Capitolul 2

Populaţie şi caracteristici. Eşantion

2.1 Definiţii şi clasificări

Pentru început, vom introduce noţiunile (termenii) de bază care se vor regăsi în toate etapele unui studiu statistic.

Def. Se numeşte *populaţie (colectivitate) statistică* orice mulţime nevidă de elemente supusă studiului în legătură cu un anumit fenomen. Elementele sale se mai numesc *indivizi* sau *unităţi statistice*.

Populaţia statistică o vom nota cu Ω, iar elementele sale cu $\omega_1, \omega_2, \ldots$

Obs. 1 Din punct de vedere matematic, populaţia este modelată prin intermediul noţiunii de mulţime.

Def. *Caracteristica statistică* reprezintă însuşirea, proprietatea sau trăsătura comună tuturor indivizilor unei populaţii, şi care stă la baza studiului acelei populaţii. Se notează cu litere mari, $X, Y, \ldots$ şi se mai numeşte *variabilă statistică* sau *variabilă aleatoare*.

Exemple

a) Pentru populaţia alcătuită din studenţii unei facultăţi, se pot studia caracteristici precum: sex, vârstă, stare civilă, notele la examene, media, etc.

b) Dacă populaţia este reprezentată de un anumit bun realizat în serie, atunci caracteristicile pot fi: calitatea (control de calitate), dimensiunea, preţul, etc.

c) În cazul în care populaţia este constituită din bunurile desfăcute pe piaţă de o firmă, caracteristicile considerate vor fi: sortimentele desfăcute, cantitatea din fiecare tip de bun, preţul la desfacere, etc.

d) Populaţia dintr-o anumită zonă geografică poate fi studiată din punct de vedere al: sex, vârstă, număr de ani de şcolarizare, stare civilă, grupa socio-profesională, stare de sănătate, culoare ochi, culoare păr, etc.

Obs. 2 Pentru fiecare populaţie se poate lua în studiu un număr mai mare sau mai mic de caracteristici, în funcţie de cerinţele analizei. Numărul caracteristicilor poate fi limitat din considerente de eficienţă şi de raţionalitate.

Caracteristicile determină descompunerea populaţiei în grupe omogene, care pot fi studiate separat. Fiecare caracteristică are mai multe forme de manifestare numite şi *modalităţi* sau *valori*, care trebuie să satisfacă următoarele principii:

- *principiul completitudinii*: fiecare individ al populaţiei aparţine unei clase definită de una dintre modalităţi;
- *principiul unicităţii*: un individ aparţine unei singure clase;
- *principiul organizării ierarhice a claselor*: clasele pot fi unificate prin mărirea gradului de generalitate al modalităţilor.

Vom nota modalităţile cu litere mici: $x, y, \dots$

Obs. 3 Din cele spuse mai sus rezultă că variabila statistică poate fi modelată matematic prin intermediul noţiunii de funcţie, având ca domeniu de definiţie mulţimea Ω, iar ca domeniu de valori o mulţime nevidă, de orice natură (numerică sau nu).

Clasificarea caracteristicilor se poate face după mai multe criterii. Prezentăm două criterii:

a) Clasificare după *forma de prezentare*:

- *caracteristici cantitative*: valorile lor sunt exprimate numeric. Exemple: nota, vârsta, greutatea, numărul de aruncări ale unei monede până la prima apariţie a stemei, etc.

- *caracteristici calitative*: valorile lor nu se exprimă numeric. Exemple: culoare ochi, profesie, sex.

Obs. 4 O caracteristică de un tip se poate transforma într-o caracteristică de celălalt tip. Cea mai frecventă transformare este din tipul calitativ în cel cantitativ, prin numerotarea modalităţilor. Spre exemplu, putem codifica sexul astfel: 1-masculin, 2-feminin. Invers, pentru a trece de la tipul cantitativ la cel calitativ, se pot defini grupe tipice de valori pentru caracteristica studiată.

b) Clasificare după *numărul modalităţilor*:

- *caracteristici discrete*: iau o mulţime cel mult numărabilă de valori. Exemple: sex, nota.

- *caracteristici de tip continuu*: iau valori într-un interval; pot fi doar de tip cantitativ. Exemplu: dimensiunile unei piese.

Def. O submulţime finită $\Omega_1 \subseteq \Omega$ se numeşte *eşantion* sau *selecţie*. Cardinalul său poartă numele de *volum de selecţie*. În general, eşantionul este alcătuit din acei indivizi ai populaţiei pentru care s-a observat efectiv valoarea caracteristicii studiate.

2.2 Elemente de teoria scalării

Aşa cum s-a văzut mai sus, caracteristicile pot fi măsurate în mai multe moduri. Există patru tipuri principale de scală: scala nominală, scala ordinală, scala de interval şi cea de raport.

2.2.1 Scala nominală

Valorile unei caracteristici măsurată pe *scala nominală* pot fi nume, numere sau simboluri, care însă **nu** se pot ordona. Ca urmare, caracteristicile exprimate în această scală sunt de tip calitativ şi trebuiesc îndeplinite condiţiile:

1. Valorile scalei nominale sunt exhaustive şi mutual exclusive; fiecare observaţie trebuie să corespună unei singure valori. Indivizii care iau aceeaşi valoare se consideră echivalenţi în raport cu caracteristica studiată.

2. Numele sau simbolurile care desemnează valorile pot fi interschimbate fără a altera informaţiile transmise de scală.

Exemple. Grupa sanguină (O, A, B, AB); culoarea ochilor, sex, etc.

2.2.2 Scala ordinală

Valorile unei caracteristici măsurată pe *scala ordinală* pot fi ordonate după un anumit criteriu, însă nu permit efectuarea de operaţii aritmetice între ele. Caracteristicile exprimate în această scală sunt tot de tip calitativ şi verifică condiţiile:

1. Diferenţele dintre valori nu sunt neapărat egale şi pot fi chiar imposibil de măsurat.

2. Simbolurile asociate valorilor nu sunt importante atâta timp cât relaţia de ordine este păstrată.

Exemplu. Starea unui pacient poate fi clasificată ca înrăutăţită, stabilă sau îmbunătăţită. Referitor la prima condiţie de mai sus, se observă că diferenţa dintre starea "înrăutăţită" şi "stabilă" nu este neapărat aceeaşi ca cea dintre "stabilă" şi "îmbunătăţită".

