# Proprietatile descriptorilor statistici pentru serii univariate

Prof.dr. Vergil VOINEAGU, conf.dr. Tudorel ANDREI
Catedra de Statistica si Previziune Economica, A.S.E. Bucuresti

*Studierea unui fenomen se realizeaza prin considerarea unui set de variabile masurate pentru unitatile unei populatii G constituite din n unitati elementare. Vom considera ca variabilele sunt în numar de p. Caracteristicile cuprinse în programul observarii statistice sunt definite prin urmatoarele elemente: 1. Spatiul observatiilor, notat prin $W=W_1 \times W_2 \times ... \times W_p$. 2. Structura algebrica introdusa în spatiul de observatii, notata prin $S=S_1 \times S_2 \times ... \times S_p$. 3. Aplicatia $v:G \to W$, care atribuie fiecarei unitati din cadrul populatiei valori pentru caracteristicile din planul observarii, sub forma vectorului $\mathbf{v}(i)=(v_1(i),v_2(i),...,v_p(i))$. S-a considerat $i \in G$.*
***Cuvinte cheie:*** *unitate, scala, variabila, observatie, descriptor statistic, serie univariata.*

În functie de caracteristicile multimii $\Omega$ si de structura S, se disting urmatoarele tipuri de variabile:

*1. Calitativ nominale.* În acest caz vom avea o multime formata dintr-un numar finit de coduri numerice. Aceste variabile sunt masurate pe o *scala nominala*. Pe aceasta scala nu se introduce structura de ordine din corpul numerelor reale si nici operatiile obisnuite de adunare, scadere etc. Scala este folosita numai pentru efectuarea operatiilor de clasificare si grupare a unitatilor populatiei.

*2. Calitativ ordinale.* În cadrul acestei variabile, masurata pe o scala ordinala, este permisa introducerea operatiei de ordine de pe axa numerelor reale. Nici în acest caz nu are sens efectuarea de calcule directe asupra valorilor variabilei, întrucât nu este permisa calcularea distantei dintre valori.

*3. Cantitative masurate pe o scala de interval.* În acest caz se introduce distanta între valori, dar nu are sens raportul dintre acestea, întrucât valoarea nula este arbitrara.

*4. Cantitative masurate pe o scala de raport.* În acest caz $\Omega=R$ este o multime înzestrata cu o structura de corp ordonat. Structura asociata multimii $\Omega$ este întotdeauna indusa de semantica subiacenta parametrului analizat, motiv pentru care o numim structura paradigmatica. De exemplu, parametrului "vârsta" i se asociaza: variabila calitativa ordinala (clase de vârsta exprimate prin cuvinte); o variabila cantitativa masurata pe o scala de raport (vârsta exacta a unei persoane masurata în ani întregi).

Pot fi utilizate doua tehnici pentru transformarea variabilelor: prin schimbarea structurii si prin diverse operatii de codificare. Fie variabila definita prin aplicatia $v:\Gamma \to \Omega$, care este înzestrata cu o structura S. În acest caz vom spune ca variabila este supusa unei transformari prin schimbarea structurii, daca aceasta se înlocuieste prin variabila $v':\Gamma \to \Omega$ care are structura $S'$, astfel încât $v(i)=v'(i)$ pentru orice $i \in \Gamma$.

Pentru a defini schimbarea de variabila prin codificare definim un alt spatiu al observatiilor $\Omega'$, înzestrat cu structura $S'$. În aceste conditii, definim aplicatia $c:\Omega \to \Omega'$. Astfel, noua variabila este în fapt obtinuta prin compunerea functiilor v si c.

Valorile celor *p* variabile observate sunt ordonate pentru cele *n* unitati ale populatiei într-o matrice de forma:

$$\mathbf{X}=\begin{pmatrix} x_1^1,...,x_1^j,...,x_1^p \\ ..... \\ x_i^1,...,x_i^j,...,x_i^p \\ ..... \\ x_n^1,...,x_n^j,...,x_n^p \end{pmatrix}$$

unde:

- $\mathbf{x}^j$ este un vector coloana de dimensiune n, continând valorile unei caracteristici din planul de observare pentru unitatile populatiei;
- $\mathbf{x}_I$ este vector linie de dimensiune p, cu valori ale caracteristicilor din planul observarii pentru o unitate statistica din cadrul populatiei Γ.

