Statistica descrittiva bivariata: correlazione, regressione, associazione

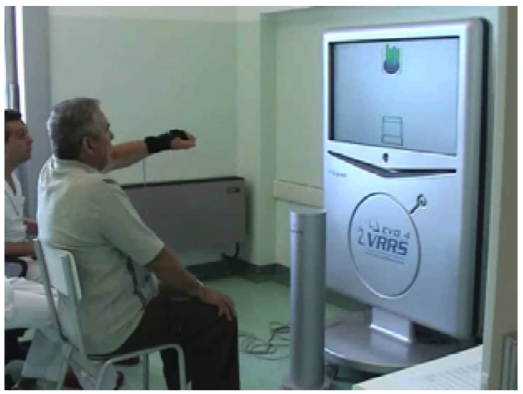
Laura Ventura
Dipartimento di Scienze Statistiche
Università degli Studi di Padova
ventura@stat.unipd.it

Materiale a cura di Laura Ventura e Alessandra Salvan
Cagliari, Novembre 2019

Ripartiamo dal Caso di Studio

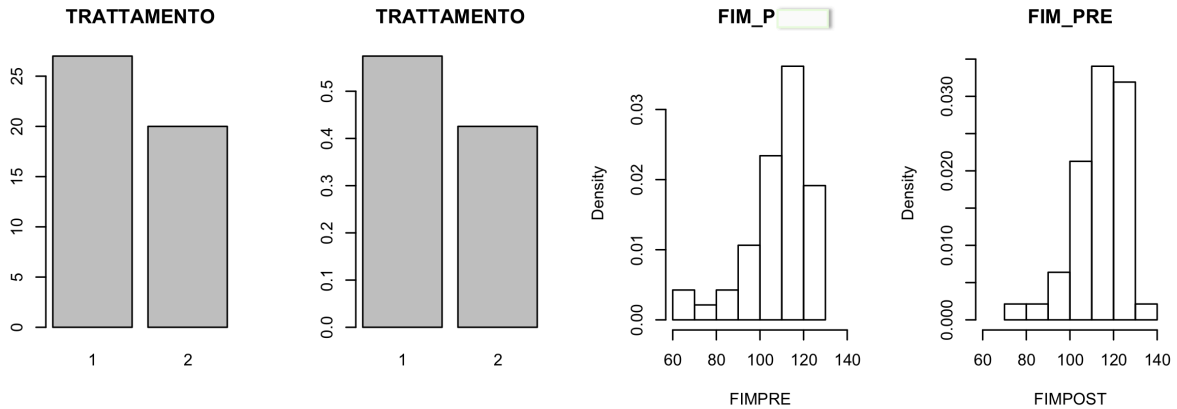
Caso di studio (i dati):
Terapie di riabilitazione per l'apprendimento motorio del braccio

- **Dataset**: misurazioni relative ad uno studio sull'apprendimento motorio di un gruppo di pazienti, esposti al trattamento con realtà virtuale (IRCCS San Camillo, Lido di Venezia).
- **Variabile di interesse**: `FIM` (*Functional Indipendence Measure*), scala dell'autonomia del paziente con valori da 0 (non autosufficienza completa) a 130 (completa autonomia).
- Si hanno anche **due trattamenti**: 27 pazienti sono stati sottoposti ad una terapia di riabilitazione in un ambiente virtuale (casi, `TRATTAMENTO=1`) e 20 pazienti sono stati sottoposti ad una terapia convenzionale (controlli, `TRATTAMENTO=2`).
- La variabile `FIM` è stata misurata sia prima (`FIMPRE`) che dopo (`FIMPOST`) la terapia ricevuta, subito dopo un infarto.

