ANALIZA PRELIMINARĂ A DATELOR STATISTICE

Concepte fundamentale ale analizei datelor

Populaţie şi eşantion. Populația sau colectivitatea generală este reprezentată de mulțimea tuturor măsurătorilor care reprezintă interes pentru cercetător sau experimentator.

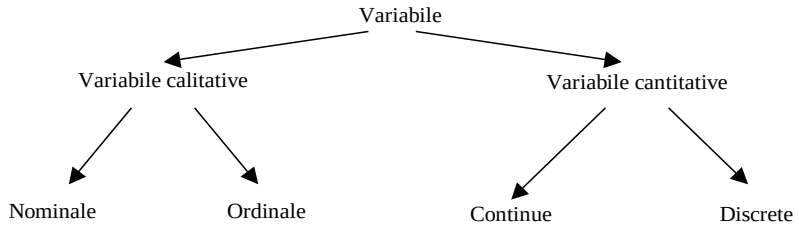
Atribut sau caracteristică. Reprezintă trăsăturile, proprietățile unităților din care este alcătuită populația.

Variabila. Variabila este un concept abstract care permite atribuirea de valori, numerice sau nenumerice, unui atribut sau caracteristici. Ea trebuie să fie înzestrată cu o sintaxă univocă şi o semantică precisă.

Variabilele pot fi de două tipuri: variabile calitative şi variabile cantitative.

Variabilele calitative sunt variabile ce diferă prin tip, se referă la proprietăți nenumerice ale unităților elementare aparținând unei populații şi nu pot fi exprimate numeric. Valorile variabilelor calitative sunt numite *modalități*.

Variabilele cantitative sunt variabile care diferă prin mărime, se referă la proprietăți numerice ale unităților elementare dintr-o populație și sunt exprimate în unități numerice. În funcție de natura valorilor pe care le iau, variabilele se împart în două categorii: variabile de tip discret şi variabile de tip continuu.



Asocierea valorilor la variabile se face în urma procesului de măsurare. Măsurarea se face prin intermediul unor repere și sisteme de referință cunoscute sub denumirea de ***scală***. ***Scala nominală*** este asociată variabilelor calitative de tip nominal. ***Scala ordinală*** este asociată variabilelor calitative de tip ordinal. Scale metrice: ***scala interval*** , ***scala raport.***

Densitate de probabilitate și funcție de repartiție

Densitatea de probabilitate măsoară posibilitatea ca o variabilă să ia o anumită valoare. Este deci o funcție definită pe mulțimea de valori posibile ale variabilei cu valori în intervalul [0,1]:

$f(x) = P(X=x)$,

unde X este variabila iar x este o valoare pe care o poate lua.

Exemplu. Să presupunem că avem o variabilă reprezentând talia (înălțimea) unor subiecți umani, exprimată în centimetrii. Probabilitatea ca variabila să ia valoarea 175 este:

$P(X = 175) = f(175)$

și exprimă probabilitatea ca un individ să aibă 175 cm înălțime.

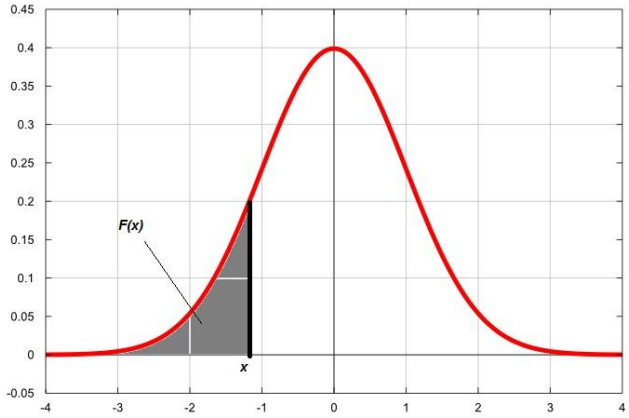
Densitatea de probabilitate ca funcție analitică, poate diferite forme particulare, specifice. Diferențele sunt de natura domeniului de definiție și al valorilor parametrilor determinate de forma analitica a funcției. De exemplu, chiar dacă avem un același domeniu de definiție, să zicem, subiecți umani (persoane), densitatea de probabilitatea a unei variabile *talia* va avea formă diferită de variabila *venit anual*.

Exemple de densitate de probabilitate: Gaussiană (normală), uniformă, Poisson etc.

