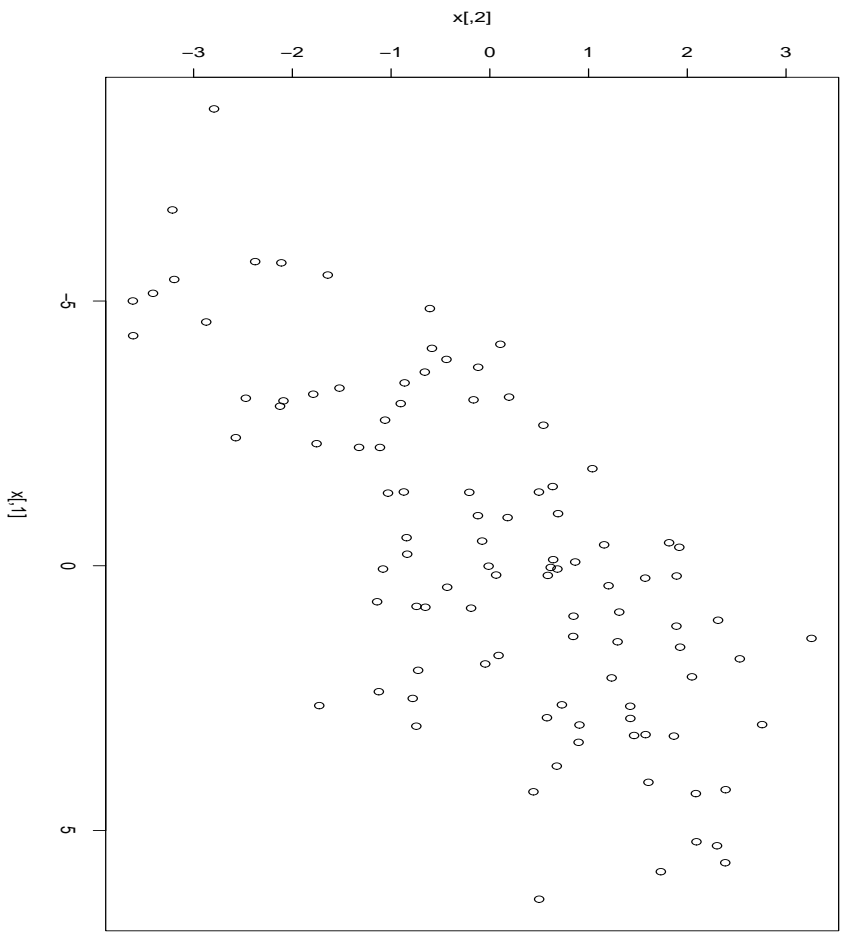
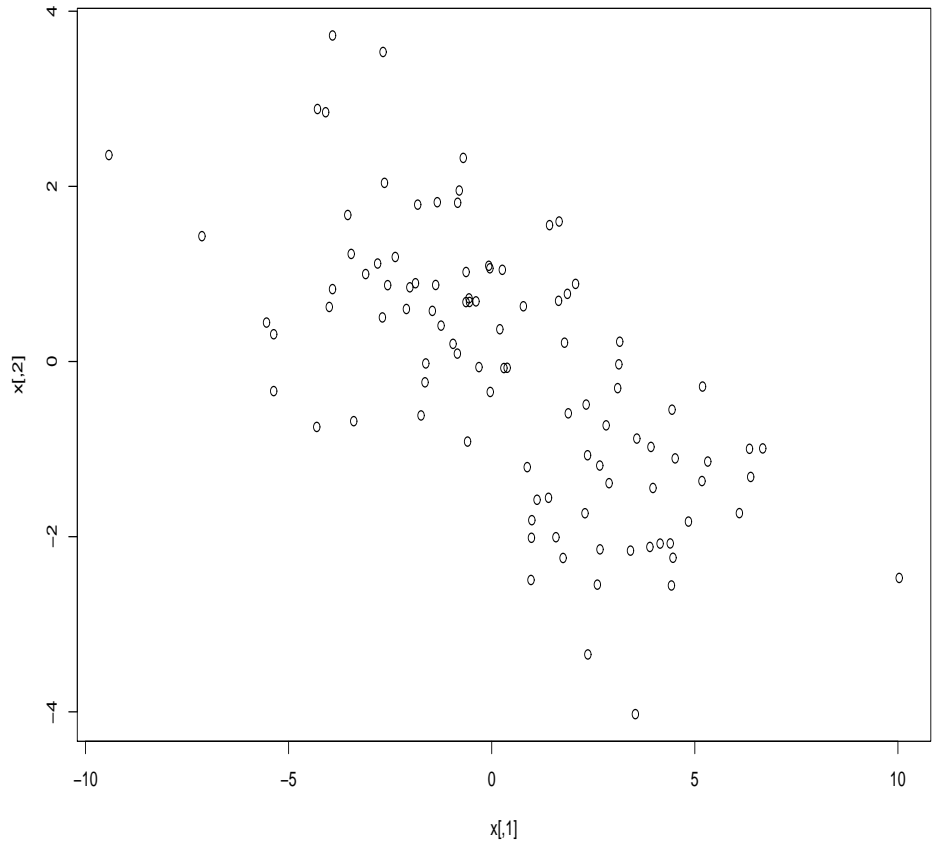


