ANALISI DEI DATI PER IL MARKETING 2014

Marco Riani
mriani@unipr.it
http://www.riani.it

ANALISI DELLE CORRISPONDENZE (cap. VII)

Problema della riduzione delle dimensioni

- L'ANALISI DELLE COMPONENTI PRINCIPALI (cap. VI) per una matrice di dati quantitativi
- L'ANALISI DELLE CORRISPONDENZE (cap. VII) per una tabella di contingenza

RIEPILOGO DELLE CP

Matrice *nxp*, con variabili tutte quantitative

OBIETTIVO: Date *p* variabili (correlate) si vogliono ottenere *k* (*k*<<*p*) indicatori di sintesi (le CP) **ortogonali** (*non correlati*)

VALIDITA' DEL MODELLO: la quota (percentuale) di varianza totale spiegata deve essere elevata (>0.95^p *soglia minima di accettabilità*)

RAPPRESENTAZIONE CONGIUNTA DI VARIABILI E UNITA': **IL BIPLOT**

- Nel caso di 2 CP estratte:
- Rappresentazione nel piano cartesiano dei *p* vettori (frecce) corrispondenti alle variabili
- Rappresentazione nel piano cartesiano degli *n* punti corrispondenti ai punteggi (*scores*) delle unità

SCOPI DELL'ANALISI DELLE CORRISPONDENZE

- Studio congiunto di righe e colonne d'una tabella di contingenza *r* x *c con variabili nominali o ordinali* (con numero di righe e colonne elevato)
- Condizione necessaria: esistenza di associazione significativa tra le 2 variabili
- Relazioni tra **trasformazioni** delle righe e della colonne della tabella:
- *profili riga e profili colonna*

Esempio

- Analisi della tabella di contingenza tra professione e tipo di acquisto prevalente per un campione di clienti di un centro commerciale (file corrisp.xlsx)

MASSE E PROFILI

- **Masse** = frequenze relative marginali:
- $f_{i.} = n_{i.} / n$
- $f_{.j} = n_{.j} / n$
- **Profilo della riga** *i*-esima: vettore di
- $f_{ij} / f_{i.}$ ➔ *matrice dei profili riga*
- **Profilo della colonna** *j*-esima: vettore di
- $f_{ij} / f_{.j}$ ➔ *matrice dei profili colonna*

V. File di Excel

- Nella matrice dei profili riga le masse di colonna ($f_{.j}$) sono interpretabili come «il profilo medio di riga»

$$\sum_{i=1}^{r} \frac{f_{ij}}{f_{i.}} f_{i.} = f_{.j} \quad j = 1,\ldots,c.$$

V. File di Excel

- Nella matrice dei profili colonna le masse di riga ($f_{i.}$) sono interpretabili come «il profilo medio di colonna»

$$f_{i.} = \sum_{j=1}^{c} \frac{f_{ij}}{f_{.j}} f_{.j} \qquad j = 1, \ldots, c.$$

PROFILI COLONNA

La massa dei profili colonna indica che la maggior parte degli acquisti effettuati nel centro commerciale dal campione in esame, viene effettuata quotidianamente, con una percentuale del 41.5%, mentre il tipo di acquisto che "pesa meno" è riferito agli acquisti in occasione di ricorrenze, che mostra una massa pari all'8.5%

Profili riga e colonna

- L'analisi dei profili riga può essere considerata come lo studio di r punti in uno spazio a c-1 dimensioni
- L'analisi dei profili colonna può essere visto come lo studio di c punti in uno spazio a r-1 dimensioni

Matrici Dr e Dc

Nel prosieguo si indicherà con $\mathbf{D}_r$ una matrice diagonale di dimensioni $r \times r$ contenente le masse di riga $r_{i.}$ lungo la diagonale principale e, analogamente, si indicherà con $\mathbf{D}_c$ una matrice diagonale di dimensione $c \times c$ contenente le masse di colonna $c_{.j}$ lungo la diagonale principale. Immaginiamo di

