# Capitolul 6

# TRATAREA SIMULTANĂ A DOUĂ VARIABILE CALITATIVE

Dependența în planul datelor a două variabile calitative, se numește **asociere**. Şi în acest caz funcționează **postulatul epistemologic**:

> "O legătură în plan fenomenologic implică asocierea celor două variabile calitative corespunzătoare în planul datelor".

Reciproca nu este valabilă şi deci, ca şi în cazul variabilelor cantitative, există **asocieri aparente** sau **false asocieri**.

Pentru măsurarea asocierii ori pentru susținerea absenței sale, şirurile bivariate de variante ale celor două variabile calitative se grupează în tabele cu dublă intrare, numite *tabele de contingență.*

*Definiție:* Se numeşte **tabel de contingență** rezultatul "ventilării unei populații după variantele a două caracteristici calitative" [3].

**Exemplul 6.**

Să presupunem că avem următorul şir bivariat de specii, respectiv parcele în care s-au întâlnit speciile respective:

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parcela: | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Specia: | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | | |

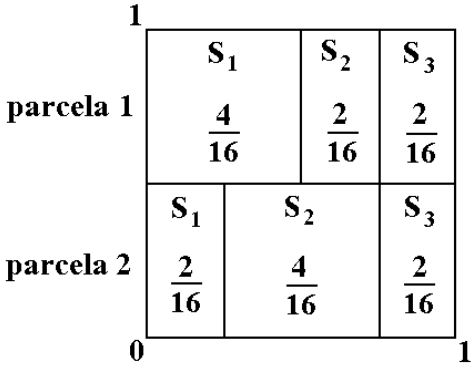
Tabelul de contingență corespunzător va fi format din 2 linii corespunzătoare celor 2 parcele şi 3 coloane corespunzătoare celor 3 specii şi va conține, în celula de pe linia *i* şi coloana *j*, numărul de perechi (*i*, *j*) din şirul bivariat. In final sunt adăugate coloana şi linia „totalurilor marginale" (coloana totalurilor pe linii şi linia totalurilor pe coloane) precum şi totalul general:

| | **specia 1** | **specia 2** | **specia3** | Totaluri pe linii: |
|---|---|---|---|---|
| ***parcela 1*** | 4 | 2 | 2 | 8 |
| ***parcela 2*** | 2 | 4 | 2 | 8 |
| *Totaluri pe coloane:* | 6 | 6 | 4 | Total general: **16** |

## § 6.1. Sinteza grafică

Un tabel de contingenţă poate fi reprezentat prin **areale în dreptunghiuri** (în cadrul unui pătrat [16]).

Alături este reprezentat tabelul de contingenţă anterior.




## § 6.2. Sinteza numerică bivariată

Pentru perechile de variabile cantitative am văzut că dependenţa statistică apare plasată între independenţă şi dependenţa funcţională. În mod analog, aici vom plasa asocierea între independenţă şi ceea ce se numeşte [32] *asocierea completă.*

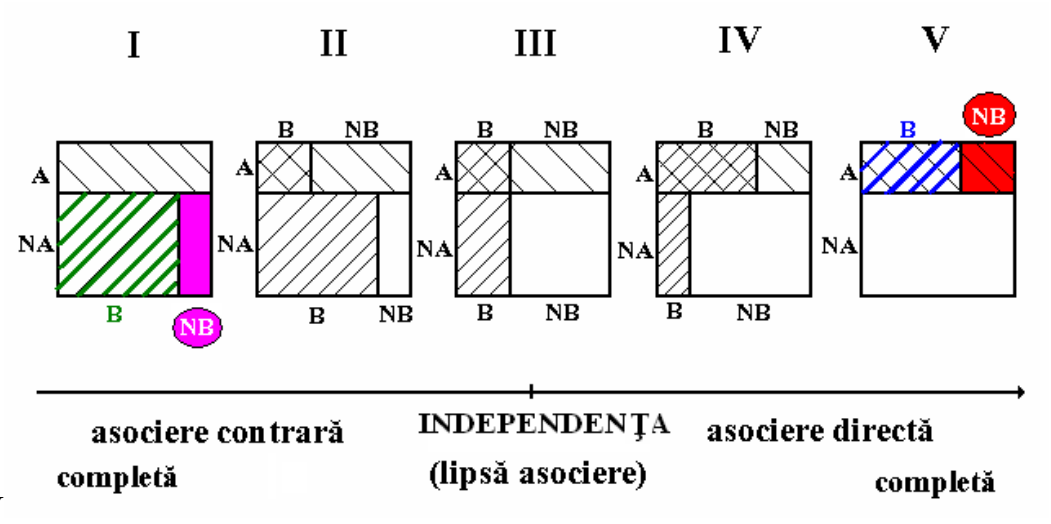
### 6.2.1. Tabele $2 \times 2$

În cazul variabilelor binare, care produc tabele de contingenţă $2 \times 2$, analogia merge mai departe şi anume cu dependenţele liniare, unde vorbim despre *corelaţie directă*, respectiv *inversă,* aici propunem să vorbim despre *asociere directă* respectiv *contrară*.