2.2.3 Scala de interval

Valorile unei caracteristici măsurată pe *scala de interval* sunt numere echidistante. Scala nu are un "punct zero" (o origine care să reprezinte absenţa caracteristicii măsurate). Cu ajutorul acestei scale se poate determina cât de mult (sau cât de puţin) din caracteristica măsurată reprezintă fiecare valoare. Sunt permise operaţii de adunare şi de scădere între valori. Caracteristicile exprimate în această scală sunt de tip cantitativ.

Exemplu. Temperatura exprimată în grade Celsius (observaţi că valoarea 0° este cea la care îngheaţă apa şi nu reprezintă absenţa temperaturii).

2.2.4 Scala de raport

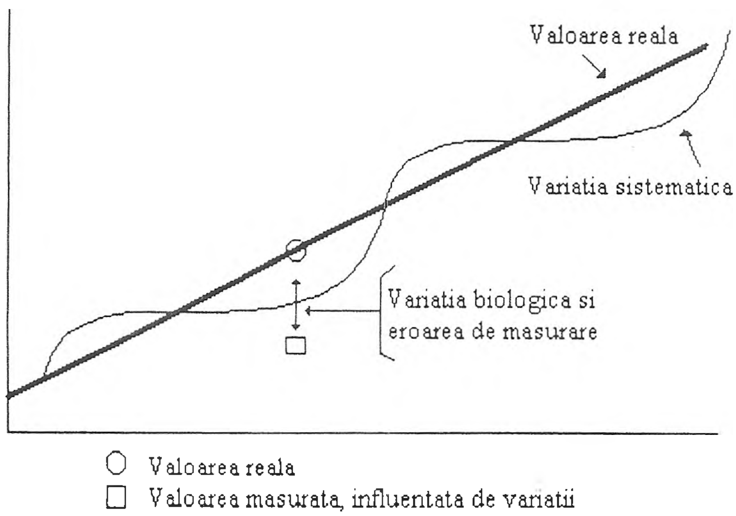
Pe *scala de raport*, măsurătorile încep dintr-un punct zero (origine) şi valorile sunt echidistante. Sunt permise toate operaţiile aritmetice, iar caracteristicile exprimate în această scală sunt de tip cantitativ.

Exemple. Temperatura exprimată în grade Kelvin, lungimea, greutatea, volumul etc.

2.3 Sursele variaţiilor din date

Fluctuaţiile ce pot apărea în timpul culegerii datelor reflectă în general combinarea efectelor mai multor fenomene. Spre exemplu, măsurătorile clinice pot fi afectate de:

- *variaţia biologică* (exemplu: valoarea tensiunii unui acelaşi pacient, în condiţii identice, poate să difere);
- *variaţia sistematică* (exemplu: valoarea tensiunii unui pacient diferă în funcţie de momentul zilei şi de poziţia în care stă);
- *eroarea de măsurare*, care poate fi atât aleatoare, cât şi sistematică (spre exemplu, datorită calibrării instrumentelor).



Capitolul 3

Tabele de distribuţie. Gruparea datelor

Având în vedere observaţia 4 din secţiunea anterioară, în cele ce urmează vom considera numai caracteristici statistice cantitative.

3.1 Serii statistice (date brute)

Forma cea mai simplă de prezentare a datelor statistice provenind dintr-o singură variabilă statistică este prin enumerarea observaţiilor efectuate asupra variabilei: $x_1, x_2, ..., x_n$. Această enumerare se numeşte *serie statistică* sau *eşantion*, n fiind volumul său. Unele valori ale eşantionului se pot repeta.

Dacă datele din eşantion se ordonează crescător, rezultă un *şir variaţional*, $x_{(1)}, x_{(2)}, ..., x_{(n)}$, unde

$$\begin{cases} 1)\ \{x_1, x_2, ..., x_n\} = \left\{x_{(1)}, x_{(2)}, ..., x_{(n)}\right\} \\ 2)\ x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)} \end{cases}$$

Exemplu. S-a măsurat înălţimea (în cm) a 20 de copii dintr-o şcoală, obţinându-se următoarea serie statistică:

133	136	120	138	133	131	127	141	127	143
130	131	125	144	128	134	135	137	133	129

Şirul variaţional corespunzător este:

120	125	127	127	128	129	130	131	131	133
133	133	134	135	136	137	138	141	143	144

Obs. 1 Deoarece numărul datelor dintr-o serie statistică poate fi foarte mare (de ordinul miilor), aceasta devine greu de prelucrat şi de studiat. Ca urmare, datele se prezintă şi sub formă de tabele de distribuţie.

3.2 Tabele de distribuţie. Tipuri de frecvenţe

3.2.1 Tabel de distribuţie a frecvenţelor absolute

Acest tip de tabel se obţine dintr-o serie statistică astfel: dacă $(x_1, ..., x_n)$ este un eşantion de volum n de observaţii făcute asupra variabilei statistice X, atunci:

1. Se formează şirul variaţional: $x_{(1)}, x_{(2)}, ..., x_{(n)}$.
2. Pe o coloană (linie) se trec doar valorile distincte din şirul variaţional, fie ele $x'_{(1)}, x'_{(2)}, ..., x'_{(k)}$, $k \leq n$.
3. Pe a doua coloană (linie), în dreptul fiecărei valori, se trece numărul său de apariţii din eşantionul iniţial: $n_1, ..., n_k$. Acest număr de apariţii al unei valori observate $x'_{(i)}$ se numeşte *frecvenţă absolută*, notată n_i.