Daca $p_I$ sunt ponderi (frecvente relative), cu $\sum_{i=1}^{n} p_i = 1$, atunci vom defini matricea ponderilor astfel:

$$\mathbf{D}=\text{diag}(\mathbf{p_I})=\begin{pmatrix} p_1 & 0...0...0 \\ 0 & p_2...0...0 \\ ..... & \\ 0 & 0.... p_i ...0 \\ ..... & \\ 0 & 0....0....p_n \end{pmatrix}$$

Se observa ca în cazul în care unitatile populatiei au aceasi pondere $p_i = \frac{1}{n}$, atunci $\mathbf{D}=\frac{1}{n}\mathbf{I}$, unde **I** este matricea unitate.

Prin metode adecvate, vom prelucra seriile de date pentru fiecare caracteristica în parte, cât si pentru fiecare unitate statistica. Determinam în egala masura indicatori pentru masurarea dependentelor dintre caracteristici, a similaritatilor unitatilor statistice etc. Pentru prelucrarea seriilor de date vom introduce:

- descriptori si metrici în spatiul unitatilor;
- descriptori si metrici în spatiul variabilelor.

**Proprietati ale descriptorului m(x)**

Vom prezenta în continuare câteva proprietati care pot fi luate în considerare pentru alegerea unei masuri pentru tendinta unei serii univariate.

Definim masura statistica a tendintei centrale prin aplicatia:

$$m: \Omega_j^n \to R \quad j = \overline{1, p}$$

unde $m(x) \leftarrow (x_1, x_2, \ldots, x_n)$.

Demonstram pentru fiecare masura statistica, ca tendinta centrala satisface una sau mai multe din aceste proprietati.

"Norul de puncte" asociat multimii unitatilor din Γ pentru o caracteristica din planul de observare se noteaza cu:

$$N=\left\{ (x_i, p_i) \mid i = \overline{1, n} \right\}$$

Pentru ansamblul caracteristicilor din planul observarii definim "norul de puncte" prin:

$$\mathbf{N}=\{(\mathbf{x}_I, \mathbf{p}_i) \mid i = \overline{1, n}\}.$$

În cazul în care doi vectori **(x, y)** sunt comparati sau sunt folositi în cadrul unei expresii algebrice, vom considera ca acest lucru este posibil atât din punct de vedere algebric, cât si economic. Pentru a asigura comparabilitatea economica pentru cei doi vectori vom folosi fie doua serii de date pentru aceeasi caracteristica, fie vom recurge la diverse transformari. Prin transformari de origine si intensitate, vom obtine un vector de valori normalizate din intervalul [0, 1] .

În cele ce urmeaza vom prezenta proprietatile mai importante pe care o masura statistica a tendintei centrale le îndeplineste. Precizam ca aceste proprietati sunt puse în discutie pentru cazul seriilor de distributie unimodale.

Proprietatea 1 [$P_1$]. Masura statistica m(x) aplicata vectorilor ansamblului $\Omega^n$ îndeplineste *proprietatea de intermediaritate* daca si numai daca $\forall \mathbf{x} \in \Omega^n$ avem ca aceasta se încadreaza între cele doua valori extreme:

$$\min(x) \le m(x) \le \max(x).$$

Toate masurile statistice ale tendintei centrale folosite în statistica descriptiva satisfac aceasta proprietate. Daca $x \in \Omega^n$, atunci nu este implicit si faptul ca $m(x) \in \Omega$. Aceasta este o restrictie robusta pentru o masura a tendintei centrale a unei serii de date.

Proprietatea 2 [$P_2$]. O masura statistica satisface *proprietatea de monotonie* daca, doua serii de date ale aceleasi caracteris-

tici, $\forall$ **x**, **y** $\in \Omega^n$ cu $x_i \geq y_i$, atunci $m(x) \geq m(y)$.

Daca prima proprietate este satisfacuta de toti indicatorii tendintei centrale folositi în statistica descriptiva, nu acelasi lucru se întâmpla si în cazul acestei proprietati. Este cazul, de exemplu, al valorii modale, care nu satisface proprietatea pentru orice serie de valori.

Daca pentru doi vectori **x**, **y** $\in \Omega^n$, $x_i \geq y_i$ $\forall$ $i = \overline{1,n}$ si în plus exista valori în **x** si **y** astfel încât $x_i \neq y_i$, atunci **x** >**y** .

Proprietatea 3 [$P_3$]. O masura statistica satisface *proprietatea de regularitate* daca si numai daca $\forall$ **x**,**y**$\in$ $\Omega^n$, cu **x** > **y** , se obtine m(x) > m(y).