Să presupunem că dorim să exprimăm dependenţa între o anumită specie şi altitudinea la care este găsită. De exemplu, se ştie că floarea de colţ creşte numai la altitudine mare. Spunem că există o asociere completă directă între altitudine şi floarea de colţ. Analog, vom spune că există o asociere completă CONTRARA[24] între altitudine şi grâu, deoarece grâul creşte numai acolo unde NU exista altitudine (mare). Prin urmare, două variabile calitative binare *a* (cu variantele *A* şi *NA* - citeşte *Non A*) şi *b* (cu variantele *B* şi *NB*) se numesc **complet asociate în mod direct**, dacă *B* are loc doar în prezenţa lui *A* sau invers. Se vor numi **complet asociate în mod contrar**, dacă *B* are loc doar în absenţa lui *A* sau invers. În exemplele anterioare *a* este variabila binară "altitudine", cu variantele *A* = "altitudine" şi *NA* = "Non altitudine", iar *b* este variabila specie cu variantele *B* = "floare de colţ" şi *NB* = "Non floare de colţ", în primul caz, respectiv *B* = "grâu" şi *NB* = "Non grâu", în al doilea caz.

Exemplul cu floarea de colţ este reprezentat în desenul V din figura următoare, iar cel cu grâul în desenul I. Desenul IV arată o situaţie în care *B* are loc în prezenţa lui A, dar şi în absenţa sa (adică prezenţa lui NA). Proporţia prezenţei lui *B* în *NA* este însă mai mică decât proporţia prezenţei în *A* (vezi aria haşurată exclusiv ascendent raportată la dreptunghiul orizontal inferior în care este inclusă, respectiv aria dublu haşurată raportată la dreptunghiul superior din care face parte). Este deci cazul unei asocieri directe între *A* şi *B* dar nu complete. În mod simetric, desenul II exprimă o asociere contrară dar nu completă.

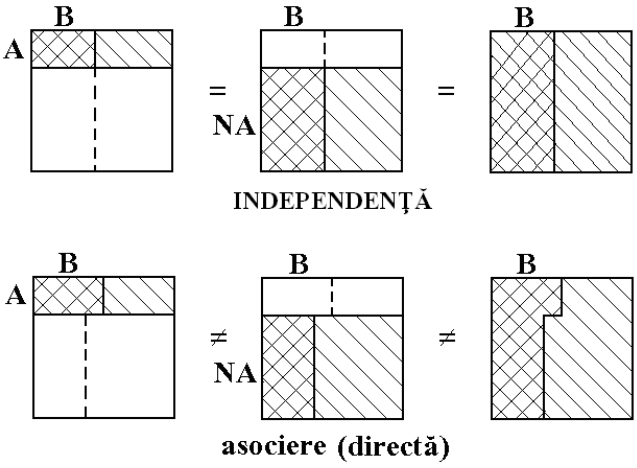
[24] În [32] se vorbeşte despre caracteristici calitative complet asociate, respectiv complet neasociate. Ultima denumire se poate, însă, confunda cu independenţa. De aceea am propus această nouă terminologie.




Y

În fine, desenul III exprimă **independența** între două va-riabile calitative: proporţia lui *B* din A este egală cu proporția lui *B* din NA (ceea ce implică şi egalitatea cu proporţia lui *B*).

În caz contrar, cele două variabile sunt asociate. (Vezi figura alăturată. O prezentare mai ri-guroasă a independenței se află în volumul următor şi în [12].)




