Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia,
Carlo Natale Lauro, Marina Marino

UTET UNIVERSITÀ

5 Analisi delle relazioni tra due caratteri

- Rappresentazione congiunta di una coppia di fenomeni statistici: Distribuzioni doppie di frequenze
- Analisi delle relazioni tra due caratteri
- Misure di dipendenza
- Le relazioni fra variabili quantitative
- Le relazioni lineari

Capitolo a cura di
M.G. Grassia, C.N. Lauro, M.Marino

Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia,
Carlo Natale Lauro, Marina Marino

5 Analisi delle relazioni tra due caratteri

Capitolo a cura di
M.G. Grassia, C.N. Lauro, M.Marino

Analisi delle relazioni

Quale indice misura la relazione
tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

Analisi delle relazioni

Che tipo di variabili?

X e Y variabili qualitative

X variabile qualitativa e Y variabile quantitativa

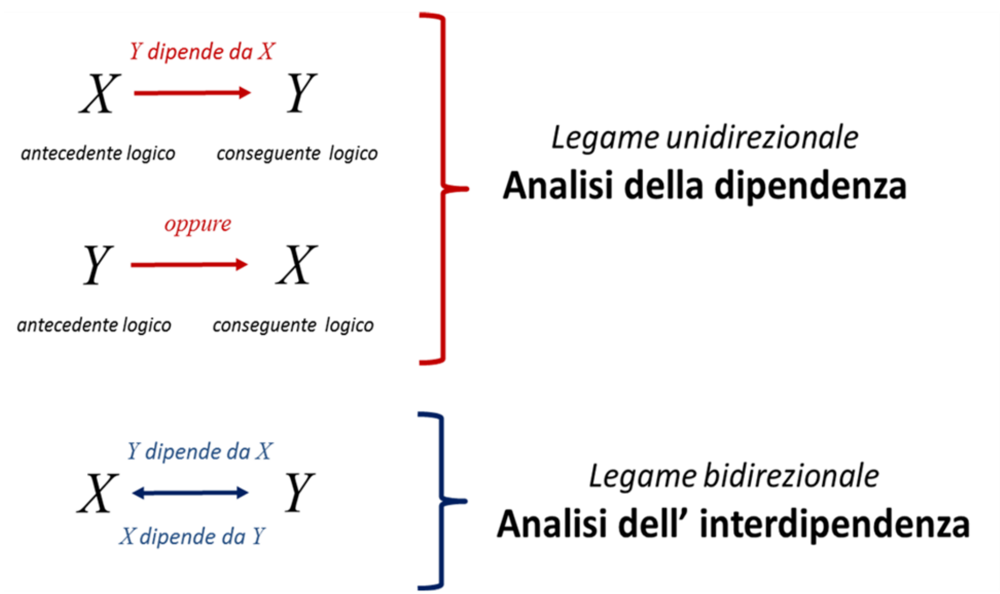
X e Y variabili quantitative

Analisi delle relazioni

Che tipo di relazione supporre ?

La scelta dell'indice di associazione dipende dal tipo di relazione logica supponibile tra i due caratteri osservati

Analisi della dipendenza e analisi dell'interdipendenza



Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

	2 mutabili	1 variabile 1 mutabile	2 variabili
Approccio simmetrico (interdipendenza)			
Approccio asimmetrico (dipendenza)			

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

	2 mutabili	1 variabile 1 mutabile	2 variabili
Approccio simmetrico (interdipendenza)	χ^2, Φ^2, V	χ^2, Φ^2, V	ρ
Approccio asimmetrico (dipendenza)	λ	η^2	Modello di Regressione

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

	2 mutabili	1 variabile 1 mutabile	2 variabili
Approccio simmetrico (interdipendenza)	χ^2, Φ^2, V	χ^2, Φ^2, V	
Approccio asimmetrico (dipendenza)			

Analisi delle relazioni

Indipendenza statistica = *la conoscenza della modalità di uno dei due caratteri non migliora la "previsione" della modalità dell'altro*

⬇

Se X è indipendente da Y allora anche Y è indipendente da X

Associazione tra variabili

Indipendenza statistica = *la conoscenza della modalità di uno dei due caratteri non migliora la "previsione" della modalità dell'altro*

