Session
Correspondence analysis

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

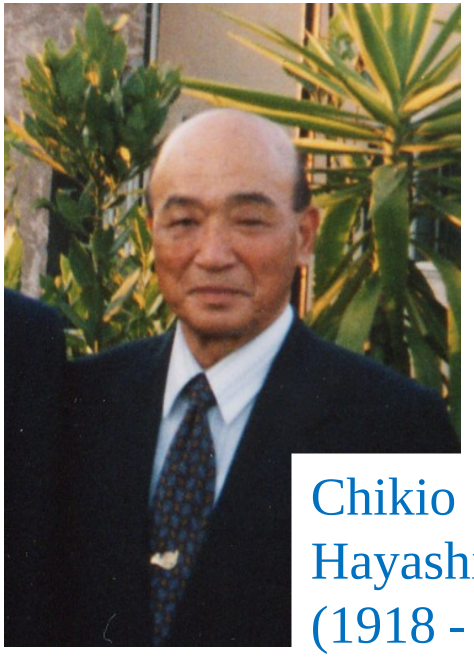
Information System tracking

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu *lidia.montero@upc.edu*

Ronald Aylmer **Fisher,**
1890 –1962

Brigitte Escofier (1941-1994)



Chikio
Hayashi,
(1918 - 2002)

Jean Paul Benzécri (1932)

1. Data and notation

2. Relationship between two categorical variables

3. CA: description of the desviation to independence

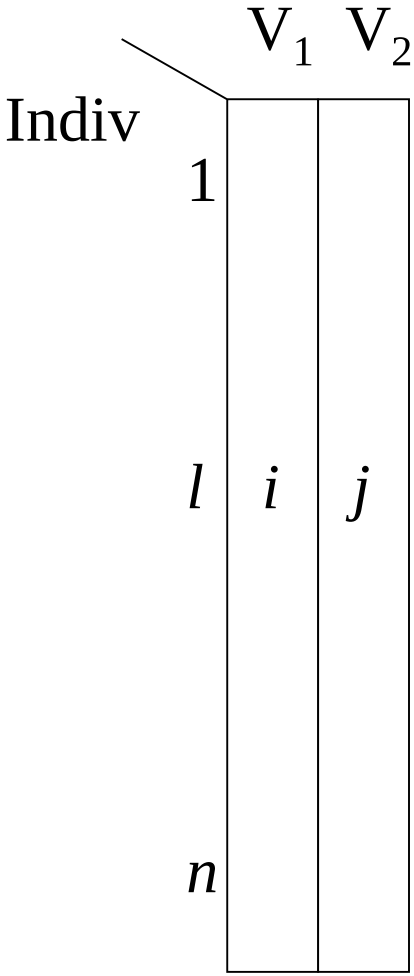
4. Geometrical approach: cloud of rows and cloud of columns. Superposition of both graphics of rows and columns

5. Helps to interpretation

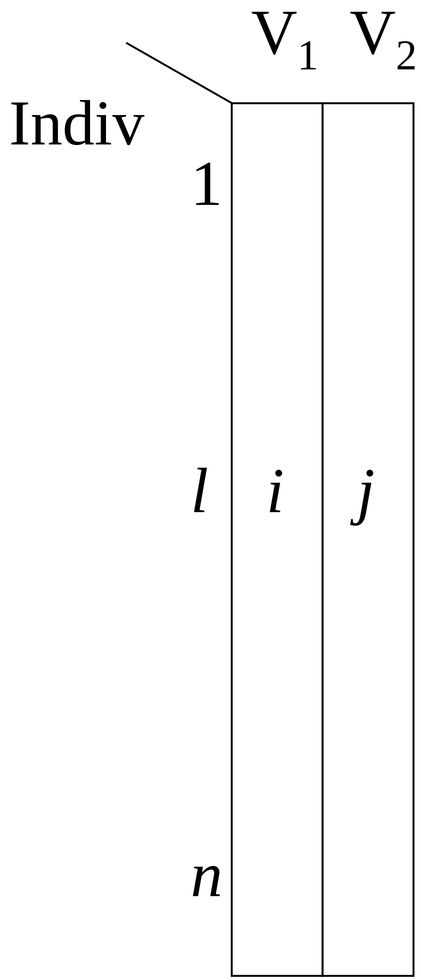
6. Transition relationships: barycentric properties

7. Complements: supplementary elements; intensity of the relationship

1. Data



Two categorical variables observed on the same individuals



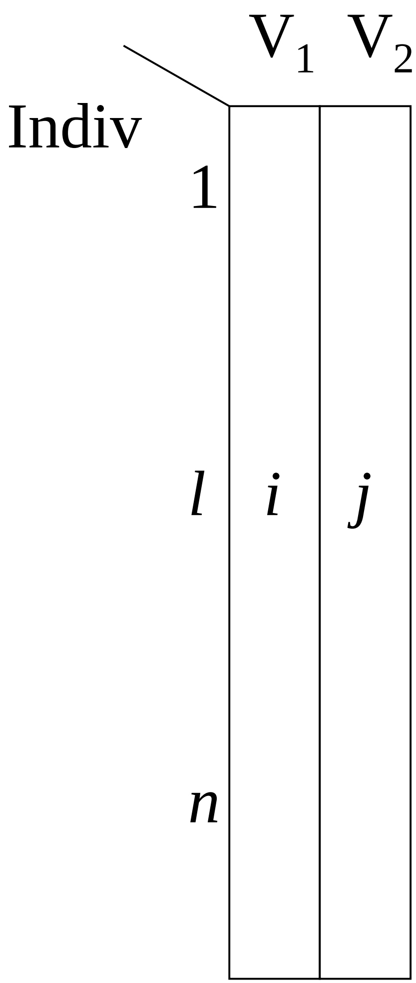
Exemple
Health survey in Croatia

Edad en clase (7 categoríes)
and
Estado de salud (5 categoríes)

