Session
Correspondence Analysis

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System tracking

Prof. Mónica Bécue Bertaut & Lidia Montero

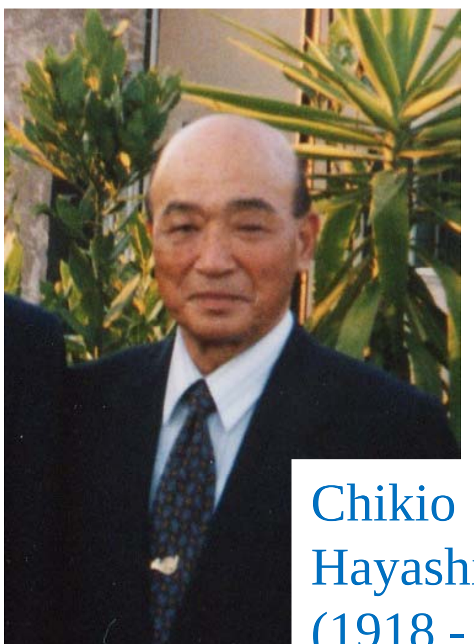
Monica.becue@upc.edu lidia.montero@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Key names in CA

Ronald Aylmer **Fisher,**
1890 –1962

Brigitte Escofier (1941-1994)

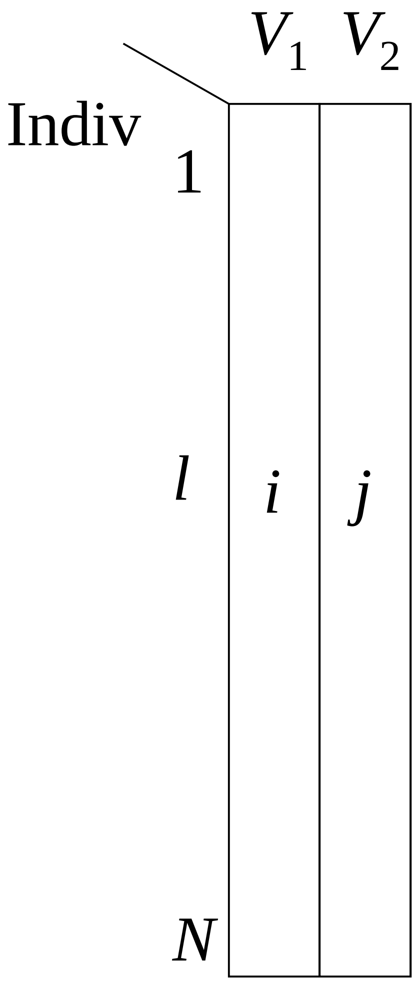


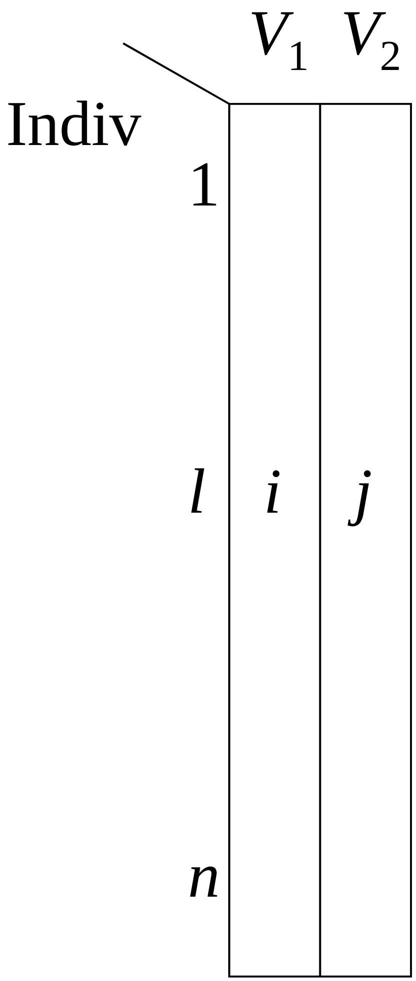
Chikio
Hayashi,
(1918 - 2002)

Jean Paul Benzécri (1932)

1. Data and notation
2. Relationships between categorical variables
3. CA: description of the deviation to independence model
4. Gemetrical view point: row and column clouds
5. Helps to the interpretation
6. Transition relationships
7. Illustrative (supplementary) elements
8. Intensidad of the relationship