Tabelul astfel obţinut se numeşte *tabel de distribuţie a frecvenţelor absolute* sau *distribuţie empirică* (*statistică, observată*) sau *seria statistică a frecvenţelor absolute*.

Este uşor de observat că, dacă avem în total k valori distincte în şirul variaţional, frecvenţele absolute verifică relaţia

$$n_1 + ... + n_k = n. \tag{3.1}$$

3.2.2 Tabel de distribuţie a frecvenţelor relative

Frecvenţele se pot exprima şi în valori relative la numărul total de observaţii. Aceste frecvenţe se numesc *relative* şi sunt definite prin

$$f_n\left(X = x'_{(i)}\right) = n'_i = \frac{n_i}{n}.$$

Uneori se exprimă în procente, sub forma $100\frac{n_i}{n}\%$. Tabelul corespunzător se numeşte *al frecvenţelor relative* sau *seria statistică a frecvenţelor relative.*

Relaţia (3.1) devine în acest caz

$$n'_1 + \ldots + n'_k = 1.$$

Obs. 2 Când n este foarte mare, $f_n\left(X = x'_{(i)}\right) = n'_i \approx P\left(X = x'_{(i)}\right)$, adică probabilitatea ca variabila statistică X rezultată în urma observaţiilor să ia a i-a valoare distinctă din şirul variaţional.

3.2.3 Frecvenţe cumulate crescător şi descrescător

În scopul reprezentării cât mai compacte, inclusiv în formă grafică, a datelor statistice, în statistică sunt utilizate frecvenţele cumulate. Acestea se obţin prin adunarea din aproape în aproape a frecvenţelor absolute sau relative.

Def. Se numeşte *frecvenţă absolută cumulată crescător (descrescător)* corespunzătoare unei valori x, suma frecvenţelor absolute ale tuturor valorilor variabilei mai mici sau egale cu x (mai mari sau egale cu x), anume

$$\sum_{y \le x} n_y \quad \left(\sum_{y \ge x} n_y\right).$$

Def. Se numeşte *frecvenţă relativă cumulată crescător (descrescător)* corespunzătoare unei valori x a variabilei, suma frecvenţelor relative ale tuturor valorilor variabilei mai mici sau egale cu x (mai mari sau egale cu x), anume

$$\sum_{y \le x} n'_y \quad \left(\sum_{y \ge x} n'_y\right).$$

Obs. 3 Se observă uşor că frecvenţa relativă cumulată este raportul dintre frecvenţa absolută cumulată şi volumul populaţiei.

Exemplu. Reluând datele de mai sus (înălţimile celor 20 de copii), avem următorul tabel cu frecvenţele absolute, relative şi cumulate:

Înăl-ţimea	Frecv. abs.	Frecv. relat.	Frec. abs. cumul. cresc.	Frec. abs. cumul. descresc.	Frec. rel. cumul. cresc.	Frecv. cumul. descresc.
120	1	0.05	1	20	0.05	1
125	1	0.05	2	19	0.1	0.95
127	2	0.1	4	18	0.2	0.9
128	1	0.05	5	16	0.25	0.8
129	1	0.05	6	15	0.3	0.75
130	1	0.05	7	14	0.35	0.7
131	2	0.1	9	13	0.45	0.65
133	3	0.15	12	11	0.6	0.55
134	1	0.05	13	8	0.65	0.4
135	1	0.05	14	7	0.7	0.35
136	1	0.05	15	6	0.75	0.3
137	1	0.05	16	5	0.8	0.25
138	1	0.05	17	4	0.85	0.2
141	1	0.05	18	3	0.9	0.15
143	1	0.05	19	2	0.95	0.1
144	1	0.05	20	1	1	0.05

3.2.4 Tabele de distribuţie grupate

Pentru un volum n foarte mare de date distincte, se recomandă gruparea lor în clase (intervale de valori), sub forma:

Intervalul	Frecv. abs.	Frecv. relat.
$[y_1, y_2)$	n_1	n'_1
$[y_2, y_3)$	n_2	n'_2
...	...	...
$[y_s, y_{s+1})$	n_s	n'_s
$\sum$	n	1

Valoarea centrală (mijlocul) unei clase este, pentru clasa $[y_i, y_{i+1})$,

$$c_i = \frac{y_i + y_{i+1}}{2}.$$

Exemplu. Să grupăm datele de mai sus în intervale de lungime 5:

Intervalul	Frecv. abs.	Frecv. relat.
$[120, 125)$	1	0.05
$[125, 130)$	5	0.25
$[130, 135)$	7	0.35
$[135, 140)$	4	0.2
$[140, 145)$	3	0.15
$\sum$	20	1

3.2.5 Tabele de distribuţie pentru două caracteristici

Dacă analiza statistică se face după două caracteristici, rezultatele obţinute se trec într-o tabelă cu două intrări.

Exemplu. Clasificarea a 1500 de persoane după culoarea ochilor şi a părului:

Culoare păr	Culoare ochi				$\sum$
	negri	căprui	verzi	albaştri	
negru	145	285	30	11	471
castaniu	62	431	87	67	647
blond	33	36	185	128	382
$\sum$	240	752	302	206	1500

Din acest tip de tabel sunt uşor de obţinut tabelele de distribuţie ale fiecărei caracteristici în parte, denumite şi *distribuţii marginale*. Pe exemplul numeric, avem:

Culoare păr	Frecv. abs.
negru	471
castaniu	647
blond	382
$\sum$	1500

Culoare ochi	Frecv. abs.
negri	240
căprui	752
verzi	302
albaştri	206
$\sum$	1500

Capitolul 4

Reprezentări grafice

4.1 Diagramele frecvenţelor necumulate

Se pleacă de la seria frecvenţelor absolute sau relative şi se pot trasa trei tipuri de diagrame: diagrama cu bastonaşe, histograma (diagrama cu dreptunghiuri) şi poligonul frecvenţelor. Pentru toate cele trei tipuri de diagrame, pe axa OX se trec valorile observate, iar dacă se lucrează cu date grupate, se trec limitele intervalelor sau centrele lor.