⬇

Indipendenza in distribuzione

Tutti i profili riga identici e uguali alla distribuzione marginale di colonna

Tutti i profili colonna identici e uguali alla distribuzione marginale di riga

Associazione tra variabili

Indipendenza in distribuzione

Tutti i profili riga identici e uguali alla distribuzione marginale di colonna

Condizione occupazionale		gruppo medico	gruppo economico	gruppo letterario	Totale (%)
		Gruppo di corsi di laurea			
Occupati stabilmente	(% riga)	35.25	22.90	41.85	100
Occupati precariamente	(% riga)	35.25	22.90	41.85	100
Non lavorano	(% riga)	35.25	22.90	41.85	100
	Totale (%)	35.25	22.90	41.85	100

$$\frac{n_{ij}}{n_{i.}} * 100 = \frac{n_{.j}}{N} * 100 \Rightarrow \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N} \Rightarrow n_{ij} = \frac{n_{i.}\, n_{.j}}{N}$$

Associazione tra variabili

Indipendenza in distribuzione

Tutti i profili colonna identici e uguali alla distribuzione marginale di riga

Condizione occupazionale		Gruppo di corsi di laurea: gruppo medico	gruppo economico	gruppo letterario	Totale (%)
Occupati stabilmente	(% colonna)	60.80	60.80	60.80	60.80
Occupati precariamente	(% colonna)	31.29	31.29	31.29	31.29
Non lavorano	(% colonna)	7.90	7.90	7.90	7.90
	Totale (%)	100	100	100	100

$$\frac{n_{ij}}{n_{.j}} * 100 = \frac{n_{i.}}{N} * 100 \Rightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N} \Rightarrow n_{ij} = \frac{n_{i.}\, n_{.j}}{N}$$

Associazione tra variabili

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

1. Due caratteri sono **indipendenti** se la generica frequenza congiunta è uguale a:

$$n_{ij}^{*} = \frac{n_{i.} n_{.j}}{n}$$

2. C'è **Dipendenza perfetta** di Y da X se ad ogni modalità di X è associata una sola modalità di Y

3. C'è **Interdipendenza perfetta** tra X e Y se ad ogni modalità di X corrisponde una sola modalità di Y e viceversa

Associazione tra variabili

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

Due caratteri sono **indipendenti** se la generica frequenza congiunta è uguale a:

$$n_{ij}^{*} = \frac{n_{i.} n_{.j}}{n}$$

Indipendenza

		Y			
		y_1	y_2	y_3	**Totale**
X	x_1	1	4	5	**10**
	x_2	2	8	10	**20**
	x_3	2	8	10	**20**
	x_4	1	4	5	**10**
Totale		**6**	**24**	**30**	**60**

Associazione tra variabili

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

C'è **Dipendenza perfetta** di Y da X se ad ogni modalità di X è associata una sola modalità di Y

Y dipende perfettamente da X

		Y			
		y_1	y_2	y_3	**Totale**
X	x_1	43	0	0	43
	x_2	0	0	15	15
	x_3	0	0	20	20
	x_4	0	52	0	52
	Totale	43	52	35	130

Associazione tra variabili

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

C'è **Interdipendenza perfetta** tra X e Y se ad ogni modalità di X corrisponde una sola modalità di Y e viceversa

Interdipendenza perfetta

		Y			
		y_1	y_2	y_3	**Totale**
X	x_1	43	0	0	43
	x_2	0	0	15	15
	x_3	0	52	0	52
	Totale	43	52	15	110

Associazione tra variabili

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

In condizioni di **indipendenza assoluta** le distribuzioni condizionate di un carattere non mutano al variare delle modalità dell'altro carattere e, quindi, le frequenze teoriche sono uguali alle frequenze osservate.

Se tale uguaglianza non è soddisfatta per tutti i valori di i e j si dice che i due caratteri sono connessi.