1. Data and notation





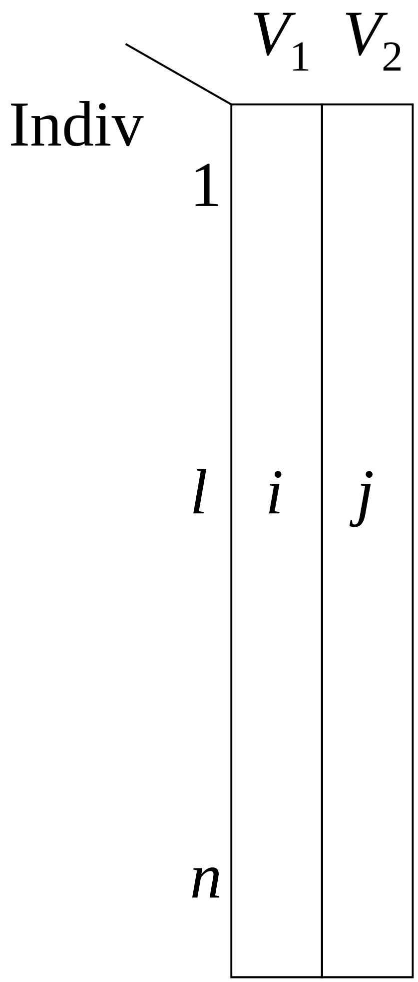
Example: Croatian survey

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

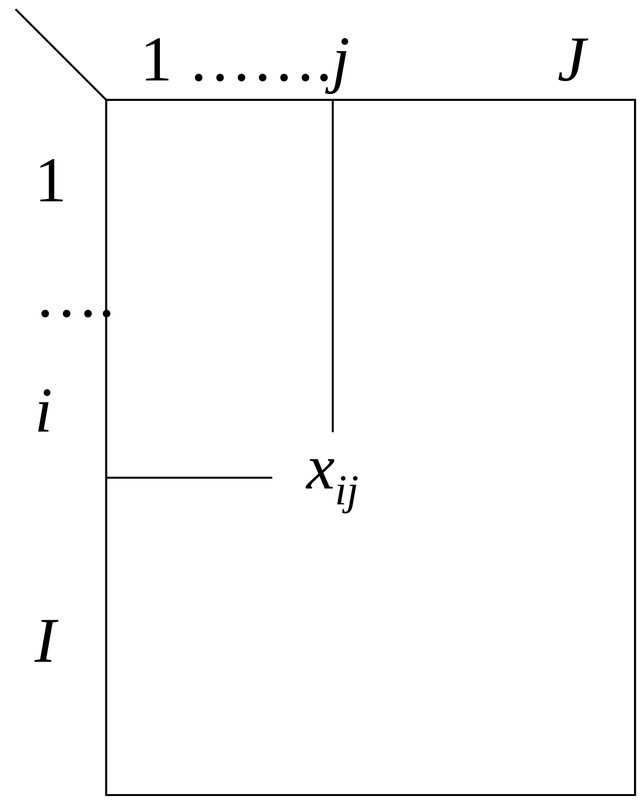
```
> summary(base$Edad_classe)
  18-25 años   26-35 años   36-45 años  46-55 años   56-65 años  66-75 años 76 y más
        639      833           766         794            798       818           389

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

2. Relationship between categorical variables



Contingency table



x_{ij} : respondents who present category i of V_1 and category j of V_2

Crossed table/ Contingency table

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

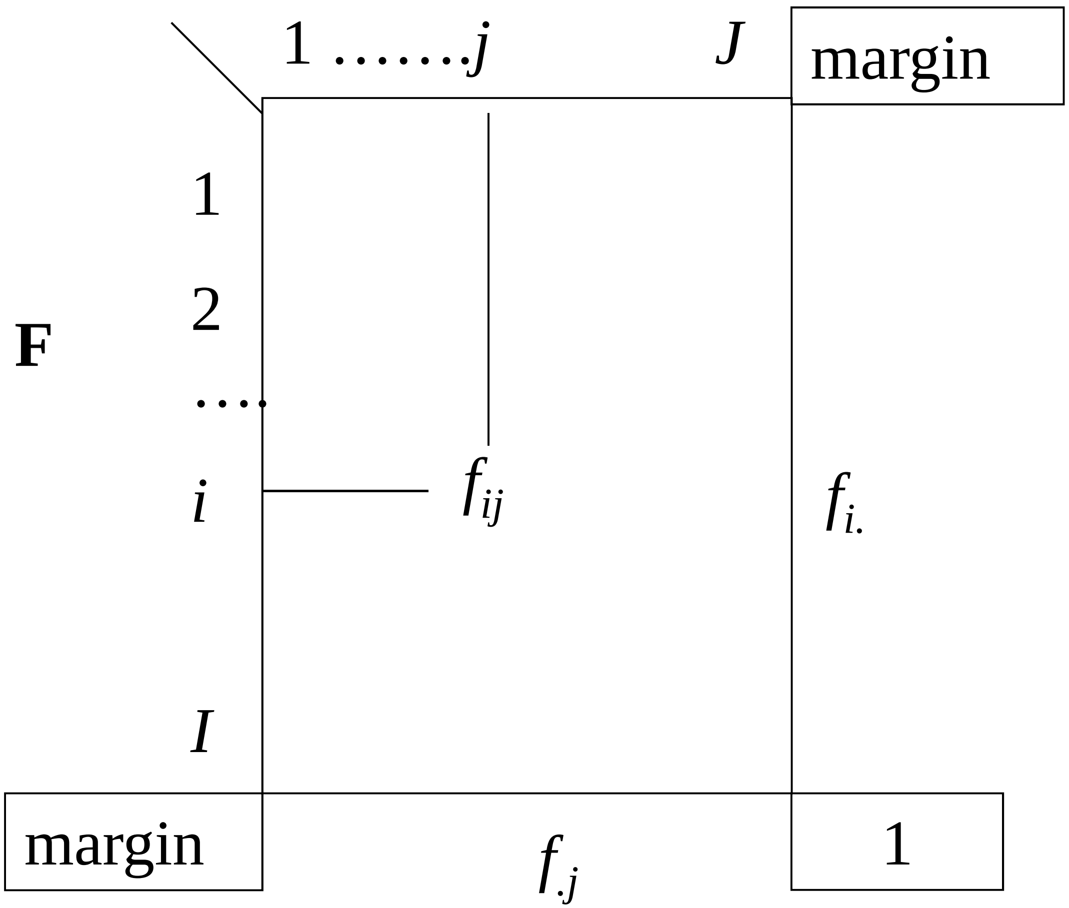
Margins?

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

Proportion table and margins

Tabla **F**



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

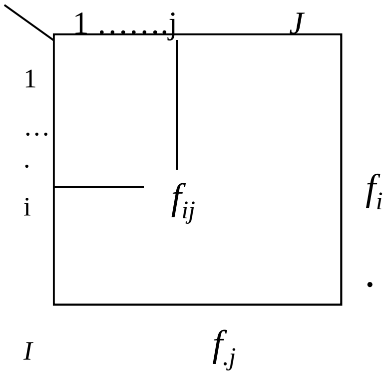
$$f_{i.} = \sum_j f_{ij}$$

