Curs 10

Elemente de statistică

10.1 Populaţii statistice

Statistica se poate defini ca fiind ştiinţa care se ocupă cu colectarea, prezentarea, clasificarea, analiza şi interpretarea cantitativă a datelor şi cu aplicarea teoriei probabilităţilor în analiza şi estimarea caracteristicilor (parametrilor) populaţiei.

Definiţia 10.1.1 Populaţia statistică *este o mulţime de elemente supusă cercetării statistice. Un element al acestei populaţii se numeşte* unitate statistică.

Definiţia 10.1.2 *O proprietate comună tuturor unităţilor statistice se numeşte* caracteristică statistică *a populaţiei.*

Caracteristicile pot fi:
- **cantitative**, dacă *pot fi cuantificate* printr-un număr,
- **calitative**, dacă *nu pot fi cuantificate* printr-un număr, ci prin aprecieri de tipul bun, foarte bun, mult, puţin.

Definiţia 10.1.3 *Valoarea numerică a unei caracteristici cantitative, care se modifică de la o unitate statistică la alta, se numeşte* variabilă statistică.

Definiţia 10.1.4 *O submulţime a unei populaţii statistice se numeşte* selecţie *sau* eşantion.

De exemplu, dacă se solicită opinia unei populaţii într-o problemă oarecare este dificil să fie consultată întreaga populaţie şi se recurge la extragerea unui eşantion.

Un eşantion ale cărui unităţi au fost alese la întâmplare se numeşte *eşantion reprezentativ* (*aleator*). Extragerea unui eşantion reprezentativ se realizează astfel încât elementele să aibă şanse egale de a fi extrase. În multe situaţii se folosesc numere aleatoare. De exemplu dacă avem o listă a populaţiei iniţiale $x_1, x_2, \ldots, x_N$ se aleg numere aleatoare $k_1, k_2, \ldots, k_n$ între 1 şi N şi un eşantion aleator este $x_{k_1}, x_{k_2}, \ldots, x_{k_n}$.

Variabilele statistice pot fi *discrete* (ex: numărul de becuri care se ard după 1000 de ore de întrebuinţare) sau *continue* (ex: timpii de defectare a unui număr fixat de n becuri). Aceste rezultate experimentale se prezintă sub forma diverselor scheme grafice care indică o anumită comportare a variabilei statistice. În cazul variabilelor discrete rezultatele sunt prezentate sub forma unui tablou de forma următoare:

X	x_1	x_2	x_3	x_4	$\cdots$	x_k
n	n_1	n_2	n_3	n_4	$\cdots$	n_k

(10.1)

cu $n = \sum_{j=1}^{k} n_j$. Numărul n se numeşte *volumul selecţiei.* Presupunem că în tabel am aranjat astfel încât $x_1 < x_2 < x_3 < \cdots < x_k$.

Frecvenţa absolută a unei valori x_j, $j = \overline{1,k}$ a unei caracteristici este numărul n_j şi reprezintă numărul de unităţi statistice ale populaţiei statistice care corespund valorii x_j a variabilei statistice X.

Frecvenţa relativă a unei valori x_j, $j = \overline{1,k}$ a unei caracteristici este raportul dintre frecvenţa absolută şi volumul selecţiei, adică $f_j = \frac{n_j}{n}$, $j = \overline{1,k}$. Suma frecvenţelor relative este egală cu 1.

Frecvenţa absolută cumulată crescător este suma frecvenţelor absolute valorilor mai mici sau egale cu valoarea x_j: $\sum_{i=1}^{j} n_i$, $j \leq k$.

Frecvenţa relativă cumulată crescător a unei valori x_j, $j = \overline{1,k}$ a unei caracteristici este suma frecvenţelor relative corespunzătoare valorilor mai mici sau egale cu x_j: $\frac{1}{n}\sum_{i=1}^{j} n_i$, $j \leq k$.

Definiţia 10.1.5 *Se numeşte* funcţie de repartiţie empirică *asociată unei v. a. X şi unei selecţii $(x_1, x_2, \ldots, x_n)$ Funcţia*

$$F_n^* : \mathbb{R} \to \mathbb{R}, F_n^*(x) = \frac{\textit{număr de valori } x_j < x}{n} = \frac{k_x}{n}.$$

Teorema de mai jos pune în evidenţă că funcţiile de repartiţie empirice aproximează oricât de bine funcţia reală de repartiţie.

