Lezione 8: Relazioni tra variabili

Corso di Statistica
Facoltà di Economia
Università della Basilicata

Prof. Massimo Aria

- aria@unina.it

Indipendenza

Lo studio delle relazioni esistenti tra due variabili statistiche parte dalla definizione del concetto di indipendenza.

In particolare, in statistica si considerano tre concetti di indipendenza tra caratteri:

- *indipendenza assoluta*
- *indipendenza in media*
- *incorrelazione lineare*

Si parla di **indipendenza assoluta** (o *indipendenza in* distribuzione) tra due caratteri quando le modalità assunte dalla X non modificano la distribuzione di Y.

In altre parole, la distribuzione condizionata della Y dato $X=x_i$ non cambia per qualunque i=1,2,…,k.
Similmente, la distribuzione condizionata della X dato $Y=y_j$ non cambia per qualunque j=1,2,…,h.

Frequenze teoriche

Quando tra due variabili X e Y vi è indipendenza assoluta, allora i profili (le distribuzioni condizionate espresse in termini di frequenze relative) riga saranno tutti uguali tra loro e pari al profilo medio.

Analogamente ciò sarà vero anche per i profili colonna.

Da questa affermazione è possibile derivare un'ulteriore definizione di indipendenza assoluta:

"X e Y si dicono indipendenti quando le frequenze osservate sono uguali alle frequenze teoriche per ogni cella (i,j) della distribuzione doppia".

Indipendenza in distribuzione

- Si ha quando le frequenze osservate in ogni cella (i,j) equivalgono alle frequenze "teoriche"

$$n^*_{ij} = \frac{n_{i+} n_{+j}}{N}$$

- Le distribuzioni condizionate relative (profili) si equivalgono
  - per riga $\frac{n_{ij}}{n_{i+}} = \frac{n_{+j}}{N}$ per $j = 1, \ldots, h$
  - per colonna $\frac{n_{ij}}{n_{+j}} = \frac{n_{i+}}{N}$ per $i = 1, \ldots, k$

Indipendenza in distribuzione

Concetto di contingenza

La costruzione di un indice che misuri il grado di connessione tra due caratteri statistici X e Y si basa sul concetto di contingenza.

Si definisce **contingenza** c_{ij} la differenza tra la frequenza osservata e la frequenza teorica di una generica cella ij.

Nel caso di indipendenza le contingenze sono tutte nulle mentre queste cresceranno, in valore assoluto, al crescere del grado di dipendenza tra i caratteri.

Concetto di contingenza

$$c_{ij} = \left(n_{ij} - n^*_{ij}\right)$$

Freq. osservate — Freq. teoriche

$$\sum_i \sum_j c_{ij} = \sum_i \sum_j \left(n_{ij} - n^*_{ij}\right) = 0$$

La somma delle contingenze è sempre nulla!!

Contingenze

Indice di connessione di Pearson

Il grado di connessione tra due caratteri statistici si misura attraverso l'**indice di connessione di Pearson** (X^2).

Esso è ottenuto come *somma delle contingenze quadratiche relative*.

L'indice assume valore pari a zero in caso di indipendenza in distribuzione e aumenta al crescere del grado di connessione.

Solitamente l'indice X^2 è impiegato per la misurare la relazione tra due mutabili.
Infatti per questo tipo di variabili l'unica informazione analizzabile riguarda le frequenze congiunte.

Indice di connessione del Pearson

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}}$$

$$0 \le \chi^2 \le N \times [\min(k,h) - 1]$$

Il più piccolo valore tra il numero di righe e il numero di colonne della tabella meno 1

Indice di connessione di Pearson

Ulteriori indici di connessione

Per svincolare l'indice X^2 dalla numerosità N della popolazione, sono state proposte numerose varianti tra cui:

- l'**indice di contingenza media quadratica** che calcola la media delle contingenze al quadrato relative.
Esso è ottenuto rapportano il X^2 a N.

- l'**indice V di Cramer**, che consiste in una versione normalizzata dell'indice di contingenza media quadratica.
Esso sarà pari a 0 nel caso di indipendenza e pari a 1 nel caso di massima connessione.

