UNIVERSITA' DEGLI STUDI DI PERUGIA
DIPARTIMENTO DI FILOSOFIA SCIENZE SOCIALI UMANE E DELLA FORMAZIONE
Corso di Laurea in Scienze per l'Investigazione e la Sicurezza

7. ASSOCIAZIONE TRA CARATTERI

Prof. Maurizio Pertichetti

Statistica sociale

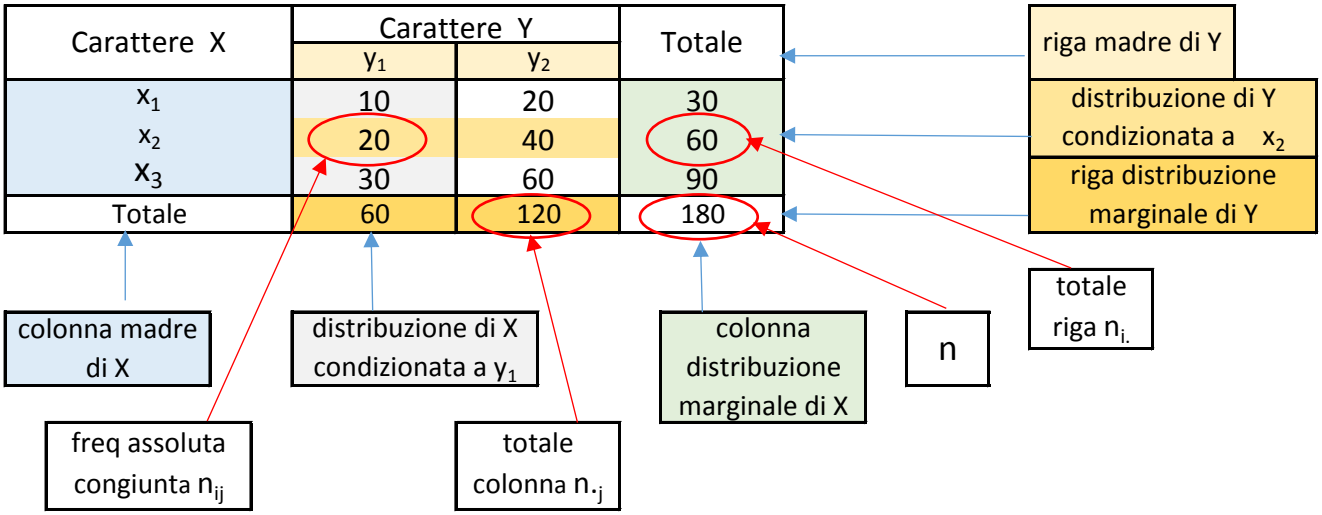
7. ASSOCIAZIONE TRA CARATTERI

Come già anticipato, nell'analisi dei dati si è sempre più spesso interessati a comprendere se tra due o più caratteri, che si presentano congiuntamente sulle unità statistiche di una popolazione, vi possa essere un qualche legame e, nel caso, quale sia il grado di di tale relazione. In questa sede, limitatamente alle modalità o classi di modalità di due caratteri, ci occuperemo delle relazioni di :

- **dipendenza** (o **indipendenza**) **assoluta** o **stocastica** attraverso l'analisi delle sole frequenze, particolarmente utile quando la distribuzione fa riferimento alle combinazioni di frequenze associate a due mutabili. Il grado di relazione fra le due variabili viene misurato con diversi indici statistici che, nel concreto, rappresentano la distanza tra la situazione effettivamente osservata e quella teorica riferita all'ipotesi di indipendenza. Gli indici in tal modo ottenuti per misurare tale legame associativo sono detti **indici di connessione**:
- **dipendenza** (o **indipendenza**) **interpolativa** attraverso l'individuazione di una funzione analitica capace di esprimere la relazione esistente tra due variabili, posta come implicita l'esistenza di una antecedenza logica di una variabile rispetto ad un'altra. Con il termine regressione si intende il modello atto a descrivere la relazione esistente tra una variabile dipendente e una o più variabili indipendenti o esplicative.

Oltre alla dipendenza tra caratteri, la teoria delle relazioni statistiche studia l'**interdipendenza**, ossia il legame reciproco tra due variabili, e il termine che sprime tale particolare relazione è quello di **correlazione**.

Riprendiamo la distribuzione doppia di frequenze e la corrispondente tabella a doppia entrata.