Funcția de repartiție reprezintă probabilitatea ca o variabilă aleatoare să ia valori dintr-un anumit interval:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy .$$

Din punct de vedere geometric $F(x)$ este aria de sub curba densității de probabilitate:



Indicatori ai variabilelor aleatoare

Există trei grupe de indicatori:
- *indicatori de poziție*: media, momentul simplu, mediana, percentilele, cuartilele și modulul;
- *indicatori de împrăștiere*: amplitudinea, varianța, abaterea medie absolută, abaterea standard și coeficientul de variație, momentele centrate;
- *indicatori de formă a repartiției*: simetria și aplatizarea.

Media.
Cazul discret:

$$E(X) = \mu = \sum_{x \in R} x \cdot f(x),$$

unde f(*x*) este probabilitatea ca variabila să ia valoarea *x* (densitatea de probabilitate).
Pentru o repartiție uniformă cu *n* subiecți:

$$E(X) = \mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Cazul continuu: $E(X) = \mu = \int_R x \cdot f(x) dx$.

Momentul simplu de ordin k
Cazul discret: $M_k = E\left(X^k\right) = \sum_{x \in R} x^k \cdot f(x).$

Pentru o distribuție uniformă $M_k = \frac{1}{n} \sum_{x \in R} x^k$.

Cazul continuu: $E(X^2) = M_k = \int_R x^k \cdot f(x) dx$.
Se poate observa că momentul de ordin 1 este media.

Mediana este acea valoare care împarte setul de valori posibile in două: 50% valori mai mici și 50% valori mai mari. Deci:
$P(X \le x_{me}) = 0.5$
unde x_{me} este mediana.
Percentila de ordinul *p* este acea valoare care are proprietatea că cel mult *p*% dintre valorile seriei sunt mai mici decât ea și cel mult (100-*p*)% dintre valori sunt mai mari.
Să presupunem că avem seria Y_i, i=1,n. Notăm cu $Y[k]$ elementul de rang *k* al seriei (cel care are *k*-1 elemente mai mici). Dacă notăm cu *y*(*p*) percentila de ordin *p*, aceasta se calculează astfel:
$y(p) = Y[k] + d \cdot (Y[k+1]-Y[k])$,
unde:
k este parte întreagă din $p \cdot (n+1)/100$ și reprezintă numărul valorilor din serie mai mici decât percentila de ordin *p*, iar *d* reprezintă $p \cdot (n+1)/100-k$ (partea zecimală a numărului real $p \cdot (n+1)/100$) și reflectă distanța procentuală la care se află percentila de elementul $Y[k]$. Valoarea *d* locul unde se află percentila față de valorile din jurul ei.
Exemplu.
Y = (25,10,1,1200,1010).

Căutăm percentila de ordin 61.
Prin sortarea crescătoare a lui *Y* obținem:
Y = (1,10,25,1010,1200).
$p\cdot(n+1)$ = 61*6/100=3.66
Rezultă: k = 3, d = 0.66.
$y(61) = Y_3+d(Y_4-Y_3) = 25 + 0.66\cdot(1010-25) = 675.1$

Cuartila inferioară, notată cu Q_1, este percentila de ordinul 25.
Cuartila de mijloc, notată cu Q_2, este percentila de prdinul 50.
Cuartila superioară, notată cu Q_3, este percentila de ordinul 75.
Indicatorul interquartile este diferența dintre cuartila superioară și cea inferioară.
Modulul este valoarea cea mai probabilă. În mod uzual modulul se determină ca valoarea cu frecvența cea mai mare.
Amplitudinea reprezintă diferența dintre valoarea cea mai mare și valoarea cea mai mică a unei variabile aleatoare:
$A = X_{max} - X_{min}$.
Abaterea medie absolută caracterizează împrăștierea valorilor unei variabile aleatoare:

$d = \sum_{x\in R}|x-\mu|\cdot f(x)$, pentru cazul discret,

$d = \int_R |x-\mu|\cdot f(x)dx$, pentru cazul continuu.

Pentru o repartiție uniformă, cu f(x) = $\frac{1}{n}$, unde *n* reprezintă numărul valorilor posibile,

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

Varianța caracterizează cel mai bine împrăștierea valorilor unei variabile aleatoare. Așa cum sugerează și numele, este o măsură a variabilității valorilor posibile luate de variabilă:

$$\sigma^2 = \sum_{x\in R}(x-\mu)^2\cdot f(x),\ \sigma^2 = \int_R (x-\mu)^2\cdot f(x)dx,\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Pentru comparabilitate cu valorile variabilei aleatoare, din punct de vedere al unităților de măsură, se utilizează abaterea medie pătratică sau ***abaterea standard***: $\sigma = \sqrt{\sigma^2}$.