4.1.1 Diagrama cu bastonaşe

Acest tip de grafic este foarte bun pentru datele negrupate. Astfel, pentru fiecare valoare observată distinctă se trasează un segment vertical de lungime egală cu frecvenţa corespunzătoare valorii respective.

4.1.2 Histograma

Se aplică pentru datele grupate în intervale. Este alcătuită din dreptunghiuri având ca bază intervalele de valori, iar aria lor este proporţională cu frecvenţele. În funcţie de lungimea intervalelor, avem două cazuri:

1. *Histograme pentru intervale de lungimi egale*, caz în care înălţimea fiecărui dreptunghi coincide cu frecvenţa intervalului care stă la baza sa.
2. *Histograme pentru intervale de lungimi inegale*; în acest caz, pentru calcularea înălţimii dreptunghiurilor se aplică următorul algoritm:

- pe o coloană separată se trec lungimile tuturor intervalelor;
- se alege o lungime standard (cea care apare mai frecvent sau care este mai mică);
- pentru un interval oarecare, dacă lungimea sa este egală cu ($\alpha\times$ lungimea standard), atunci înălţimea dreptunghiului corespunzător va fi $\left(\frac{1}{\alpha}\times \text{frecvenţa corespunzătoare acelui interval}\right)$.

Exemplul 1. Considerăm datele anterioare (înălţimile celor 20 de copii) grupate în clase de lungime 5. Histograma este cea din fig. 1.

Exemplul 2. Considerăm masele a 35 de obiecte, măsurate în kg. Datele sunt grupate ca în tabelul de mai jos, iar în fig. 2 este trasată histograma corespunzătoare.

Masa	Frecvenţă	Lungime interval	Înălţime dreptunghi
$[6,9)$	4	3 standard	4
$[9,12)$	6	3 standard	6
$[12,18)$	10	6 $2\times$ standard	$10/2=5$
$[18,21)$	3	3 standard	3
$[21,30)$	12	9 $3\times$ standard	$12/3=4$

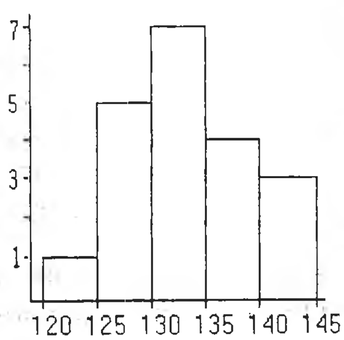


Fig. 1

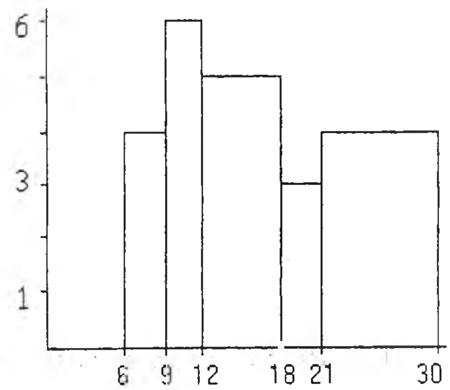


Fig. 2

4.1.3 Poligonul frecvenţelor

Se obţine unind prin linii (frânte) extremităţile superioare ale bastonaşelor din diagrama cu bastonaşe sau mijloacele extremităţilor superioare ale dreptunghiurilor din histogramă.

4.2 Diagramele frecvenţelor cumulate

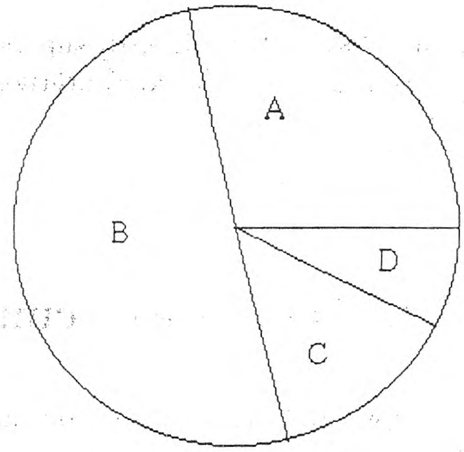
Frecvenţele cumulate se reprezintă ca mai înainte, dar numai prin histograme sau prin poligoanele frecvenţelor.

4.3 Diagrame circulare ("pie" sau "plăcintă")

Ariile sectoarelor de cerc dintr-o asemenea diagramă sunt proporţionale cu frecvenţele.

Exemplu. Considerăm vânzările de benzină de la patru benzinării, notate A, B, C, D.

Benzi-nărie	Vânzări (sute litri)	Unghiul sectorului
A	90	$360° \cdot 90/280 = 115.7°$
B	140	$360° \cdot 140/280 = 180°$
C	30	$360° \cdot 30/280 = 38.6°$
D	20	$360° \cdot 20/280 = 25.7°$
$\sum$	280	



Aceste diagrame se pot utiliza pentru a compara două sau mai multe seturi de date similare. Spre exemplu, dacă se reprezintă structura culturilor agricole pentru un teren într-o diagramă circulară, putem compara diagramele a două zone diferite.

Capitolul 5

Valori caracteristice

5.1 Parametri de poziţie

Parametrii de poziţie sunt valori de referinţă la care se raportează toate valorile unui eşantion. În cele ce urmează vom prezenta principalii asemenea parametri, anume media, mediana şi moda (modul).

5.1.1 Media (de selecţie), $\bar{x}$

Este vorba despre media aritmetică obişnuită. Calculul său depinde însă de modul de grupare al datelor. Astfel, cu notaţiile din secţiunile anterioare, avem:

- pentru un eşantion (date brute) $x_1, x_2, \ldots, x_n$, media este

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,;$$

- pentru datele dintr-un tabel de distribuţie a frecvenţelor absolute sau relative, media se calculează prin

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x'_{(i)} n_i = \sum_{i=1}^{k} x'_{(i)} n'_i\,;$$

- pentru datele grupate în intervale,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} c_i n_i .$$

Reamintim şi definiţia *mediei ponderate* a eşantionului $x_1, x_2, ..., x_n$, cu ponderile $f_i > 0, i = \overline{1,n}$,

$$\bar{x}_f = \frac{\sum\limits_{i=1}^{n} f_i x_i}{\sum\limits_{i=1}^{n} f_i}.$$

Interpretare. $\bar{x}$ este valoarea în jurul căreia se grupează valorile caracteristicii studiate.