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

Relationship between V_1 and V_2: deviation from the independence model

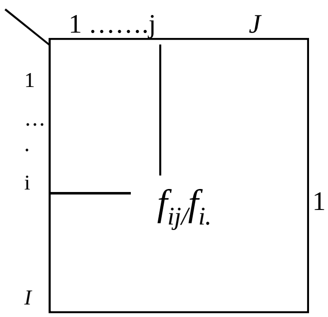
Proportion table and margins

	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**



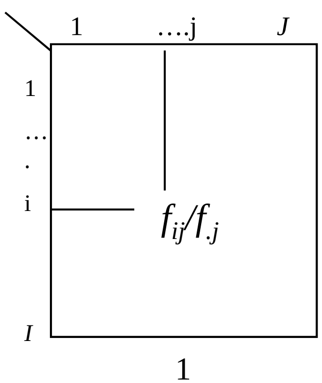
In the case of independence

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



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

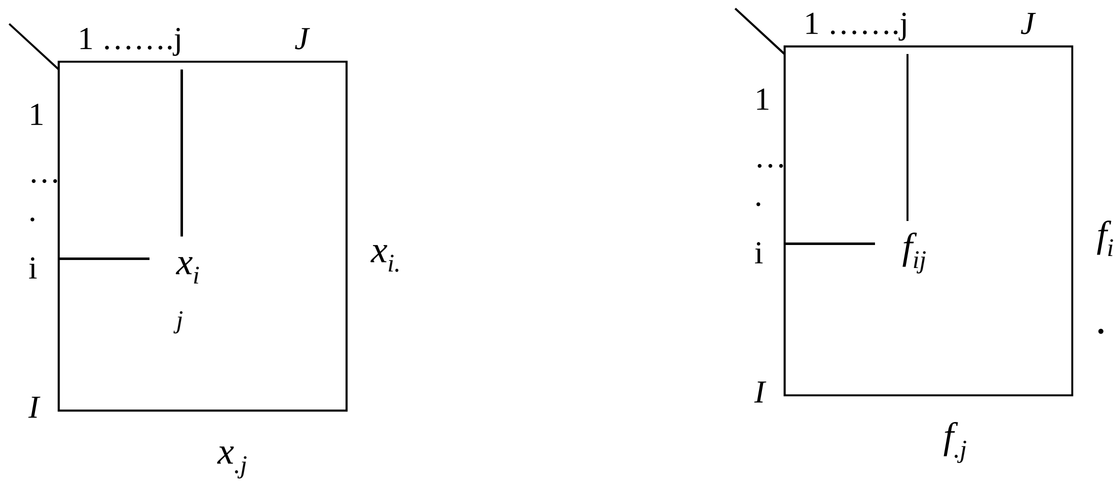
Row-profile table $\mathbf{D}_{\mathbf{I}}^{-1}\mathbf{F}$



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

Columns-profile table $\mathbf{F}\mathbf{D}_{\mathbf{J}}^{-1}$

Observed data



Estimation of the independence model

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

Expected counts, under the hypothesis of independence

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

Significance of the relationship between the variables

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