La connessione risulta tanto più marcata quanto maggiori sono le differenze tra le frequenze osservate e le frequenze teoriche

$$c_{ij} = n_{ij} - n_{ij}^*$$

Se la connessione è positiva

$$n_{ij} > n_{ij}^*$$

Se la connessione è negativa

$$n_{ij} < n_{ij}^*$$

Associazione tra variabili
Indice del Chi quadrato

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

- Si considera una tabella a doppia entrata
- Si calcolano le **contingenze** $c_{ij} = n_{ij} - n_{ij}^{*}$
- Si calcolano i quadrati delle contingenze e si rapportano alla rispettiva frequenza teorica

$$\frac{{c_{ij}}^{2}}{{n_{ij}}^{*}} = \frac{(nij - {n_{ij}}^{*})^{2}}{{n_{ij}}^{*}}$$

- Si sommano le contingenze al quadrato rapportate alle frequenze teoriche

$$\chi^{2} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{c_{ij}^{2}}{{n_{ij}}^{*}} \geq 0$$

Associazione tra variabili
Proprietà dell'Indice del Chi quadrato

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

- L'indice χ^2 ha un limite inferiore che è pari a 0 (caso di indipendenza). Non ammette un limite superiore e pertanto non è possibile quantificare la dipendenza

- L'indice χ^2 è un indice simmetrico nel senso che misura contemporaneamente la dipendenza tra X, Y

- **Può essere calcolato per tutti i tipi di tabella, contingenza, correlazione, mista (per variabili di qualsiasi natura). E' l'unico indice calcolabile con variabili entrambe qualitative**

Associazione tra variabili
Indice del Chi quadrato

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

Tuttavia, l'indice χ2 dipende fortemente dalla numerosità del collettivo, per cui tende ad assumere valori tanto più grandi quanto maggiore è la numerosità del collettivo.

L'indice del Chi-quadrato

- ✓ È uguale o maggiore di 0
- ✓ Cresce al crescere di N
- ✓ Cresce al crescere delle modalità

Associazione tra variabili
Indice relativo (V di Cramer)

Indipendenza – dipendenza statistica

(assoluta o in distribuzione)

Indice di **contingenza quadratica**:

$$\Phi^2 = \chi^2 / N$$

Indice **V di Cramer**:

$$V = \sqrt{\frac{\Phi^2}{\min[(h-1),(k-1)]}}$$

- Varia tra 0 e 1
- Vale 0 nel caso di indipendenza.
- Vale 1 se:
  - i due caratteri sono perfettamente associati e il numero di righe della tabella è uguale al numero di colonne ($\boldsymbol{k} = \boldsymbol{h}$);
  - X dipende perfettamente da Y ed il numero di righe della tabella è maggiore del numero di colonne ($\boldsymbol{k} > \boldsymbol{h}$) ;
  - Y dipende perfettamente da X ed il numero di righe della tabella è minore del numero di colonne ($\boldsymbol{k} < \boldsymbol{h}$).

La seguente tabella riporta la distribuzione degli occupati per settori di attività economica e per posizione professionale.

Settori	**Posizione Professionale**	
	Dipendenti	Autonomi
Agricoltura	485	776
Industria	4147	956
Altre attività	4941	2546

Si vuole calcolare il grado di associazione tra i due caratteri mediante un indice appropriato.

Associazione tra variabili
Esempio

La seguente tabella riporta la distribuzione degli occupati per settori di attività economica e per posizione professionale.

Settori	Posizione Professionale	
	Dipendenti	Autonomi
Agricoltura	485	776
Industria	4147	956
Altre attività	4941	2546

Si vuole calcolare il grado di associazione tra i due caratteri mediante un indice appropriato.

Settore è una variabile qualitativa sconnessa

Posizione professionale è una variabile qualitativa sconnessa

Test del chi-quadro χ^2

Associazione tra variabili
Esempio

n_{ij}

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura	485	776	1261
Industria	4147	956	5103
Altre attività	4941	2546	7487
Totale	9573	4278	13851

n_{ij}^*

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura			1261
Industria			5103
Altre attività			7487
Totale	9573	4278	13851

$$n_{ij}^* = \frac{n_{i.}n_{.j}}{n}$$

Associazione tra variabili
Esempio

n_{ij}

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura	485	776	1261
Industria	4147	956	5103
Altre attività	4941	2546	7487
Totale	9573	4278	13851

n_{ij}^*

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura	871,5		1261
Industria			5103
Altre attività			7487
Totale	9573	4278	13851

$$n_{ij}^* = \frac{n_{i.} n_{.j}}{n}$$

Associazione tra variabili
Esempio

n_{ij}

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura	485	776	1261
Industria	4147	956	5103
Altre attività	4941	2546	7487
Totale	9573	4278	13851