Proprietate.
a) Media verifică relaţia

$$x_{(1)} \leq \bar{x} \leq x_{(n)}.$$

b) Dacă valorile x_i din eşantionul iniţial sunt transformate în $y_i = a + bx_i$, atunci $\bar{y} = a + b\bar{x}$.
c) Dacă se combină (concatenează, comasează) două eşantioane de volume m, n şi de medii $\bar{x}$, respectiv $\bar{y}$, atunci media eşantionului rezultat este

$$\bar{z} = \frac{m\bar{x} + n\bar{y}}{m+n}.$$

5.1.2 Mediana

Mediana este acea valoare care are proprietatea că jumătate dintre observaţiile din eşantion sunt mai mici sau egale cu ea, cealaltă jumătate dintre observaţii fiind mai mari sau egale decât ea. Cu alte cuvinte, mediana este acea valoare numerică ce împarte şirul variaţional corespunzător eşantionului dat în două părţi egale. Calculul său se face tot în funcţie de modul de grupare al datelor:
- pentru un eşantion $x_1, x_2, ..., x_n$, mediana se calculează astfel:

1. se ordonează crescător eşantionul;

2. dacă n este impar, atunci mediana este observaţia de rang $\frac{n+1}{2}$, iar dacă n este par, mediana este $\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)/2$.

Exemplu. Pentru eşantionul $7, 7, 2, 3, 4$, mediana este 4, iar pentru $36, 41, 27, 32$, este $(32 + 36)/2 = 34$.

- pentru datele dintr-un tabel de distribuţie: presupunem că se cunoaşte n, deci putem afla frecvenţele absolute (inclusiv pe baza frecvenţelor relative). Mediana rezultă astfel:

  1. se calculează frecvenţele absolute cumulate crescător;

  2. se calculează $(n+1)/2$;

  3. se determină prima linie din tabel pentru care frecvenţa absolută cumulată crescător este $\geq (n+1)/2$, valoarea similară de pe linia precedentă fiind $< (n+1)/2$. Valoarea caracteristicii corespunzătoare acestei linii este mediana.

  O altă metodă care se poate aplica în acest caz, dacă n nu este prea mare, constă în scrierea valorilor caracteristicii în ordine crescătoare (şirul variaţional) până la cea de-a $(n+1)/2$ -a valoare, care va fi chiar mediana.

- pentru datele grupate în intervale, aflarea medianei se face astfel:

  1. se trasează grafic curba frecvenţelor absolute cumulate crescător;

  2. pentru $y = (n+1)/2$, se află prin interpolare x-ul corespunzător de pe curba trasată. Acesta este mediana căutată.

Proprietate. Fie $\alpha \in \left\{x'_{(1)}, x'_{(2)}, ..., x'_{(k)}\right\}$. Atunci

$$\min_{\alpha} \left(\sum_{i=1}^{k} \left| x'_{(i)} - \alpha \right| n_i \right) = \sum_{i=1}^{k} \left| x'_{(i)} - mediana \right| n_i.$$

5.1.3 Moda (modul)

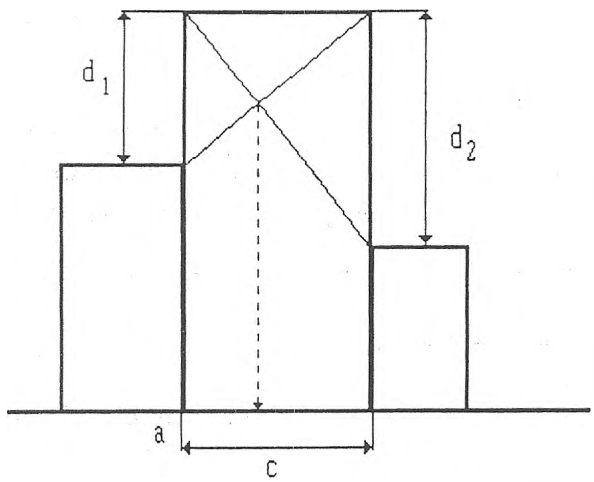
Moda (modul) este valoarea din eşantion căreia îi corespunde cea mai mare frecvenţă absolută (relativă).

- pentru un eşantion (date brute) $x_1, x_2, ..., x_n$, se construieşte un tabel de distribuţie şi se procedează ca mai jos;

- pentru datele dintr-un tabel de distribuţie a frecvenţelor absolute sau relative, se determină $n_x = \max\limits_y n_y$, x fiind chiar moda;

- pentru datele grupate în intervale, se parcurg paşii:

1. se trasează histograma;
2. se alege cel mai înalt dreptunghi din histogramă;
3. cu notaţiile din figură, $moda = a + \dfrac{d_1}{d_1 + d_2} \cdot c$.



5.1.4 Cuartile şi percentile

Se numesc *cuartile* cele trei valori care împart şirul variaţional în patru părţi egale. Cele 99 de valori care împart şirul variaţional într-o 100 de părţi egale se numesc *percentile*. Astfel, considerând n elemente aranjate crescător, avem:

Cuartila inferioară	Q_1	valoarea a $\frac{1}{4}(n+1)$-a
Mediana	Q_2	valoarea a $\frac{1}{2}(n+1)$-a
Cuartila superioară	Q_3	valoarea a $\frac{3}{4}(n+1)$-a
A 10-a percentilă	P_{10}	valoarea a $\frac{10}{100}(n+1)$-a
A 90-a percentilă	P_{90}	valoarea a $\frac{90}{100}(n+1)$-a

etc.