Si sa dalla teoria che una variabile Y si dice **indipendente** da una variabile X se la prima rimane costante al variare dei valori assunti dalla seconda. In caso contrario si dice che Y è **funzione** di X. L'assenza di una qualsiasi relazione tra due caratteri X e Y desumibili da una distribuzione doppia di frequenza è detta **indipendenza assoluta**, e si evince esaminando le distribuzioni condizionate.
Più precisamente, riprendendo la tabella precedente:

Indipendenza di Y da X

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	10	20	30
x_2	20	40	60
x_3	30	60	90
Totale	60	120	180

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	0,333	0,667	1,000
x_2	0,333	0,667	1,000
x_3	0,333	0,667	1,000
Totale	0,333	0,667	1,000

il carattere Y si dirà indipendente dal carattere X se le frequenze relative delle distribuzioni condizionate di Y risultano uguali tra loro e uguali alle frequenze marginali relative, per cui al variare della modalità X la distribuzione relativa di Y è la medesima.

Indipendenza di X da Y

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	10	20	30
x_2	20	40	60
x_3	30	60	90
Totale	60	120	180

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	0,167	0,167	0,167
x_2	0,333	0,333	0,333
x_3	0,500	0,500	0,500
Totale	1,000	1,000	1,000

analogamente il carattere X si dirà indipendente dal carattere Y, se le frequenze relative delle distribuzioni condizionate di X risultano uguali tra loro e uguali alle frequenze marginali relative, per cui al variare della modalità Y la distribuzione relativa di X è la medesima.

Il **concetto di indipendenza è simmetrico** per cui, se il carattere Y è indipendente dal carattere X, allora vale anche la relazione contraria, ovvero anche il carattere X è indipendente dal carattere Y.

Pertanto due caratteri X e Y si diranno indipendenti se è verificata l'uguaglianza:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$ **da cui si ottengono le frequenze teoriche di indipendenza** $$n_{ij}' = \frac{n_{i.} \times n_{.j}}{n}$$

Come si può evincere la tabella utilizzata si riferisce a due caratteri tra loro indipendenti, in quanto per ognuna delle frequenze assolute congiunte vale la suddetta uguaglianza:

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	10	20	30
x_2	20	40	60
x_3	30	60	90
Totale	60	120	180

$$n_{11}' = \frac{n_{1.} \times n_{.1}}{n} = \frac{30 \times 60}{180} = 10$$

$$n_{32}' = \frac{n_{3.} \times n_{.2}}{n} = \frac{90 \times 120}{180} = 60$$

Per contro la mancata validità per le frequenze assolute congiunte dell'uguaglianza di cui sopra, implica l'esistenza di una situazione di dipendenza.

La **dipendenza perfetta** è naturalmente l'antitesi della indipendenza. In particolare:

- Il carattere Y dipende perfettamente dal carattere X se ad ogni modalità del carattere X è associata una ed una sola modalità del carattere Y:

Carattere X	Carattere Y		Totale
	y_1	y_2	
x_1	0	20	20
x_2	20	0	20
x_3	0	60	60
Totale	20	80	100

tale relazione di dipendenza non è biunivoca.

- Il carattere X dipende perfettamente dal carattere Y se ad ogni modalità del carattere Y è associata una ed una sola modalità del carattere X:

Carattere X	Carattere Y				Totale
	y_1	y_2	y_3	y_4	
x_1	20	0	0	0	20
x_2	0	20	0	0	20
x_3	0	0	30	60	90
Totale	20	20	30	60	130

La **perfetta interdipendenza**, o se vogliamo la dipendenza reciproca, può essere raggiunta solo nel caso di tabella quadrata:

Carattere X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	25	0	0	25
x_2	0	0	30	30
x_3	0	45	0	45
Totale	25	45	30	100

e si ha quando: ad ogni modalità del carattere X corrisponde una e una sola modalità di Y e, concomitantemente, ad ogni modalità del carattere Y corrisponde una ed una sola modalità di X. Ovvero quando per ogni riga e colonna si ha un solo valore.

Gli **indici statistici** in grado di evidenziare l'indipendenza di un carattere statistico da un altro sono basati sul confronto (sulla distanza) tra le **frequenze osservate e quelle teoriche**, sotto l'ipotesi di indipendenza, e sono denominati **indici di connessione**. Tali indici assumono valori tanto più piccoli, quanto più esiste indipendenza tra i caratteri studiati.

Un indicatore in grado di misurare l'associazione tra due caratteri è dato dall'indice **chi-quadrato**, χ^2 , la cui espressione analitica è:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - n_{ij}' \right)^2}{n_{ij}'}$$

La differenza $(n_{ij} - n_{ij}')$ tra la frequenza osservata e la frequenza teorica è denominata **contingenza**.

Si tratta di un indice assoluto che ammette **valore minimo 0** se $\mathbf{n_{ij} = n_{ij}'}$, ossia se esiste indipendenza tra i caratteri , ma non ammette valore massimo in caso di dipendenza, quando $\mathbf{n_{ij} \neq n_{ij}'}$.