Intensity of the relationship

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

El AC does not say anyting about the significance of the relationship between the variables, only about the intensity

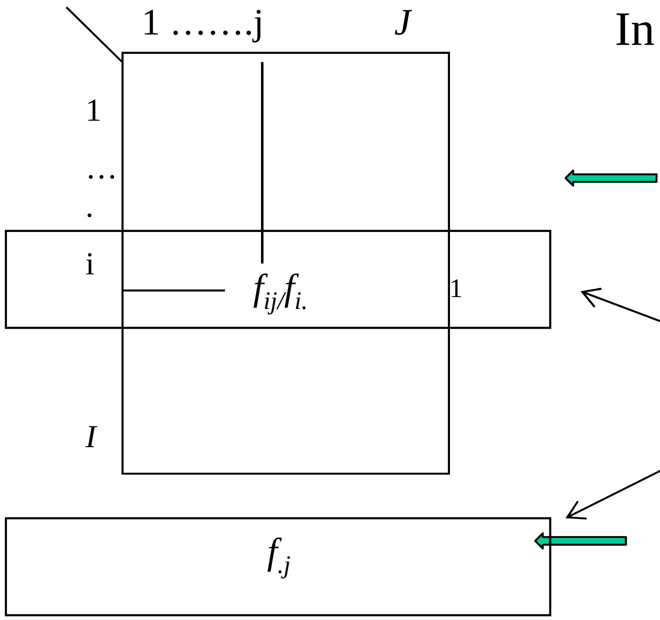
and visualises the structure of the relationship

In the example

Pearson's Chi-squared test

data: tablo
X-squared = 1582.633, df = 24, p-value < 2.2e-16

3. CA: Description of the deviation to independence



In the case of independence $\dfrac{f_{ij}}{f_{i.}} = f_{.j}$

Row-profiles table

We have to compare and describe the differences

Mean- profile (=margin of the proportion table)

	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**

Do the 36-45 have a profile close to the mean profile?

And the youngest class?

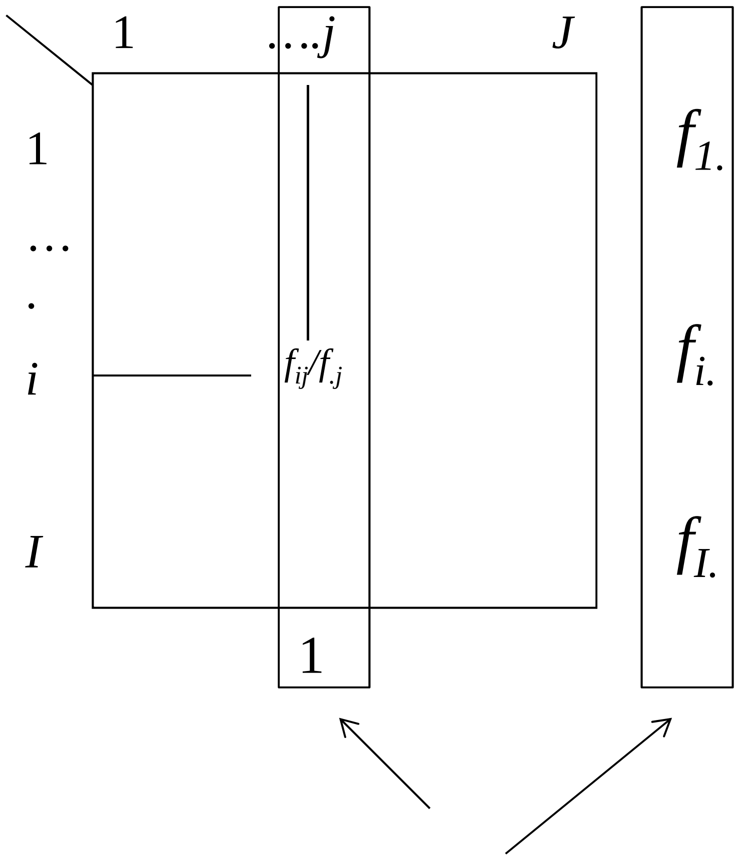
And the oldest class?

….

Column-profile table

In the case of independence $\frac{f_{ij}}{f_{.j}} = f_{i.}$

Mean column-profile



To compare and describe the differences