**Exemplele 6.2.1.**

a. Să presupunem că am împărțit în două parcele o anumită suprafață de teren şi am numărat pentru fiecare parcelă exemplarele care aparțin, respectiv nu aparțin unei anumite specii *S*, obținând următorul tabel de contingență la care am adăugat totalurile pe linii, pe coloane (denumite totaluri marginale) şi totalul general.

| | **"specia *S*"=*B*** | **"specie ≠ *S*"=*NB*** | |
|---|---|---|---|
| ***"parcela 1"=A*** | 8 | 12 | 20 |
| ***"parcela 2"=NA*** | 22 | 33 | 55 |
| | 30 | 45 | 75 |

Se observă că prin introducerea parcelelor, proporția speciei *S* nu se modifică, adică proporția speciei *S* în parcela 1 (8 / 20 = 2 / 5) este egală cu proporția în parcela 2 (22 / 55 = 2 / 5) şi cu proporția în tot arealul (30 / 75 = 2 / 5). Prin urmare, cele două variabile sunt independente. Altfel spus cele două parcele împart același biotop, doar în mod convențional.

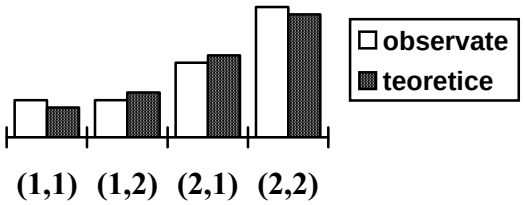
b. Să repetăm problema anterioară schimbând datele din interiorul tabelului, dar păstrând totalurile marginale pentru comparabilitate.

| | **"specia S"=*B*** | **"specie ≠ S"=NB** | |
|---|---|---|---|
| ***"parcela 1"=A*** | 10 | 10 | 20 |
| ***"parcela 2"=NA*** | 20 | 35 | 55 |
| | 30 | 45 | 75 |

În acest caz, observăm deosebiri între proporțiile speciei *S* în cele două parcele: 10 / 20 (= 1 / 2 ) este diferit de 20 / 55 (= 4 / 11), proporțiile fiind diferite, evident şi de proporția speciei *S* în tot arealul (30 / 75 = 2 / 5). Există, deci, o asociere directă (vezi fig. IV de mai sus) între *A* şi *B* (1 / 2 fiind mai mare decât 4 / 11). În consecință, judecând în cadrul statisticii descriptive, cele două parcele delimitează două biotopuri ce conțin biocenoze diferite.

✓ Deoarece totalurile marginale sunt aceleaşi, iar proporțiile din primul tabel (a) respectă condiția de independență, acesta poate fi considerat pentru cel de-al doilea (b), tabel teoretic în ipoteza de independență. Modul în care se calculează, în general, pentru un tabel dat, tabelul cu frecvențele teoretice în ipoteza de independență, este explicat la subpunctul următor.

Judecând astfel, putem măsura asocierea prin depărtarea de la independență. Aceasta, deoarece reperul central al problemei asocierii este independența, aşa cum se observă şi din figurile I-V. Prin urmare, va trebui să apelăm la o măsură a depărtării dintre cele două distribuții, cea de frecvențe observate (*o*) şi cea de frecvențe teoretice în ipoteza de independență (*t*).




Măsura clasică de comparare a două distribuții a fost prezentată în 3.7.7. şi este:

$$\chi^2=\boxed{\frac{\sum(o-t)^2}{t}}\ \text{(formula teoretică)} = \boxed{\sum\frac{o^2}{t}-N}\ \text{(formula de calcul)},$$

unde *N* este totalul general (volumul seriei bivariate).

**Exemplul 6.2.1'.**

În cazul exemplului 5.2.1. vom avea:

$$\chi^2 = \Sigma\, o^2 / t - N = 10^2 / 8 + 10^2 / 12 + 20^2 / 22 + 35^2 / 33 - 75 \cong$$
$$12{,}5 + 8{,}3 + 18{,}2 + 37{,}1 - 75 = 76{,}1 - 75 = 1{,}1.$$

Problema rămasă nerezolvată este cum calculăm tabelul în ipoteza de independență. Vom prezenta acest lucru în continuare, pe cazul general al unui tabel de contingență cu *p* linii şi *q* coloane.

