Metode numerice și statistică
Curs 7

lect. Ciprian Deliu
✉ cdeliu@tuiasi.ro
moodle.deliu.ro

Universitatea Tehnică "Gh. Asachi" Iași
Facultatea de Hidrotehnică, Geodezie și Ingineria Mediului

2019

Introducere

Statistica descriptivă este ramura statisticii care se ocupă cu prezentarea, organizarea și interpretarea unei colecții de date. Descrierea acestor informații se poate face grafic (prin liste, grafice liniare, de distribuție, etc.), sau prin indicatori statistici (medie, mediană, abatere, etc.)
Analiza statistică a unui fenomen începe cu statistica formală (culegerea datelor despre fenomenul respectiv și înregistrarea datelor). Datele sunt apoi analizate și intepretate cu ajutorul *statisticii matematice*.

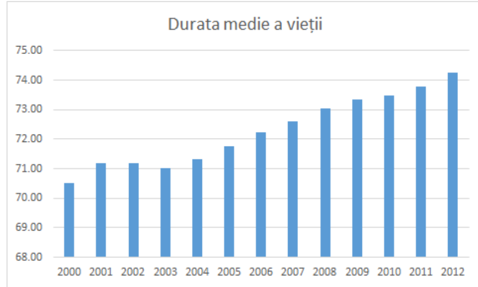
Definiție

*Prin **populație** statistică se înțelege orice mulțime care formează obiectul unei analize statistice. Elementele unei populații statistice se numesc unități statistice sau indivizi.*

- Prin **caracteristică** a unei populații statistice înțelegem o trăsătură comună unităților acelei populații.
- Caracteristicile pot fi calitative sau cantitative. Caracteristicile cantitative pot fi măsurate folosind numere reale. Valoarea numerică a unei caracteristici se numește **variabilă aleatoare**.

Exemplu: durata medie a vieții în Romania în perioada 2000-2012; observăm că această valoare crește începând cu anul 2003, după o scădere nesemnificativă.

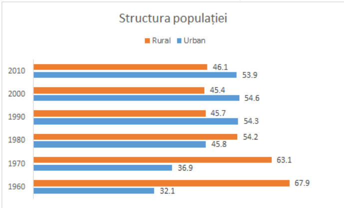
Anul	Durata
2000	70.53
2001	71.19
2002	71.18
2003	71.01
2004	71.32
2005	71.76
2006	72.22
2007	72.61
2008	73.03
2009	73.33
2010	73.47
2011	73.77
2012	74.26



- Reprezentarea grafică realizată pentru studierea schimbărilor sau pentru comparerea variabilelor statistice se numește grafic.
- Reprezentarea cu batoane folosește batoane orizontale sau verticale, ale căror lungimi sunt chiar valorile variabilei statistice. Batoanele verticale se folosesc de obicei pentru caracteristici care variază în timp.

Reprezentarea cu batoane orizontale prezintă variante adaptate, de exemplu reprezentarea pe componente, fără realizarea unei comparații cu întregul.

Anul	Urban	Rural
1960	32.1%	67.9%
1970	36.9%	63.1%
1980	45.8%	54.2%
1990	54.3%	45.7%
2000	54.6%	45.4%
2010	53.9%	46.1%



Graficul liniar pe porțiuni este format din segmente de dreaptă ce se obțin prin unirea perechilor de valori corespunzătoare ale unei perechi de variabile diferite.

Anul	Total imigranți
2003	3267
2004	2987
2005	3704
2006	7714
2007	9575
2008	10030

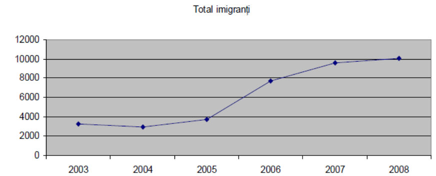


Diagrama circulară arată descompunerea unui întreg în părțile sale componente. Ele se exprimă ca procente din total și sunt reprezentate prin segmente de cerc, unghiurile la centru având măsuri egale cu procentul corespunzător din 360^0.

