Descrierea unei variabile calitative

Principalii indicatori care sunt calculați pentru variabilele calitative sunt:
- frecvența absolută care reprezintă numărul de indivizi la care se înregistrează o anumită modalitate
- frecvența relativă care reprezintă frecvența absolută raportată la numărul de indivizi.

Legătura dintre două variabile calitative. Testul de independență $\chi 2$

Fie două distribuții $X = \{x_1, x_2, \dots, x_T\}$ și $Y = \{y_1, y_2, \dots, y_T\}$. Variabila X are p modalități iar variabila Y, q modalități. Frecvențele încrucișate sunt memorate în tabelul N:

$N = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1q} \\ n_{21} & n_{22} & \dots & n_{2q} \\ \dots & & & \\ n_{p1} & n_{p2} & \dots & n_{pq} \end{bmatrix}$. Un element oarecare, n_{ij}, reprezintă numărul de instanțe la care variabila X are modalitatea i iar variabila Y are modalitatea j. Frecvențele cumulate pe linii și coloană sunt definite astfel:

$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$, $i = 1, p$ - numărul de instanțe la care se întâlnește modalitatea i pentru variabila X;

$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$, $j = 1, q$ - numărul de instanțe la care se întâlnește modalitatea j pentru variabila Y.

Testul $\chi 2$ este utilizat pentru a stabili dacă există o legătură între cele două variabile calitative (nominale). Ipoteza nulă specifică faptul că nu există o relație între cele două variabile, adică:
H0: Cele două variabile sunt independente
H1: Cele două variabile sunt dependente
Pașii aplicării testului sunt:
1. Se calculează frecvențele medii estimate:

$ne_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{T}$, $i = 1,p$, $j = 1,q$

2. Se calculează statistica testului:

$$\chi^2_{Calculat} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - ne_{ij}\right)^2}{ne_{ij}}$$

3. Se calculează valoarea critică a testului: $\chi^2_{Critic}(\alpha; r)$ unde α este pragul de semnificație al testului iar r este numărul gradelor de libertate, în acest caz $r = (p\text{-}1)\cdot(q\text{-}1)$.
4. Decizia asupra acceptării sau respingerii ipotezei H0 se ia astfel:

dacă $\chi^2_{Calculat} > \chi^2_{Critic}(\alpha; r)$ atunci se respinge ipoteza nulă cu un nivel de încredere 1-α, deci cele două variabile se influențează reciproc.

În cele mai multe situații se utilizează frecvențe relative. Acestea se determină prin raportarea frecvențelor absolute la numărul de instanțe. Frecvențele relative se memorează într-un tabel $F = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1q} \\ f_{21} & f_{22} & \dots & f_{2q} \\ \dots & & & \\ f_{p1} & f_{p2} & \dots & f_{pq} \end{bmatrix}$.

Statistica testului se va calcula astfel:

$$\chi^2_{Calculat} = T \cdot \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)^2}{f_{i\bullet} f_{\bullet j}},$$

unde $f_{ij} = \frac{n_{ij}}{T}$, $f_{i\bullet} = \frac{n_{i\bullet}}{T}$, $f_{\bullet j} = \frac{n_{\bullet j}}{T}$.

Descrierea indivizilor

Un individ este descris prin mulțimea de valori luate de un grup de variabile pentru individul respectiv.
Se notează cu X matricea valorilor luate de n instanțe (indivizi) pentru m variabile:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & & \dots & \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

Fiecărui individ i se asociază o ***pondere***, p_i , i=1,n. Suma ponderilor este 1. Dacă indivizilor li se asociază aceeași pondere atunci $p_i = \frac{1}{n}$.

Mulțimea punctelor date de cei n indivizi care formează colectivitatea studiată corespunde unui nor de puncte m-dimensionale (m fiind numărul variabilelor).

Se poate defini baricentrul punctelor sau ***centrul de greutate*** al norului , vectorul mediilor aritmetice ale celor m variabile care descriu colectivitatea:

$$g = \sum_{i=1}^{n} p_i w_i = \sum_{i=1}^{n} p_i \begin{bmatrix} x_{i1} \\ \dots \\ x_{im} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ \dots \\ \sum_{i=1}^{n} p_i x_{im} \end{bmatrix},$$

unde $w_i = \begin{bmatrix} x_{i1} \\ \dots \\ x_{im} \end{bmatrix}$ este vectorul coloană reprezentând valorile luate de cele m variabile pentru individul i.