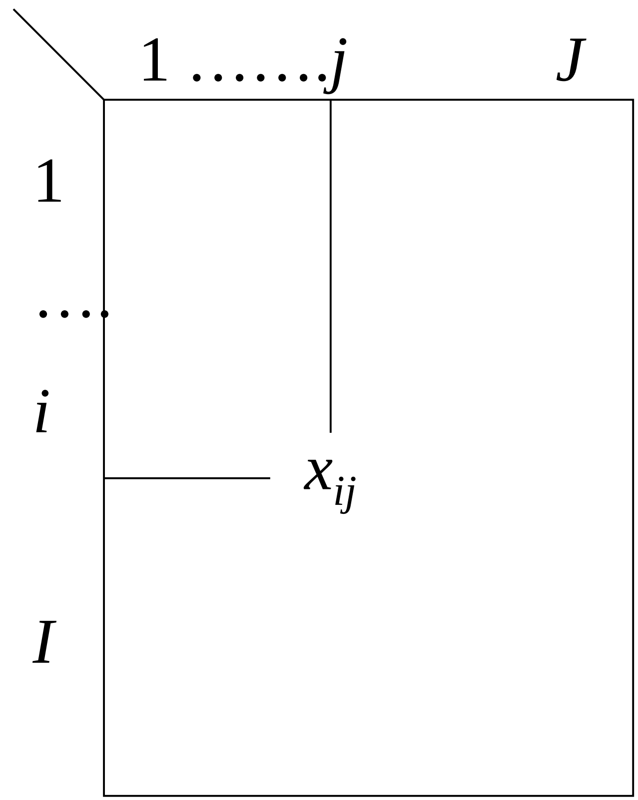
on n=5037 individuals

```
> summary(base$Edad_classe)
  18-25 años   26-35 años   36-45 años  46-55 años   56-65 años  66-75 años 76 y más
        639      833            766        794            798       818           389

> summary(base$B1)
health-excellent health-very good      health-good      health-fair    health-poor
             472              833             1367             1322           1043
```



Contingency table



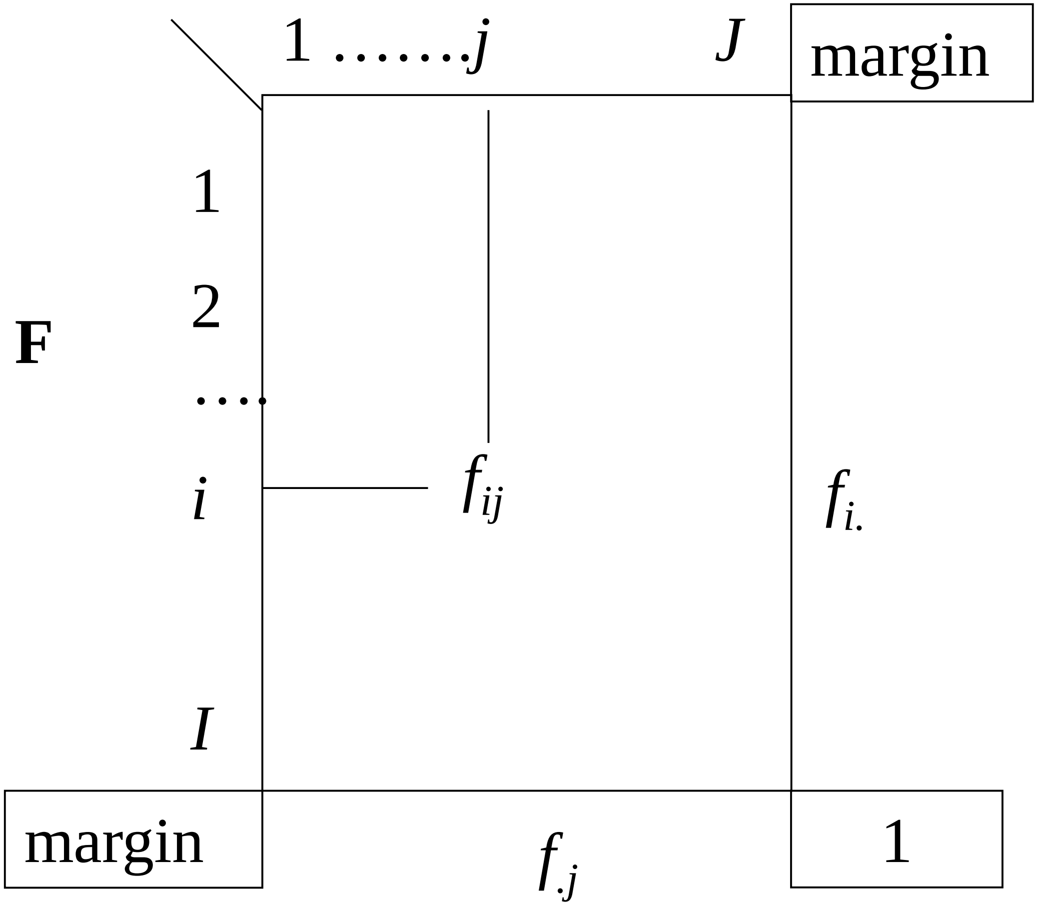
x_{ij} = count of individuals who present at the same time category i of V_1 and category j of V_2

Crossed table and margins

	health-excellent	health-very good	health-good	health-fair	health-poor	
18-25 años	181	216	161	69	12	639
26-35 años	144	263	259	129	38	833
36-45 años	62	150	266	201	87	766
46-55 años	35	105	260	239	155	794
56-65 años	26	43	190	281	258	798
66-75 años	17	38	166	283	314	818
76 y más años	7	18	65	120	179	389
	472	833	1367	1322	1043	5037

