

# RELAZIONI STATISTICHE

## Analisi delle relazioni tra due variabili osservate X, Y

E' necessario identificare:

- il **tipo di relazione** che si vuole studiare
- **la natura delle variabili** di partenza
- l'**indice** con cui misurare la relazione in esame
  
- In base al tipo di relazione che si vuole studiare distinguiamo relazioni di:
  - indipendenza assoluta
  - indipendenza in media
  - dipendenza lineare
  - interdipendenza lineare
  
- X e Y possono essere:
  - entrambe qualitative (tabella di contingenza)
  - una qualitativa e l'altra quantitativa (tabella mista)
  - entrambe quantitative (tabella di correlazione)
  
- Gli indice principali proposti in letteratura sono:
  - L'indice  $\chi^2 \Rightarrow$  misura tipica dell'indipendenza assoluta
  - L'indice  $\eta^2 \Rightarrow$  misura tipica dell'indipendenza in media
  - Il coefficiente di regressione lineare  $\beta \Rightarrow$  misura tipica della dipendenza lineare
  - Il coefficiente di correlazione  $\rho \Rightarrow$  misura tipica della interdipendenza

# INDIPENDENZA ASSOLUTA

| X/Y            | y <sub>1</sub>  | y <sub>2</sub> | .. | y <sub>j</sub>  | .. | y <sub>h</sub>  | Totale          |
|----------------|-----------------|----------------|----|-----------------|----|-----------------|-----------------|
| x <sub>1</sub> | n <sub>11</sub> |                |    |                 |    |                 | n <sub>1.</sub> |
| x <sub>2</sub> |                 |                |    |                 |    |                 |                 |
| ..             |                 |                |    |                 |    |                 |                 |
| x <sub>i</sub> |                 |                |    | n <sub>ij</sub> |    |                 | n <sub>i.</sub> |
| ..             |                 |                |    |                 |    |                 |                 |
| x <sub>s</sub> |                 |                |    |                 |    | n <sub>sh</sub> |                 |
| Totale         | n <sub>.1</sub> |                |    | n <sub>.j</sub> |    |                 | N               |

$$\frac{\sum_{i=1}^s n_{ij}}{N} = \frac{n_{.j}}{N} \quad \text{Frequenza marginale relativa di colonna}$$

$$\frac{\sum_{j=1}^h n_{ij}}{N} = \frac{n_{i.}}{N} \quad \text{Frequenza marginale relativa di riga}$$

$$\frac{n_{ij}}{n_{.j}} = \text{Frequenza condizionata alla j-colonna}$$

$$\frac{n_{ij}}{n_{i.}} = \text{Frequenza condizionata alla i-esima riga}$$

**in condizioni di indipendenza assoluta (stocastica):**

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N} \quad \text{ed} \quad \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$

**le distribuzioni condizionate di un carattere non mutano al variare delle modalità dell'altro carattere**

conseguentemente:

**le frequenze teoriche sono uguali alle frequenze osservate**

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

se tale uguaglianza non è soddisfatta per tutti i valori di  $i$  e di  $j$ , si dice che i due caratteri  $X$ ,  $Y$  sono **connessi**  
la connessione risulta tanto più marcata quanto maggiori sono le differenze tra le frequenze osservate e le frequenze teoriche (contingenze)

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

**connessione positiva**

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

**connessione negativa**

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

**L'INDICE  $\chi^2$**

L'indice più utilizzato per misurare la connessione tra due variabili è l'indice  $\chi^2$

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^h \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

$$0 \leq \chi^2 < \infty$$

## PROPRIETÀ

L'indice  $\chi^2$  ha un limite inferiore che è pari a 0 (caso di indipendenza). Non ammette però un massimo assoluto e, quindi, non è possibile quantificare la dipendenza esistente tra i due caratteri.

L'indice  $\chi^2$  è un indice **simmetrico** nel senso che misura contemporaneamente la dipendenza tra X,Y.

L'indice  $\chi^2$  può essere calcolato per tutti i tipi di tabella, contingenza, correlazione, mista, (e quindi per variabili di qualsiasi natura). E' l'unico indice che misura la dipendenza tra due caratteri entrambi qualitativi.