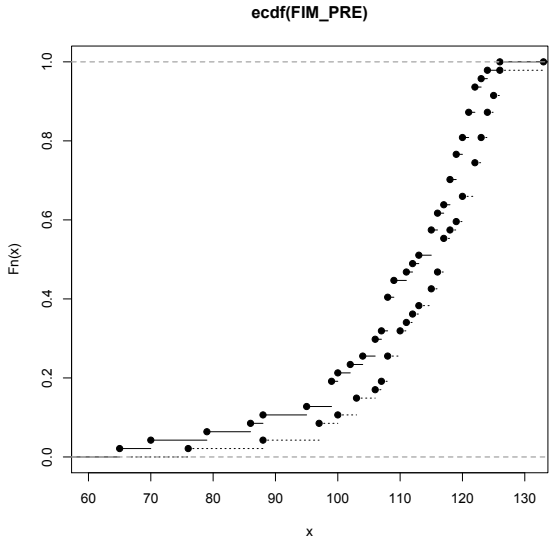


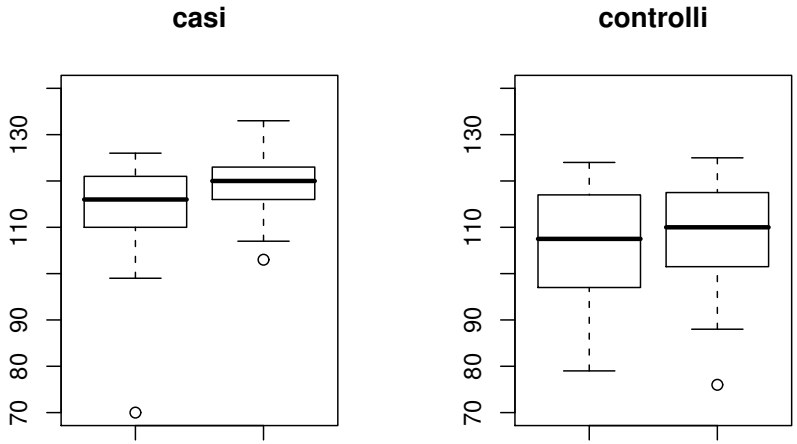
Riassunto della "lezione" precedente: Analisi esplorativa

- TRATTAMENTO (variabile qualitativa): i casi sono 27 (57.4%) e i controlli sono 20 (42.6%).
- FIM (variabile quantitativa): La media di FIMPRE è di 109.3 (sd=13.8) e la media di FIMPOST è di 114.6 (sd=10.9).
  La media di FIMPRE è di 113.3 per i casi (sd=11.4) e 103.95 per i controlli (sd=15.14).
  La media di FIMPOST è 118.9 per i casi (6.81) e 108.65 per i controlli (12.6).
  La mediana di FIMPRE è di 116 per i casi e 107.5 per i controlli.
  La mediana di FIMPOST è 120 per i casi e 110 per i controlli.



Riassunto della "lezione" precedente: altri grafici utili





Un passo avanti:
analisi esplorativa bivariata

Dai dati univariati ai dati bivariati

- In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità. Esempi:
  – *"Le misurazioni della FIM prima della terapia sono in relazione con le misurazioni dopo la terapia?"*
  – o *"il voto di maturità è in relazione con la performance universitaria?"*
- Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un'altra. Esempi:
  – *"conoscendo il valore della FIMPRE, si può stimare il valore della FIMPOST?"*
  – *"conoscendo l'età del paziente, è possibile prevedere il miglioramento nella FIM?"*

- La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame. A tale scopo, per variabili quantitative, si tratteranno:
  – La CORRELAZIONE, che misura la dipendenza lineare tra due variabili;
  – La REGRESSIONE, che valuta la relazione lineare tra due variabili.