Output ACP

- Analogamente a quanto prodotto dall'analisi in componenti principali, l'analisi delle corrispondenze fornisce:
- autovalori, che sono il quadrato di valori singolari (la cui somma è chiamata "inerzia");
- percentuale di varianza spiegata (detta "percentuale di inerzia spiegata");
- *factor loadings* (correlazione tra i profili di riga e colonna con gli assi principali);
- *scores* dei profili di riga e colonna sugli assi principali;
- comunalità (percentuale di inerzia spiegata dai profili riga e profili colonna).

PRINCIPI FONDAMENTALI

- L'operazione di centratura non fa altro che rimuovere i centroidi delle righe e delle colonne e deve essere interpretata alla stregua dell'operazione di calcolo degli scostamenti dalla media aritmetica per un insieme di variabili quantitative

$$f_{ij} - f_{i.} \times f_{.j}$$

- Le frequenze relative marginali di riga (colonna) possono essere interpretate come i centroidi (profili medi) di colonna (riga).

PRINCIPI FONDAMENTALI

- Centratura e normalizzazione della tabella di contingenza di dimensioni *r* x *c* (p.274):

$$S = [s_{ij}] = \frac{f_{ij} - f_{i.} \times f_{.j}}{\sqrt{f_{i.} \times f_{.j}}}$$

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2}.$$

INERZIA

- Inerzia totale = varianza totale della tabella di contingenza =
- = somma dei quadrati di s_{ij}
- = $\chi^2 / n = \Phi^2$
- Max inerzia = min (*r*-1, *c*-1)
- Richiamo: **V di Cramer (p. 114)**
- V = √ inerzia / √ min (*r*-1, *c*-1)

$$V = \frac{\phi}{\sqrt{\min[(r-1),(c-1)]}} \quad \begin{cases} 0 \\ 1 \end{cases}$$

- Il massimo valore che può assumere l'inerzia è dato da min(r-1,c-1)
- Nella nostra tabella 13 x 4 che stiamo analizzando, il valore massimo possibile dell'inerzia è uguale a 3

DISTANZA TRA PROFILI

- Per studiare le relazioni congiunte tra profili riga e colonna è necessario introdurre una distanza tra le due nuvole di punti
- La distanza euclidea tra i profili riga (colonna) non è soddisfacente poiché attribuisce un peso uguale a tutte le colonne (righe)

$$d_E(i,i^*) = \sqrt{\sum_{j=1}^{c}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i^*j}}{n_{i^*.}}\right)^2}$$

Metrica (distanza) del chi-quadrato

- Si vuole invece attribuire un peso rilevante alle colonne in cui vi è una grande distanza tra due punti riga, anche se il totale di colonna è piccolo

- Tra due profili riga: formula (7.6)

$$d_{\chi^2}(i,i^*) = \sqrt{\sum_{j=1}^{c} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i^*j}}{n_{i^*.}} \right)^2} .$$

Metrica (distanza) del chi-quadrato

- Tra due profili colonna: formula (7.7).

$$d_{\chi^2}(j,j^*) = \sqrt{\sum_{i=1}^{r} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij^*}}{n_{.j^*}} \right)^2} ,$$

Indicando con j e j^* due generiche colonne,

Distanza dal profilo medio (di riga)

- Distanza (con la metrica del chi-quadrato) di ogni profilo di riga dal profilo **medio** (ponderato) di riga

$$d_{\chi^2}(i,\bar{i}) = \sqrt{\sum_{j=1}^{c} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{\bar{i}j}}{n_{\bar{i}.}} \right)^2} = \sqrt{\sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{\bar{i}j}}{f_{\bar{i}.}} \right)^2} = \sqrt{\sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2}$$

- misura di quanto i profili individuali si scostano dal profilo medio

- Nel nostro esempio i profili riga che si discostano di più dalla media sono
- Imprenditore commerciante e studente