La COVARIANZA

Quando i caratteri sono entrambi QUANTITATIVI oltre all'indice di connessione posso valutare se c'è la tendenza di modalità "elevate" di un carattere ad associarsi con modalità "elevate" dell'altro carattere (**CONCORDANZA**) o viceversa (**DISCORDANZA**)

Una valutazione di questo tipo può essere fatta in prima istanza osservando i grafici (diagramma a bolle o diagramma di dispersione):





INDICE ASSOLUTO DI CONCORDANZA

Per una distribuzione doppia **unitaria**

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)$$

per una distribuzione di frequenza usando le **frequenze assolute**

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^h (x_i - \mu_X)(y_j - \mu_Y) n(x_i, y_j)$$

usando le **frequenze relative**

$$\text{Cov}(X, Y) = \sum_{j=1}^k \sum_{i=1}^h (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j)$$

ESEMPIO

X = PESO

Y=ALTEZZA

media di X = $(50+55+60+57+65)/5 = 57.4$

media di Y = $(1.6+1.7+1.68+1.65+1.75)/5 = 1.676$

$$\begin{aligned} cov(x, y) &= \frac{1}{5}[(50 - 57.4)(1.6 - 1.676) + \\ &\quad + (55 - 57.4)(1.7 - 1.676) + \\ &\quad + (60 - 57.4)(1.68 - 1.676) + \\ &\quad + (57 - 57.4)(1.65 - 1.676) \\ &\quad + (65 - 57.4)(1.75 - 1.676)] = 0.272 \end{aligned}$$

ESEMPIO

X = NUM DIPENDENTI

Y = FATTURATO MENSILE (in migliaia di Euro)

X	1	1	2	2
Y	2	3	4	1

$$E(X) = 6/4 = 1.5$$

$$E(Y) = 10/4 = 2.5$$

$$\text{Cov}(X, Y) = (0.25 - 0.25 + 2.25 - 2.25)/4 = 0$$

i due caratteri sono correlativamente indipendenti.

Se per l'ultima impresa il fatturato fosse stato di 2000E mensili (invece che di 1000E) avremmo che la covarianza sarebbe stata positiva, indicando che i due caratteri sono ora concordanti.

Attenzione però : se cambia l'ultimo valore di Y cambia anche la media di Y che diventa 2.75 quindi devo rifare tutti i conti. In particolare avremmo che:

$$\text{Cov}(X,Y) = 0.1666$$

Se per l'ultima impresa il fatturato fosse stato di 3000E mensili la covarianza resta positiva e sale a:

$$\text{Cov}(X,Y) = 0.3333$$

Ma quale è il valore massimo che la covarianza può assumere?