Indice di contingenza media quadratica

$$\phi^2 = \frac{\chi^2}{N} = \frac{1}{N} \sum_i \sum_j \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}}$$

$$0 \le \phi^2 \le \min(k,h) - 1$$

Indice V di Cramer

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

$$0 \le V \le 1$$

Ulteriori indici di connessione

Esempio di misurazione della connessione

Uno studioso vuole stabilire se c'è dipendenza tra il reddito e il numero di viaggi effettuati in un anno, considerando un collettivo di 100 individui descritto nella tabella seguente:

	0 Viaggi	1 Viaggio	2 Viaggi	Più di 2 Viaggi	
Basso	12	12	4	1	29
Medio	10	14	10	6	40
Alto	4	6	12	9	31
	26	32	26	16	100

1 ° passo: calcolo delle frequenze teoriche:

$n_{ij}^* = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}$	0 Viaggi	1 Viaggio	2 Viaggi	Più di 2 Viaggi
Basso	7,54	9,28	7,54	4,64
Medio	10,4	12,8	10,4	6,4
Alto	8,06	9,92	8,06	4,96

2° passo: calcolo delle contingenze

$C_{ij} = \left(n_{ij} - n_{ij}^*\right)$	0 Viaggi	1 Viaggio	2 Viaggi	Più di 2 Viaggi
Basso	4,46	2,72	-3,54	-3,64
Medio	-0,4	1,2	-0,4	-0,4
Alto	-4,06	-3,92	3,94	4,04

3° passo: calcolo delle contingenze al quadrato rapportate alle frequenze teoriche

$\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$	0 Viaggi	1 Viaggio	2 Viaggi	Più di 2 Viaggi
Basso	2,638	0,797	1,662	2,856
Medio	0,015	0,113	0,015	0,025
Alto	2,045	1,549	1,926	3,291

4° passo: calcolo degli indici di connessione Chi Quadrato e Phi Quadrato

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*} = 16,932 \qquad \phi^2 = \frac{\chi^2}{N} = \frac{16,932}{100} = 0,169$$

Indipendenza in media

Sia X una mutabile e Y una variabile quantitativa e sia (X,Y) la variabile doppia generata dall'osservazione congiunta di X e Y.

In questo caso, nello studio della relazione doppia è possibile considerare un diverso concetto di dipendenza che coinvolge anche i valori assunti dalla variabile quantitativa.

Si diche che **Y è indipendente in media da X**, se al variare delle modalità della X, le medie delle distribuzioni condizionate di Y (*medie condizionate*) rimangono costanti.

Indipendenza in media → $M(Y|x_1)=M(Y|x_2),\ldots, M(Y|x_i),\ldots, M(Y|x_k)=M(Y)$

L'**indipendenza in distribuzione implica quella in media** ma non è vero il contrario.

Esempio di distribuzione doppia mista

Nella tabella seguente si riporta la distribuzione doppia dei prezzi di un certo prodotto venduto in tre diverse città italiane.

X=Città **Y**=Prezzo del prodotto

		Y		
		80--|100	100--|140	
	Val.cent.	90	120	
X	Roma	3	3	6
	Napoli	7	2	9
	Firenze	2	3	5
		12	8	20

La media generale è: $\mu_Y = \frac{(90 \cdot 12) + (120 \cdot 8)}{20} = 102$

Medie condizionate

- La media per la città di Roma è: $\mu_{Y|x=RO} = \frac{(90 \cdot 3) + (120 \cdot 3)}{6} = 105$
- La media per la città di Napoli è: $\mu_{Y|x=NA} = \frac{(90 \cdot 7) + (120 \cdot 2)}{9} = 96{,}67$
- La media per la città di Firenze è: $\mu_{Y|x=FI} = \frac{(90 \cdot 2) + (120 \cdot 3)}{5} = 108$

Le medie condizionate sono diverse tra loro e quindi diverse dalla media generale

Non vi è indipendenza in media!

Scomposizione della devianza

Partendo da una distribuzione mista, è possibile scomporre la variabilità complessiva della variabile Y (carattere quantitativo) rispetto alle modalità della variabile X (mutabile).

Questo importante risultato prende il nome di **scomposizione della devianza**.

La devianza totale della Y è scomposta in due quantità:

- **Devianza Interna ai gruppi –** *Devianza Within*
- **Devianza Tra i gruppi** – *Devianza Between*

Tale che

Dev(Tot)=Dev(W)+Dev(B)

Devianza e sua scomposizione

$$Dev(Y) = \sum_i \sum_j (y_j - \mu_Y)^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \mu_{Y|x_i} + \mu_{Y|x_i} - \mu_Y)^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \mu_{Y|x_i})^2 n_{ij} + \sum_i \sum_j (\mu_{Y|x_i} - \mu_Y)^2 n_{ij} +$$

$$+ 2\sum_i \sum_j (y_j - \mu_{Y|x_i})(y_j - \mu_{Y|x_i}) n_{ij}$$

Sviluppando il quadrato si dimostra che il doppio prodotto è nullo!