```
> 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
```

Has “health-poor” a profile which differs from the others? From the mean-profile?

Are the profiles of “very good health” y “excellent health” very different?

4. CA: Geometrical approach

CA= Analysis of the cloud of rows

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

Weights of the rows $f_{i.}$ stored into the diagonal matrix $\mathbf{D}_{\mathbf{I}}$

Metric $\mathbf{D}_{\mathbf{J}}^{-1}$ with generric term $\dfrac{1}{f_{.j}}$

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

$\longrightarrow$ distributional equivalence

CA= Analysis of the cloud of columns

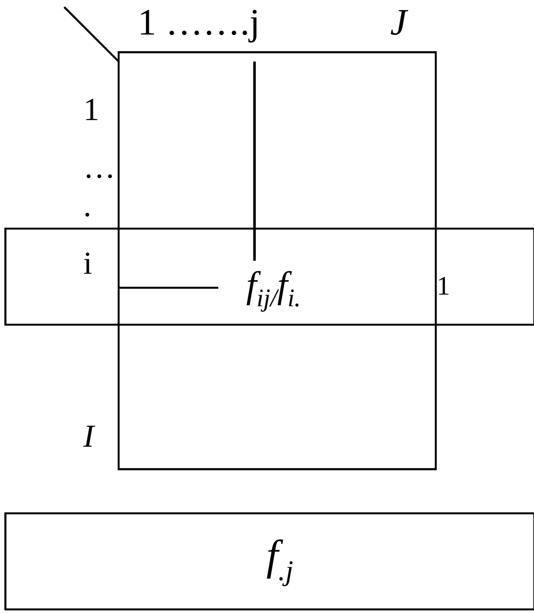
Cloud of rows described by their profile $\dfrac{f_{ij}}{f_{i.}}$ Matrix $\mathbf{D}_{\mathbf{J}}^{-1}\mathbf{F}'$

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

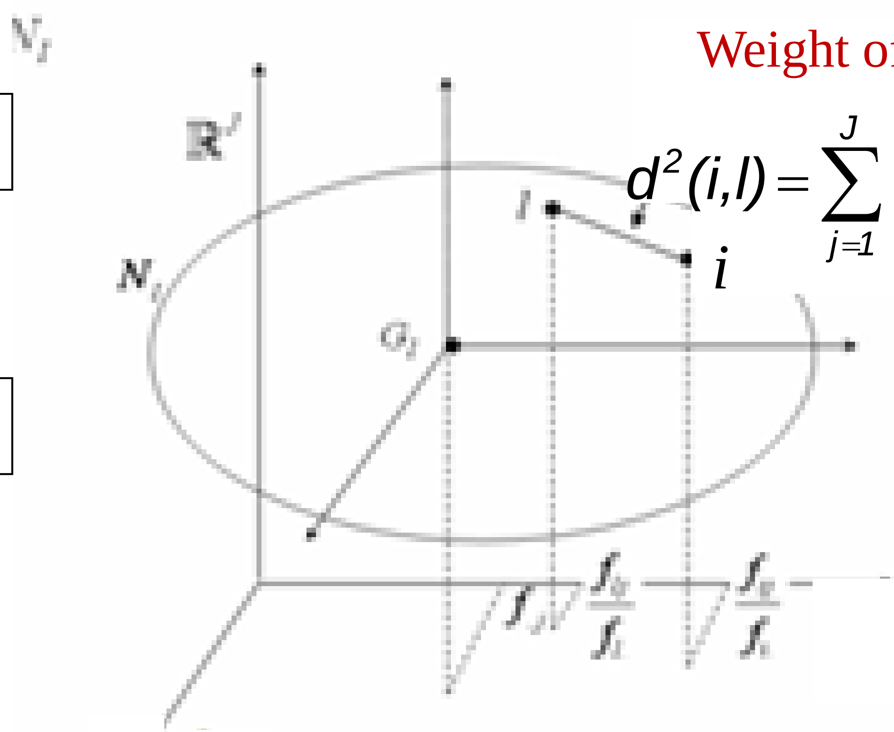
Métrica del chi.2 $\mathbf{D}_{\mathbf{I}}^{-1}$ with generic term $\dfrac{1}{f_{i.}}$