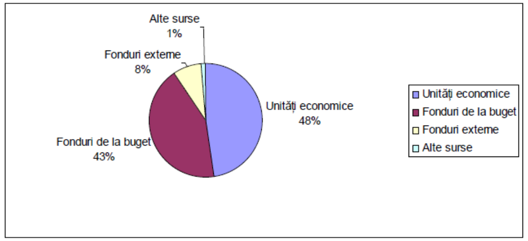


Figura arată structura cheltuielilor din domeniul cercetare-dezvoltare, din punctul de vedere al surselor de finanțare, în România în 2001.

Definiție

*O variabilă statistică se numește **discretă** dacă nu poate lua decât valori izolate în intervalul său de variație, și se numește **continuă** dacă poate lua toate valorile posibile în intervalul său de variație.*

- Exemple de variabile discrete: numărul capitolelor unei cărți, numărul articolelor produse într-o fabrică, etc.
- Exemple de variabile continue: înălțimea unei persoane, ora sosirii unui tren, etc.

Exemplu: Considerăm un număr de 40 de angajați al căror salariu exprimat în mii de lei este dat în tabelul următor:

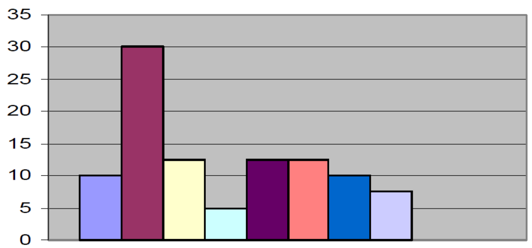
0.831	0.904	0.896	0.961	0.981
0.956	1.705	1.591	1.156	1.221
1.587	0.991	1.981	1.459	1.861
0.82	1.141	1.452	1.344	1.42
1.805	1.052	1.731	1.75	0.976
1.091	1.201	1.895	0.972	1.071
1.605	0.989	1.858	1.081	1.492
1.594	1.354	1.946	1.671	1.057

O descriere a seriei statistice obținute se realizează prin construirea unui tabel al frecvențelor, în care observațiile sunt clasificate în funcție de numărul unităților statistice care se află între anumite limite.

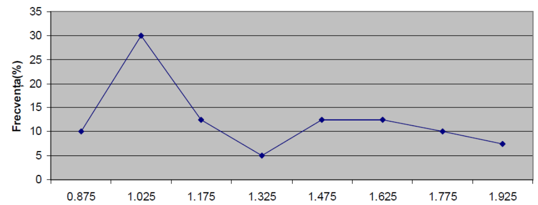
Limitele clasei	Mijlocul clasei	Frecvența absolută	Frecvența relativă(%)	Frecvența cumulată absolută	Frecvența cumulată relativă(%)
[0.8,0.95)	0.875	4	10	4	10
[0.95,1.1)	1.025	12	30	16	40
[1.1,1.25)	1.175	5	12.5	21	52.5
[1.25,1.4)	1.325	2	5	23	57.5
[1.4,1.55)	1.475	5	12.5	28	70
[1.55,1.7)	1.625	5	12.5	33	82.5
[1.7,1.85)	1.775	4	10	37	92.5
[1.85,2)	1.925	3	7.5	40	100

- Media aritmetică a limitelor unei clase se numește *mijlocul* sau *valoarea centrală* a clasei.
- Diferența dintre cea mai mare și cea mai mică margine se numește *domeniu* sau *amplitudine*.
- *Frecvența absolută* este dată de numărul unităților statistice aflate între limitele unei clase.
- *Frecvența relativă* este raportul dintre frecvența absolută și numărul total al unităților statistice.

- *Frecvența cumulată absolută* a unei clase este suma frecvențelor până la clasa respectivă.
- *Frecvența cumulată relativă* este raportul dintre frecvența cumulată absolută și numărul total al unităților statistice.
- *Histograma* este o reprezentare cu batoane fără spațiu între acestea, având pe axa orizontală marginile claselor și frecvențele pe cea verticală.



- *Poligonul frecvențelor* este un grafic liniar pe porțiuni, mijloacele claselor fiind reprezentate pe axa orizontală și frecvențele pe cea verticală.
- Fiecare mijloc are o frecvență corespunzătoare marcată printr-un punct, iar punctele consecutive se unesc prin segmente de dreaptă, rezultând o linie poligonală.