n_{ij}^*

Settori	Posizione Professionale		Totale
	Dipendenti	Autonomi	
Agricoltura	871,5	389,5	1261
Industria	3527	1576	5103
Altre attività	5174,5	2312,5	7487
Totale	9573	4278	13851

$$n_{ij}^* = \frac{n_{i.} n_{.j}}{n}$$

Valori in caso di indipendenza

Associazione tra variabili
Esempio

n_{ij}	n_{ij}^*	$c_{ij}= n_{ij}- n_{ij}^*$	$c_{ij}= (n_{ij}- n_{ij}^*)^2$	$c_{ij}= (n_{ij}- n_{ij}^*)^2/n_{ij}^*$
485	871,5			
4147	3527			
4941	5174,5			
776	389,5			
956	1576			
2546	2312,5			
Totale				

Associazione tra variabili
Esempio

n_{ij}	n_{ij}^*	$c_{ij} = n_{ij} - n_{ij}^*$	$c_{ij} = (n_{ij} - n_{ij}^*)^2$	$c_{ij} = (n_{ij} - n_{ij}^*)^2 / n_{ij}^*$
485	871,5	-386,5		
4147	3527	620		
4941	5174,5	-233,5		
776	389,5	386,5		
956	1576	-620		
2546	2312,5	233,5		
Totale				

Associazione tra variabili
Esempio

n_{ij}	n_{ij}^*	$c_{ij}= n_{ij}- n_{ij}^*$	$c_{ij}= (n_{ij}- n_{ij}^*)^2$	$c_{ij}= (n_{ij}- n_{ij}^*)^2/n_{ij}^*$
485	871,5	-386,5	149382,25	
4147	3527	620	384400	
4941	5174,5	-233,5	54522,25	
776	389,5	386,5	149382,25	
956	1576	-620	384400	
2546	2312,5	233,5	54522,25	
Totale				

Associazione tra variabili
Esempio

n_{ij}	n_{ij}^*	$c_{ij}= n_{ij}- n_{ij}^*$	$c_{ij}= (n_{ij}- n_{ij}^*)^2$	$c_{ij}= (n_{ij}- n_{ij}^*)^2/n_{ij}^*$
485	871,5	-386,5	149382,25	171,4
4147	3527	620	384400	109
4941	5174,5	-233,5	54522,25	10,54
776	389,5	386,5	149382,25	383,5
956	1576	-620	384400	243,9
2546	2312,5	233,5	54522,25	23,6
Totale				**941,94**

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{c_{ij}^2}{n_{ij}^*} \geq 0$$

$$\chi^2 = 941,94$$

Associazione tra variabili
Esempio

χ^2 = 941,94

$\chi^2 > 0$

$\downarrow$

associazione
tra le due variabili

Ma quanto è forte questa associazione?

$$V = \sqrt{\frac{\chi^2}{\text{n}\,[\min(H-1),(K-1)]}}$$

$$V = \sqrt{\frac{941{,}94}{13.851}} = \mathbf{0{,}26}$$

Dall'analisi della V di Cramér emerge che i due caratteri non sono completamente indipendenti, ma esiste una debole associazione.

Associazione tra variabili
Proviamo insieme

Si vuole calcolare il livello di associazione esistente tra scelta del *mezzo di trasporto* e *condizioni meteorologiche* della giornata.

		mezzo di trasporto			
		bicicletta	autobus	automobile	Totale
condizioni metereologiche	sereno	84	26	11	**121**
	variabile	29	98	29	**156**
	pioggia	7	26	55	**88**
	Totale	**120**	**150**	**95**	**365**

Associazione tra variabili
Proviamo insieme

Si vuole calcolare il livello di associazione esistente tra scelta del *mezzo di trasporto* e *condizioni meteorologiche* della giornata.

		mezzo di trasporto			
		bicicletta	autobus	automobile	Totale
condizioni metereologiche	sereno	84	26	11	**121**
	variabile	29	98	29	**156**
	pioggia	7	26	55	**88**
	Totale	**120**	**150**	**95**	**365**

$$\chi^2 = 169{,}18$$

$$\Phi^2 = 169{,}18/365 = 0{,}4635$$

$$V = \sqrt{\frac{0{,}4635}{\min(3-1; 3-1)}} = 0{,}481$$

Tra i due caratteri sussiste un livello medio di associazione

Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET Università