From the count table to the proportion table

Table **F**



$$f_{ij} = \frac{x_{ij}}{n}$$

$$f_{i.} = \sum_j f_{ij}$$

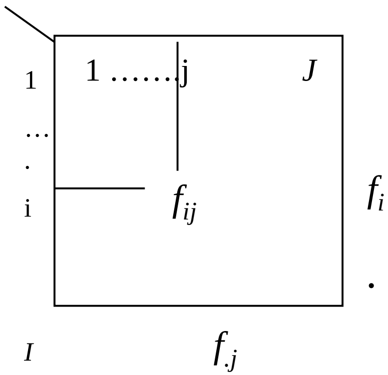
$$f_{.j} = \sum_i f_{i.}$$

Relationship between V_1 and V_2: desviation between data and independence model

Proportion table and margins

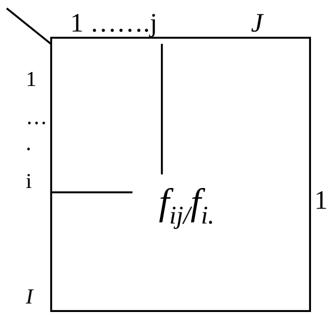
	health-excellent	health-very good	health-good	health-fair	health-poor	
18-25 años	0.036	0.043	0.032	0.014	0.002	**0.127**
26-35 años	0.029	0.052	0.051	0.026	0.008	**0.166**
36-45 años	0.012	0.030	0.053	0.040	0.017	**0.152**
46-55 años	0.007	0.021	0.052	0.047	0.031	**0.158**
56-65 años	0.005	0.009	0.038	0.056	0.051	**0.159**
66-75 años	0.003	0.008	0.033	0.056	0.062	**0.162**
76 y más años	0.001	0.004	0.013	0.024	0.036	**0.078**
	0.093	**0.167**	**0.272**	**0.263**	**0.207**	**1.000**

2. Relationship/independence between two categorical variables



It there were independence

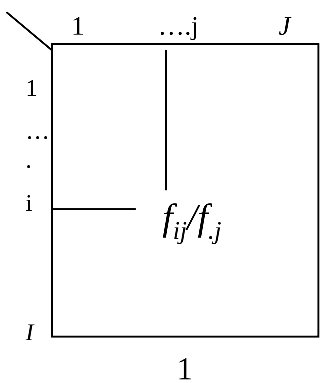
$$f_{ij} = f_{i.} \cdot f_{.j}$$



$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

Row-profiles table

Table $\mathbf{D}_{\mathbf{I}}^{-1}\mathbf{F}$

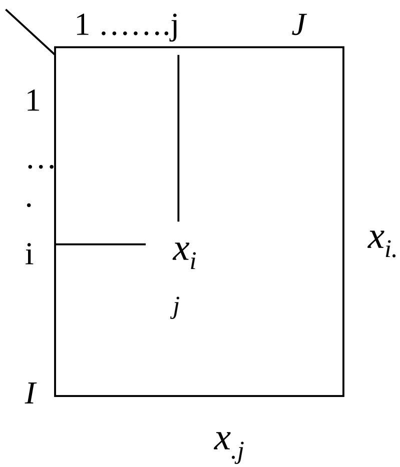


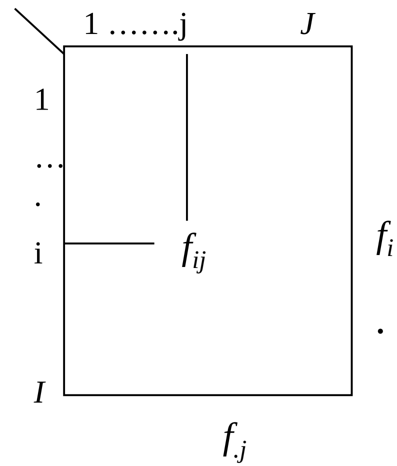
$$\frac{f_{ij}}{f_{.j}} = f_{i.}$$

Column-profiles table

Tabla $\mathbf{F}\mathbf{D}_{\mathbf{J}}^{-1}$

Observed data





Estimation of the independence model

$$\hat{f}_{ij} = f_{i.} \cdot f_{.j}$$

Expected counted under independence hypothesis

$$\hat{x}_{ij} = n \cdot f_{i.} \cdot f_{.j}$$

Significance of the relationship between the two variables: χ^2 test

$$\chi^2 = \sum_{i,j} \frac{\left(x_{ij} - \hat{x}_{ij}\right)^2}{\hat{x}_{ij}}$$

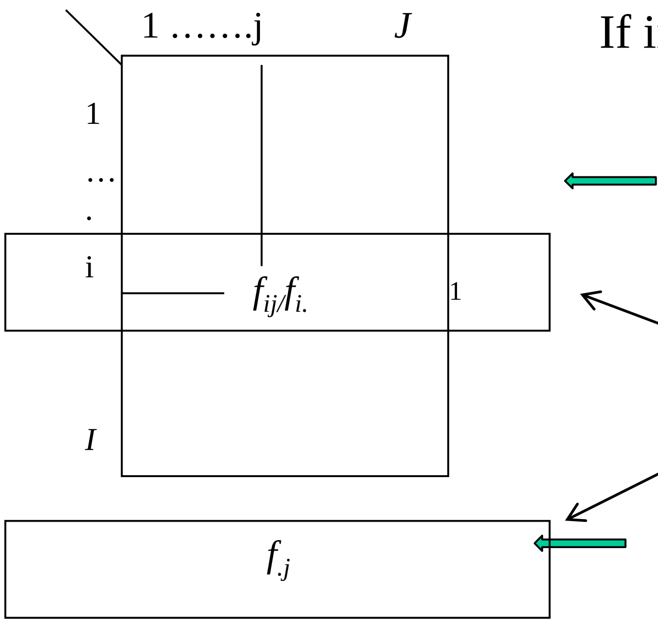
… distribution…*p*-value

Intensity of the relationship between the two variables

$$\Phi^2 = \sum_{i,j} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}} = \frac{\chi^2}{n}$$

CA simultaneously analyzes the row and column profiles tables. It does not inform about the significance of the relationship but about the intensity of the relationship and visualizes the structure of the relatonship.