Definim şi *distanţa intercuartile* $= Q_3 - Q_1$, iar *semi-distanţa intercuartile* $= (Q_3 - Q_1)/2$. Avantajul acestor distanţe este acela că ele depind exclusiv numai de valorile din mijlocul şirului variaţional, nefiind afectate de valorile extreme.

Aflarea cuartilelor şi percentilelor se face similar cu aflarea medianei.

Exemplu. Să se afle cuartilele Q_1, Q_3 şi semi-distanţa intercuartilelor pentru următorul eşantion:

$$3, 2, 3, 9, 6, 6, 12, 11, 8, 2, 3, 5, 7, 5, 4, 4, 5, 12, 9$$

Soluţie. Avem 19 numere, pe care le ordonăm crescător:

$$2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9, 9, 11, 12, 12$$

Cuartila inferioară este valoarea a 5-a, deci $Q_1 = 3$. Cuartila superioară este valoarea a 15-a, deci $Q_3 = 9$. Rezultă că semi-distanţa intercuartilelor $= (Q_3 - Q_1)/2 = 3.\square$

5.2 Parametri de împrăştiere

Aceşti parametri măsoară gradul de împrăştiere al valorilor observate faţă de parametrul de poziţie dat.

5.2.1 Amplitudinea

Amplitudinea absolută este diferenţa dintre valoarea cea mai mare şi valoarea cea mai mică din eşantion, deci

$$A = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Amplitudinea relativă este raportul dintre amplitudinea absolută şi medie, anume $A/\bar{x}$.

Amplitudinea nu se defineşte pentru datele grupate în intervale.

Obs. Spre deosebire de amplitudinea absolută, amplitudinea relativă poate fi utilizată pentru compararea gradului de împrăştiere a două eşantioane diferite, chiar şi atunci când valorile acestora au unităţi de măsură diferite.

5.2.2 Abateri medii

Acestea pot fi de mai multe tipuri, în funcţie de valoarea de referinţă:

- *abaterea medie liniară absolută faţă de medie* este, pentru eşantionul $x_1, ..., x_n$,

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|\ ;$$

- *abaterea medie liniară absolută faţă de mediană* este

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - mediana|\ ;$$

- *abaterea medie liniară absolută faţă de modă* este

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - moda|\ .$$

5.2.3 Dispersia şi abaterea medie pătratică

Dispersia unui eşantion se notează cu s^2 (uneori cu σ^2) şi se calculează astfel:

- pentru un eşantion (date brute) $x_1, x_2, ..., x_n$,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\ ;$$

- pentru datele dintr-un tabel de distribuţie a frecvenţelor absolute sau relative,

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}\left[x'_{(i)} - \bar{x}\right]^2 n_i = \sum_{i=1}^{k}\left[x'_{(i)} - \bar{x}\right]^2 n'_i\ ;$$

- pentru datele grupate în intervale,

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(c_i - \bar{x})^2 n_i\ .$$

Interpretare. Dispersia măsoară gradul de împrăştiere al valorilor observate în jurul mediei. Cu cât dispersia este mai mare, cu atât împrăştierea valorilor din eşantion faţă de medie este mai mare, şi reciproc.

Abaterea medie pătratică se defineşte ca fiind radical din dispersie, deci

$$s = \sqrt{s^2}.$$

Interpretare. Abaterea medie pătratică este o măsură foarte utilă a gradului de împrăştiere. Pentru cele mai multe distribuţii, majoritatea observaţiilor se situează în intervalul $(\bar{x} - 3s,\ \bar{x} + 3s)$ (regula celor "3σ").

Proprietăţi.

a) Dacă $x_1 = x_2 = ... = x_n$ atunci $s^2 = 0$.

b) O formulă alternativă pentru calculul dispersiei este

$$s^2 = \overline{x^2} - (\bar{x})^2.$$

c) Dacă valorile x_i din eşantionul iniţial sunt transformate în $y_i = a + bx_i$, atunci

$$s_y^2 = b^2 s_x^2.$$

d) Considerând funcţia

$$f(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2,$$

atunci

$$\min_{a\in\mathbb{R}} f(a) = f(\bar{x}) = s^2.$$

e) Dacă se combină două eşantioane de volume n, m, de medii $\bar{x}'$, respectiv $\bar{x}''$, şi de dispersii s_1^2, respectiv s_2^2, atunci media eşantionului rezultat o notăm cu $\bar{x}$ (formula sa este dată la media de selecţie), iar dispersia este

$$s^2 = \frac{ns_1^2 + ms_2^2}{n+m} + \frac{n(\bar{x}' - \bar{x})^2 + m(\bar{x}'' - \bar{x})^2}{n+m}.$$

Primul termen al sumei reprezintă dispersia în interiorul "grupelor" (eşantioanelor), iar al doilea termen reprezintă dispersia dintre "grupe" (eşantioane).

Capitolul 7

Eşantionul. Repartiţii de selecţie

În acest capitol, vom prezenta mai întâi câteva elemente despre selecţie şi eşantion, în special selecţia aleatoare simplă. Prin studiul principalelor caracteristici de selecţie, vom introduce şi conceptul de repartiţie de selecţie.

7.1 Noţiuni de selecţie

7.1.1 Populaţia şi eşantionul

Am văzut anterior că prin *populaţie* se înţelege în general mulţimea de indivizi care se studiază, iar *eşantionul* este o submulţime a acestei mulţimi, alcătuită din indivizii care s-au observat efectiv în timpul unui studiu neexhaustiv.

Ca exemple de populaţii amintim populaţia umană dintr-o anumită regiune geografică, o mulţime de bunuri etc. În astfel de cazuri, conceptul de populaţie este uşor de înţeles, fiind vorba despre *populaţii finite*, ai căror indivizi pot fi număraţi fără dificultate. Noţiunea de populaţie nu este însă întotdeauna atât de simplă. Există numeroase populaţii, adesea ipotetice, care sunt *infinite*. Spre exemplu, insectele ce trăiesc într-o anumită regiune formează o populaţie greu de numărat; mulţimea tuturor experimentelor posibile de un anumit tip sau mulţimea tuturor variantelor unui joc sunt alte două asemenea exemple.