Correlazione

- La correlazione misura l'associazione tra due variabili quantitative. È lo strumento che si utilizza quando si hanno a disposizione coppie di valori di variabili ⇒ permette di valutare come variano i valori di una variabile al variare dell'altra e viceversa.
- Esempi:
  – Numero di sigarette fumate in gravidanza e tasso di crescita del feto ⇒ all'aumentare del numero di sigarette fumate diminuisce il tasso di crescita (correlazione negativa).
  – Livello di colesterolo e BMI (Body Mass Index = peso (kg)/altezza^2 (m^2)) ⇒ tanto è maggiore il BMI quanto è maggiore il livello di colesterolo (correlazione positiva).
  – Il valor medio della temperatura (ambiente) e il BMI ⇒ non c'è motivo di pensare che la temperatura influenzi il BMI delle persone (assenza di correlazione).
- La relazione può essere valutata tramite:
  – Un grafico (grafico di dispersione)
  – Un indice che quantifica il grado di correlazione (coefficiente di correlazione)

Diagramma di dispersione

- Nello studio dell'associazione tra due variabili quantitative misurate sulle stesse unità statistiche, indicate con X e Y, è molto utile disegnare un grafico, il diagramma di dispersione, prima di procedere con altre analisi formali.

Nel grafico di dispersione le coppie

$$(x_1, y_1)\ (x_2, y_2)\ \ldots\ (x_n, y_n)$$

di valori di due variabili quantitative misurate sulle n unità sono rappresentati come punti di un piano cartesiano, i cui assi corrispondono alle due variabili.

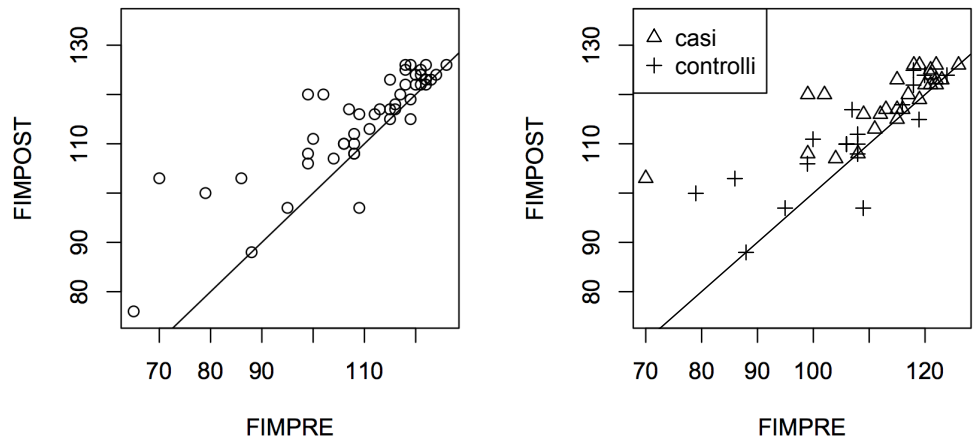


Diagramma di dispersione

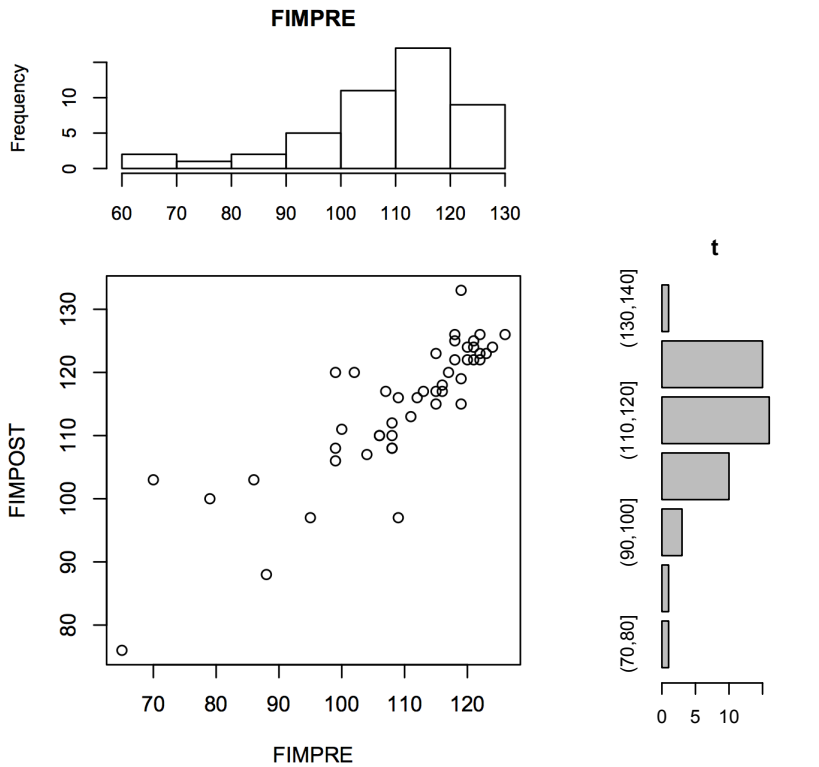
– Ogni punto del grafico rappresenta una unità.