Distanza dal profilo medio (di riga)

Una media ponderata di scostamenti al quadrato dalla media, rappresenta fondamentalmente la varianza totale della tabella di contingenza

$$inerzia = \sum_{i=1}^{r} (\text{massa di riga}_i) \times \left[d_{\chi^2}(i,\bar{i}) \right]^2 = \sum_{i=1}^{r} f_{i.} \times \left[d_{\chi^2}(i,\bar{i}) \right]^2 = \frac{\chi^2}{n}.$$

- **Inerzia totale** = misura del grado di dispersione attorno al profilo medio

Distanza dal profilo medio (di colonna)

- **Tutto quello che abbiamo visto prima per le righe vale anche per le colonne**
- **V. Implementazione file Excel**
- **Inerzia totale** = misura del grado di dispersione attorno al profilo medio

Obiettivo dell'analisi

$$S = [s_{ij}] = \frac{f_{ij} - f_{i.} \times f_{.j}}{\sqrt{f_{i.} \times f_{.j}}}$$

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2}.$$

- I nostro obiettivo è cercare la matrice **X** che rende minima la somma dei quadrati della matrice **S**-**X**, in simboli.

$$\min_{\mathbf{X}} \| \mathbf{S} - \mathbf{X} \|^2$$

Obiettivo dell'analisi

- Dato che la somma dei quadrati della matrice **S** (inerzia totale), può essere scritta in funzione della matrice di corrispondenza **P** come media ponderata nelle metriche definite da funzioni delle matrici $\mathbf{D}_r$ e $\mathbf{D}_c$ occorre generalizzare la scomposizione in valori singolari al caso in cui le righe e le colonne della matrice presentino un diverso peso

Passi della procedura

- Svd di $\mathbf{D}_{\mathbf{r}}^{-1/2}(\mathbf{P}-\mathbf{rc}')\mathbf{D}_{\mathbf{c}}^{-1/2}=\mathbf{S}$

$$\mathbf{S}=\sum_{h=1}^{k}\gamma_h \mathbf{u}_h \mathbf{v}_h',$$

- Coordinate di riga

$$\mathbf{y}_h^{(r)}=\mathbf{D}_r^{-1/2}\gamma_h \mathbf{u}_h$$

- Coordinate di colonna

$$\mathbf{y}_h^{(c)}=\mathbf{D}_c^{-1/2}\gamma_h \mathbf{v}_h$$

BONTA' DELL'ANALISI (p. 282)

- Contributo della *i*-esima componente all'inerzia totale: (λ_i / inerzia) . 100
- Le prime due componenti devono spiegare una percentuale elevata dell'inerzia

INTERPRETAZIONE DEL RIASSUNTO DEI PROFILI RIGA

- Masse (di riga)
- Punteggio (*scores*) dei profili (saranno rappresentati nel biplot)
- Contributo di ogni riga all'inerzia totale
- Contributo del punto all'inerzia della dimensione (servono per interpretare gli assi) $\Rightarrow$ PUNTI DOMINANTI
- Contributo della dimensione all'inerzia del punto = quota di spiegazione del punto passando dallo spazio originario a R^2.

Varianza ponderata della dimensione h-esima

- coincide con l'autovalore h-esimo = inerzia della dimensione h-esima

$$\lambda_h = \sum_{i=1}^{r} \left(y_{ih}^{(r)}\right)^2 f_{i.} = \sum_{j=1}^{c} \left(y_{jh}^{(c)}\right)^2 f_{.j}$$

- Il contributo di ogni punto (riga) all'inerzia della dimensione h-esima è dato da