- Poligonul frecvențelor cumulate este un grafic liniar pe porțiuni care se realizează similar cu poligonul frecvențelor, singura schimbare fiind aceea că în locul frecvențelor apar frecvențele cumulate.

Caracteristici numerice

Datele statistice pot fi descrise cu ajutorul unor indicatori statistici. În acest sens, există două mari categorii:

- *măsuri ale tendinței centrale*: media, mediana, moda, etc.
- *măsuri ale variației* sau împrăștierii: amplitudinea, abaterea, etc.

Pentru o variabilă discretă, **moda** este valoarea cu frecvența maximă.
Pentru o variabilă continuă, clasa cu frecvența maximă se numește **clasa modală**, iar mijlocul acesteia este **moda** variabilei.

Definiție

Pentru cazul discret, **mediana** *unei mulțimi de date ordonate crescător* $x_1 \leq x_2 \leq \dots \leq x_m$ *este valoarea de mijloc* $x_{\frac{m+1}{2}}$ *dacă* m *este impar, sau media celor două valori de mijloc* $\frac{1}{2}\left(x_{\frac{m}{2}} + x_{\frac{m}{2}+1}\right)$ *dacă* m *este par.*

Exemple:

- mediana mulțimii $\{5, 6, 8, 9, 12\}$ este 8
- mediana mulțimii $\{15, 18, 20, 24, 28, 30\}$ este $\frac{20+24}{2} = 22$.

Definiție

Pentru o variabilă continuă, prima clasă a cărei frecvență cumulată asociată este mai mare decât $\frac{1}{2}$ *se numește* **clasa medianei**

Pentru clasa $[x_{i-1}, x_i]$, notăm cu f_i frecvența clasei, F_i frecvența cumulată și $h_i = x_i - x_{i-1}$ lungimea clasei. Efectuând o interpolare în clasa medianei $[x_{k-1}, x_k)$ obținem valoarea medianei

$$m_d = x_{k-1} + h_k \frac{0.5 - F_{k-1}}{f_k}$$

Definiție

Media *(de selecție) a unei mulțimi* $x_1, x_2, \dots, x_m$ *se definește prin*

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

Exemplu: Considerăm următoarea listă de prețuri (în lei) pentru centrale termice: 9900, 10300, 11200, 12500, 7600, 17500. Costul mediu al unei centrale este

$$\bar{x} = \frac{1}{6}(9900 + 10300 + 11200 + 12500 + 7600 + 17500) = 11500$$

Dacă $x_1, x_2, \dots, x_k$ sunt valorile distincte din selecție și notăm cu n_i frecvența lui x_i pentru $i = 1, 2, \dots, k$, atunci media de selecție se rescrie

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i},$$

și notând cu $f_i = \frac{n_i}{m}$, $i = 1, \dots, k$ rezultă

$$\bar{x} = \sum_{i=1}^{k} f_i x_i.$$

Definiție

Considerăm un tabel al frecvențelor cu k *clase. Dacă* $x_1^*, x_2^*, \dots, x_k^*$ *sunt mijloacele claselor,* $n_1, n_2, \dots, n_k$ *frecvențele lor absolute și* $f_1, f_2, \dots, f_k$ *frecvențele relative, atunci* **media** *distribuției este*

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i^*}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} f_i x_i^*.$$

Pentru datele din tabelul cu salarii anterior media este
$\bar{x} = \frac{4\cdot 0.875+12\cdot 1.025+5\cdot 1.175+2\cdot 1.375+5\cdot 1.475+5\cdot 1.625+4\cdot 1.775+3\cdot 1.925}{4+12+5+2+5+5+4+3} = 1.3175$

Se observă că media nu dă o imagine completă a datelor de selecție sau a distribuției. De exemplu, mulțimile $\{2, 2, 2, 5, 8, 8, 8\}$, $\{3, 3, 5, 5, 5, 7, 7\}$, $\{4, 4, 4, 5, 6, 6, 6\}$ au aceeași medie, dar au structuri diferite. Acesta este motivul pentru care sunt introduse măsuri ale variației, care indică gradul de împrăștiere a datelor în jurul mediei.

