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. RELAZIONE TRA CARATTERI

## LA CONNESSIONE

Prof. Maurizio Pertichetti

Statistica sociale

# 7. RELAZIONE TRA CARATTERI: LA CONNESSIONE

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 tale relazione. In questa sede limiteremo l'analisi alle relazioni tra due caratteri. In termini tecnici si parla di dipendenza logica tra due caratteri quando tra questi vi è conoscenza a priori dell'esistenza di una relazione di causa ed effetto.

Se due variabili sono dipendenti, perché a questo induce il ragionamente sul loro rapporto di causa ed effetto, si può ritenere che esse lo siano anche in termini statistici. Diversamente, si può affermare che vi è indipendenza quando tra i due caratteri non si evince nessuna relazione di causa ed effetto, e che quindi lo siano anche in termini statistici. La dipendenza implica che vi sia una direzione nel legame tra i due caratteri, nel senso che se intervengono mutamenti in uno di essi, di conseguenza mutamenti si devono avere nell'altro. Come dire che nella relazione che presumibilmente li lega uno deve essere interpretato come l'antecedente logico e l'altro come il conseguente logico. Il legame in sostanza deve intendersi come unidirezionale e asimmetrico.

Ciò premesso:

- quando l'indipendenza è studiata attraverso l'analisi delle sole frequenze di una distribuzione doppia si parla di **connessione** tra i due caratteri. E il grado di relazione fra le due variabili, in assenza di indipendenza, 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**. E poiché sono ottenuti utilizzando la distribuzione delle frequenze e non le modalità, si deve sottolineare che essi sono gli unici indici calcolabili per misurare l'associazione tra caratteri qualitativi non ordinati;
- quando l'indipendenza è studiata attraverso l'analisi delle modalità assunte dai due caratteri si parla di **dipendenza funzionale** tra i due caratteri. E il grado di relazione fra le due variabili viene misurato mediante l'individuazione di una funzione analitica. Con il termine **regressione** si intende il modello atto a descrivere la relazione.

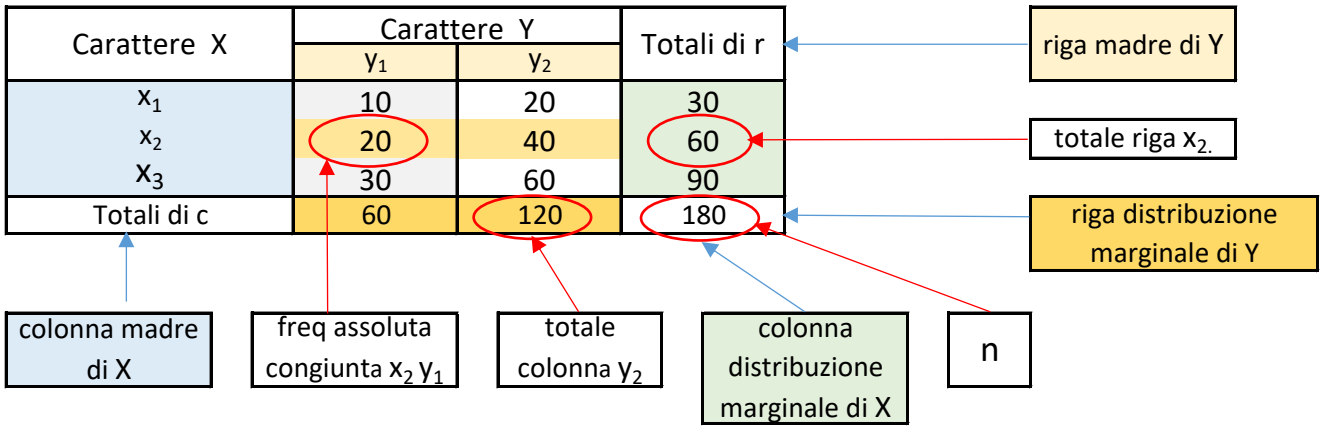
Oltre alla dipendenza tra caratteri, la teoria delle relazioni statistiche studia l'**interdipendenza**, ossia il legame reciproco tra due variabili, e il termine che sprime tale particolare relazione è quello di **correlazione**.

