarizate relativ la distribuţia sa de probabilitate.

Definiţia 1.4.1 *Fie X o variabilă aleatoare discretă de valori x_i, şi $p_i = P(X = x_i)$, $i = \overline{1,n}$, probabilităţile cu care X ia aceste valori. Valoarea medie a variabilei X este numărul notat $M(X)$,*

$$M(X) = \sum_{i=1}^{n} x_i p_i. \tag{17}$$

Pentru o variabila aleatoare X ce ia o mulţime infinit numărabilă de valori, x_i, $i \in I$, valoarea medie este $M(X) = \sum_{i \in I} x_i p_i$, dacă seria $\sum_{i \in I} x_i p_i$ este absolut convergentă (adică seria $\sum_{i \in I} |x_i p_i|$ este convergentă) şi X nu are valoare medie dacă seria nu este absolut convergentă.

Cu alte cuvinte media unei variabile aleatoare discrete este media ponderată a valorilor sale, ponderile fiind probabilităţile $p_i = P(X = x_i)$.

În cazul unei variabile aleatoare discrete, uniform distribuite:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ \dfrac{1}{n} & \dfrac{1}{n} & \dots & \dfrac{1}{n} \end{pmatrix}$$

valoarea medie, $M(X) = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.

Exemplu de variabilă aleatoare discretă ce nu are valoare medie: Fie variabila aleatoare:

$$X = \begin{pmatrix} k \\ a/k^2 \end{pmatrix}, k = 1, 2, \ldots, n \ldots \tag{18}$$

unde $a \in \mathbb{R}$, este astfel încât $\displaystyle\sum_{k=1}^{\infty} \frac{a}{k^2} = 1$. Atunci $\displaystyle\sum_{k=1}^{\infty} k\frac{a}{k^2} = \sum_{k=1}^{\infty} \frac{a}{k}$ este o serie divergentă (seria armonică) şi deci variabila aleatoare nu are valoare medie.

În cele ce urmează definim şi prezentăm rezultate relativ la variabile discrete cu un număr finit de valori. Corespondentele acestora pentru variabile cu o mulţime infinit numărabilă de valori se extind în mod evident, dacă acestea au valori medii.

O variabilă aleatoare discretă X ce ia o singură valoare, $a \in \mathbb{R}$, are distribuţia de probabilitate:

$$X = \begin{pmatrix} a \\ 1 \end{pmatrix} \tag{19}$$

şi valoarea medie $M(X) = a$. Notăm în cele ce urmează o astfel de variabilă aleatoare, cu a, adică ilustrăm direct că este o constantă.

Definiţia 1.4.2 *Fie X o variabilă aleatoare discretă,*

$$X = \begin{pmatrix} x_i \\ p_i \end{pmatrix}, i \in I$$

având valoarea medie m. Dispersia variabilei aleatoare X, notată $D^2(X)$ sau $\sigma^2(X)$, este numărul real definit prin:

$$\sigma^2(X) = \sum_{i=1}^{n} (x_i - m)^2 p_i$$

Abaterea standard a variabilei X este $\sqrt{\sigma^2(X)}$ şi se notează prin $\sigma(X)$.

Remarcăm că dacă valorile x_i ale variabilei aleatoare X sunt mai apropiate de media m, atunci pătratele diferenţelor $(x_i - m)$ sunt mici şi deci dispersia sau împrăştierea valorilor în jurul mediei m este redusă. Se spune că media este în acest caz reprezentativă pentru valorile variabilei. Dacă însă diferenţele respective sunt mai mari, dispersia este mai mare şi deci media mai puţin reprezentativă. În concluzie, dispersia unei variabile aleatoare este o măsură a gradului de împrăştiere a valorilor variabilei aleatoare în jurul mediei sale.

O variabilă aleatoare constantă, $X = a$, are dispersia $\sigma^2(X) = 0$.

Valoarea medie a unei variabile aleatoare se foloseşte în analiza algoritmilor la studiul performanţei algoritmilor.