### 6.2.2. $\chi^2$ pentru independență versus asociere, în tabele de contingență $p \times q$

Forma generală a unei tabele de contingență este:

| $y$: $x$ = | **Variantele lui $x$** $x_1\ x_2\ \ldots x_j\ \ldots\ x_q$ | **Sume pe linii** $l(i)$: |
|---|---|---|
| $y_1$ | . | |
| $y_2$ | . | |
| . | . | |
| . | . | |
| . | . | |
| $y_i$ | $\ldots\ o(i,j)\ \ldots$ | $l(i)=\sum_{j=1}^{q} o(i,j)$ |
| . | . | |
| . | . | |
| . | . | |
| $y_p$ | . | |
| **Sume pe coloane** $c(j)$ : | $c(j)=\sum_{i=1}^{p} o(i,j)$ | $N=\sum_{i=1}^{p}\sum_{j=1}^{q} o(i,j)=\sum_{i=1}^{p} l(i)=\sum_{j=1}^{q} c(j)$ |

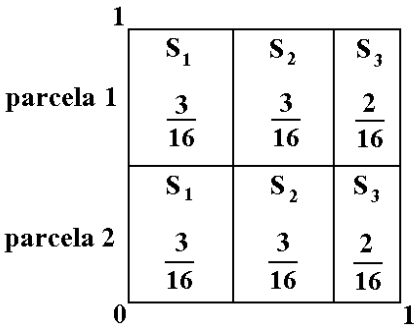
Prin $l(i)$ s-a notat totalul marginal al elementelor de pe linia $i$, iar prin $c(j)$ totalul marginal al elementelor de pe coloana $j$. $N$ este totalul general.

Tabelul de frecvențe teoretice în ipoteza de independență va fi alcătuit din elementele:

$t(i, j) = l(i) \cdot c(j) / N$.

Formula de calcul rezultă din caracterizarea independenței de mai sus[25].

Cititorul va face două exerciții utile calculând frecvențele teoretice[26] pentru tabelul din exemplul 5.2.1.b. (obținând tabelul 5.2.1.a) și pentru tabelul $2 \times 3$ din exemplul 5 (pentru care va obține tabelul din desenul alăturat).




În cazul general formula teoretică, respectiv cea de calcul pentru $\chi^2$ vor fi:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(o(i,j)-t(i,j))^2}{t(i,j)} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{o(i,j)^2}{t(i,j)} - N.$$

Reținem, deci, că $\chi^2$ măsoară depărtarea de la independență, respectiv intensitatea absolută a asocierii.

---

[25] în care înlocuim $A$ cu *varianta i a lui y* și $B$ cu *varianta j a lui x*. Atunci independență înseamnă că proporția variantei $j$ a lui $x$ din întreaga variantă $i$ a lui $y$ ($t(i, j) / l(i)$) este egală cu proporția întregii variante $j$ ($c(j) / N$), adică: $t(i, j) / l(i) = c(j) / N$, de unde rezultă egalitatea de demonstrat.

[26] Frecvențele teoretice pot fi și numere cu zecimale, nu numai numere întregi, ca în exemplele de aici.

**1° Coeficientul de contingență al lui Ciuprov**

Pentru că $\chi^2$ prezintă dezavantajul creşterii o dată cu numărul de linii *p*, de coloane *q* şi numărul de unități statistice *N*, fără a exprima în acelaşi timp o creştere a asocierii, s-a introdus o măsură relativă a asocierii şi anume *coeficientul de contingență al lui Ciuprov*, notat *T*, coeficient al cărui pătrat este dat de formula:

$$T^2 = \frac{\chi^2}{N \cdot \sqrt{(p-1)\cdot(q-1)}}.$$

*T* se obține extrăgându-se radicalul de ordinul doi din $T^2$. Se consideră numai rezultatul pozitiv[27], deoarece în cazul acestui coeficient nu mai există avantajul oferit în cazul variabilelor cantitative, de coeficientul de corelație liniară al lui Pearson, care prin semnul său indică sensul legăturii (corelației). Este natural să fie aşa, deoarece în cazul variabilelor calitative nu mai dispunem de o ordine pentru fiecare variabilă şi, deci, nu putem vorbi de un sens al asocierii decât în cazul variabilelor binare, aşa cum am arătat mai sus în 5.2.1.

**Exemplul 5.2.1''.**

Continuând exemplul numeric de mai sus (5.2.1'.),

$$T^2 = \frac{\chi^2}{N \cdot \sqrt{(p-1)\cdot(q-1)}} = \frac{1{,}1}{75 \cdot \sqrt{(2-1)\cdot(2-1)}} = \frac{1{,}1}{75} \cong 0{,}01,$$

de unde rezultă că $T = \sqrt{0{,}01} = 0{,}1$. Deoarece *T* este diferit de zero, există o asociere între cele două variabile, de intensitate slabă însă, *T* fiind apropiat de 0.

[27] *T* variază între 0 şi 1 dacă $p = q$ [32].