De altfel, în general, termenii de populaţie şi de eşantion sunt utilizaţi referitor la mulţimi de măsuri şi nu la mulţimile de indivizi asupra cărora aceste măsuri au fost aplicate. Se consideră, spre exemplu, populaţia înălţimilor sau vârstelor locuitorilor unei anumite regiuni, precizând deci care este caracteristica particulară care prezintă interes în legătură cu indivizii studiaţi.

7.1.2 Selecţia aleatoare simplă

Pentru a cunoaşte în totalitate caracteristicile unei anumite populaţii, ar trebui ca fiecare element al populaţiei respective să poată fi studiat. O asemenea operaţie se mai numeşte şi *recensământ*. Din păcate, o asemenea situaţie este foarte rară, devenind imposibilă în cazul populaţiilor numeroase (din motive precum costul şi rapiditatea culegerii datelor) sau infinite. Ca urmare, se recurge la studiul unui eşantion reprezentativ pentru populaţie.

Def. Prin *selecţie* sau *sondaj* vom înţelege mulţimea operaţiilor de alegere sau selectare, dintr-o populaţie, a indivizilor care vor constitui eşantionul.

Există mai multe metode de construire a unui eşantion, dintre care ne oprim asupra selecţiei aleatoare simple.

Def. Un eşantion se numeşte *aleator* când probabilitatea ca un individ al populaţiei să facă parte din eşantion este aceeaşi, indiferent de individ. Eşantionul se numeşte *aleator şi simplu* dacă este aleator şi, în plus, selecţiile indivizilor ce vor constitui eşantionul se fac independent una de alta.

Unul dintre procedeele cele mai utilizate pentru obţinerea unor asemenea eşantioane din populaţii finite sau infinite, este *metoda numerelor aleatoare*.

Pentru a asigura caracterul aleator şi simplu al unui eşantion cu ajutorul numerelor aleatoare, se asociază mai întâi un grup de cifre (un număr) fiecărui individ din populaţie. Se generează apoi un şir de numere aleatoare sau se ia din tabelele de numere aleatoare, pornind de la o anumită valoare, şi se includ în eşantion indivizii ai căror numere apar în şir. În cazul populaţiilor finite trebuie procedat cu grijă, astfel încât să nu se selecteze un acelaşi individ de mai multe ori (selecţiile trebuiesc efectuate "fără revenire").

7.2 Caracteristici de selecţie

Pentru a studia valorile unui eşantion sunt necesare cel puţin două caracteristici: o valoare centrală şi o măsură a împrăştierii valorilor eşantionului în jurul acestei valori centrale.

opulaţia înălţimilor
d deci care este ca-
u indivizii studiaţi.

te populaţii, ar tre-
tudiat. O asemenea
menea situaţie este
neroase (din motive
ite. Ca urmare, se
ulaţie.

mea operaţiilor de
vor constitui eşan-

ion, dintre care ne

tatea ca un individ
diferent de individ.
şi, în plus, selecţiile
una de alta.

erea unor asemenea
numerelor aleatoare.
eşantion cu ajutorul
(un număr) fiecărui
e aleatoare sau se ia
valoare, şi se includ
l populaţiilor finite
n acelaşi individ de
e").

puţin două caracte-
ilor eşantionului în

[illegible] unei valori centrale este o soluţie a problemei următoare: să se în-[illegible] valorile observate $(x_1, ..., x_n)$ cu o unică valoare c, cât mai apropiată [illegible] valori ale eşantionului. Aceasta presupune şi alegerea unei măsuri [illegible] dintre c şi x_i, deci a unei măsuri a împrăştierii. Altfel spus, con-[illegible] spaţiul eşantioanelor de volum n, $\mathbb{R}^n$, se caută o distanţă pe acest [illegible] pentru a măsura abaterea dintre $(x_1, ..., x_n)$ şi eşantionul constant [illegible]

[illegible] se foloseşte distanţa $\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$, minimul acesteia va fi atins [illegible] media $c = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, iar măsura gradului de împrăştiere devine [illegible] dispersia $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Se poate folosi şi o altă distanţă, pre-[illegible] $\sum |x_i - c|$, caz în care minimul va fi atins pentru valoarea mediană a [illegible]ului, iar măsura gradului de împrăştiere va fi abaterea medie faţă de [illegible]. Cuplul $(\bar{x}, s^2)$ este însă cel mai utilizat.

[illegible] ce urmează, vom presupune pentru simplificare că populaţia stu-[illegible] este infinită, iar selecţia este aleatoare şi simplă.

[illegible].1 Media de selecţie

[illegible] un prim eşantion de n observaţii, $(x_1, ..., x_n)$, pentru care s-a [illegible] media $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Dacă se selectează în condiţii similare un al [illegible] eşantion de acelaşi volum, $(x_1', ..., x_n')$, media corespunzătoare [illegible] $\frac{1}{n}\sum x_i'$, va fi în general diferită de prima medie observată. Acest lucru [illegible] valabil şi pentru mediile altor eşantioane prelevate în condiţii similare: [illegible] $(x_1''', ..., x_n''')$ *etc.*

[illegible] considera şirul infinit al observaţiilor de rang i din fiecare eşantion [illegible] x_i''', ... ca fiind observaţii efectuate asupra unei aceleaşi variabile [illegible] X_i, $i = \overline{1,n}$. Valorile medii observate $\bar{x}, \bar{x}', \bar{x}'', \bar{x}''', ...$ devin astfel şi [illegible] ale unei variabile aleatoare $\bar{X}$, care depinde de $X_1, ..., X_n$ astfel

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

[illegible] $\bar{X}$ se numeşte *medie de selecţie*. Fiind variabile aleatoare, $X_1, ..., X_n$ şi

$\bar{X}$ au repartiţii denumite *repartiţii de selecţie.*

Selecţia fiind aleatoare şi simplă, variabilele $X_1, ..., X_n$ sunt independente şi identic repartizate. Să notăm cu m media lor şi cu σ^2 dispersia.