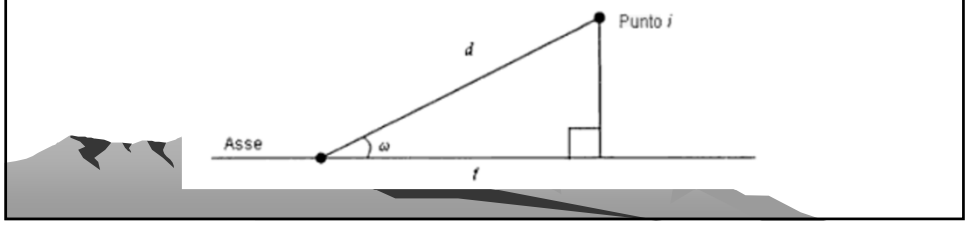
$$\frac{\left(y_{ih}^{(r)}\right)^2 f_{i.}}{\lambda_h}$$

I contributi più alti vengono chiamati punti dominanti

Contributo dei punti all'inerzia delle dimensioni analizzate

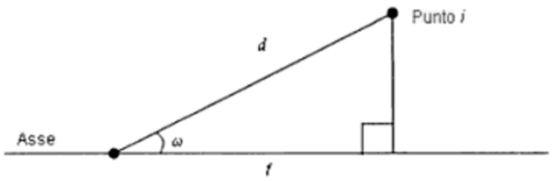
- Totale dei contributi della dimensioni all'inerzia del punto ➔ comunalità

$$\frac{\left(y_{ih}^{(r)}\right)^2}{d_i^2} = \cos^2 \omega$$



Contributo dei punti all'inerzia delle dimensioni analizzate

$$\frac{\left(y_{ih}^{(r)}\right)^2}{d_i^2} = \cos^2 \omega$$



- Contributo per una dimensione alto ➔ l'angolo tra il vettore del punto e l'asse è piccolo ➔ il punto è situato nella direzione dell'asse ➔ alta correlazione con la dimensione

PROCEDURA DI SPSS

- ANALIZZA – RIDUZIONE DEI DATI – ANALISI DELLE CORRISPONDENZE
- Le variabili devono essere codificate come numeriche (CODICI NUMERICI)
- Occorre definire l'intervallo per i CODICI delle categorie (min e max)
- Inserire le etichette per i codici

ESEMPI DI ANALISI DELLE CORRISPONDENZE

Studenti in biblioteca

- TABELLA 3X4

Tavola di contingenza provenienza * motivo

			motivo				
			studio	internet	consultazione e prestito	incontrare amici	Totale
provenienza	in sede	Conteggio	50	13	3	6	72
		Conteggio atteso	47.5	14.4	6.5	3.6	72.0
	fuorisede	Conteggio	57	14	6	3	80
		Conteggio atteso	52.8	16.0	7.2	4.0	80.0
	pendolare	Conteggio	25	13	9	1	48
		Conteggio atteso	31.7	9.6	4.3	2.4	48.0
Totale		Conteggio	132	40	18	10	200
		Conteggio atteso	132.0	40.0	18.0	10.0	200.0

SIGNIFICATIVITA' DELL'ASSOCIAZIONE

- CHI QUADRATO significativo al 5%

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)
Chi-quadrato di Pearson	13.268[a]	6	.039
N. di casi validi	200		

a. 4 celle (33.3%) hanno un conteggio atteso inferiore a 5 Il conteggio atteso minimo è 2.40.