– Permette di verificare visivamente se le coppie di punti presentano una qualche forma di regolarità e per vedere come i punti si disperdono intorno a un particolare punto di riferimento: il baricentro della nuvola dei punti, ossia il punto di coordinate (m_x, m_y).

– La nuvola di punti ha una forma allungata verso l'alto $\Rightarrow$ a modalità crescenti della X corrispondono più frequentemente modalità crescenti della Y.

– Si possono considerare convenzioni grafiche per punti ripetuti.

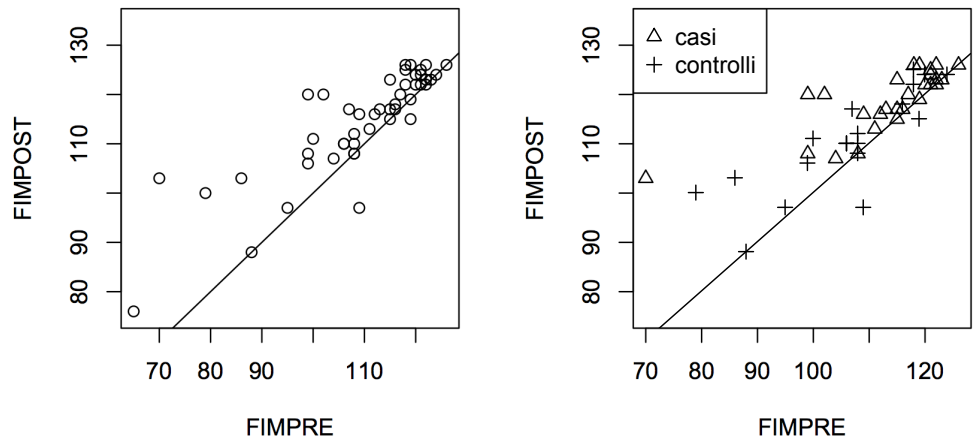
P.s. La media aritmetica e la varianza di X *sono* $m_x = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *e* $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_x^2$. *Analogamente, si indicano con* m_y *e* S_y^2 *media e varianza di* Y.



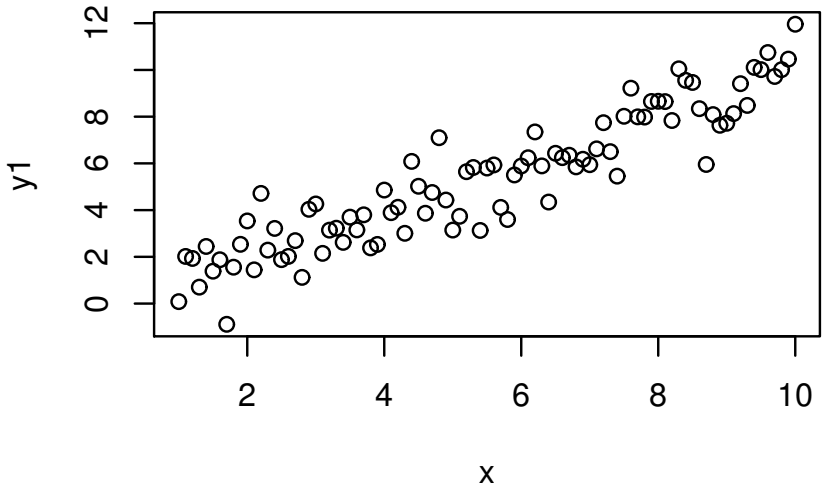
Il ruolo delle variabili X e Y è simmetrico?