Comparing profiles



If independence: $\frac{f_{ij}}{f_{i.}} = f_{.j}$

Row profiles table

Comparar y describir las desviaciones

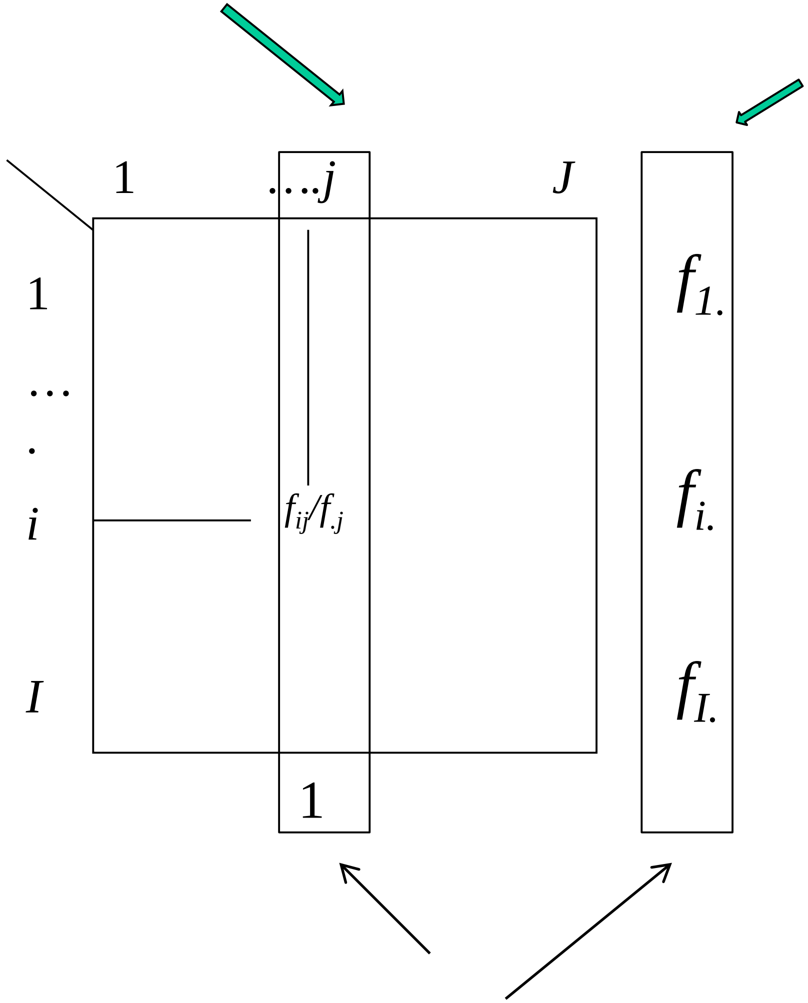
Row margin ot the proportion table
“average profile”

	health-excellent	health-very good	health-good	health-fair	health-poor
18-25 años	0.283	0.338	0.252	0.108	0.019
26-35 años	0.173	0.316	0.311	0.155	0.046
36-45 años	0.081	0.196	0.347	0.262	0.114
46-55 años	0.044	0.132	0.327	0.301	0.195
56-65 años	0.033	0.054	0.238	0.352	0.323
66-75 años	0.021	0.046	0.203	0.346	0.384
76 y más años	0.018	0.046	0.167	0.308	0.460
Perfil-medio	**0.093**	**0.167**	**0.272**	**0.263**	**0.207**

Column profiles table

If there were independence $\frac{f_{ij}}{f_{i.}} = f_{i.}$

Column-margin of the proportion table
“Average column profile”



To compare and describe the deviations

```
> profil.col

             health-excell health-very good health-good health-fair health-poor
  18-25 años         0.383            0.259       0.118       0.052       0.012   0.127
  26-35 años         0.305            0.316       0.189       0.098       0.036   0.166
  36-45 años         0.131            0.180       0.195       0.152       0.083   0.152
  46-55 años         0.074            0.126       0.190       0.181       0.149   0.158
  56-65 años         0.055            0.052       0.139       0.213       0.247   0.159
  66-75 años         0.036            0.046       0.121       0.214       0.301   0.162
  76 y más años      0.015            0.022       0.048       0.091       0.172   0.078
```

3. CA

CA= First, analysis of the cloud of rows

Cloud of rows described by their profile $\dfrac{f_{ij}}{f_{i.}}$ Matrix $\mathbf{D_I^{-1}F}$

Weights of rows $f_{i.}$ stored into diagonal matrix $\mathbf{D_I}$

chi.2 metric with generic term $\dfrac{1}{f_{.j}}$ $\mathbf{D_J^{-1}}$

$$d^2(i,l) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$$

$\longrightarrow$ distributional equivalence

CA= Then, analysis of the cloud of columns

Cloud of columns described by their profile $\dfrac{f_{ij}}{f_{.j}}$ Matrix $\mathbf{D}_{\mathbf{J}}^{-1}\mathbf{F}'$

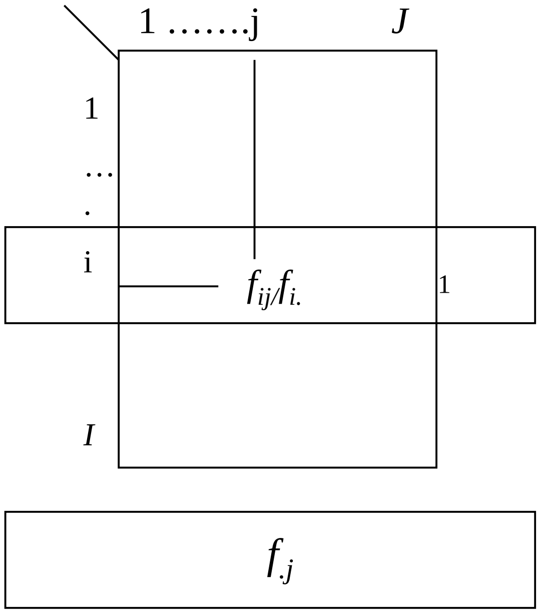
Weighted of the columns $f_{.j}$ stored into diagonal matrix $\mathbf{D}_{\mathbf{J}}$