Teorema 10.1.6 *Fie o populatie statistică şi X variabila aleatoare ataşată ei cu functia de repartitie F. Pentru o selectie de volum n: $(x_1, x_2, \ldots, x_n)$ construim ca mai sus funcţia de repartiţie empirică F_n^*. Atunci*

$$\forall \varepsilon > 0 : P\left\{|F(x) - F_n^*| \geq \varepsilon\right\} \overset{n\to\infty}{\to} 0, \quad \varepsilon \textit{ fixat.}$$

Demonstraţie. Să notăm cu $p = P(X < x) = F(x)$, şi cu $F_n^*(x) = \frac{k_x}{n}$ funcţia de repartiţie empirică. Definim v. a. $Y_1, Y_2, \ldots, Y_n$, astfel încât Y_j are valoarea 1 dacă $x_j < x$ si 0 în caz contrar. Variabilele $Y_1, Y_2, \ldots, Y_n$ sunt independente (selecţiile fiind independente) şi au distribuţia

$$\begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}.$$

Avem $M[Y_j] = p$, $D[Y_j] = p(1-p)$. Evident v. a. $Z_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ are $M[Z_n] = p$ şi dispersia $D[Z_n] = \frac{p(1-p)}{n}$.

Aplicăm inegaliatea lui Cebâşev lui Z_n şi găsim

$$P\left\{|F(x) - F_n^*| \geq \varepsilon\right\} = P\left\{|Z_n - p| \geq \varepsilon\right\} \leq \frac{D[Z_n]}{\varepsilon^2} = \frac{p(1-p)}{n} \overset{n\to\infty}{\to} 0.$$

□

10.2 Reprezentarea grafică a datelor statistice

O reprezentare a datelor statistice este sub forma histogramei. *Histograma* se construieşte sub forma unor dreptunghiuri lipite, cu baza pe Ox, mărimea lor fiind egală la bază cu

mărimea intervaluilui de variaţie respectiv. Înălţimea dreptunghiului va fi dată de frecvenţa corespunzătoare fiecărui interval de variaţie. Histograma arată forma de repartiţie, densitatea de repartiţie a frecvenţelor şi gradul de asimetrie.

Exemplul 10.2.1 Punctele obţinute de studenţi care au promovat examenul de matematică şi care cuantifică cunoştinţele lor sunt:

{64, 62, 76, 82, 66, 76, 72, 71, 74, 72, 71, 73, 70, 75, 77, 84, 92, 86, 62, 58, 78, 80, 79, 84, 83, 82, 66, 68, 68, 82, 84, 78, 76, 69, 77, 58, 62, 82, 85, 58, 78, 84, 94, 88, 77, 78, 88, 91, 70, 71, 78, 58, 65, 53, 60, 49, 68, 74, 71, 66, 68, 71, 73, 70, 85, 78, 65, 54, 51, 78, 89, 66, 68, 95, 94, 99, 81, 81, 92, 88, 99, 81, 81}

O reprezentare a acestor date este:

X	49	51	53	54	58	60	62	64	65	66	68	69
n_i	1	1	1	1	4	1	3	1	2	4	5	1
X	70	71	72	73	74	75	76	77	78	79	80	81
n_i	3	5	2	2	2	1	3	3	7	1	1	4
X	82	83	84	85	86	88	89	91	92	94	95	99
n_i	4	1	4	2	1	3	1	1	2	2	1	2

În cazul variabilelor continue, datele se grupează într-un tablou de forma:

X	$[t_1, t_2)$	$[t_2, t_3)$	$\cdots$	$[t_m, t_{m+1})$
n	n_1	n_2	$\cdots$	n_m

.

În acest caz X este variabila statistică, iar n_j este numărul valorilor variabilei statistice aflate în intervalul $[t_j, t_{j+1})$ şi se numeşte *frecvenţa absolută*. La fel $n = \sum_{j=1}^{m} n_j$ este *volumul selecţiei*.

Frecvenţa relativă corespunzătoare întervalului $[t_j, t_{j+1})$ este $\frac{n_j}{n}$ şi reprezintă numărul valorilor variabilei statistice aflate în intervalul $[t_j, t_{j+1})$ împărţit la volumul selecţiei.

Analog se introduc şi celelalte noţiuni de frecvenţa absolută cumulată, frecvenţa relativă cumulată etc.

10.2.1 Gruparea pe clase

În urma oricărei selecţii de volum n dintr-o populaţie de numere se obţine un şir finit de n numere numit *serie statistică* (de volum n). Cum construim o densitate de probabilitate empirică? Pentru a răspunde la această întrebare grupăm termenii unei serii statistice în intervale disjuncte: I_1, I_2,..., I_r, după criterii mai mult sau mai puţin subiective.