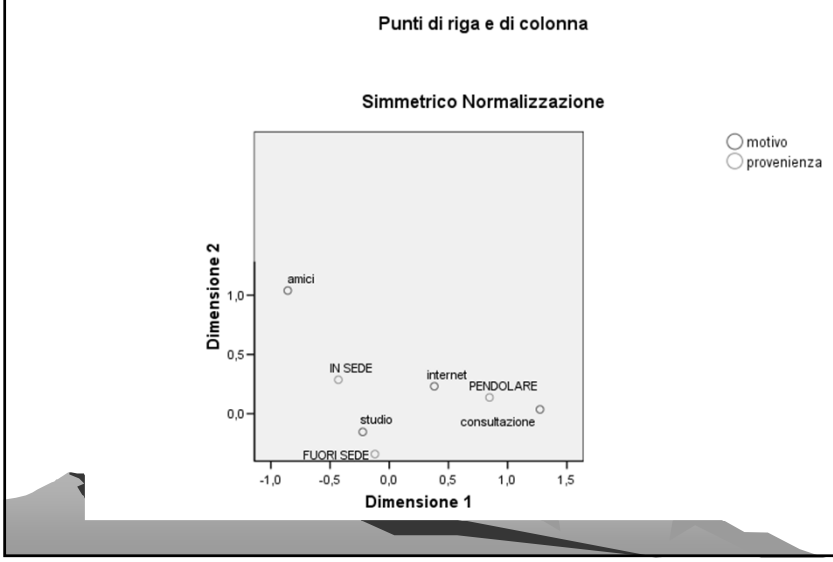
INDICI DI ASSOCIAZIONE

- V DI CRAMER

Misure simmetriche

		Valore	Sig. appross.
Nominale per nominale	Phi	.258	.039
	V di Cramer	.182	.039
	Coefficiente di contingenza	.249	.039
N. di casi validi		200	

GRAFICO DELLE 2 DIMENSIONI (BIPLOT)



INTERPRETAZIONE DEL BIPLOT

- L'origine (0, 0) è il baricentro = profili medi di riga e colonna
- Il punto 1=in sede ha un profilo dei motivi molto simile a quello medio
- Il punto 1= studio ha un profilo delle provenienze molto simile a quello medio
- Punti dominanti del primo asse sono 3=pendolare e 3=consultazione e del secondo asse 4=amici.
- I pendolari hanno una frequenza di consultazione molto maggiore di quella attesa nell'ipotesi di indipendenza tra provenienza e motivi

GIUDIZIO E VOTO SULLA LAUREA IN TRADE

Tavola di contingenza GIUDIZIOlaureaTRADE * VOTOlaureaTRADE

Conteggio

		VOTOlaureaTRADE					
		6,00	7,00	8,00	9,00	10,00	Totale
GIUDIZIOlaureaTRADE	2	2	0	0	1	1	4
	3	4	3	1	1	0	9
	4	0	20	88	9	2	119
	5	0	0	14	11	10	35
Totale		6	23	103	22	13	167

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)
Chi-quadrato di Pearson	131,250[a]	12	,000
N. di casi validi	167		

a. 14 celle (70,0%) hanno un conteggio atteso inferiore a 5. Il conteggio atteso minimo è ,14.

ANALISI CORRISPONDENZE

Riepilogo

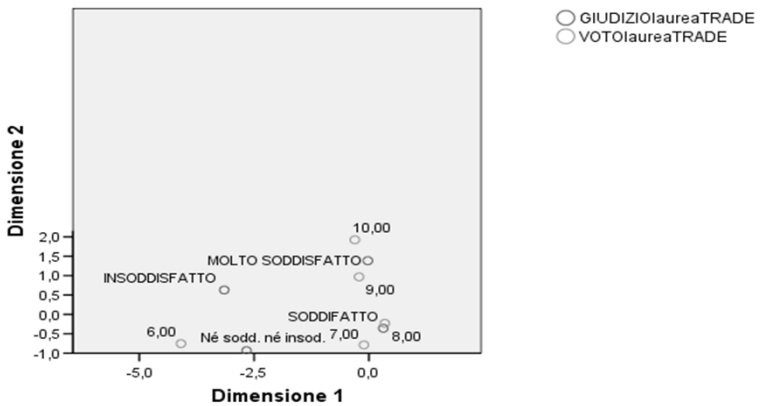
Dimensione	Valore singolare	Inerzia	Chi-Quadrato	Sig.	Proporzione di inerzia		Confidenza del valore singolare	
					Spiegata	Cumulata	Deviazione standard	Correlazione 2
1	,691	,477			,607	,607	,086	-,014
2	,550	,303			,385	,992	,069	
3	,079	,006			,008	1,000		
Totale		,786	131,250	,000[a]	1,000	1,000		

a. 12 gradi di libertà

BIPLOT

Punti di riga e di colonna

Simmetrico Normalizzazione



Esempio da CORBELLINI A., RIANI M., DONATINI A. (2008). Multivariate Data Analysis Techniques to Detect Early Warnings of Elderly Frailty. STATISTICA APPLICATA. vol. 20, pp. 159-178
PDF http://www.riani.it/pub/CRD10.pdf