Proprietate.

$$\text{a) } \mathbb{E}\bar{X} = m,\ Var\left(\bar{X}\right) = \frac{\sigma^2}{n}.$$

b) Din legea numerelor mari, $\bar{X} \xrightarrow[n\to\infty]{a.s.} m.$

c) Din teorema limitei centrale,

$$\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{repart.} \text{o v.a. repartizată } N(0,1).$$

Dem.

$$\text{a) } \mathbb{E}\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = m,\ Var\left(\bar{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n}Var\left(X_i\right) = \frac{\sigma^2}{n}.$$

7.2.2 Dispersia de selecţie

Procedând ca în cazul mediei de selecţie, putem calcula dispersiile pentru diferitele eşantioane:

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

aceste dispersii fiind considerate ca valori observate asupra variabilei aleatore

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Def. S^2 se numeşte *dispersie de selecţie.*

Proprietate.

$$a)\ \mathbb{E}S^2 = \frac{n-1}{n}\sigma^2,\ Var\left(S^2\right) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right]$$

unde μ_4 este momentul centrat de ordin 4 al v.a. X_i.

$$b)\ S^2 = \overline{X^2} - \bar{X}^2.$$

c) $S^2 \xrightarrow[n\to\infty]{a.s.} \sigma^2.(LNM)$

d) $\dfrac{\left(S^2-\sigma^2\right)\sqrt{n}}{\sqrt{\mu_4-\sigma^4}} \xrightarrow[n\to\infty]{repart.}$ o v.a. repartizată $N(0,1).(TLC)$

Dem.

a) $$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m + m - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2\left(m-\bar{X}\right)}{n}\sum_{i=1}^{n}(X_i -$$

$$+\left(m-\bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2 - 2\left(m-\bar{X}\right)^2 + \left(m-\bar{X}\right)^2 \Rightarrow$$

$$ES^2 = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) - Var(\bar{X}) = \frac{1}{n}n\sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2.\square$$

Obs. Considerând variabila transformată

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2,$$

[illegible] $ES^{*2} = \sigma^2$.

7.2.3 Funcţia de repartiţie empirică

Notăm cu F funcţia de repartiţie comună variabilelor $X_1, ..., X_n$.

Def. Se numeşte *funcţie de repartiţie empirică* (sau *funcţie de repartiţi de selecţie*) $F_n^* : \mathbf{R}\times\Omega \to [0,1]$ definită prin

$$F_n^*(x,\omega) = \frac{\left|\left\{i\in\overline{1,n}\,|\, X_i(\omega) < x\right\}\right|}{n} = \frac{\text{nr. de observaţii } < x}{n}.$$

Proprietate. Funcţia de repartiţie empirică are următoarele proprietăţi:

1. Pentru $\forall\omega\in\Omega$ fixat, $F_n^*(\cdot,\omega)$ este o funcţie în scară.

2. Pentru $\forall x\in\mathbf{R}$ fixat, $F_n^*(x,\cdot)$ este o v.a. simplă cu proprietatea că

$$nF_n^*(x,\cdot) \sim Binomial\left(n, F(x)\right).$$

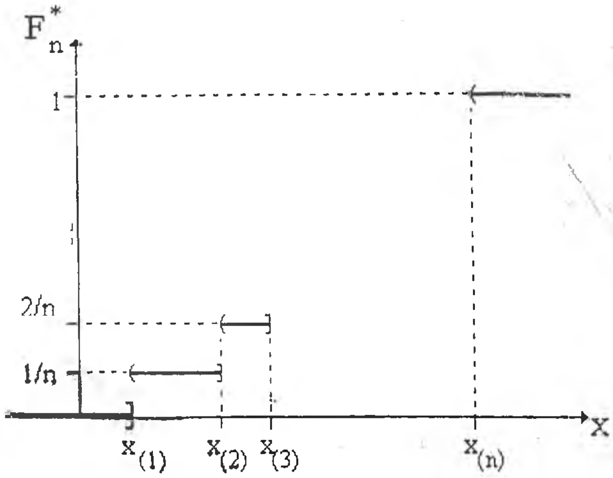
c) $\mathbb{E}\left(F_n^*(x,\cdot)\right) = F(x)$, $Var\left(F_n^*(x,\cdot)\right) = \frac{1}{n}F(x)\left[1 - F(x)\right]$.

d) Pentru $\forall x \in \mathbf{R}$ fixat, $F_n^*(x,\cdot) \xrightarrow[n\to\infty]{a.s.} F(x)$.

e) **Teorema Glivenko-Cantelli:** $\sup\limits_{x\in\mathbf{R}} \left|F_n^*(x,\cdot) - F(x)\right| \xrightarrow[n\to\infty]{a.s.} 0 \Leftrightarrow$

$$\Leftrightarrow P\left(\lim_{n\to\infty}\sup_{x\in\mathbf{R}} \left|F_n^*(x,\cdot) - F(x)\right| = 0\right) = 1$$ (convergenţa lui F_n^* către F este uniform a.s.).

Dem. a)



b) Se consideră v.a. auxiliare $Z_k(\omega) = \begin{cases} 1, & X_k(\omega) < x \\ 0, & altfel \end{cases}$, $k = \overline{1,n} \Rightarrow$ $Z_k \sim Bernoulli\left(F(x)\right)$. Deci

$$nF_n^*(x,\cdot) = \sum_{k=1}^{n} Z_k \sim Binomial\left(n, F(x)\right).$$

c) Rezultă imediat din proprietăţile repartiţiei Binomiale.

$$d)\quad F_n^*(x,\cdot) = \frac{1}{n}\sum_{k=1}^{n} Z_k \xrightarrow[n\to\infty]{a.s.} \mathbb{E}Z_k = F(x)\ (LNM).\square$$