Dacă avem o selecţie $x_1, x_2, \ldots, x_n$ de volum n obţinute în urma măsurării unor mărimi fizice, tehnice, economice de care depinde evoluţia unor procese sau fenomene, notăm cu $m = \min_i x_i$ şi $M = \max_i x_i$. Organizăm datele selecţiei în grupe sau subintervale obţinute divizând intervalul $[m, M]$ într-un număr r de subintervale congruente (de aceeaşi lungime). Pentru determinarea numărului de subintervale de grupare poate fi folosită *formula lui H. A. Sturges*:

$$r = [1 + 3.322 \lg n]. \tag{10.2}$$

Definiţia 10.2.2 *Se numeşte* pasul de histogramă *numărul* $h = \frac{M - m}{r}$.

Intervalele de grupare sunt astfel

$$I_1 = [m, m+h)\,, \quad I_2 = [m+h, m+2h)\,, \ldots, \quad I_r = [m+(r-1)h, M]\,.$$

Formula lui Sturges (10.2) iniţial a fost scrisă sub forma echivalentă $r = 1 + \log_2 n$. Într-adevăr, avem $\log_2 n = \dfrac{\lg n}{\lg 2} = \log_2 10 \lg n \approx 3.322 \lg n$.

Notăm cu n_1 numărul acelor valori x_i care sunt situate în intervalul I_1, n_2 numărul acelor valori x_i care sunt situate în intervalul I_2, ..., n_r numărul acelor valori x_i care sunt situate în intervalul I_r. Histograma asociată selecţiei $x_1, x_2, \ldots, x_m$ este reuniunea dreptunghiurilor $D_k = I_k \times [0, n_k]$ (produs cartezian), pentru $1 \leq k \leq r$. *Moda* selecţiei este subgrupa (intervalul I_k) ce corespunde dreptunghiului cel mai înalt.

Deci datele selecţiei $x_1, x_2, \ldots, x_n$ se grupează în r grupe şi există posibilitatea de vizualizare: se alege un sistem ortogonal de axe xOy, cu intervalul $[m, M]$ plasat pe axa Ox (cu originea O eventual în punctul m). Pe intervalul I_1 se construieşte dreptungiul de înălţime n_1 şi baza de lungime h, pe intervalul I_2 se construieşte dreptungiul de înălţime n_2 şi baza de lungime h, etc. Intervalul I_k pentru care n_k este maxim este moda selecţiei. Dacă două subgrupe sunt la fel de înalte şi detaşate de celelalte, se spune că selecţia este bimodală (ex: figura 6.2).

10.2.2 Algoritmul pentru construcţia histogramelor

Date iniţiale: $x_1, x_2, \ldots, x_n$.

Pasul I. Se calculează $m = \min\limits_i x_i$, $M = \max\limits_i x_i$, $r = [1 + 3.322 \lg n]$ şi $h = \dfrac{M-m}{r}$.

Pasul II. Se determină intervalele $I_k = [m + (r-1)\,h, m + rh)$ şi frecvenţele absolute n_k, $1 \leq k \leq r$.

Pasul III. Se constuiesc dreptunghiurile ce formează histograma $D_k = I_k \times [0, n_k]$, $1 \leq k \leq r$.

Reluăm Exemplul 10.2.1.

Pasul I. $M = \max\limits_i x_i = 99$, $m = \min\limits_i x_i = 49$, $n = 83$,

$$r = [1 + 3.322 \lg 83] = 7$$
$$h = \frac{M-m}{r} = \frac{99-49}{7} = 7.14286$$

Pasul II. Se determină intervalele I_k şi frecvenţele n_k, $1 \leq k \leq 7$.

I_k	$[49, 56.1429)$	$[56.1429, 63.2857)$	$[63.2857, 70.4286)$
n_k	3	8	16
I_k	$[70.4286, 77.5714)$	$[77.5714, 84.7143)$	$[84.7143, 91.8571)$
n_k	18	22	8
I_k	$[91.8571; 99]$		
n_k	7		

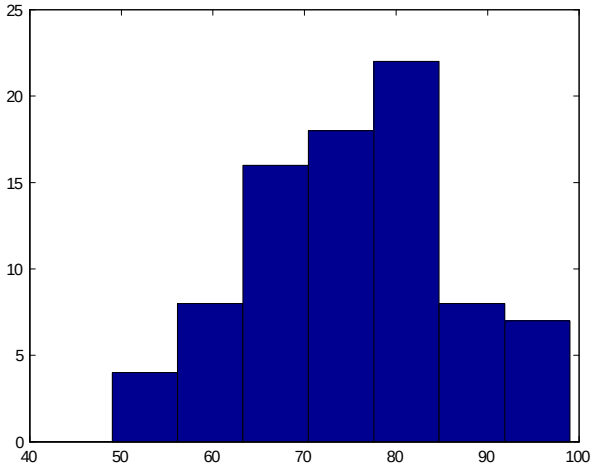
Pasul III. Desenăm histograma.