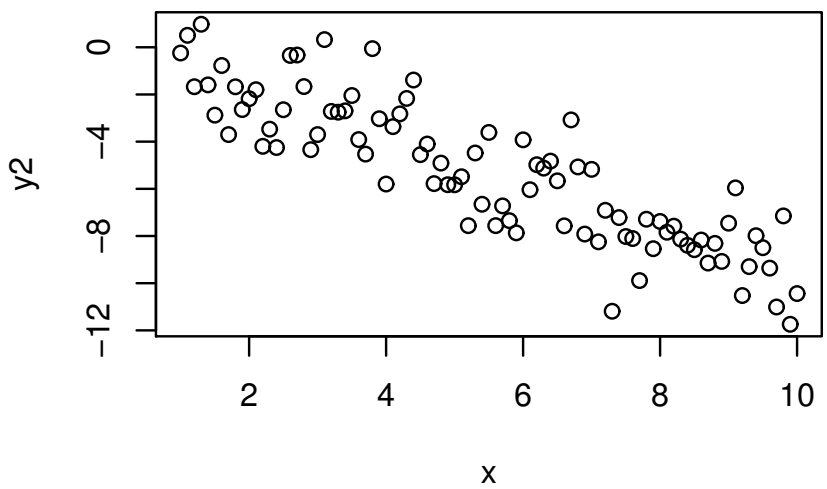
- A volte può essere importante spiegare una delle due variabili in funzione dell'altra. Si avrà quindi una VARIABILE ESPLICATIVA X e una VARIABILE RISPOSTA Y.
- Ma a volte non ha importanza quale sia l'una e quale sia l'altra.

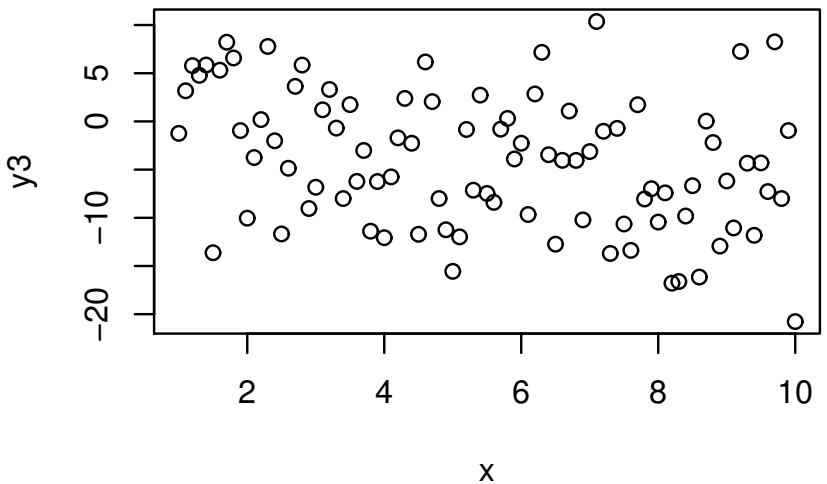
Nell'ESEMPIO della FIM è ragionevole voler esprimere la FIMPOST (Y) a partire dalla FIMPRE (X), misurabile a inizio trattamento. Dal grafico di dispersione si vede che, in generale, nei pazienti con FIMPRE elevata anche la FIMPOST è elevata $\Rightarrow$ correlazione positiva.