Definiție

Pentru o variabilă discretă, diferența dintre cea mai mare și cea mai mică valoare a selecției se numește **amplitudine**.
Pentru o variabilă continuă, **amplitudinea** *este diferența dintre limita superioară a clasei cu cele mai mari margini și limita inferioară a clasei cu cele mai mici margini.*

Definiție

Fie datele de selecție $x_1, x_2, \dots, x_m$ *având media* $\bar{x}$. **Abaterea medie** *se definește prin relația*

$$a.m. = \frac{1}{m} \sum_{i=1}^{m} |x_i - \bar{x}|.$$

Exemplu: Considerăm datele de selecție $\{12, 15, 13, 20, 13\}$. Media lor este $\bar{x} = \frac{1}{5}(12 + 15 + 13 + 20 + 13) = 14.6$, iar abaterea medie este

$$a.m. = \frac{1}{5}\left(|12 - 14.6| + |15 - 14.6| + |13 - 14.6| + |20 - 14.6| + |13 - 14.6|\right) = 2.32$$

așadar valorile de selecție diferă în medie cu 2.32 față de media 14.6.
Fie $x_1, x_2, \dots, x_k$ valorile distincte ale unei selecții X având media $\bar{x}$, iar n_i frecvența lui x_i pentru $i = 1, \dots, k$. Atunci

$$a.m. = \frac{\sum_{i=1}^{k} n_i |x_i - \bar{x}|}{\sum_{i=1}^{k} n_i}$$

Notând cu $f_i = \frac{n_i}{m}$, $i = 1, \dots, n$ frecvențele relative, obținem

$$a.m. = \sum_{i=1}^{k} f_i |x_i - \bar{x}|.$$

Definiție

Fie o variabilă continuă cu un tabel al frecvențelor cu k *clase. Dacă* $x_1^*, x_2^*, \dots, x_k^*$ *sunt mijloacele claselor,* $n_1, n_2, \dots, n_k$ *frecvențele lor absolute și* $f_1, f_2, \dots, f_k$ *frecvențele relative, atunci* **abaterea medie** *este*

$$a.m. = \frac{\sum_{i=1}^{k} n_i |x_i^* - \bar{x}|}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} f_i |x_i^* - \bar{x}|.$$

Pentru datele din tabelul cu salarii anterior abaterea medie este $\frac{119.85}{40} = 2.9962$.

Definiție

Fie datele de selecție $x_1, x_2, \dots, x_m$ *având media* $\bar{x}$. **Dispersia** *se definește ca fiind*

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})^2.$$

Valoarea $\sigma = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})^2}$ *se numește* **abaterea standard** *de selecție (empirică).*

Fie $x_1, x_2, \dots, x_k$ valorile distincte ale unei selecții X având media $\bar{x}$, iar n_i frecvența lui x_i pentru $i = 1, \dots, k$. Atunci dispersia este

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} n_i}$$

Notând cu $f_i = \frac{n_i}{m}$, $i = 1, \dots, n$ frecvențele relative, obținem

$$\sigma^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2.$$

Definiție

Fie o variabilă continuă cu un tabel al frecvențelor cu k *clase. Dacă* $x_1^*, x_2^*, \dots, x_k^*$ *sunt mijloacele claselor,* $n_1, n_2, \dots, n_k$ *frecvențele lor absolute și* $f_1, f_2, \dots, f_k$ *frecvențele relative, atunci* **dispersia** *distribuției este*

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i (x_i^* - \bar{x})^2}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} f_i (x_i^* - \bar{x})^2.$$

Pentru datele din tabelul cu salarii anterior, dispersia este $\sigma^2 = \frac{445.275}{40} = 1.114$, iar abaterea standard este $\sigma = 1.055$.
Exercițiu: O grupă formată din 20 studenți susțin un test la matematică și se obțin rezultatele din tabelul următor:

Nota	Frecvența absolută	Frecvența relativă (%)
3	1	5
4	1	5
5	2	10
6	2	10
7	5	25
8	6	30
9	2	10
10	1	5

Să se calculeze media, mediana, moda, amplitudinea, abaterea medie, dispersia și abaterea standard.