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

$\longrightarrow$ distributional equivalence



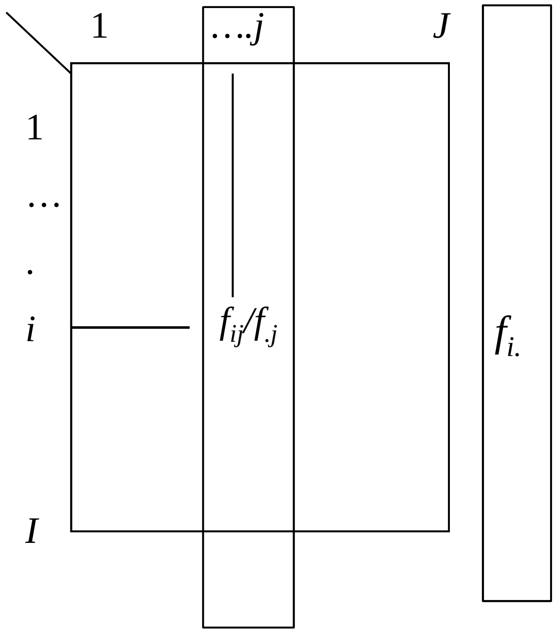
Cloud of rowa



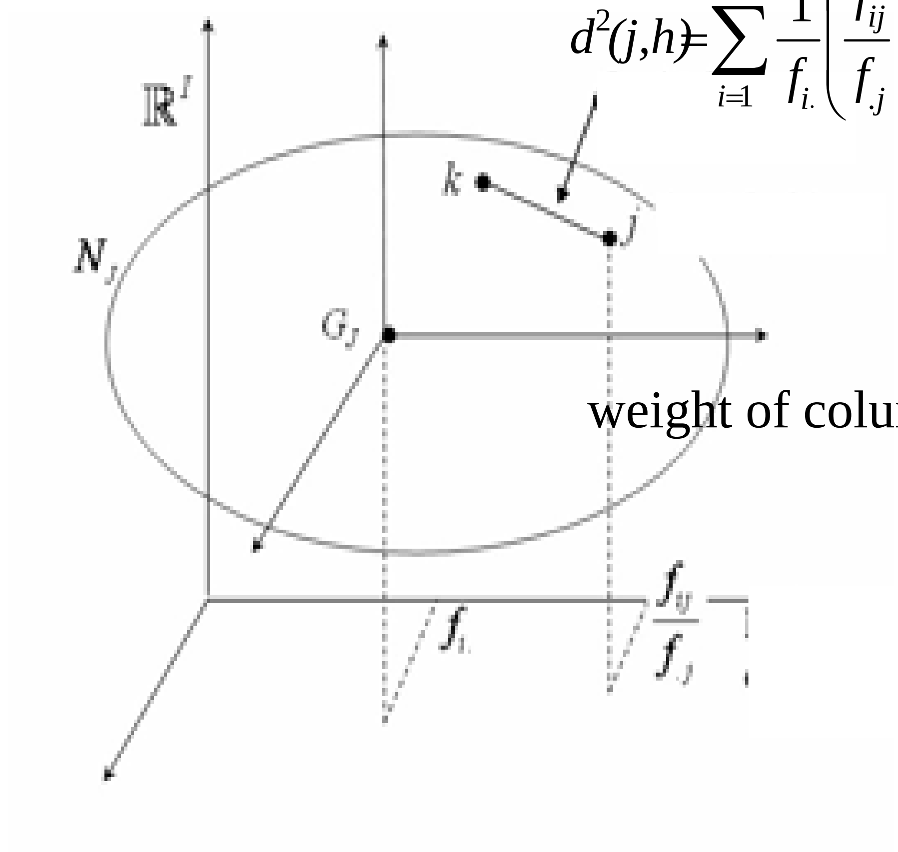
Weight of point i: $f_{i.}$

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

Cloud of columns



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



weight of columns j: $f_{.j}$

If independence exists?

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

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

Both clouds have a null inertia

$$Inercia\left(N_I \middle| G_I\right) = Inercia\left(N_J \middle| G_J\right) = 0$$

If not, the relationship is greater as so far the inertia is greater

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

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

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

Representation in a low-dimension space