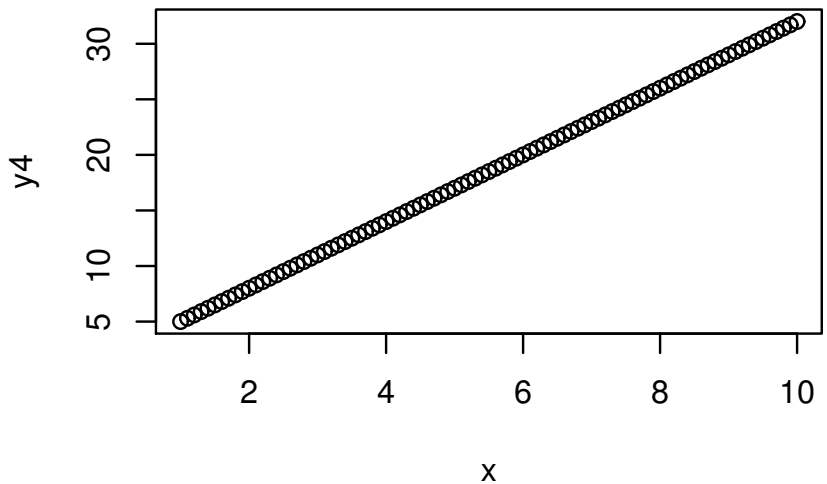


. . . qualche situazione tipo . . . guess the correlation!



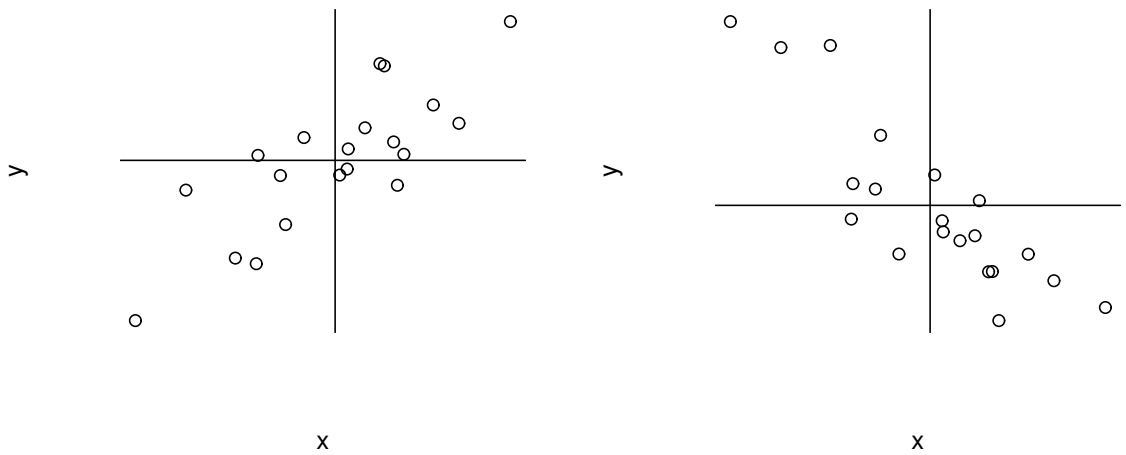






La covarianza

- Per avere una valutazione analitica del grado di associazione tra due variabili quantitative, esiste un indice che misura la dispersione nel piano dei punti dal proprio centro: la COVARIANZA.
- Il nome lascia intuire che si tratta di un'estensione al caso di due variabili della varianza. La covarianza si basa infatti sugli scarti delle x_i dalla propria media, $(x_i - m_x)$, e delle y_i dalla propria media, $(y_i - m_y)$.
- La covarianza, a differenza della varianza che è sempre positiva, misura l'eventuale direzione del legame, ovvero se le due variabili si muovono nella stessa direzione o in direzioni opposte. Il segno della covarianza riflette il senso crescente o decrescente dell'allineamento tendenziale.



La covarianza

- La covarianza segnala una concordanza (sia che X e Y decrescono o crescono) con un segno $+$ e una discordanza (quando X cresce e Y decresce, o viceversa) con il segno -. Formalmente, l'indice è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) \ .$$

- Una formula alternativa per il calcolo della covarianza è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_x m_y$$

- Si noti che $S_{xx} = S_x^2$, ossia la covarianza tra X e X coincide con la varianza di X.