Métrica del chi.2 with generic term $\dfrac{1}{f_{i.}}$ $\mathbf{D}_{\mathbf{I}}^{-1}$

$$d^2(j,h) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ih}}{f_{.h}} \right)^2$$

$\longrightarrow$ distributional equivalence

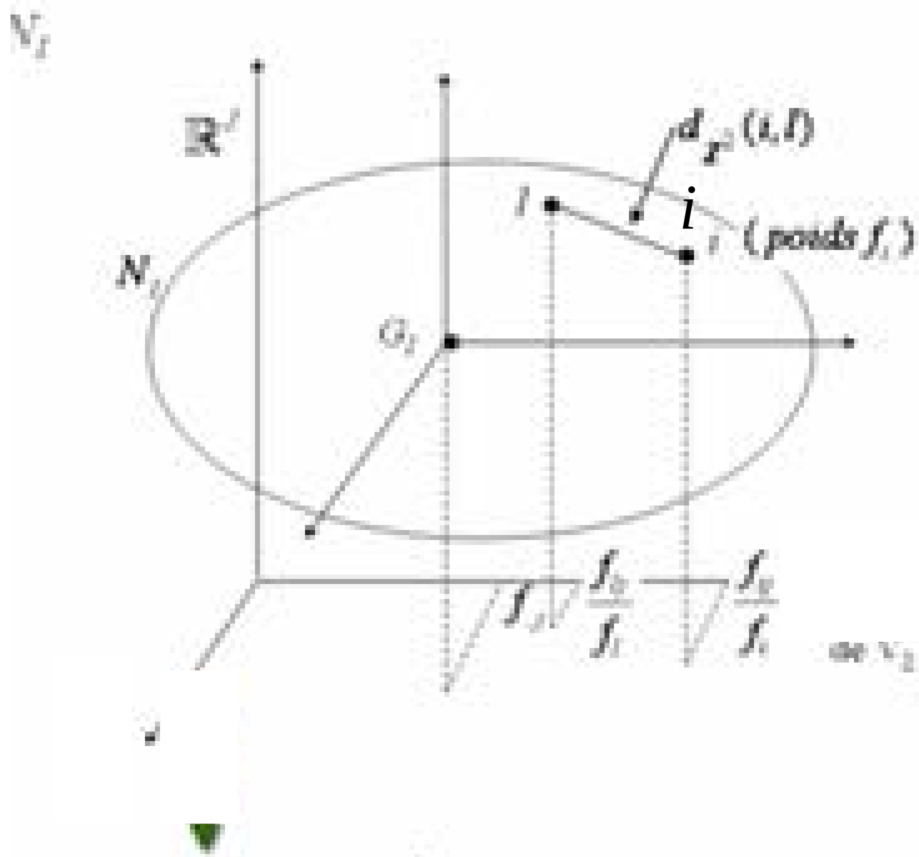
4. Geometrical approach: cloud of rows and cloud of columns