Vom numi ***inerţie totală*** a norului de puncte , media ponderată a pătratelor distanțelor punctelor față de centrul de greutate:

$I_g = \sum_{i=1}^{n} p_i (w_i - g)^t M (w_i - g)$, unde $(w_i\text{-}g)^t$ este un vector linie, transpusul vectorului coloană w_i-g, iar M este metrica utilizată (tipul de distanță).

Inerția norului în raport cu un punct oarecare din spaţiu, h, este:

$$I_h = \sum_{i=1}^{n} p_i (w_i - h)^t M (w_i - h).$$

Relația dintre cele doua valori ale inerției este:

$I_h = I_g + (g-h)^t M (g-h)$. (***Relaţia lui Huygens***)

Inerția în raport cu centrul de greutate este minimă.

Măsuri de asemănare

Se numește *măsură de asemănare* (similaritate/disimilaritate) orice aplicație cu valori numerice care permite să se exprime o legătură între indivizi, sau între variabile. Dacă se notează cu Ω mulțimea indivizilor sau variabilelor, un *indice de similaritate* pe mulțimea Ω este o aplicație s care verifică următoarele trei proprietăți:

1. s este o aplicație a lui $\Omega \times \Omega$ în $\mathbf{R}^+$;
2. s este simetrică: $\forall (w,w') \in \Omega\times\Omega$: $s(w,w') = s(w',w)$;
3. $(w,w') \in\Omega\times\Omega$ cu $w \neq w'$: $s(w,w) = s(w',w') > s(w,w')$.

Un *indice de disimilaritate* este o aplicație s' care satisface primele două condiții din definiția indicelui de similaritate, iar condiţia 3 este înlocuită prin cerinţa ca:

3'. $\forall w\in\Omega$: $s'(w,w) = 0$.

O ***distanţă,*** notată cu d^2, este un indice de disimilaritate care verifică în plus următoarele două proprietăți:

4. $d^2(w,w') = 0 \Leftrightarrow w = w'$;
5. $d^2(w,w') \leq d^2(w,w'') + d^2(w'',w')$, $\forall w,w',w'' \in\Omega$ - *inegalitatea triunghiului*.

Dacă luam în considerare doi indivizi din tabelul de observații $w_i = \begin{bmatrix} x_{i1} \\ \dots \\ x_{im} \end{bmatrix}$ și $w_k = \begin{bmatrix} x_{k1} \\ \dots \\ x_{km} \end{bmatrix}$, ($m$ este numărul de variabile) distanța dintre cei doi indivizi se poate defini sub forma:

$d^2(w_i,w_k) = (w_i - w_k)^t Q (w_i - w_k)$

unde Q este o matrice simetrică pozitiv definită numită metrică.
Mai des utilizate sunt următoarele tipuri de distanțe:
1. *Distanţa euclidiană simplă*. Q este matricea unitate.
2. *Distanţa lui Mahalanobis*, aplicată în analiza discriminantă, unde Q reprezintă matricea de covarianţă.

$$V = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_m) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \dots & Cov(X_2, X_m) \\ & \dots & & \\ Cov(X_m, X_1) & Cov(X_m, X_2) & & Cov(X_m, X_m) \end{bmatrix},$$

unde $X_j = \begin{bmatrix} x_{1j} \\ \dots \\ x_{nj} \end{bmatrix}$ sunt înregistrările pentru variabila j la cei n indivizi.

3. *Distanţa* χ^2 este în mod particular bine adaptată tabelelor de frecvenţe. Se aplică în analiza factorială a corespondenţelor iar Q este o matrice a indicatorilor χ^2 calculați pentru fiecare pereche de variabile.

Tabele de date

Un tabel de date este o matrice care se construiește din ansamblul de indivizi şi variabile.