Per definire un indice che abbia un massimo assoluto ci si riconduce all'indice:

$$\phi^2 = \frac{\chi^2}{N}$$

che ammette come limite superiore (valore massimo) :

$$\phi^2 \leq \min \{ (s-1); (h-1) \}$$

Si può quindi costruire un indice normalizzato che assuma valori tra (0,1)

$$V - \text{Cramér} = \frac{\phi^2}{\min \{ (s-1); (h-1) \}}$$

## Problemi

Data la seguente tabella riportante il numero di matrimoni per grado di istruzione del marito e della moglie, calcolare il grado di connessione tra i due caratteri

| Grado di istruzione marito | Grado di istruzione moglie |         |            |                 |               | <i>Totale</i> |
|----------------------------|----------------------------|---------|------------|-----------------|---------------|---------------|
|                            | laurea                     | diploma | lic. media | lic. elementare | nessun titolo |               |
| laurea                     | 236                        | 327     | 119        | 48              | 2             | 732           |
| diploma                    | 134                        | 1038    | 701        | 430             | 20            | 2323          |
| lic.media                  | 30                         | 492     | 1199       | 1222            | 110           | 3053          |
| lic.elementare             | 9                          | 218     | 666        | 4523            | 951           | 6367          |
| nessun titolo              | 3                          | 20      | 33         | 405             | 1195          | 1643          |
| <i>Totale</i>              | 412                        | 2082    | 2718       | 6628            | 2278          | 14118         |

## Soluzione

### frequenze teoriche

| Grado di istruzione marito | Grado di istruzione moglie |         |            |                 |               | <i>Totale</i> |
|----------------------------|----------------------------|---------|------------|-----------------|---------------|---------------|
|                            | laurea                     | diploma | lic. media | lic. elementare | nessun titolo |               |
| laurea                     | 21                         | 108     | 141        | 344             | 118           | 732           |
| diploma                    | 68                         | 342     | 447        | 1091            | 375           | 2323          |
| lic.media                  | 89                         | 450     | 588        | 1433            | 493           | 3053          |
| lic.elementare             | 186                        | 939     | 1226       | 2989            | 1027          | 6367          |
| nessun titolo              | 48                         | 243     | 316        | 771             | 265           | 1643          |
| <i>Totale</i>              | 412                        | 2082    | 2718       | 6628            | 2278          | 14118         |

## contingenze

| Grado di istruzione marito | Grado di istruzione moglie |         |            |                 |               | <i>Totale</i> |
|----------------------------|----------------------------|---------|------------|-----------------|---------------|---------------|
|                            | laurea                     | diploma | lic. media | lic. elementare | nessun titolo |               |
| laurea                     | 215                        | 219     | -22        | -296            | -116          | 0             |
| diploma                    | 66                         | 696     | 254        | -661            | -355          | 0             |
| lic.media                  | -59                        | 42      | 611        | -211            | -353          | 0             |
| lic.elementare             | -177                       | -721    | -560       | 1534            | -76           | 0             |
| nessun titolo              | -45                        | -236    | -283       | -366            | 930           | 0             |
| <i>Totale</i>              | 0                          | 0       | 0          | 0               | 0             | 0             |

Commento:

si osserva l'esistenza di una marcata attrazione dei due coniugi tra livelli prossimi e repulsione tra livelli di istruzione molto diversi

$$\chi^2 = \frac{215^2}{21} + \frac{219^2}{108} + \dots + \frac{930^2}{265} = 11961$$

$$\phi^2 = \frac{11961}{14118} = 0,847$$

$$V = \frac{0,847}{4} = 0,21$$

Data la seguente tabella riportante le preferenze di 40 consumatori rispetto a due tipi di bibita ed al tasso di zucchero, calcolare il grado di connessione tra i due caratteri

|                  | preferenza legata al tasso di zucchero |              |                |        |
|------------------|--|--------------|----------------|--------|
| bibita preferita | zuccherata                             | indifferente | non zuccherata | Totale |
| coca cola        | 8                                      | 9            | 7              | 24     |
| pepsi cola       | 6                                      | 4            | 6              | 16     |
| <i>Totale</i>    | 14                                     | 13           | 13             | 40     |

### Soluzione

#### frequenze teoriche

|                  | preferenza legata al tasso di zucchero |              |                |        |
|------------------|--|--------------|----------------|--------|
| bibita preferita | zuccherata                             | indifferente | non zuccherata | Totale |
| coca cola        | 8,4                                    | 7,8          | 7,8            | 24     |
| pepsi cola       | 5,6                                    | 5,2          | 5,2            | 16     |
| <i>Totale</i>    | 14                                     | 13           | 13             | 40     |