Per cui, nella misura di associazione si fa riferimento all'**indice normalizzato** V di **Cramer** dato da:

$$V = \sqrt{\frac{\chi^2}{n \times \min[(r-1);(c-1)]}}$$

Dove **n x min[(r-1);(c-1)]** sta a significare che il totale delle osservazioni **n** va moltiplicato per il valore più piccolo tra **r**, numero delle righe, e **c**, numero delle colonne detratto 1.

Tale indice varia tra **0**, nel caso di indipendenza, e **1**, nel caso di massima dipendenza.

Esempio di calcolo dell'indice chi-quadrato e dell'indice normalizzato di Cramer.

Data la tabella di frequenze osservate

Caratt X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	2	5	15	22
x_2	4	14	10	28
x_3	7	6	12	25
Totale	13	25	37	75

Sulla base dell'uguaglianza $n_{ij}' = \frac{n_{i.} \times n_{.j}}{n}$ si procede alla costruzione di una nuova tabella dove, fermi restando i valori delle righe e colonne marginali, al posto delle frequenze congiunte osservate si sostituiscono le frequenze congiunte teoriche nell'ipotesi di indipendenza dei due caratteri.

Caratt X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	22*13/75	22*25/75	22*37/75	22
x_2	28*13/75	28*25/75	28*37/75	28
x_3	25*13/75	25*25/75	25*37/75	25
Totale	13	25	37	75

Caratt X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	3,813	7,333	10,853	22
x_2	4,853	9,333	13,813	28
x_3	4,333	8,333	12,333	25
Totale	13	25	37	75

Si prosegue poi con l'elaborazione della tabella delle contingenze $(n_{ij} - n_{ij}')$.

Carattere X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	-1,813	-2,333	4,147	0
x_2	-0,853	4,667	-3,813	0
x_3	2,667	-2,333	-0,333	0
Totale	0	0	0	0

Ed infine tenuto conto dell'espressione $\frac{(n_{ij} - n_{ij}')^2}{n_{ij}'}$ si perviene al calcolo del chi-quadrato

Carattere X	Carattere Y			Totale
	y_1	y_2	y_3	
x_1	0,862	0,742	1,584	3,189
x_2	0,150	2,333	1,053	3,536
x_3	1,641	0,653	0,009	2,303
Totale	2,653	3,729	2,646	**9,028**

$\chi^2 = \mathbf{9,028}$

e quindi della V di Cramer:

$$V = \sqrt{\frac{\chi 2}{n \times \min[(r-1);(c-1)]}} = \sqrt{\frac{9,028}{75 \times 2}} = \sqrt{0,060} = \mathbf{0,245}$$

Dal risultato di V si evince che tra i due caratteri vi è una bassa connessione .

Nel caso di una tabella quadrata e di caratteri che presentano solo due modalità

	y_1	y_2	totale
x_1	a	b	a+b
x_2	c	d	c+d
totale	a+c	b+d	a+b+c+d

l'indice chi-quadrato e l'indice normalizzato di Cramer possono essere calcolati ricorrendo anche alle seguenti espressioni:

$$\chi^2 = \frac{(a \times d - b \times c)^2 \times n}{(a+b) \times (c+d) \times (a+c) \times (b+d)} \qquad V = \frac{(a \times d - b \times c)}{\sqrt{(a+b) \times (c+d) \times (a+c) \times (b+d)}}$$

Esempio dei diversi modi di calcolare l'indice Chi-quadro e l'indice normalizzato di Cramer.

Data la tabella di frequenze osservate

	y_1	y_2	TOT
x_1	17	9	26
x_2	11	15	26
TOT	28	24	52

y_1	y_2	TOT
14,000	12,000	26,000
14,000	12,000	26,000
28,000	24,000	52,000

y_1	y_2	TOT
0,6429	0,7500	1,3929
0,6429	0,7500	1,3929
1,2857	1,5000	**2,7857**

$\chi^2 = \sum\sum(Oss-Teo)^2/Teo = \mathbf{2,786}$

n x min tra (r-1);(c-1) = 52 x (2-1) = **52**

$$V = \sqrt{\frac{\chi^2}{52}} = \sqrt{\frac{2,786}{52,000}} = \sqrt{0,0536} = \mathbf{0,231}$$

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

17	9	26
11	15	26
28	24	52

a x d = 255
b x c = 99

$$\chi^2 = \frac{(a\times d-b\times c)^2 \times n}{(a+b)\times(c+d)\times(a+c)\times(b+d)} = \frac{(255-99)^2 \times 52}{26 \times 26 \times 28 \times 24} = \frac{1.265.472}{454.272} = \mathbf{2,786}$$

$$V = \frac{(a\times d-b\times c)}{\sqrt{(a+b)\times(c+d)\times(a+c)\times(b+d)}} = \frac{(255-99)}{\sqrt{26 \times 26 \times 28 \times 24}} = \frac{156}{\sqrt{454.272}} = \mathbf{0,231}$$