Coeficientul de variație se calculează ca raport între abaterea standard și media variabilei: $C_v = \frac{\sigma}{\mu}$. Prin natura calculului, coeficientul de variație este standardizat, nedepinzând de unitățile de măsură ale variabilelor. O variabilă este cu atât mai omogenă cu cât coeficientul de variație este mai apropiat de 0.
Momentele centrate de ordin k au în plus față de momentele simple, diferența față de medie, astfel:

$MC_k(X) = MC_k = \sum_{x\in R}(x-\mu)^k\cdot f(x)$ - pentru cazul discret;

$MC_k(X) = MC_k = \int_R (x-\mu)^k\cdot f(x)dx$ - pentru cazul continuu.

Momentul centrat de ordin doi este varianța.

Asimetria măsoară gradul în care valorile sunt distribuite de o parte sau de alta a valorii centrale:

$$S = \frac{MC_3}{\sigma^3}$$

Cu cât valoarea lui S este mai apropiată de 0 cu atât distribuția este mai simetrică. Valorile negative indică asimetrie stânga în timp ce valorile pozitive indică asimetrie dreapta.

Aplatizarea:

$K = \frac{MC_4}{\sigma^4}$ sau $K = \frac{MC_4}{\sigma^4} - 3$

Cu cât valoarea lui K este mai apropiată de 0 cu atât distribuția va fi mai aplatizată. A doua formulă are ca punct de referință repartiția normală. Astfel, repartițiile mai aplatizate au valori negative pentru K.

Distribuții empirice

Indicatorii prezentați se referă la nivelul întregii populații studiate. În majoritatea situațiilor, comportamentul unei variabile aleatoare la nivelul întregii populații nu poate fi studiat din cauza problemelor de obținere completă a informațiilor. Studiul efectiv al comportamentului unei variabile se face pe mulțimea observațiilor aparținând unor eșantioane ale colectivității generale. Eșantionul este fomat din mulțimea observațiilor $\{x_1, x_2, ..., x_T\}$ unde T reprezintă volumul eșantionului. Prin distribuție empirică se înțelege mulțimea valorilor observate aparținând eșantionului.

În cadrul unui eșantion densitatea de probabilitate are forma: $f_T(X) = \frac{1}{T}$ și se numește densitatea de probabilitate empirică. Prin urmare, media și varianța acestei distribuții sunt: $\bar{x} = \frac{1}{T}\sum_{i=1}^{T} x_i$, $\sigma^2 = \frac{1}{T}\sum_{i=1}^{T}\left(x_i - \bar{x}\right)^2$ sau $\sigma^2 = \frac{1}{T-1}\sum_{i=1}^{T}\left(x_i - \bar{x}\right)^2$ pentru T-1 grade de libertate.

Teste de concordanță

Un test de concordanță este o ipoteză statistică, o presupunere cu privire la caracteristicile unei repartiții, existența unei legi de repartiție. Ca în orice test statistic sunt definite două alternative:
- ipoteza nulă sau H0 constând în afirmația făcută;
- ipoteza alternativă sau H1 care constă în non-afirmație.

Un test statistic este o procedură specifică în urma căreia se trage o concluzie logică privind afirmația din ipoteza nulă: este adevărată sau falsă. Această procedură este una probabilistică. Testul are asociat un grad de încredere. În cazul testelor de concordanță este verificată ipoteza că o distribuție empirică este distribuită după o lege de probabilitate specificată, sau că două distribuții empirice urmăresc aceeași lege. Utilizarea clasică este cea legată de "concordanța" dintre modelul empiric și modelul teoretic considerat adecvat pentru populația din care provin datele statistice. În orice test sunt calculate două mărimi:
- valoarea calculată a testului sau ***valoarea critică***,
- ***valoarea efectivă*** a testului sau statistica testului.

Valoarea critică depinde de gradul în care sunt acceptate valori marginale, caracterizate prin densități mici de probabilitate. Acesta este ***pragul de semnificație*** și reflectă zona de respingere a ipotezei nule. Complementar, ***gradul de încredere*** reflectă zona de acceptare. Dacă valoarea efectivă este mai mică sau egală decât valoarea critică, ipoteza H0 este acceptată, altfel este respinsă.

Metodele de analiză a datelor adeseori fac presupuneri cu privire la distribuții, prepuneri care trebuie verificate. Din multitudinea de teste de concordanță, două se detașează ca frecvență de utilizare: testul $\chi 2$ și testul Smirnov-Kolmogorov.