#### contingenze

|                  | preferenza legata al tasso di zucchero |              |                |        |
|------------------|--|--------------|----------------|--------|
| bibita preferita | zuccherata                             | indifferente | non zuccherata | Totale |
| coca cola        | -0,4                                   | 1,2          | -0,8           | 0      |
| pepsi cola       | 0,4                                    | -1,2         | 0,8            | 0      |
| <i>Totale</i>    | 0                                      | 0            | 0              | 0      |

$$\chi^2 = \frac{0,4^2}{8,4} + \frac{1,2^2}{7,8} + \dots + \frac{0,8^2}{5,2} = 0,71$$

$$\phi^2 = V = \frac{0,71}{40} = 0,01$$

## INDIPENDENZA IN MEDIA

| <b>X/Y</b>    | $y_1$    | $y_2$ | .. | $y_j$    | .. | $y_h$    | <b>Totale</b> |
|---------------|----------|-------|----|----------|----|----------|---------------|
| $x_1$         | $n_{11}$ |       |    |          |    |          | $n_{1.}$      |
| $x_2$         |          |       |    |          |    |          |               |
| ..            |          |       |    |          |    |          |               |
| $x_i$         |          |       |    | $n_{ij}$ |    |          | $n_{i.}$      |
| ..            |          |       |    |          |    |          |               |
| $x_s$         |          |       |    |          |    | $n_{sh}$ |               |
| <b>Totale</b> | $n_{.1}$ |       |    | $n_{.j}$ |    |          | <b>N</b>      |

$$M(X) = \frac{\sum_{i=1}^s x_i n_{i.}}{N} \quad \text{media generale di X}$$

$$M(Y) = \frac{\sum_{j=1}^h y_j n_{.j}}{N} \quad \text{media generale di Y}$$

$$M(X / y_j) = \frac{\sum_{i=1}^s x_i n_{ij}}{n_{.j}} \quad j=(1, \dots, h)$$

media condizionata di X rispetto a Y

$$M(Y / x_i) = \frac{\sum_{j=1}^h y_j n_{ij}}{n_{i.}} \quad i=(1, \dots, s)$$

media condizionata di Y rispetto a X

in condizioni di indipendenza in media di Y da X

$$M(Y / x_1) = M(Y / x_2) = \dots\dots\dots = M(Y / x_s) = M(Y)$$

le medie condizionate di Y rispetto a X sono tutte uguali tra loro ed uguali alla media generale

in condizioni di indipendenza in media di X da Y

$$M(X / y_1) = M(X / y_2) = \dots\dots\dots = M(X / y_h) = M(X)$$

le medie condizionate di X rispetto a Y sono tutte uguali tra loro ed uguali alla media generale di X

### PROPRIETÀ

l'indipendenza in media non è una proprietà simmetrica, se Y è indipendente in media da X non è detto l'inverso;

l'indipendenza assoluta implica l'indipendenza in media, non è vero l'inverso

l'indipendenza in media di Y da X può essere calcolata solo se Y è quantitativa, qualsiasi sia la natura di X

## L'INDICE $\eta^2$

L'indice più utilizzato per misurare la dipendenza in media tra due variabili è l'indice  $\eta^2$

$$\eta_{y/x}^2 = \frac{\sum_{i=1}^s (M_{y/x_i} - M(Y))^2 n_i}{\sum_{j=1}^h (y_j - M(Y))^2 n_j}$$

$$\text{DEV}(B) = \sum_{i=1}^s (M_{y/x_i} - M(Y))^2 n_i$$

Devianza tra gruppi

$$\text{DEV}(T) = \sum_{j=1}^h (y_j - M(Y))^2 n_j$$

Devianza totale

$$\text{DEV}(T) - \text{DEV}(B) = \text{DEV}(W)$$

$$\text{DEV}(W) = \sum_{i=1}^s \sum_{j=1}^h (y_j - M(Y/x_i))^2 n_{ij}$$

Devianza nei gruppi

$$0 \leq \eta^2 \leq 1$$

## PROPRIETÀ

L'indice  $\eta^2$  assume valore 0 quando c'è indipendenza in media ed 1 quando c'è massima dipendenza in media (il che si realizza quando nella tabella ad un solo valore della variabile condizionata corrisponde un unico valore della variabile condizionante)

L'indice  $\eta^2$  è un indice asimmetrico, in presenza di due variabili quantitative se ne devono calcolare due per valutare la dipendenza di una dall'altra

L'indice  $\eta^2$  può essere calcolato per tabelle di correlazione e per tabelle miste, in quest'ultimo caso misura la bontà della divisione in gruppi definita dalla variabile qualitativa.