- ***Tabelul de observații***. Un astfel de tabel se obține atunci când elementele matricei sunt valori numerice oarecare. Pe linii sunt așezați indivizii iar pe coloane sunt așezate variabilele urmărite. Fie n numărul de instanțe și m numărul de variabile. Tabelul de observații este de obicei notat:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

- ***Tabelul de contingenţă***. Pentru date calitative. X şi Y două variabile calitative

$Z = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1q} \\ f_{21} & f_{22} & \dots & f_{2q} \\ \dots & & & \\ f_{p1} & f_{p2} & \dots & f_{pq} \end{bmatrix}$, cu frecvenţe relative, unde p este numărul de modalități al variabilei X şi q este numărul de modalităţi al variabilei Y.

O alta formă a tabelului de contingenţă este următoarea:

$$Z = \begin{bmatrix} \text{Modalitatea 1 a variabilei X} & \text{Modalitatea 1 a variabilei Y} & f_{11} \\ \text{Modalitatea 2 a variabilei X} & \text{Modalitatea 1 a variabilei Y} & f_{21} \\ \dots & \dots & \dots \\ \text{Modalitatea } p \text{ a variabilei X} & \text{Modalitatea 1 a variabilei Y} & f_{p1} \\ \text{Modalitatea 1 a variabilei X} & \text{Modalitatea 2 a variabilei Y} & f_{12} \\ \dots & \dots & \dots \\ \text{Modalitatea } p \text{ a variabilei X} & \text{Modalitatea } q \text{ a variabilei Y} & f_{pq} \end{bmatrix}.$$

Acest tabel are $p \cdot q$ linii şi 3 coloane. Primele două coloane cuprind valori nominale pentru variabilele calitative iar a treia coloană cuprinde frecvenţele.

- ***Tabelul disjunctiv complet***. Este utilizat pentru variabile calitative.

$D = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & 1 & \dots \\ & \dots & & \\ 1 & 0 & \dots & 0 \end{bmatrix}$. Are n linii şi p coloane.

n – numărul de instanțe
p – numărul de modalități pentru variabila calitativă, X.
D^tD este matricea diagonală a frecvenţelor absolute.
Pentru v variabile calitative:

$$D = \begin{bmatrix} \overbrace{0 \ \dots \ 1 \ \dots \ 0}^{m_1} & \overbrace{0 \ \dots \ 1 \ \dots \ 0}^{m_2} & \dots & \overbrace{0 \ \dots \ 1 \ \dots \ 0}^{m_v} \\ & & & \end{bmatrix}$$

- ***Tabelul de preferinţe***. Este un tip de tabel pentru variabile calitative, şi exprimă preferinţele unui grup de indivizi faţă de valorile unei caracteristici. Foloseşte o scală de tip ordinal. Tabelele de acest tip sunt frecvent utilizate în studii de marketing. De exemplu, în tabelul următor sunt prezentate preferinţele a patru persoane pentru cinci mărci de parfum:

Persoane interogate	Mărci de parfumuri				
	M_1	M_2	M_3	M_4	M_5
w_1	1	2	5	4	3
w_2	4	2	3	1	5
w_3	5	4	3	1	2
w_4	1	2	3	5	4

Preferinţele sunt exprimate prin note de la 1 loa 5.

- ***Tabelul binar***. Este tabelul care nu conţine decât valori 0 sau 1. Este folosit, ca şi tabelul de preferinţe, în anchete statistice. În exemplul de mai jos se găsesc răspunsurile unor indivizi la întrebarea *Citiţi ziarul Z ?*.

Indivizi	Ziare			
	Z_1	Z_2	Z_3	Z_4
w_1	1	0	0	0
w_2	1	1	0	0
w_3	0	0	1	1

Răspunsurile pot fi *Da* sau *Nu* şi sunt codificate cu 1, respectiv 0.