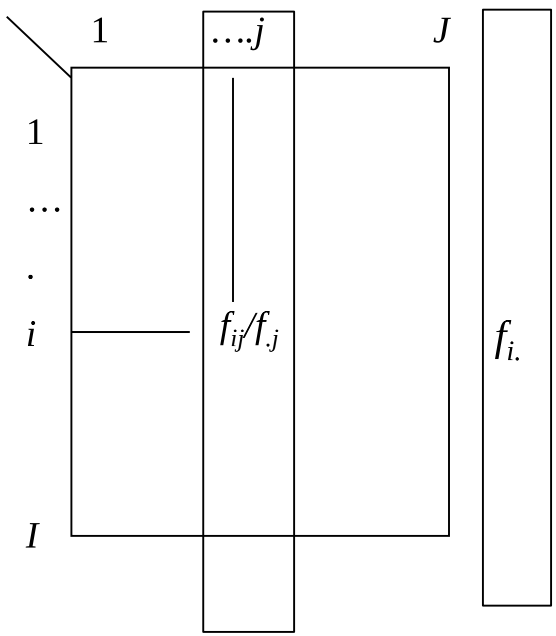
Cloud of rows

$$d^2(i,l) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$$

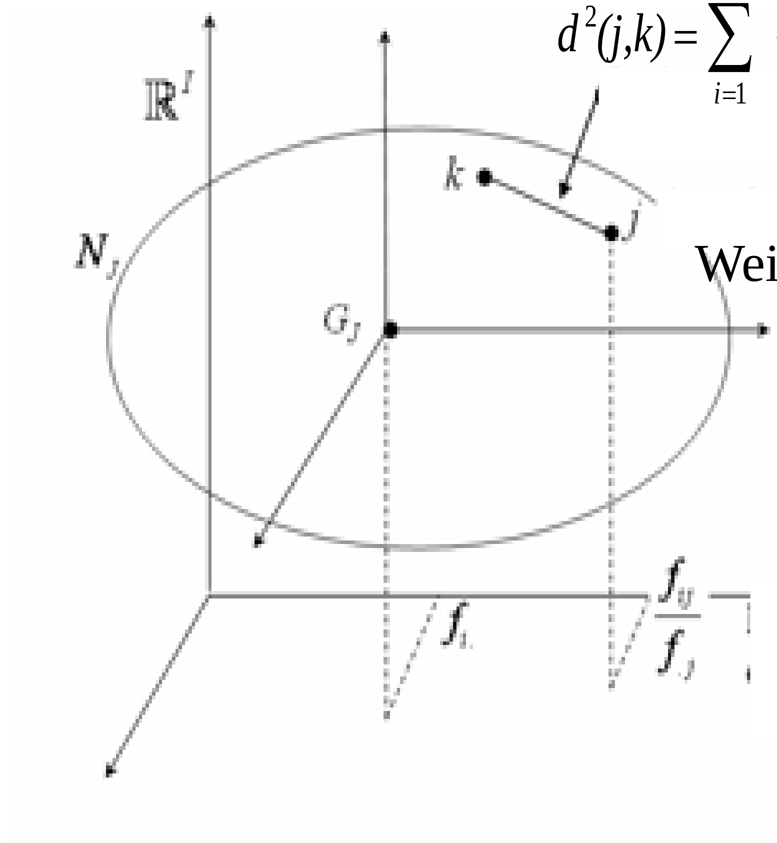
Weight of point i: $f_{i.}$



Cloud of column profiles



$$d^2(j,k) = \sum_{i=1}^{I} \frac{1}{f_{.i}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2$$



Weight of point j: $f_{.j}$

If there were independent

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

Both clouds have a null inertia

$$Inercia\left(N_I \middle| G_I\right) = Inercia\left(N_J \middle| G_J\right)$$

The intensity of the relationship is higher insofar as the inertia is higher

$$Inercia\left(N_I \middle| G_I\right) = \sum_i Inercia\left(i \middle| G_I\right) = \sum_i f_{i.} d^2\left(i, G_I\right) = \sum_j f_{.j} d^2\left(j, G_J\right) =$$

$$= \sum_i \sum_j \frac{1}{f_{i.} f_{.j}} \left(f_{ij} - f_{i.} \cdot f_{.j}\right)^2 =$$

$$= \Phi^2 = \frac{\chi^2}{n} = Inercia\left(N_J \middle| G_J\right)$$

Representation is a low dimension space

Find the subspace which better sums up the data

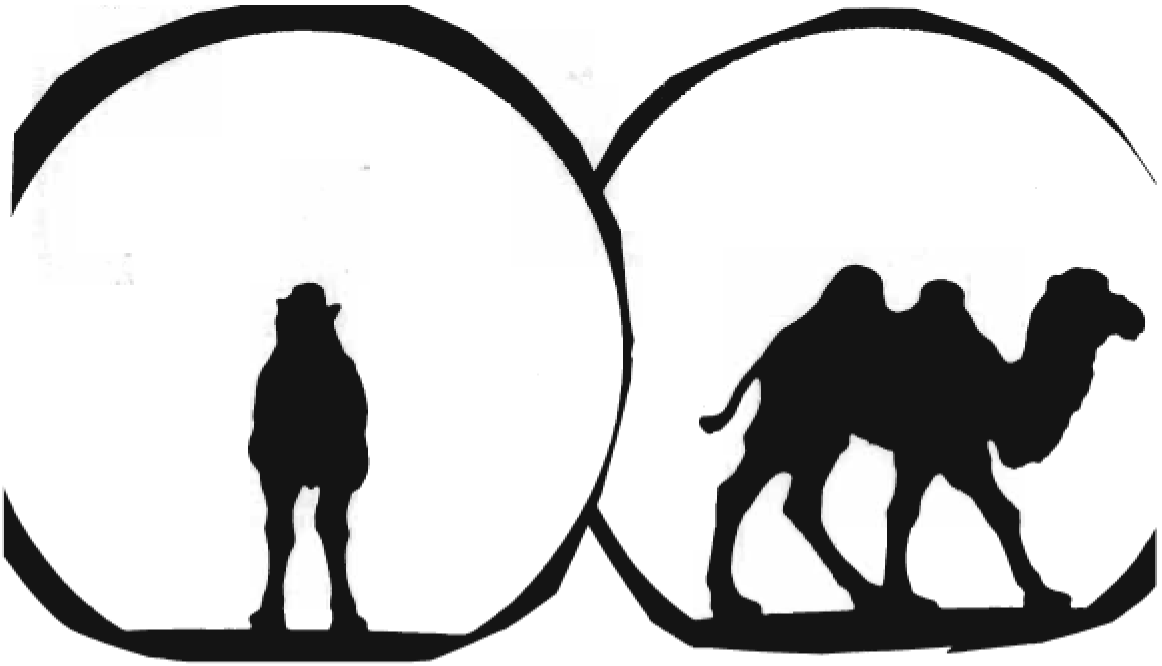
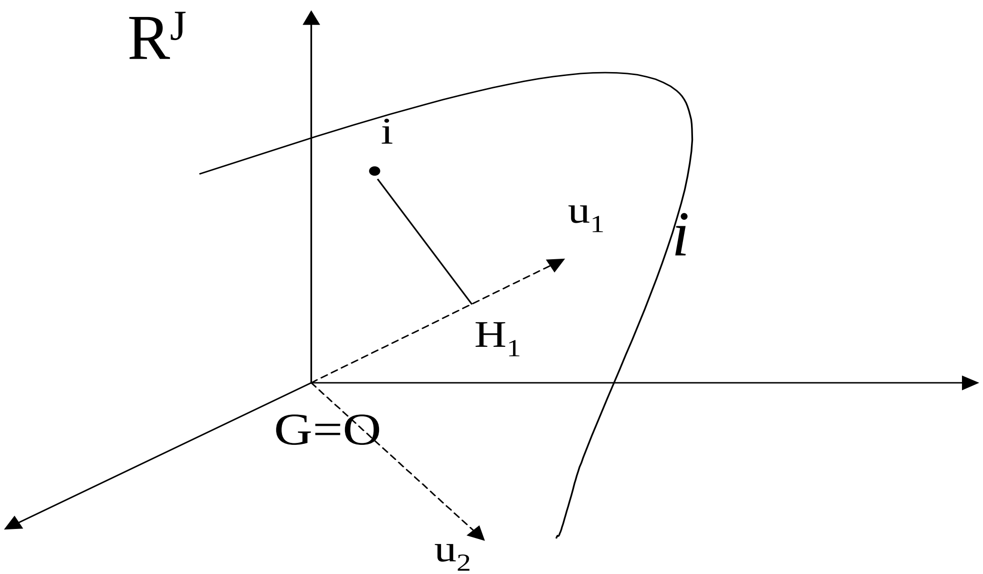


Figure: Camel vs dromedary?