Find the subspace which better sums up the data

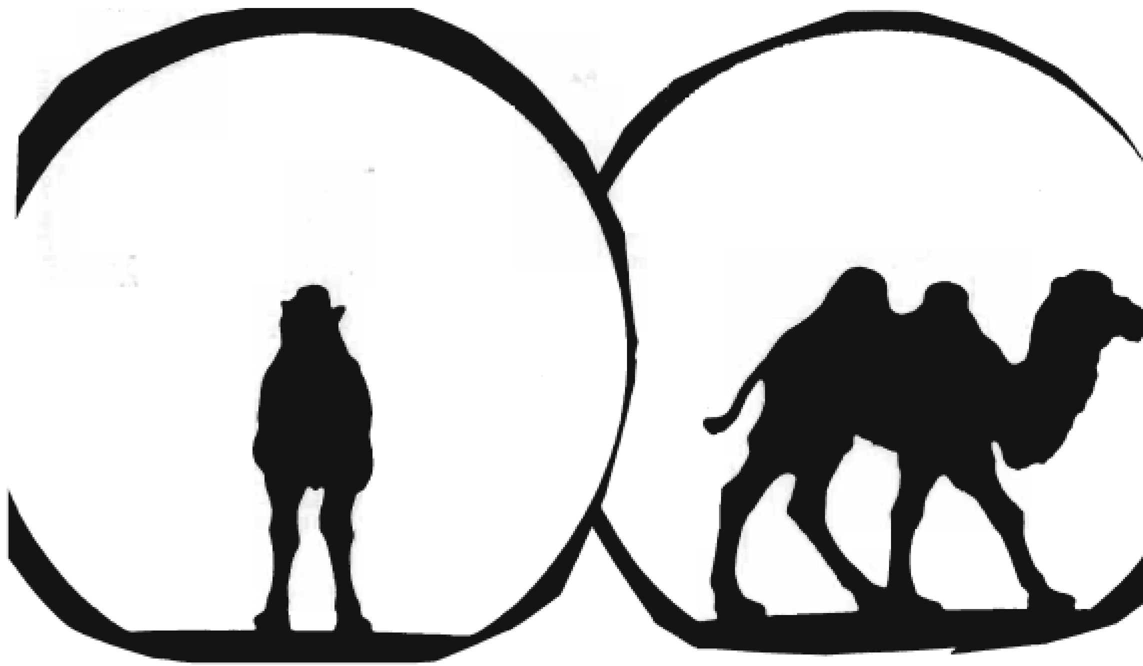
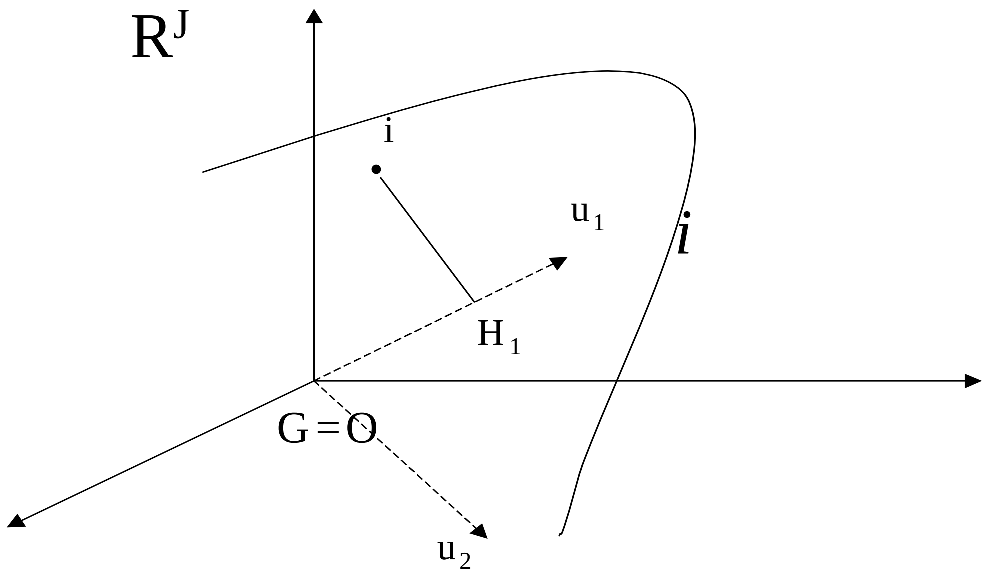


Figure: Camel vs dromedary?

Same rationale as in PCA



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

$u_1 \quad \lambda_1$

$u_2 \quad \lambda_2$

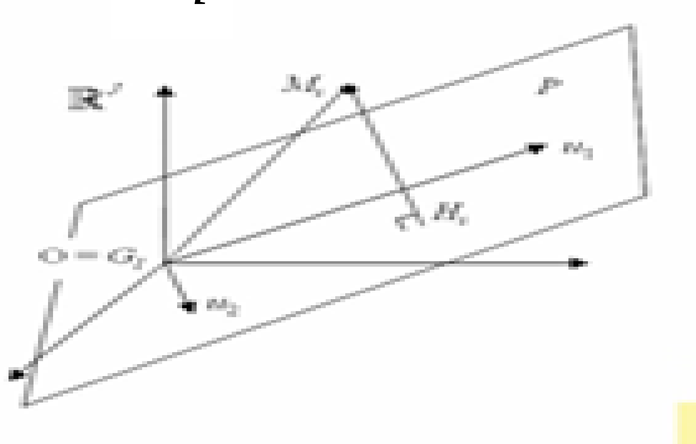
$u_3 \quad \lambda_3$

$\ldots \quad \ldots$

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

-

i



De forma simétrica en el otro espacio…..

v_1 λ_1

v_2 λ_2

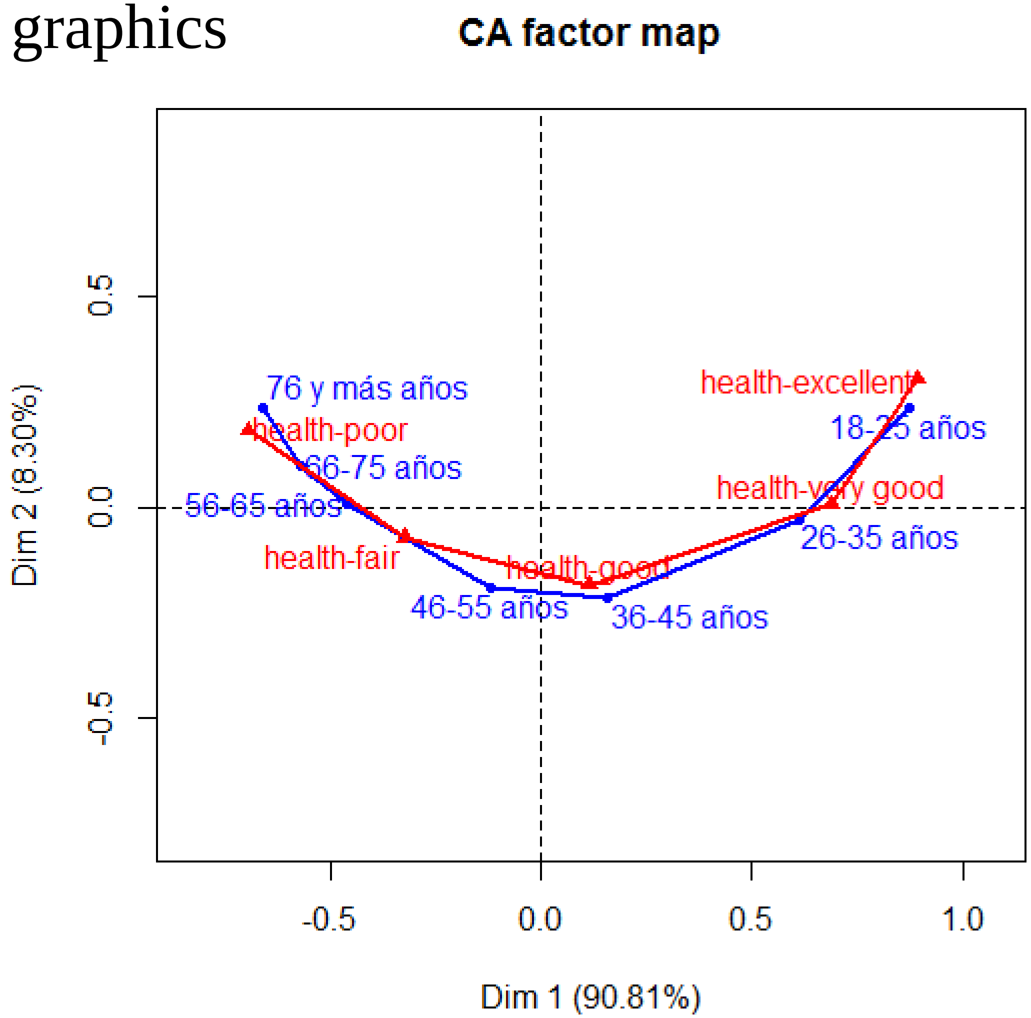
v_3 λ_3

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

-

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

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



Guttman effect

5. Helps to interpretation

$$\text{Kept inertia} \Big/ \text{Total Inertia}$$

In the example, we are interested by the first two axes

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