## Problemi

Data la seguente tabella che riporta la distribuzione di 700 studenti secondo il voto conseguito all'esame di statistica e il sesso, calcolare quanto incide la differenza tra maschi e femmine rispetto al voto finale

|         | Voto  |       |       |       |       | Totale |
|---------|-------|-------|-------|-------|-------|--------|
| Sesso   | 18-20 | 21-23 | 24-25 | 26-28 | 29-30 |        |
| maschi  | 11    | 23    | 38    | 116   | 75    | 263    |
| femmine | 55    | 163   | 107   | 48    | 64    | 437    |
| Totale  | 66    | 186   | 145   | 164   | 139   | 700    |

## Soluzione

$$\begin{aligned} M(\text{Voto})/\text{maschio} &= \frac{(19 \cdot 11) + (22 \cdot 23) + (24,5 \cdot 38) + (27 \cdot 116) + (29,5 \cdot 75)}{263} = \\ &= 26,58 \end{aligned}$$

$$\begin{aligned} M(\text{Voto})/\text{femmine} &= \frac{(19 \cdot 55) + (22 \cdot 163) + (24,5 \cdot 107) + (27 \cdot 48) + (29,5 \cdot 64)}{437} = \\ &= 23,882 \end{aligned}$$

$$\begin{aligned} M(\text{Voto}) &= \frac{(19 \cdot 66) + (22 \cdot 186) + (24,5 \cdot 145) + (27 \cdot 164) + (29,5 \cdot 139)}{700} = \\ &= 24,896 \end{aligned}$$

$$\begin{aligned} \text{Dev}(\text{Voto tra i due sessi}) &= (26,58 - 24,896)^2 \cdot 263 + (23,882 - 24,896)^2 \cdot 437 \\ &= 1194,883 \end{aligned}$$

$$\begin{aligned} \text{Dev}(\text{Voto}) &= (19 - 24,896)^2 \cdot 66 + (22 - 24,896)^2 \cdot 186 + \\ &+ (24,5 - 24,896)^2 \cdot 145 + (27 - 24,896)^2 \cdot 164 + (29,5 - 24,896)^2 \cdot 139 \\ &= 7549,387 \end{aligned}$$

$$\eta^2 = \frac{1194,883}{7549,387} = 0,158$$

Il 15,8% della variabilità del voto è spiegato dal sesso, debole dipendenza in media

Si osservano 13 studenti in base al voto conseguito all'esame di statistica e la frequenza al corso, calcolare quanto la variabilità dei voti dipende dalla frequenza.

- studenti frequentanti: 23,25,25,27,28,29,29,30;
- studenti non frequentanti: 18,19,20,22,30

$$M(\text{Voto})/\text{freq} = \frac{23 + (2 \cdot 25) + 27 + 28 + (29 \cdot 2) + 30}{8} = 27$$

$$M(\text{Voto})/\text{non freq} = \frac{18 + 19 + 20 + 22 + 30}{5} = 21,8$$

$$M(\text{Voto}) = \frac{(27 \cdot 8) + (21,8 \cdot 5)}{13} = 25$$

$$\begin{aligned} \text{Dev}(\text{Voto tra gruppi}) &= (27 - 25)^2 \cdot 8 + (21,8 - 25)^2 \cdot 5 \\ &= 83,2 \end{aligned}$$

$$\begin{aligned} \text{Dev}(\text{Voto}) &= (18 - 25)^2 + (19 - 25)^2 + (20 - 25)^2 + (22 - 25)^2 + \\ &+ (23 - 25)^2 + (25 - 25)^2 \cdot 2 + (27 - 25)^2 + (28 - 25)^2 + \\ &+ (29 - 25)^2 \cdot 2 + (30 - 25)^2 \cdot 2 = \\ &= 218 \end{aligned}$$

$$218 = 134,8 + 83,2$$

Il 38% della differenza tra i voti dipende dalla differenza tra gruppi

$$\text{Dev(Voto / freq)} = (23 - 27)^2 + (25 - 27)^2 \cdot 2 + (27 - 27)^2 + (28 - 27)^2 + \\ + (29 - 27)^2 \cdot 2 + (30 - 27)^2 = 42$$

$$\text{Dev(Voto / nonfreq)} = (18 - 21,8)^2 + (19 - 21,8)^2 + (20 - 21,8)^2 + (22 - 21,8)^2 + \\ + (30 - 21,8)^2 = 92,8$$

$$134,8 = 42 + 92,8$$