Testul $\chi 2$

Testul $\chi 2$ este un test general, care poate fi aplicat oricărei distribuții empirice căreia putem sa îi calculăm funcția de repartiție. Testul $\chi 2$ se aplică datelor grupate (sau datelor de frecvență). Algoritmic, testul se aplică astfel:

1. Fie distribuția empirică $X = \{x_1, x_2, ..., x_T\}$. Vor fi împărțite observațiile în m grupe și se vor determina frecvențele absolute ale grupelor:
fa_i, $i = 1,m$
2. Se calculează frecvențele medii estimate prin funcția de repartiție testată:
$fe_i = T\cdot(F(l_{i+1})-F(l_i))$, $i = 1, m$,
unde F este funcția de repartiție testată iar l_i, $i = 1,m+1$ sunt limitele grupelor
3. Se calculează valoarea efectivă a testului sau statistica testului:

$$\chi^2_{Calculat} = \sum_{i=1}^{m}\frac{(fa_i - fe_i)^2}{fe_i}$$

4. Se determină valoarea critică a testului $\chi^2_{Critic}(\alpha ; m - c + 1)$

unde:
- α este nivelul (pragul) de semnificație al testului;
- c este numărul de parametri ai distribuției F (distribuția normală-gaussiană are doi parametrii, media și abaterea standard);
- $m-c+1$ numărul de grade de libertate ale distribuției $\chi 2$.

Această valoare se calculează aplicând funcția de repartiție a distribuției $\chi 2$ pentru parametrii specificați.
5. Sunt testate ipotezele:
H0 - distribuția X urmează legea de repartiție F

H1 - distribuția X nu urmează legea de repartiție F
Decizia asupra acceptării sau respingerii ipotezei H0 se ia astfel:
dacă $\chi^2_{Calculat} \leq \chi^2_{Critic}$ atunci se acceptă ipoteza nulă, respectiv datele provin din distribuția testată
altfel se respinge ipoteza nulă, respectiv datele nu provin din distribuția testată.

Testul Smirnov-Kolmogorov

Este utilizat pentru testarea ipotezei de normalitate. Etapele algoritmului:
1. Fie distribuția empirică $X = \{x_1, x_2, ..., x_T\}$. Se calculează media distribuției și abaterea standard, μ și σ.
2. Se ordonează crescător valorile eșantionului şi se obţine eşantionul ordonat:
$x_{(1)}, x_{(2)}, ..., x_{(T)}$
3. Se calculează funcția de repartiție normală pentru valorile ordonate:
$F(x_{(1)}), F(x_{(2)}), ... , F(x_{(T)})$
4. Se calculează funcția de repartiție empirică:
$Fe(x_{(j)}) = \frac{j}{T}$, j=1,T, deoarece densitatea de probabilitate pentru repartiția empirică este $\frac{1}{T}$
4. Se calculează valoarea efectivă a testului sau statistica testului:
$D = \underset{j}{Max} \left| Fe\left(x_{(j)}\right) - F\left(x_{(j)}\right) \right|$
5. Se determină valoarea critică a testului, $d_{1-\alpha,T}$, unde 1-α este gradul de încredere
6. Se ia decizia astfel:
-dacă $D \leq d_{1-\alpha,T}$ se acceptă ipoteza normalității cu un grad de încredere 1-α
- dacă $D > d_{1-\alpha,T}$ se respinge ipoteza normalității cu un grad de încredere 1-α

Relația dintre două variabile cantitative. Legătura liniară simplă

Dacă se notează cu X şi cu Y două variabile cantitative şi cu x_i şi y_i valorile luate de variabile pentru individul i, legătura liniară simplă dintre cele două variabile este dată de relația:

$$y_i = ax_i + b + e_i, \quad i = 1,n$$

unde e_i este un termen rezidual.

Problema care se pune este de a măsura ***intensitatea legăturii*** dintre cele două variabile deoarece legătura nu este de regulă absolută. De exemplu, dacă urmărim variabilele *greutate* şi *talie* la un grup de persoane vom observa că ele variază în general împreună şi în același sens. Există însă situații în care indivizi cu talie mai mică pot avea greutăți mai mari decât indivizi cu talie mai mare.

Relația dintre variabilele X şi Y va fi cu atât mai intensă cu cât valorile reziduale e_i vor fi mai mici. Din punct de vedere matematic vom determina parametrii a și b astfel încât $\sum_{i=1}^{n} e_i^2$ să fie minimă.