Aceasta proprietate este mult mai robusta decât $P_2$. Astfel, majoritatea indicatorilor calculati pe baza structurilor de ordine satisfac $P_2$, dar nu si proprietatea $P_3$. Este cazul, spre exempu, cel al medianei care verifica $P_2$ dar nu si proprietatea de regularitate.

Proprietatea 4 [$P_4$]. Masura statistica a tendintei centrale îndeplineste *proprietatea de omogenitate* daca si numai daca $\forall$ **x** $\in$ $\Omega^n$ si $\forall$ $\lambda \in R^*$ este satisfacuta inegalitatea:

$$m(\lambda \mathbf{x}) = \lambda\, m(\mathbf{x})$$

Se demonstreaza fara dificultate ca aceasta proprietate este îndeplinita de media aritmetica si de indicatorii medii de pozitie. De regula pentru a asigura $\lambda \mathbf{x} \hat{I} \Omega^n$ vom considera cazul particular $\lambda \hat{I} R_+^*$

Proprietatea 5 [$P_5$]. Pentru un ansamblu $\Omega^n$ vom spune ca masura statistica este *simetrica* daca $\forall \mathbf{x} \in \Omega^n$ , atunci:

$$m(-\mathbf{x}) = -\, m(\mathbf{x}).$$

În fapt, aceasta proprietate este un caz particular al proprietatii $P_4$, situatie în care vom lua $\lambda = -1$. Proprietatea este deosebit de restrictiva daca, pe lânga egalitatea de mai sus, se adauga si conditia de apartenenta $-\mathbf{x} \in \Omega^n$.

Proprietatea 6 [$P_6$]. Masura m(x) satisface *proprietatea de aditivitate* daca $\forall$ **x**,**y**$\in$ $\Omega^n$ doi vectori pentru care are sens sa definim operatia de adunare, atunci:

$$m(\mathbf{x}+\mathbf{y}) = m(\mathbf{x}) + m(\mathbf{y}).$$

Vom arata ca nu toate masurile care vor fi utilizate pentru analiza tendintei centrale a seriilor de date satisfac proprietatea de aditivitate. De accea, vom introduce conceptele de masura a tendintei centrale subaditiva si supraaditiva.

O masura a tendintei centrale este sub-aditiva daca $\forall$ **x**,**y**$\in$ $\Omega^n$, atunci $m(\mathbf{x}+\mathbf{y}) \leq m(\mathbf{x}) + m(\mathbf{y})$. În mod asemanator definim ca o masura este supraaditiva daca $\forall$ **x**,**y**$\in$ $\Omega^n$, avem ca $m(\mathbf{x}+\mathbf{y}) \geq m(\mathbf{x}) + m(\mathbf{y})$.

Pentru a prezenta proprietatea urmatoare se noteaza **u** = (1, 1, …,1).

Proprietatea 7 [$P_7$]. O masura statistica accepta o *schimbare de origine*, de amplitudine h$\in$ R, daca $\forall$ **x**$\in$ $\Omega^n$ si $\forall$ h$\in$ R avem ca $m(\mathbf{x}+h\mathbf{u}) = h + m(\mathbf{x})$.

Daca pentru o masura a tendintei centrale sunt verificate simultan $P_4$ si $P_7$, atunci $\forall$ **x**$\in$ $\Omega^n$ si $\forall$ h, $\lambda \in$ R, avem $m(\lambda\mathbf{x}+h\mathbf{u}) = h + \lambda m(\mathbf{x})$. De cele mai multe ori indicatorii tendintei centrale folositi în statistica descriptiva au aceasta proprietate. În cazul în care $\lambda$=0 vom obtine m(h**u**) = h, deci centrul de greutate al unui sir de valori constante este însasi valoarea care constituie seria de date.

**Descriptori ai tendintei centrale**

Vom prezenta în cele ce urmeaza câteva masuri ale tendintei centrale posibil de utilizat pentru o serie univariata de valori $(x_i)_{i=\overline{1,n}}$, utilizând ponderile $(p_i)_{i=\overline{1,n}}$. Sistemul de ponderare, care poate fi constituit pe baza frecventelor relative sau probabilitatilor, este definit prin vectorul **p** = $(p_1, p_2, \ldots, p_n)$ cu $p_i \geq 0$ si $\sum_i p_i = 1$.