FORMMULA SEMPLIFICATA per il CALCOLO della COVARIANZA

Si dimostra che:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

dove, per una **distribuzione unitaria**:

$$E(XY) = \frac{1}{n} \sum_{i=1}^n X_i Y_i$$

per una distribuzione di frequenza usando le **frequenze congiunte relative**:

$$E(XY) = \sum_{i=1}^h \sum_{j=1}^k X_i Y_j f(x_i, y_j)$$

o, equivalentemente, usando le **frequenze assolute congiunte**,

$$E(XY) = \frac{1}{n} \sum_{i=1}^h \sum_{j=1}^k X_i Y_j n(x_i, y_j)$$

Supponiamo di aver osservato il numero di figli (X) e il numero di auto (Y)

X ha assunto valori 0,1,0,1,1,2,2,2,3,1

Y ha assunto valori 1,0,1,1,2,1,1,2,2,1

abbiamo che la media di X = $E(X) = 1.3$

abbiamo che la media di Y = $E(Y) = 1.2$

la media di XY = $E(XY) = 1.8$

quindi la covarianza è

$$C(X,Y) = E(XY) - E(X) E(Y) = \\ = 1.8 - 1.2 * 1.3 = 0.24$$

in alternativa possiamo calcolare la covarianza nel modo seguente:

$$\begin{aligned} C(X,Y) = & \\ & (0-1.3)*(1-1.2)*0.2 \\ & + (1 - 1.3)*(0 -1.2)*0.1 \\ & + (1 - 1.3)*(1 -1.2)*0.2 \\ & + (1- 1.3)*(2 -1.2)*0.1 \\ & + (2 - 1.3)*(1 -1.2)*0.2 \\ & + (2 - 1.3)*(2 -1.2)*0.1 \\ & +(3 - 1.3)*(2 -1.2)*0.1 = 0.24 \end{aligned}$$

La covarianza è un indice assoluto quindi possiamo interpretarne solo il segno e non il valore.

Una covarianza **positiva**
indica **concordanza**

Una covarianza **negativa**
indica **discordanza**

Un valore della covarianza **nullo**
indica **indipendenza correlativa**

L'indipendenza correlativa è
un **concetto simmetrico**:

$$Cov(X, Y) = 0 \Leftrightarrow Cov(Y, X) = 0$$

se X è correlativamente indipendente da Y anche Y è correlativamente indipendente da X

INDICE RELATIVO DI CONCORDANZA

per passare da un indice assoluto ad un indice relativo lo si divide per il valore massimo che l'indice assoluto può assumere.

Si dimostra che:

$$-\sigma_X \times \sigma_Y \leq \text{Cov}(X, Y) \leq \sigma_X \times \sigma_Y$$

quindi, un indice relativo di concordanza è il **coefficiente di correlazione lineare**:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Notiamo che $\rho(X, Y) = \rho(Y, X)$.

Risulta inoltre che

$$-1 \leq \rho(X, Y) \leq 1$$

Se $\rho = 1$ si dice che fra i due caratteri esiste perfetta relazione lineare **positiva**

Se $\rho = -1$ si dice che fra i due caratteri esiste perfetta relazione lineare **negativa**

Se $\rho = 0$ si dice che i due caratteri sono **correlativamente indipendenti**

Se $\rho \neq \pm 1$ allora fra i due caratteri non esiste una perfetta relazione lineare ma può esistere altra **relazione funzionale** (non lineare ma per esempio quadratica)

L'indipendenza statistica implica l'indipendenza correlativa (non è vero il contrario)

Quindi:

$$\boxed{\chi^2 = 0 \Leftrightarrow \tilde{\chi}^2 = 0} \rightarrow \boxed{Cov(X, Y) = 0 \Leftrightarrow \rho(X, Y) = 0}$$

Se però c'è perfetta relazione lineare (positiva o negativa) cioè se $\rho = \pm 1$ allora c'è anche massima connessione e l'indice relativo basato sulle contingenze $\tilde{\chi}^2 = 1$

Mentre l'assenza di una relazione lineare, ($\rho \neq \pm 1$), non implica l'assenza di altri tipi di relazione. Due caratteri con covarianza nulla possono essere comunque connessi ($\chi^2 \neq 0$)

L'analisi della correlazione (solo x caratteri quantitativi) è un arricchimento rispetto all'analisi della connessione.