- ***Tabelul de modalităţi***. Atunci când fiecare întrebare a unei anchete statistice presupune mai multe răspunsuri, ne găsim în faţa unui tabel de modalităţi. Astfel, dacă la întrebarea din exemplul anterior s-ar putea da trei răspunsuri: *Niciodată* - răspuns codificat cu valoarea naturală 1, *Câteodată* - codificat cu 2, *Deseori* - codificat cu 3, tabelul de modalităţi asociat ar putea fi următorul:

Indivizi	Ziare			
	Z_1	Z_2	Z_3	Z_4
w_1	3	2	2	1
w_2	3	3	1	1
w_3	1	2	3	3

- ***Tabelul de proximităţi***. Atunci când se evaluează asemănările sau diferenţele între fiecare cuplu de indivizi, se construieşte un tabel de proximităţi. De exemplu, putem considera patru mărci de autoturisme, pe care să le comparăm unele cu altele. Apropierea dintre o marcă *j* şi o marcă *i* poate fi făcută printr-o notă de la 1 la 10, sau media unor note de la 1 la 10, date, eventual, de un grup de specialişi în domeniu. Se poate observa în Tabelul următor că se foloseşte o scală de tip raport.

Mărci autoturisme	Mărci autoturisme			
	M1	M2	M3	M4
M1	10	4.3	9.3	2.3
M2	4.3	10	7.6	9.3
M3	9.3	7.6	10	3.6
M4	2.3	9.3	3.6	10

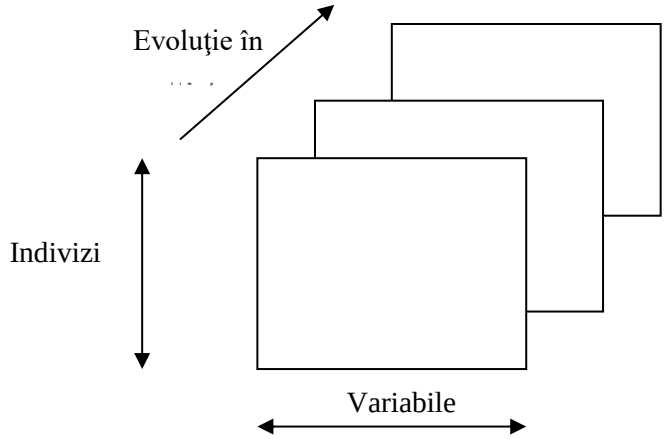
- ***Tabele eterogene***. Sunt tabele în care variabilele sunt de diferite tipuri. Un exemplu de tabel eterogen este următorul:

Produse	Variabile		
	Preţ	Punct de lucru	Mod de transport
p_1	3.5	1	3

p_2	4	3	2
p_3	10	2	1

Parametrului *preţ* i s-a asociat o variabilă cantitativă măsurabilă, parametrului *Punct de lucru* i s-a asociat o variabilă calitativă ordinală, iar parametrului *Mod de transport* i s-a asociat tot o variabilă calitativă ordinală.

- ***Tabele tridimensionale***. Pot fi oricare din tipurile de tabele deja prezentate, la care se adaugă o a treia dimensiune, de exemplu, timpul. Astfel, pentru tabelul eterogen de mai sus, dacă evaluarea se face la unumite intervale de timp, se obţine un tabel cu trei dimensiuni conform figurii:



Schimbarea de variabilă

Necesitate:
a) Când datele sunt grupate ***în tabele eterogene*** şi se doreşte exprimarea unei variabile descriptive cu ajutorul alteia, astfel încât toate variabilele să devină de acelaşi tip.
b) ***Pentru a putea aplica o anumită metodă de analiză*** a datelor în situaţia în care aceasta este incompatibilă cu tipul datelor. De exemplu, prin schimbare de variabilă se poate transforma un tabel de date calitative într-un tabel de modalităţi, putând astfel aplica într-o manieră mai eficientă analiza factorială a corespondenţelor multiple.
c) ***Pentru a sintetiza informaţia*** conţinută într-un tabel de date, reducându-i astfel mărimea. Se poate, de exemplu, înlocui mulţimea de variabile prin care se descrie o colectivitate printr-o combinaţie liniară de aceste variabile

Modalităţi de schimbare a variabilelor:
- Schimbarea de variabilă prin standardizare
- Schimbarea de variabila prin normalizare
- Schimbarea de variabilă prin codificare
    - Codificarea unei variabile cantitative prin grupare
        - Gruparea pe efective egale
        - Gruparea în intervale egale
        - Gruparea prin minimizarea inerţiei totale
    - Codificarea variabilelor calitative
        - Codificarea cu structură de ordine
        - Codificarea fără structură de ordine
        - Codificarea prin rangul mediu

Codificarea disjunctivă completă a variabilelor nominale