Pentru a defini un descriptor statistic peste ansamblu de valori din V=$\Omega^n$ vom considera, în cazul în care $\Omega \subseteq$ R, functia d(**x**,**y**) de doua variabile vectoriale **x** si **y** care satisface proprietatile:

i) d(**x**,**y**)>0, **x** $\neq$ **y**;

ii) d(**x,y**)=0, **x** =**y**;
iii) d(**x,y**)= d(**y,x**);
iv) d(**x,z**) ≤ d(**x,y**)+ d(**y,z**), **x,y,z**∈ $\Omega^n$.

Functia cu proprietatile de mai sus se numeste metrica sau distanta euclidiana. Considerând aceasta functie vom cauta sa determinam vectorul **b**=(b,b,…,b)∈ $R^n$ astfel încât d(**x,b**) sa fie minima. Vom arata ca $\min_b d(x,b) = d(x,m)$., unde prin **m**=(m, m,…,m) s-a definit un vector n dimensional, cu *m* o valoare obtinuta prin aplicarea asupra seriei a unei masuri a tendintei centrale.

Pentru $p \geq 1$ consideram ca spatiul vectorial $R^n$ este înzestrat cu norma:

$$L_p(x) = \|x\|_p = \left( \sum_{i=1}^{n} p_i |x_i|^p \right)^{\frac{1}{p}} .$$

Vom estima tendinta centrala prin intermediul descriptorului $m_p(x)=tu$. Vom determina t∈ R din conditia de minim a functiei.

$$L_p(x-tu) = \|x-tu\|_p = \left( \sum_{i=1}^{n} p_i |x_i - t|^p \right)^{\frac{1}{p}}$$

Daca p=1, atunci minimul functiei $L_p(x-tu)$ se situeaza în intervalul median; pentru p>1 valoarea minima a functiei este unica. Pentru valoarea optima a lui t definim atunci distanta dintre vectorul observatiilor si vectorul $m_p(x)$, prin:

$$d_p(x, m_p(x)) = \left\| x - u m_p(x) \right\|$$

În practica sunt utilizate urmatoarele cazuri particulare pentru parametrul p∈ R:

i) p=1, atunci valoarea minima a expresiei este mediana seriei de valori, iar indicatorul pentru masurarea dispersiei seriei de valori este:

$$e_1(x) = \sum_i p_i |x_i - x_m|.$$

ii) Pentru p=2 vom obtine metrica euclidiana. În acest caz din conditia de minim a functiei $L_2(x-tu)$ vom determina:

$$m_2 = \sum_i p_i x_i$$

iar indicatorul pentru caracterizarea gradului de dispersare este abaterea standard.

iii) Daca $p \to \infty$ atunci, dispunem de metrica convergentei uniforme, iar:

$$m_\infty = \tfrac{1}{2}(\min(x) + \max(x))$$

este centrul de greutate al intervalului în care sunt plasate observatiile caracteristicii. Pentru caracterizarea dispersiei se va utiliza masura:

$$e_\infty(x) = \frac{1}{2}(\max(x) - \min(x)).$$

Pentru vectorul **x** format din numere reale definim urmatorii descriptori statistici:

i) media de ordinul p:

$$m_p(x) = L_p(x) = \left( \sum p_i x_i^{p}{}_i \right)^{\frac{1}{p}}, \; p \in R^*;$$

Pentru p=0 vom obtine formula mediei geometrice:

$$m_0 = \lim_{p \to 0} m_p(x) = \prod_i x_i^p .$$

În tabelul 1 vor fi trecute diverse particularizari ale mediei de ordinul p.

**Tabelul 1.**

| α | Media obtinuta | Formula de calcul |
|---|---|---|
| -∞ | Minimul seriei | $m_{-\infty}(x) = \min(x)$ |
| -1 | Media armonica | $m_{-1}(x) = \dfrac{1}{\sum_i \frac{p_i}{x_i}}$ |
| 1 | Media aritmetica | $m_1(x) = \sum p_i x_i$ |
| 2 | Media patratica | $m_2(x) = \left(\sum p_i x_i^2\right)^{1/2}$ |
| ∞ | Maximul seriei | $m_\infty(x) = \max(x)$ |

Se demonstreaza ca media de ordinul p, în raport cu ordinul, este o functie monoton crescatoare. Pentru cazurile particulare din tabelul 1, se verifica relatia de ordine:

$$m_{-\infty}(x) = \min(x) \leq m_{-1}(x) \leq m_0(x) \leq$$
$$\leq m_1(x) \leq m_2(x) \leq m_\infty(x) = \max(x)$$