Se ho due caratteri di cui uno continuo ed uno discreto tipicamente ad ogni modalità del carattere continuo corrisponde un'unica modalità del carattere discreto quindi potrebbe sembrare che il carattere che presenta un numero di modalità minore (discreto) sia funzione del carattere con un numero di modalità maggiore (continuo) e quindi $\tilde{\chi}^2 = 1$ quindi bisogna stare attenti ad interpretare bene un valore di $\tilde{\chi}^2$ uguale ad uno

ESEMPIO: Per la seguente tabella a doppia entrata calcolate la covarianza:

X/Y	1	2
0	0.3	0.1
3	0	0.3
5	0.2	0.1

RISPOSTA: per calcolare la covarianza si può procedere in due modi diversi

Prima procedura.

La prima procedura parte direttamente dalla definizione di covarianza ossia:

$$\text{Cov}(X, Y) = \sum_{j=1}^k \sum_{i=1}^h (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j)$$

dove h è il numero di modalità che il carattere X ha assunto, nel nostro caso $h = 3$ e k è il numero di modalità che il carattere Y ha assunto, nel nostro caso $k = 2$.

Dopo aver determinato le due distribuzioni marginali di X ed Y possiamo calcolare la media di X che è pari a

$$E(X) = \mu_X = 3 * 0.3 + 5 * 0.3 = 2.4$$

e la media di Y è pari a

$$E(Y) = \mu_Y = 1 * 0.5 + 2 * 0.5 = 1.5.$$

Abbiamo quindi che la covarianza è pari a

$$\begin{aligned} \text{Cov}(X, Y) &= (0 - 2.4) * (1 - 1.5) * 0.3 + \\ &+ (0 - 2.4) * (2 - 1.5) * 0.1 + \\ &+ (3 - 2.4) * (1 - 1.5) * 0 + \\ &+ (3 - 2.4) * (2 - 1.5) * 0.3 + \\ &+ (5 - 2.4) * (1 - 1.5) * 0.2 + \\ &+ (5 - 2.4) * (2 - 1.5) * 0.1 = 0.2 \end{aligned}$$

Seconda procedura.

La seconda procedura sfrutta il fatto che:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Procediamo prima al calcolo del momento primo misto:

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j f(x_i, y_j) \\ &= 3 * 2 * 0.3 + 5 * 0.2 + 10 * 0.1 = 3.8 \end{aligned}$$

abbiamo quindi che la covarianza è pari a

$$\text{Cov}(X, Y) = 3.8 - 1.5 * 2.4 = 0.2$$

Il valore della covarianza risulta **positivo** e questo indica che i due caratteri sono correlati positivamente cioè , in media, a valori di X più grandi della media corrispondono valori di Y più grandi della media.

Calcoliamo ora il coeff. di correlazione lineare $\rho(X, Y)$

Devo trovare la varianza di X e di Y

$$V(X) = E(X^2) - E(X)^2$$

$$V(X) = E(X^2) - 2.4^2$$

$$E(X^2) = 0^2 * 0.4 + 3^2 * 0.3 + 5^2 * 0.3 = 10.2$$

$$V(X) = 10.2 - 2.4^2 = 4.44$$

$$\sigma(X) = \sqrt{4.44} = 2.1$$

$$V(Y) = E(Y^2) - E(Y)^2$$

$$V(Y) = E(Y^2) - 1.5^2$$

$$E(Y^2) = 1^2 * 0.5 + 2^2 * 0.5 = 2.5$$

$$V(Y) = 2.5 - 1.5^2 = 0.25$$

$$\sigma(Y) = \sqrt{0.25} = 0.5$$

Abbiamo quindi che:

$$\rho(X, Y) = \frac{0.2}{\sqrt{4.44 * 0.25}} = \frac{0.2}{2.1 * 0.5} = 0.18$$

valore vicino a zero quindi il legame è debole