Row cloud



$$Max \sum_{i} f_{i.} OH_i^{\,2}$$

u_1 λ_1

u_2 λ_2

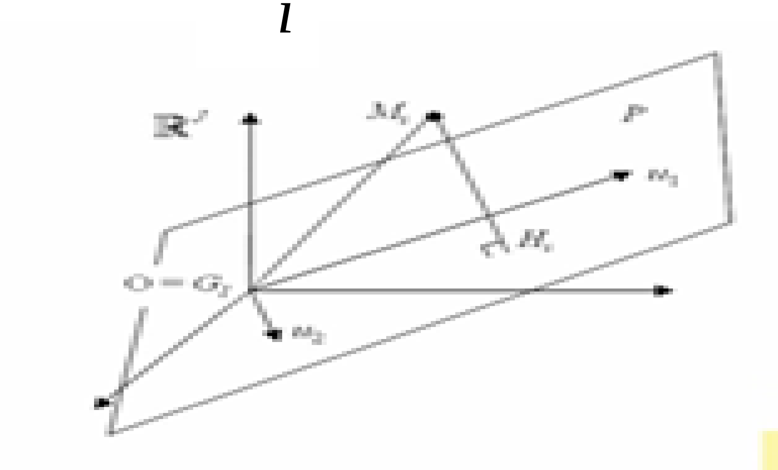
u_3 λ_3

… …

$u_{\min(I\text{-}1,J\text{-}1)}$ $\lambda_{\min(I\text{-}1,J\text{-}1)}$

-

i



In the column cloud…..

$v_1 \qquad \lambda_1$

$v_2 \qquad \lambda_2$

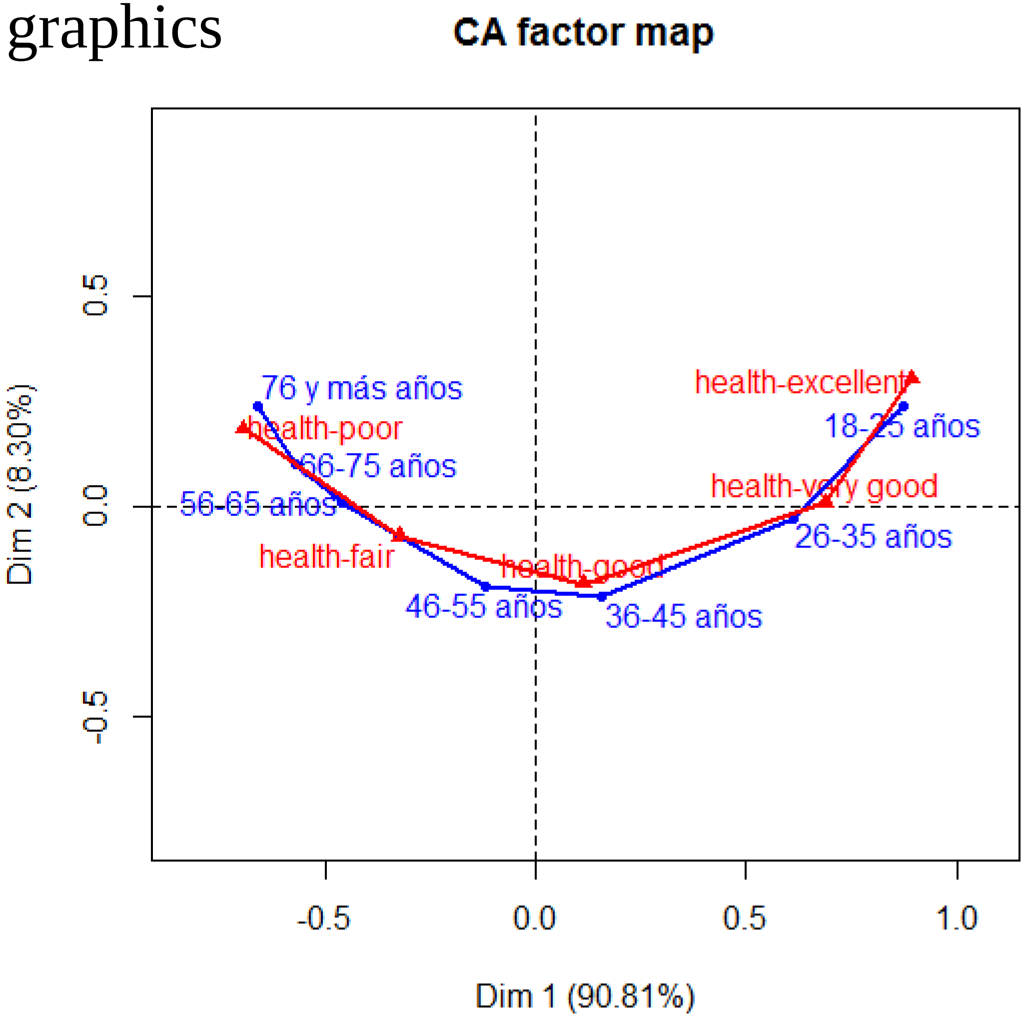
$v_3 \qquad \lambda_3$

$v_{\min(I\text{-}1,J\text{-}1)} \quad \lambda_{\min(I\text{-}1,J\text{-}1)}$

-

$$\Phi^2 = \sum_i \lambda_i = \sum_i f_{i.} d^2(i, G_I) = \sum_j f_{.j} d^2(j, G_J)$$

Graphical results: in this case, it is legitimous to superpose the row and column graphics