Soluția acestei probleme obținută aplicând regula celor mai mici pătrate este:

$$\begin{cases} a = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{Var(X)} \\ b = \bar{y} - a\bar{x} \end{cases}$$

Dacă se notează covarianța dintre cele două variabile cu $\text{Cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ rezultă:

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var(X)} \\ b = \bar{y} - a\bar{x} \end{cases}.$$

Fluctuațiile variabilei Y măsurate prin variantă, Var(y) reprezintă ***varianța totală***. Fluctuațiile valorilor calculate pentru Y, care depind de X, sunt măsurate prin varianța Var(ax+b) și reprezintă ***varianța explicată***. Fluctuațiile valorilor reziduale , Var(e), reprezintă ***varianța reziduală***. Relația dintre cele trei varianțe este următoarea:

Varianța totală =	Varianța explicată +	Varianța reziduală
Var(y) =	Var(ax+b) +	Var(e)

$$\text{Var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Deoarece, $y_i = ax_i + b + e_i$ si $b = \bar{y} - a\bar{x}$, rezultă: $y_i - \bar{y} = a(x_i - \bar{x}) + e_i$. Înlocuind în relația variantei se obține:

$$\text{Var(y)} = \frac{1}{n}\sum_{i=1}^{n}(a(x_i - \bar{x}) + e_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(a^2(x_i - \bar{x})^2 - 2a(x_i - \bar{x})e_i + e_i^2) =$$

$$\frac{a^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2 Var(x) = Var(ax) = Var(ax + b)$$

$$\frac{2a}{n}\sum_{i=1}^{n}(x_i - \bar{x})e_i = 2aCov(x, e) = 2aCov(x, y - ax - b) =$$

$$2a(Cov(x, y) - Cov(x, ax + b))$$

$$=2a(\text{Cov}(x,y) - a\ \text{Cov}(x,x)) = 2a\ (\ \text{Cov}(x, y) - \frac{Cov(x, y)}{Var(x)}Var(x)) = 0.$$

$$\frac{1}{n}\sum_{i=1}^{n}e_i^2 = Var(e)$$

Se măsoară intensitatea legăturii dintre *X* şi *Y* prin raportul dintre varianţa explicată şi varianţa totală. Acest raport , numit ***raport de corelație*** (sau ***coeficient de determinare***) este notat $R^2(x,y)$:

$$R^2(x, y) = \frac{Var(ax + b)}{Var(y)} = a^2\frac{Var(x)}{Var(y)} = \frac{Cov(x, y)^2}{Var(x)Var(y)}.$$

Rădăcina din R^2 este numit coeficient de corelație liniară și este:

$$R = \frac{Cov(x, y)}{\sigma_x \sigma_y}.$$

Interpretarea geometrică a coeficientului de corelație

O variabilă *X* luând *n* valori poate fi reprezentată printr-un vector în spațiul R^n, numit și spațiul variabilelor. În spaţiul R^n produsul scalar simplu dintre doi vectori *X* și *Y* de coordonate $(x_1,...,x_n)$ și $(y_1,...,y_n)$ este:

$$<x,y> = \sum_{i=1}^{n}x_i y_i$$

iar normele celor doi vectori sunt :

$$\|x\| = \sqrt{< x, x >} = \sqrt{\sum_{i=1}^{n}x_i^2}\ .$$

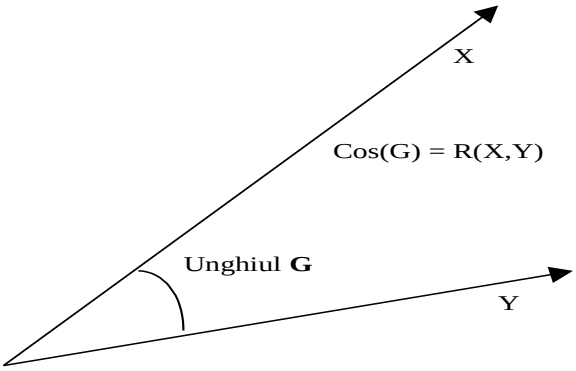
$$\|y\| = \sqrt{\sum_{i=1}^{n}y_i^2}\ .$$

Cosinusul unghiului dintre cei doi vectori este :

$$Cos(x, y) = \frac{< x, y >}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2}}.$$

Dacă se consideră vectorii *X* şi *Y* două variabile centrate, din relaţia anterioară obţinem:

$$\text{Cos}(X,Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \text{R}(X,Y)$$



Când coeficientul de corelaţie este egal cu 1 cei doi vectori sunt coliniari. Absenţa corelaţiei se traduce printr-o valoare nulă pentru R, deci între cei doi vectori este un unghi de 90 de grade.