*ii)* media logaritmica de ordinul *f*, este definita pentru orice vector $\mathbf{x} \in R$ care are ponderile $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, prin relatia:

$$l_f(x) = \frac{1}{f}\log\left(\sum_{i=1}^{n} p_i e^{f \cdot x_i}\right), \quad f \neq 0.$$

Daca $f$=0, atunci $l_0(x) = \lim_{f \to 0} l_f(x) = m_1(x)$

În fapt, între ultimele doua medii prezentate se verifica relatia:

$$l_f(x) = \log\left(m_f\left(e^x\right)\right),$$
$$\text{unde } e^e = \left(e^{x_1}, e^{x_2}, \ldots, e^{x_n}\right).$$

**Descriptori generali**

Pentru norul de puncte **N** definim *centrul de greutate*, care coincide cu vectorul mediilor aritmetice calculate pentru cele p variabile:

$$\mathbf{G} = \sum_{i=1}^{n} \mathbf{p}_I \mathbf{x}_I = \sum_{i=1}^{n} \mathbf{p}_i \begin{bmatrix} x_i^1 \\ . \\ x_i^p \end{bmatrix} = \begin{bmatrix} \overline{x_1} \\ . \\ \overline{x_p} \end{bmatrix} = \mathbf{x}^t \mathbf{D}_1.$$

Pentru a defini distanta dintre doua unitati înzestram spatiul cu o metrica specificata de matricea **M**, simetrica si pozitiv definita. Produsul scalar al doi vectori se exprima prin $\langle \mathbf{x}_I, \mathbf{x}_j \rangle = \mathbf{x}_i' \mathbf{M} \mathbf{x}_j$, iar distanta dintre doua unitati este data de :

$$d^2(\mathbf{x}_I, \mathbf{x}_j) = \|\mathbf{x}_I - \mathbf{x}_j\|^2_M = (\mathbf{x}_I - \mathbf{x}_j)' \mathbf{M} (\mathbf{x}_I - \mathbf{x}_j).$$

În practica, cele mai utilizate metrici sunt:

- **M**=**I,** situatie în care se utilizeaza produsul scalar obisnuit;
- **M**=diag($1/s_j^2$), $j = \overline{1, p}$, ceea ce revine la a diviza valorile observate ale variabilelor prin abaterile lor standard. Avantajul acestei metrici este ca distanta dintre doua unitati nu depinde de unitatea de masura, variabilele fiind echivalente ca importanta, independent de amplitudinea sau gradul de dispersare a valorilor.

Se defineste astfel inertia totala a norului de puncte prin media ponderata a patratelor distantelor punctelor fata de centrul lor de greutate:

$$\mathbf{I}_g = \sum_{1}^{n} \mathbf{p}_i (\mathbf{x}_I - \mathbf{g})' \mathbf{M} (\mathbf{x}_I - \mathbf{g}).$$

În general în raport cu un punct oarecare **h** inertia $\mathbf{I}_g$ se defineste prin:

$$\mathbf{I}_h = \sum_{1}^{n} \mathbf{p}_i (\mathbf{x}_I - \mathbf{h})' \mathbf{M} (\mathbf{x}_I - \mathbf{h}).$$

Între $\mathbf{I}_g$ si $\mathbf{I}_h$ exista relatia:

$\mathbf{I}_h = \mathbf{I}_g + (\mathbf{g}-\mathbf{h})' \mathbf{M} (\mathbf{g}-\mathbf{h})$.

Din ultima relatie obtinem ca inertia în raport cu centrul de greutate este minima.

În cazul în care $\mathbf{g} = \mathbf{0}$, atunci

$$\mathbf{I}_g = \sum_{1}^{n} \mathbf{p}_i (\mathbf{x}_I)' \mathbf{M} (\mathbf{x}_I).$$

**Bibliografie**

Benzecri, J-P., *Histoire et prehistoire de l'analyse des donnees*, Dunod, Paris, 1983.

Isaic-Maniu, Al.; Mitrut, C.; Voineagu, V., *Statistica pentru managementul afacerilor*, Editia a II-a, Editura Economica, Bucuresti, 1999.

Saporta, G., *Probabilites, analyse des donnees et statistique*, Technip, Paris, 1990.

Spircu, L, Calciu, M., Spircu, T., *Analiza datelor de marketing*, ALL, Bucuresti, 1994.