Algoritmul MATLAB pentru realizarea histogramei este următorul:

```
x=[64, 62, 76, 82, 66, 76, 72, 71, 74, 72, 71, 73, 70, 75, 77, 84, 92, 86, ...
   62, 58, 78, 80, 79, 84, 83, 82, 66, 68, 68, 82, 84, 78, 76, 69, 77, 58, ...
   62, 82, 85, 58, 78, 84, 94, 88, 77, 78, 88, 91, 70, 71, 78, 58, 65, 53, ...
```

```
 60, 49, 68, 74, 71, 66, 68, 71, 73, 70, 85, 78, 65, 54, 51, 78, 89, 66, ...
 68, 95, 94, 99, 81, 81, 92, 88, 99, 81, 81];
n=length(x); m=min(x); M=max(x); k=fix(1+3.322*log10(n));
h=(M-m)/k;
grupe=zeros(k,1);
for i=1:n
 if x(i)==M
 grupe(k)=grupe(k)+1;
 else
 t=fix((x(i)-m)/h)+1;
 grupe(t)=grupe(t)+1;
 end
end
grupe
xprim=zeros(k,1);
for i=1:k
 xprim(i)=m+(2*i-1)/2*h;
end
bar(xprim,grupe,1)
```

Histograma obţinută este următoarea:



Exemplul 10.2.3 Presupunem că într-un atelier lucrează 225 de muncitori şi că s-a analizat într-o perioadă de timp modul în care îşi realizează sarcinile. S-a constatat că 10 dintre ei realizează 101%, etc. conform tabelului de mai jos

10	15	40	23	17	18	36	45	8	5	8
101%	89%	85%	106%	127%	95%	105%	112%	117%	96%	80%

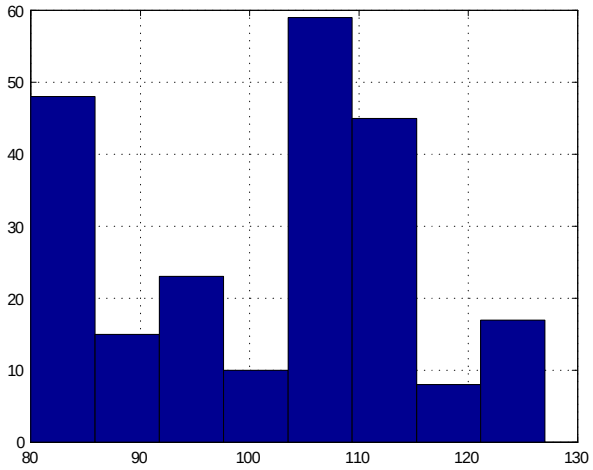
Construim histograma.

În acest caz $m = 80$, $M = 127$, $r = [1 + 3.322 \log_{10} 225] = 8$, $h = \dfrac{127 - 80}{8} = 5.875$. Vom împărţi intervalul în 8 clase de lungime 5.875 fiecare

t	$[80, 85.875)$	$[85.875, 91.75)$	$[91.75, 97.625)$
n_j	48	15	23
t	$[97.625, 103.5)$	$[103.5, 109.375)$	$[109.375, 115.25)$
n_j	10	59	45
t	$[115.25, 121.125)$	$[121.125, 127]$	
n_j	8	17	

Avem $I_1 = [80, 85.4)\,, n_1 = 48, I_2 = [85.4, 90.8)\,, n_1 = 15$, etc. Desenăm histograma

```
x=[10 15 40 23 17 18 36 45 8 5 8];
y=[101 89 85 106 127 95 105 112 117 96 80];
n=length(y); m=min(y); M=max(y); s=sum(x);
k=fix(1+3.322*log10(s));
h=(M-m)/k;
grupe=zeros(k,1);
for i=1:n
 if y(i)==M
     grupe(k)=grupe(k)+x(i);
 else
     t=fix((y(i)-m)/h)+1
     grupe(t)=grupe(t)+x(i)
 end
end
grupe
xprim=zeros(k,1);
for i=1:k
 xprim(i)=m+(2*i-1)/2*h;
end
bar(xprim,grupe,1)
grid
```



Ce concluzii se pot trage din analiza histogramei? În primele patru grupe sunt persoanele care nu-şi realizează sarcinile. Caracteristica numerică numită modă este, pentru atelierul respectiv, "să se realizeze sarcinile între 103.5% şi 109.375%". 59 din cei 225 de muncitori sunt în modă.