**Indipendenza, dipendenza e interdipendenza in distribuzione**

Riprendiamo la distribuzione doppia di frequenze e la corrispondente tabella a doppia entrata nella sua formulazione generale:

| | $y_1$ | $y_2$ | … | $y_j$ | … | $y_c$ | **Totali di riga** |
|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | … | $n_{1j}$ | … | $n_{1c}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | … | $n_{2j}$ | … | $n_{2c}$ | $n_{2.}$ |
| … | … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | … | $n_{ij}$ | … | $n_{ic}$ | $n_{i.}$ |
| … | … | … | … | … | … | … | … |
| $x_r$ | $n_{r1}$ | $n_{r2}$ | … | $n_{rj}$ | … | $n_{rc}$ | $n_{r.}$ |
| **Totali di colonna** | $n_{.1}$ | $n_{.2}$ | … | $n_{.j}$ | … | $n_{.c}$ | **n** |

E riprendiamo quella riferita ad un caso concreto:




Occupati secondo i settori e la posizione professionale

| Settori X | Posizione professionale Y | | Totali di r |
|---|---|---|---|
| | Dipendenti $Y_1$ | Autonomi $Y_2$ | |
| Agricoltura $X_1$ | 10 | 20 | 30 |
| Industria $X_2$ | 20 | 40 | 60 |
| Altre attività $X_3$ | 30 | 60 | 90 |
| Totali di c | 60 | 120 | 180 |

| | Dipendenti $Y_1$ |
|---|---|
| Agricoltura $X_1$ | 10 |
| Industria $X_2$ | 20 |
| Altre attività $X_3$ | 30 |

distribuzione di X condizionata a $y_1$

| | Dipendenti $Y_1$ | Autonomi $Y_2$ |
|---|---|---|
| Agricoltura $X_1$ | 10 | 20 |

distribuzione di Y condizionata a $x_1$

Due grandezze sono tra loro statisticamente indipendenti, quanro la legge di distribuzione di una delle due grandezze è uguale per qualunque valore dell'altra grandezza. Pertanto una variabile Y si assume come **indipendente** da una variabile X se, al variare dei valori assunti da questa, Y rimane costante. Se invece questa regolarità non si riscontra si dice allora che Y deve ritenersi come **funzione** di X. La mancanza di una qualsiasi relazione tra due caratteri X e Y deducibile da una distribuzione doppia di frequenza è detta **indipendenza assoluta**, e si accerta analizzando le distribuzioni condizionate.

Più precisamente, riprendendo la tabella precedente, riferita come esempio a due caratteri qualitativi:

Indipendenza di Y da X

| Carattere X | Carattere Y | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 10 | 20 | 30 |
| $x_2$ | 20 | 40 | 60 |
| $x_3$ | 30 | 60 | 90 |
| Totali di c | 60 | 120 | 180 |

| Carattere X | Carattere Y | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 0,333 | 0,667 | 1,000 |
| $x_2$ | 0,333 | 0,667 | 1,000 |
| $x_3$ | 0,333 | 0,667 | 1,000 |
| Totali di c | 0,333 | 0,667 | 1,000 |

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

Indipendenza di X da Y

| Carattere X | Carattere Y | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 10 | 20 | 30 |
| $x_2$ | 20 | 40 | 60 |
| $x_3$ | 30 | 60 | 90 |
| Totali di c | 60 | 120 | 180 |

| Carattere X | Carattere Y | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 0,167 | 0,167 | 0,167 |
| $x_2$ | 0,333 | 0,333 | 0,333 |
| $x_3$ | 0,500 | 0,500 | 0,500 |
| Totali di c | 1,000 | 1,000 | 1,000 |

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

Il **concetto di indipendenza** risulta essere **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.

E quindi due caratteri X e Y si diranno statisticamente indipendenti se sono verificate le uguaglianze:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \qquad \text{e} \qquad \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

ovvero ricordando la tabella a doppia entrata nella sua formulazione generale:

| | $y_1$ | $y_2$ | ... | $y_j$ | ... | $y_c$ | T rig |
|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1c}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2c}$ | $n_{2.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{ic}$ | $n_{i.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $x_r$ | $n_{r1}$ | $n_{r2}$ | ... | $n_{rj}$ | ... | $n_{rc}$ | $n_{r.}$ |
| T col | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.c}$ | n |

| | $y_1$ | $y_2$ | ... | $y_j$ | ... | $y_c$ | T rig |
|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1c}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2c}$ | $n_{2.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{ic}$ | $n_{i.}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $x_r$ | $n_{r1}$ | $n_{r2}$ | ... | $n_{rj}$ | ... | $n_{rc}$ | $n_{r.}$ |
| T col | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.c}$ | n |

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

ovvero le frequenze che si dovrebbero avere nel caso di indipendenza assoluta tra i caratteri X e Y.