Guttman effect

5. Helps to interpretation
Global quality of the representation

$$\frac{\text{Conserved inertia}}{\text{Total inertia}}$$

On the first plane:

$$\frac{\lambda_1 + \lambda_2}{\sum_{s=1}^{S} \lambda_s}$$

```
> round(res.ca$eig,2)
      eigenvalue percentage of variance cumulative percentage of variance
dim 1       0.29                  90.81                             90.81
dim 2       0.03                   8.30                             99.11
dim 3       0.00                   0.83                             99.94
dim 4       0.00                   0.06                            100.00
dim 5       0.00                   0.00                            100.00
```

```
> FI2
[1] 0.3142015=chi2/n
```

V de Cramer

```
> sqrt(sum(res.ca$eig[,1])/4)
[1] 0.2802684
```

Particularities of the eigenvalues in CA $0 \leq \lambda_s \leq 1$

What does it mean to observe an eigenvalue equal to 1?

Maximum number of axes?

How many axes we have to interpret?

Contributions and quality if representation

= what we have seen in PCA
But do not forget that the elements (rows and columns) are endowed with weights

6. Transition relationships also called barycentrical relationships

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} \cdot F_s(i)$$

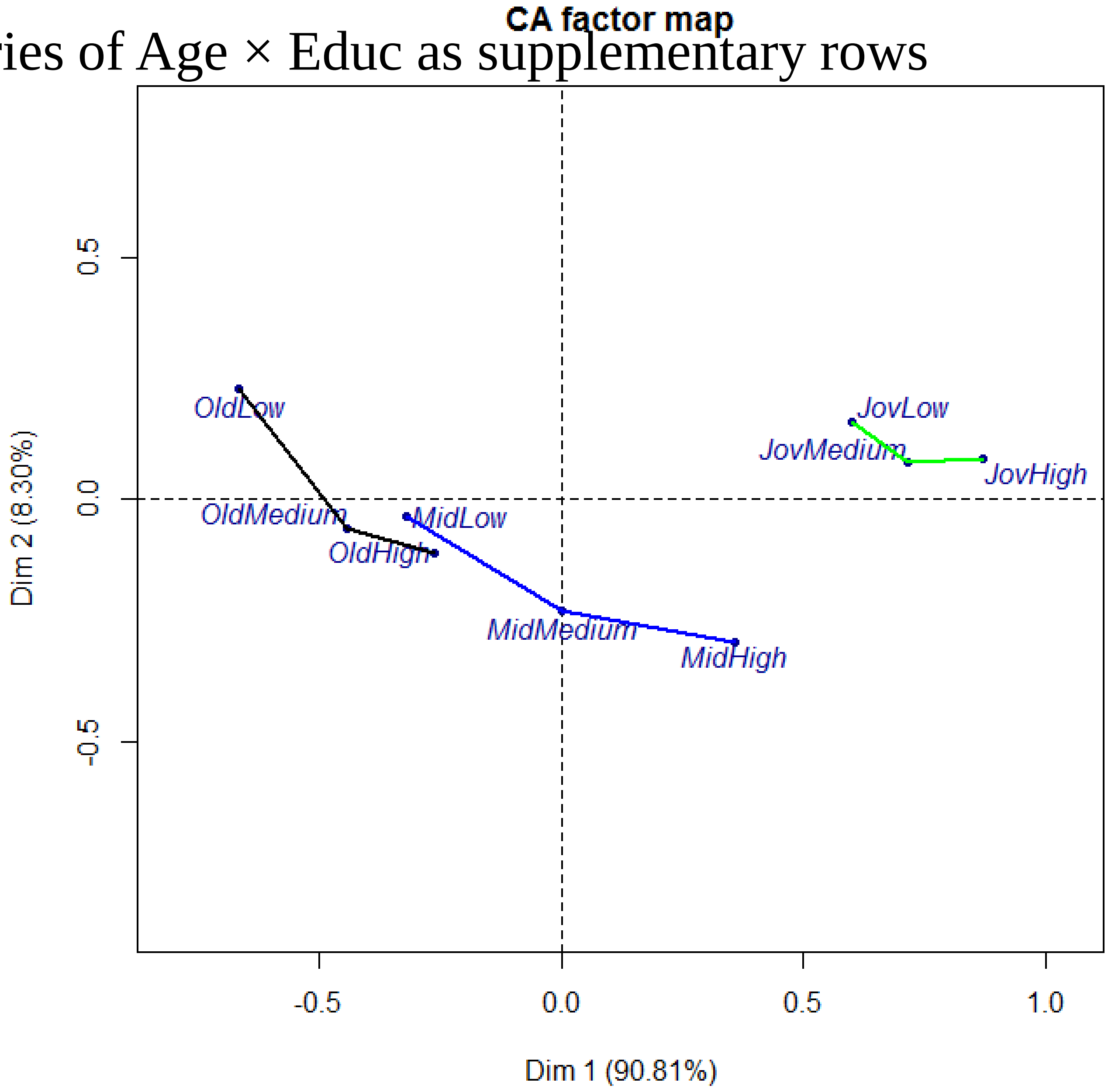
6. Complements

6.1 Supplementary columns and/or rows

$$F_s(i^+) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{i^+ j}}{f_{i^+ .}} G_s(j)$$

$$G_s(j^+) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij^+}}{f_{.j^+}} \cdot F_s(i)$$

Categories of Age × Educ as supplementary rows



6.2 Nature of the relation. Intensity of the relation. Cramer V

The graph informs about the nature of the relationship between the variables through the visualisation of the associations between the categories of one variable and these of the other

The eigenvalues– and their sum- inform about the intensity of the relationship.

The Carmer *V* allows for comparing the intensity of the relationship with its maximum (and therefore, between cross-tables with different dimensions)

$$V=\sqrt{\frac{\phi^2}{Max\left(\phi^2\right)}}=\sqrt{\frac{\phi^2}{Min\left(I-1,J-1\right)}}$$