- Campione di 2374 anziani della provincia di Parma
- Obiettivo: analizzare le relazione tra la variazione dello stato di salute e il grado di «isolamento» dell'anziano

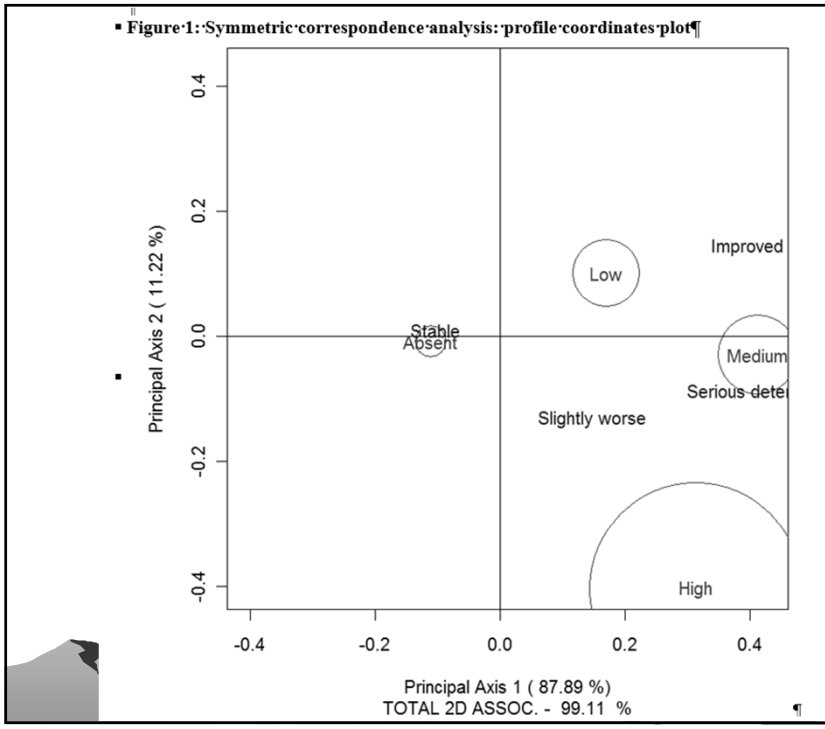
Tabella di contingenza di partenza

	Improved	Stable	Slightly worse	Serious deterioration	RowTotal
Absent	112 0.07	1318 0.78	187 0.11	69 0.04	1686 0.71
Low	55 0.14	258 0.68	45 0.12	24 0.06	382 0.16
Medium	43 0.16	152 0.57	43 0.16	30 0.11	268 0.11
High	3 0.08	21 0.55	10 0.26	4 0.11	38 0.016
ColTotal	213 0.09	1749 0.74	285 0.12	127 0.05	2374 1

Analisi preliminare

Both the two chi-square indexes (traditional chi-squared and Goodman-Kruskal τ index) have a *p*-value smaller than 0.001. This clearly suggests that the relationship between "degree of isolation" and "variation of the health conditions" is significant. If we give a closer look to the table, we can see that the relative frequencies inside the profile "Stable" increase monotonically passing from "High isolation" to "Absence of isolation" (third column of Table 4). We can see the opposite phenomenon inside the profiles "Slightly worse" and "Serious deterioration" (fourth and fifth columns of Table 4).

Figure 1: Symmetric correspondence analysis: profile coordinates plot



Asymmetric correspondence analysis: profile coordinates plot