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 | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 10 | 20 | 30 |
| $x_2$ | 20 | 40 | 60 |
| $x_3$ | 30 | 60 | 90 |
| Totali di c | 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 la negazione della indipendenza. In particolare:

- il carattere Y si dice dipende perfettamente dal carattere X se a ciascuna delle modalità del carattere X è associata una ed una sola modalità del carattere Y, ovvero quando per ogni riga si ha un solo valore diverso da zero :

| Carattere X | Carattere Y | | Totali di r |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | 0 | 20 | 20 |
| $x_2$ | 20 | 0 | 20 |
| $x_3$ | 0 | 60 | 60 |
| Totali di c | 20 | 80 | 100 |

**tale relazione di dipendenza non è biunivoca**:

- il carattere X dipende perfettamente dal carattere Y se a ciascuna delle modalità del carattere Y è associata una ed una sola modalità del carattere X, ovvero quando per ogni colonna si ha un solo valorediverso da zero :

| Carattere X | Carattere Y | | | | Totali di r |
|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | |
| $x_1$ | 20 | 0 | 0 | 0 | 20 |
| $x_2$ | 0 | 20 | 0 | 0 | 20 |
| $x_3$ | 0 | 0 | 30 | 60 | 90 |
| Totali di c | 20 | 20 | 30 | 60 | 130 |

La **perfetta interdipendenza**, o se vogliamo la interdipendenza reciproca, può essere raggiunta solo nel caso di tabella quadrata, cioè con stesso numero di righe e colonne:

| Carattere X | Carattere Y | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | 25 | 0 | 0 | 25 |
| $x_2$ | 0 | 0 | 30 | 30 |
| $x_3$ | 0 | 45 | 0 | 45 |
| Totali di c | 25 | 45 | 30 | 100 |

e si riscontra quando in corrispondenza di ciascuna modalità del carattere X si ha una ed una sola modalità di Y e, al tempo stesso, quando in corrispondenza di ciascuna modalità del carattere Y si ha una ed una sola modalità di X, ovvero quando per ciascuna delle righe e ciascuna delle colonne si ha un solo valore diverso da zero.

**Misure del legame associativo in distribuzione doppie di frequenza per caratteri qualitativi**

In una distribuzione doppia di frequenza, una volta accertata l'assenza di indipendenza o di dipendenza perfetta tra i caratteri, l'ipotesi evidentemente non può che essere quella della presenza di un qualche livello di connessione tra i due suddetti estremi, che pertanto dovrà essere misurato.

Come già anticipato, gli **indici statistici** in grado di misurare l'indipendenza di un carattere statistico da un altro sono basati sul confronto (o meglio 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** $\chi^2$, un indice assoluto 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**.

Il $\chi^2$ è sempre non negativo, ovvero assume valori sempre maggiori o uguali a zero. Ammette il **valore minimo 0** se $\mathbf{n_{ij} = n_{ij}'}$, ossia se esiste indipendenza tra i caratteri , e risulta tanto più grande quanto più ci si allontana dalla situazione di indipendenza. A parità di associazione l'indice aumenta al crescere di **n**.

Più opportuno, per la misura del legame associativo, è l'**indice normalizzato** chiamato 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.

Poiché **chi-quadrato** $\chi^2$ e **V** di **Cramer** dipendono dalla distribuzione di frequenze e non dalle modalità, ne consegue che tali indici sono gli unici utilizzabili per misurare l'associazione tra due caratteri qualitativi sconnessi.

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

Data la tabella di frequenze osservate

| Caratt X | Carattere Y | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | 2 | 5 | 15 | 22 |
| $x_2$ | 4 | 14 | 10 | 28 |
| $x_3$ | 7 | 6 | 12 | 25 |
| Totali di c | 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 | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | 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 |
| Totali di c | 13 | 25 | 37 | 75 |

| Caratt X | Carattere Y | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | 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 |
| Totali di c | 13 | 25 | 37 | 75 |

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

| Carattere X | Carattere Y | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | -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 |
| Totali di c | 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 | | | Totali di r |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | |
| $x_1$ | 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 |
| Totali di c | 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 con caratteri che presentano solo due modalità

| | $y_1$ | $y_2$ | totali di r |
|---|---|---|---|
| $x_1$ | a | b | a+b |
| $x_2$ | c | d | c+d |
| totali di c | 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:

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

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 |

$$n_{ij}' = \frac{n_{i.} \times n_{.j}}{n}$$

| $y_1$ | $y_2$ | TOT |
|---|---|---|
| 14,000 | 12,000 | 26,000 |
| 14,000 | 12,000 | 26,000 |
| 28,000 | 24,000 | 52,000 |

$$\frac{\left( n_{ij} - n_{ij}' \right)^2}{n_{ij}'}$$

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

$\chi^2$ = $\sum\sum(Oss-Teo)^2/Teo$ = **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 | 17 | 9 | 26 | a x d = 255 |
| c | d | c+d | 11 | 15 | 26 | b x c = 99 |
| a+c | b+d | a+b+c+d | 28 | 24 | 52 | |

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

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