Scomposizione della devianza

Devianza interna ed esterna

In questo modo è possibile spiegare la variabilità complessiva del carattere quantitativo attraverso le due componenti:

- **Devianza Within** è pari alla somma delle devianze delle singole distribuzioni condizionate della Y ottenute dalle modalità della X.
  Essa rappresenta la parte di variabilità di Y che non dipende dagli attributi assunti dalla X.

- **Devianza Between** è pari alla devianza delle medie condizionate rispetto alla media generale della Y.
  Essa rappresenta la parte della variabilità di Y che dipende, è generata, dalle modalità assunte dalla variabile X.

Devianza e sua scomposizione

$$Dev(Y) = \sum_i \sum_j \left(y_j - \mu_{Y|x_i}\right)^2 n_{ij} + \sum_i \left(\mu_{Y|x_i} - \mu_Y\right)^2 \sum_j n_{ij}$$

$$= \sum_i \left(Dev(Y \mid X = x_i)\right) + \sum_i \left(\mu_{Y|x_i} - \mu_Y\right)^2 n_{i+}$$

$$Dev(Y) = Dev(W) + Dev(B)$$

Devianza interna (Within) — Devianza esterna (between)

Componenti della devianza

Interpretazione della scomposizione

La **devianza "Between"** descrive la variabilità "tra" i gruppi, ossia la variabilità delle medie parziali di Y rispetto alla media generale.
La **devianza "Within"** descrive la variabilità "interna" ai gruppi, ossia la somma delle variabilità della Y in ciascun gruppo.

Quanto più i gruppi sono ben discriminati tanto maggiore è la componente di variabilità esterna rispetto a quella interna. Ciò implica che la variabile X "spiega" il comportamento della Y.

La variabile X è detta di stratificazione in quanto dalle sue modalità si determinano gli strati o gruppi parziali del collettivo.

Rapporto di correlazione di Pearson

Il **rapporto di correlazione η^2 di Pearson** descrive quanta parte della devianza totale è spiegata dalla variabilità delle medie parziali rispetto alla media generale.

Esso rappresenta una misura normalizzata della devianza between in quanto è ottenuto come rapporto tra questa e la devianza totale (che ne costituisce il massimo).

Il rapporto è pari a 0 quando c'è indipendenza in media ed è pari a 1 in assenza di variabilità interna ai gruppi cioè nel caso di massima dipendenza in media.

Rapporto di correlazione del Pearson

$$\eta^2_{Y/X} = \frac{Dev(B)}{Dev(Y)} = 1 - \frac{Dev(W)}{Dev(Y)} \qquad 0 \le \eta^2_{Y/X} \le 1$$

$\eta^2_{Y/X} \neq \eta^2_{X/Y}$ → Il rapporto di correlazione è un indice NON SIMMETRICO

Rapporto di Correlazione di Pearson

Esempio di calcolo del Rapporto di Correlazione

… continuando l'esempio visto in precedenza

La **Devianza totale** è pari a: $Dev(Y) = (90-102)^2 \cdot 12 + (120-102)^2 \cdot 8 = 1728 + 2592 = 4320$

La Devianza interna per la città di **Roma** è: $Dev(Y \mid x = RO) = (90-105)^2 * 3 + (120-105)^2 * 3 = 1350$

La Devianza interna per la città di **Napoli** è: $Dev(Y \mid x = NA) = (90-96{,}67)^2 * 7 + (120-96{,}67)^2 * 2 = 1400$

La Devianza interna per la città di **Firenze** è: $Dev(Y \mid x = FI) = (90-108)^2 * 2 + (120-108)^2 * 3 = 1080$

La **Devianza interna ai gruppi** è pari a:

$$Dev(Within) = Dev(Y \mid x = RO) + Dev(Y \mid x = NA) + Dev(Y \mid x = FI) = 3830$$

La **Devianza tra i gruppi** è pari a:

$$Dev(Between) = (105-102)^2 \cdot 6 + (96{,}67-102)^2 \cdot 9 + (108-102)^2 \cdot 5 = 490$$

La Devianza Totale è pari alla somma della Devianza *Between* e della Devianza *Within*

$$Dev(T) = Dev(B) + Dev(W) = 490 + 3830 = 4320$$

Il **rapporto di correlazione** è: $\eta^2_{Y|X} = \frac{dev(B)}{dev(T)} = \frac{490}{4320} = 0{,}1134$ Esiste una scarsa dipendenza in media!

Nella prossima lezione

Nella prossima lezione si affronteranno i seguenti argomenti:

- incorrelazione
- misure di correlazione lineare