Calculul performanţei medii a unui algoritm. Fie $\mathcal{A}$ un algoritm. Pentru a determina performanţa medie a acestuia se stabilesc următoarele:

• Se defineşte o variabilă aleatoare discretă a intrărilor, adică v.a. ce indică toate intrările posibile de volum n:

$$\mathbb{I} = \begin{pmatrix} I_1 & I_2 & \dots & I_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

Posibilele intrări (date de intrare) pot apare în situaţii practice cu probabilităţile $p_1, p_2, \dots p_k$.

• Se defineşte apoi variabila aleatoare $Y = h(X)$, unde funcţia h asociază unei intrări I_j, numărul operaţiilor pe care algoritmul le efectuează asupra datelor de intrare.

• Numărul mediu de operaţii pe care algoritmul $\mathcal{A}$ le execută ca o funcţie de volumul intrărilor, n, $M(h(\mathbb{I}) = h(I_1)p_1 + h(I_2)p_2 + \cdots + h(I_k)p_k$, se numeşte performanţa medie a algoritmului pentru intrări de volum n. Se mai notează $\mathrm{Avg}_{\mathcal{A}}(n)$.

Un algoritm $\mathcal{A}$ este optimal în cazul mediu pentru o anumită problemă, dacă:

$$\mathrm{Avg}_{\mathcal{A}}(n) \leq \mathrm{Avg}_{\mathcal{B}}(n),$$

oricare ar fi $n \geq 0$ şi oricare ar fi algoritmul $\mathcal{B}$ ce rezolvă problema în discuţie.

Exemplul 6. *Analiza căutării secvenţiale.* Presupunem că algoritmul $\mathcal{A}$, caută un element `x` în tabloul T ale cărui locaţii sunt indexate de la 0 la $n-1$. Dacă x este stocat în tabloul T, algoritmul returnează indicele celei mai din dreapta apariţii a lui x, iar dacă x nu este în T, atunci returnează -1:

```
i ← n − 1;
while(i ≥ 0 şi x ≠ T[i])
   i ← i − 1;
```

Considerăm seturile de date (intrări) $I_0, I_1, \dots, I_{n-1}$. I_k este setul de date în care locaţia cea mai din dreapta (cu cel mai mare indice) în care este stocat x este k. Notăm cu I_n setul de date corespunzător cazului în care x nu este stocat în tablou. Notăm cu $h(I_k)$ numărul de comparaţii pe care le efectuează algoritmul, când intrarea este I_k.

Observăm că $h(I_k) = n-1-k+1 = n-k$, pentru $k \in \{0, 1, 2 \dots, n-1\}$, şi $h(I_n) = n$.

Notăm cu p probabilitatea ca x să fie stocat în tablou. Astfel $1-p = P(x \notin T)$. Presupunem că dacă $x \in T$, atunci poziţiile din tablou în care este stocat x sunt echiprobabile. Atunci:

$$P(I_k) = \frac{p}{n}, \text{pt } k = \overline{0, n-1}, \quad \text{şi} P(I_n) = 1 - p$$

şi astfel performanţa medie a algoritmului pentru intrări de volum n este:

$$\mathrm{Avg}_{\mathcal{A}}(n) = \frac{p}{n}(n - 0 + (n-1) + \cdots + 1) + (1-p)1 = \frac{p(n+1)}{2} + 1 - p$$

• Dacă ştim că $x \in T$, atunci $p = 1$ şi $\mathrm{Avg}_{\mathcal{A}}(n) = \dfrac{n+1}{2}$.

• Dacă $x \notin T$ $p = 0$ şi $\mathrm{Avg}_{\mathcal{A}}(n) = n$;

• Dacă x este stocat în $T[0]$ (cazul cel mai defavorabil) atunci algoritmul efectuează n comparaţii.

În concluzie cazul cel mai defavorabil (cazul în care pt o intrare de volum n se efectuează cel mai mare număr de operaţii) corespunde intrărilor I_0 şi I_n.

Există numeroşi algoritmi pentru care este imposibil să găsim expresia analitică a funcţiei h ce dă numărul de operaţii pe care algoritmul le efectuează asupra unei intrări de volum n. În asemenea cazuri, performanţa medie se deduce prin simulare, generând intrări aleatoare şi estimând performanţa medie, statistic.