```
> 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
```

In CA $0 \leq \lambda_s \leq 1$

What doest it mean to have an eigenvalue equal to 1?

Maximum number of axes

How many axes we have to keep and interpret?

Contribution y calidad de representación de los elementos filas o columnas

= Same rules as in PCA

BUT, taking into account the weights are non-uniform

6. Transition relationships
also called baricentric relationships

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

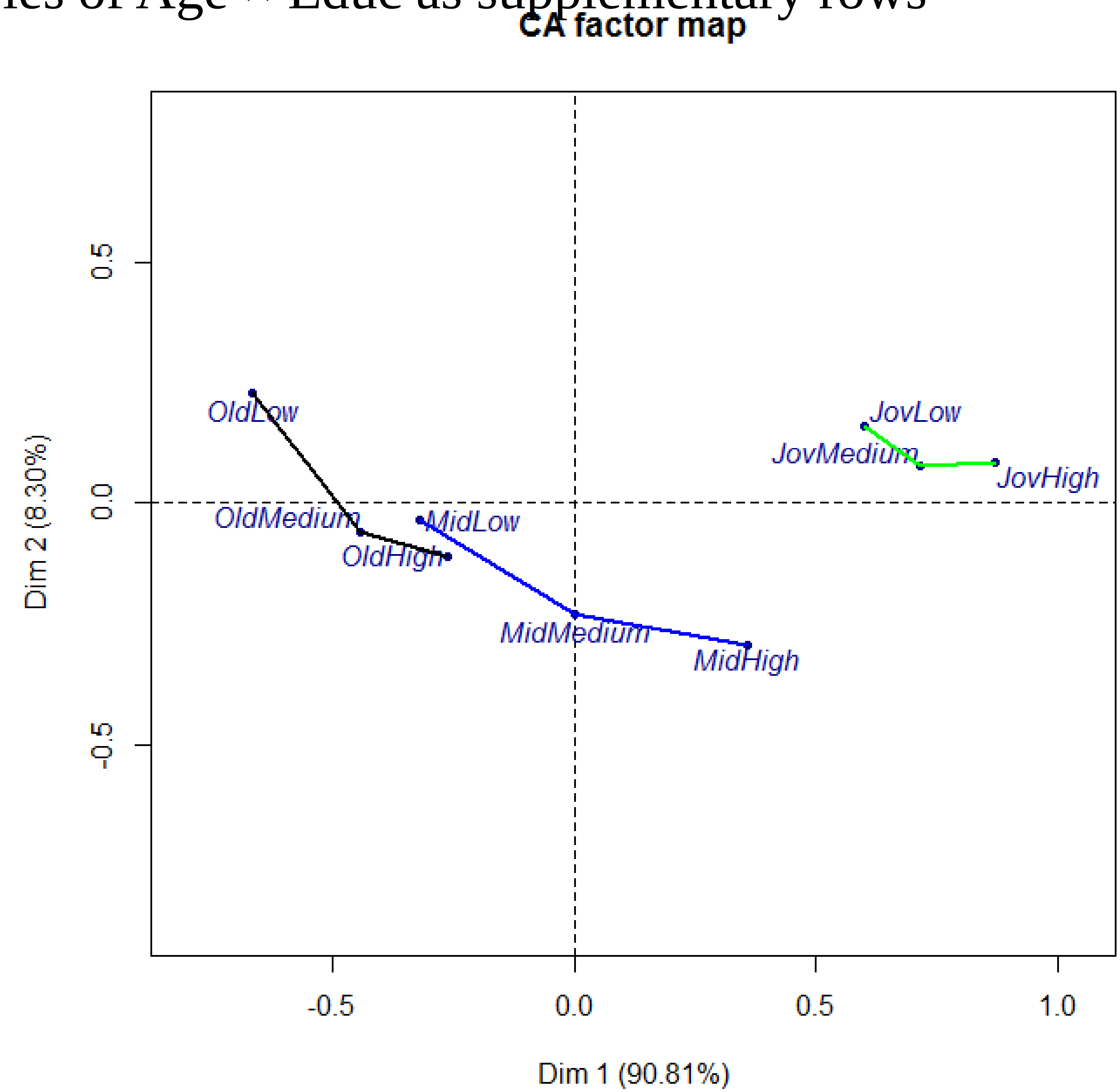
$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} \cdot F_s(i)$$

7. Illustrative elements

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

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

Categories of Age × Educ as supplementary rows



8. Intensity of the relationship- Cramer V

El gráfico nos informa de la naturaleza de la relación entre las variables, mediante la visualización de las asociaciones entre las categorías de una variables y la categorías de la otra variable

Los valores propios – y su suma- nos informan de la intensidad de la relación

El V de Cramer permite comparar la intensidad de la relación con la intensidad máxima posible

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