10.3 Caracteristici statistice ale datelor experimentale

Definiţia 10.3.1 *Dacă avem o selecţie $x_1, x_2, \ldots, x_n$ de volum n, definim* media de selecţie (empirică) *ca fiind $\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$, adică media aritmetică a celor n valori.*

Dacă datele sunt organizate sub forma unui tabel (10.1) atunci media de selecţie *este $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$, unde n_i reprezintă frecvenţa absolută de apariţie a valorii x_i a variabilei în selecţia considerată şi $\sum_{i=1}^{r} n_i = n$.*

În cazul datelor grupate x_i se înlocuieşte cu x_i^* care este valoarea din mijloc a intervalului $[x_i, x_{i+1})$ deci $\overline{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i^*$, unde r este numărul intervalelor.

Teorema 10.3.2 *Au loc relaţiile: $\overline{x + c} = \overline{x} + c$, $\overline{kx} = k\overline{x}$, $\overline{x + y} = \overline{x} + \overline{y}$.*

Mediana este valoarea $\alpha \in \mathbb{R}$ astfel ca numărul de valori $x_j \leq \alpha$ este egal cu numătul de valori $x_j \geq \alpha$. Deci mediana împarte şirul ordonat de date în două părţi egale. Dacă şirul are $2k+1$ unităţi, atunci mediana coincide cu unitatea de ordin $k+1$, dacă şirul are $2k$ unităţi, mediana este media aritmetică a unităţilor de ordin k şi $k+1$. dacă există mai multe asemenea valori pentru α, atunci ele formează un interval şi mediana este prin definiţie mijlocul acestui interval.

Exemplul 10.3.3 Pentru şirul de date: 2.5, 3.7, 1.4, 0.2, 5.4, 8.9, 4.2 şirul ordonat este 0.2, 1.4, 2.5, 3.7, 4.2, 5.4, 8.9. Avem $2k+1 = 7$, apoi $k = 3$ şi unitatea de ordin $k+1 = 4$ este 3.7 care este mediana. ◊

Dacă $x_{me} = \overline{x}$ atunci repartiţia este simetrică.

Modul selecţiei. Pentru datele negrupate este valoarea observată care are frecvenţa maximă. Pentru date grupate este o valoare din clasa cu cel mai mare număr de observaţii.

Dispersia de selecţie (empirică) este $\overline{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Dispersia de selecţie modificată este $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Observăm că dacă s^2 este mic atunci datele diferă unele de altele foarte puţin, dacă s^2 este mare, datele diferă foarte mult.

Pentru date negrupate avem

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2}{n-1} =$$
$$= \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{n^2\overline{x}^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}$$
$$= \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\overline{x}^2,$$

iar pentru date grupate

$$s^2 = \frac{\sum_{i=1}^{r} n_i x_i^{*2} - \frac{1}{n}\left(\sum_{i=1}^{r} n_i x_i^*\right)^2}{n-1} = \sum_{i=1}^{r} \frac{n_i}{n-1}(x_i^* - \overline{x})^2.$$

Motivaţia introducerii acestei dispersii o vom vedea ulterior în legătură cu conceptul de estimator nedeplasat.

$\overline{x}$ se mai numeşte *speranţa matematică* a selecţiei, dispersia de selecţie se mai numeşte *varianţa* selecţiei.

Amplitudinea unei serii statistice este diferenţa dintre cea mai mare şi cea mai mică valoare a variabilei $\mathcal{A} = x_{\max} - x_{\min}$.

În cazul v. a. continue, amplitudinea este diferenţa dintre limita superioară a ultimului interval şi limita inferioară a primului.

Abaterea medie pătratică de selecţie este $\overline{\sigma} = \sqrt{\overline{\sigma^2}}$.

Abaterea medie pătratică de selecţie modificată este $s = \sqrt{s^2}$.

Coeficientul de variaţie este $v = \dfrac{\sqrt{s^2}}{\overline{x}}$. O selecţie de valori pozitive se consideră omogenă dacă $v < \frac{1}{3}$.