Campo di variazione della covarianza

La covarianza può assumere valori sia positivi sia negativi. In particolare, vale

$$-S_x S_y \leq S_{xy} \leq S_x S_y$$

Dimostrazione.
La varianza della combinazione $aX - bY$ (Appendice), per a e b costanti, è $a^2 S_x^2 + b^2 S_y^2 - 2abS_{xy}$.

Si consideri ora la variabile T definita come $T = S_y^2 X - S_{xy} Y$. Allora, la variabile T ha varianza

$$\begin{aligned} S_T^2 &= S_y^4 S_x^2 + S_{xy}^2 S_y^2 - 2S_y^2 S_{xy} S_{xy} \\ &= S_y^4 S_x^2 - S_{xy}^2 S_y^2 \end{aligned}$$

Ma poiché vale $S_T^2 \geq 0$, deve valere la diseguaglianza

$$S_y^4 S_x^2 - S_{xy}^2 S_y^2 \geq 0$$

ossia, dividendo per S_y^2,

$$S_{xy}^2 \leq S_y^2 S_x^2$$

da cui segue la tesi.

La correlazione

Il coefficiente di correlazione

- Dalla proprietà $-S_x S_y \leq S_{xy} \leq S_x S_y$, può essere costruito un indice relativo semplicemente dividendo S_{xy} per il prodotto degli scarti quadratici medi di X e Y. L'indice così ottenuto prende valori in [-1,1] e viene detto coefficiente di correlazione:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad -1 \leq r_{xy} \leq 1$$

- La formula del coefficiente di correlazione non è poi così terribile come appare!! Può solo essere noioso calcolarla a mano. In genere si usa un software opportuno.
- Un modo di procedere può essere il seguente:

  – Per le due variabili si calcolano le medie $m_x = \frac{1}{n}\sum x_i$ e $m_y = \frac{1}{n}\sum y_i$
  – Si calcola la media dei prodotti $\frac{1}{n}\sum x_i y_i$
  – Si calcolano le medie dei quadrati $\frac{1}{n}\sum x_i^2$ e $\frac{1}{n}\sum y_i^2$
  – Si calcola la covarianza $S_{xy} = \frac{1}{n}\sum x_i y_i - m_x m_y$
  – Si calcolano $S_x = [\frac{1}{n}\sum x_i^2 - m_x^2]^{1/2}$ e $S_y = [\frac{1}{n}\sum y_i^2 - m_y^2]^{1/2}$
  – Queste sono le grandezze che servono per calcolare r_{xy}

- In sintesi: come si interpreta il valore trovato di r_{xy}?

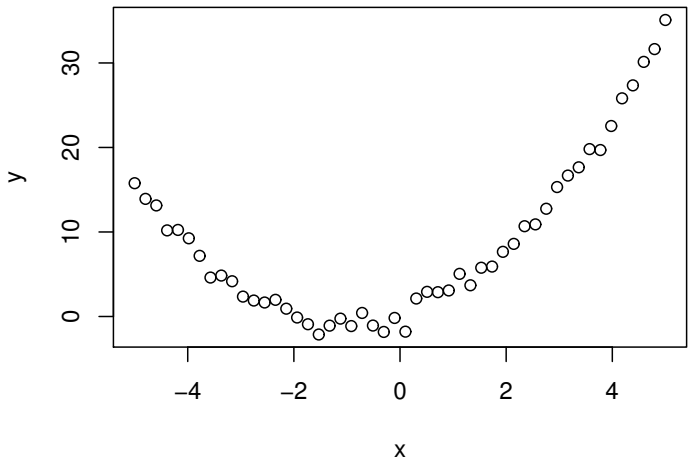
Guida all'interpretazione di r_{xy}

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = +1$: correlazione positiva perfetta (tutti i punti su una retta: concordi)
- $r_{xy} = -1$: correlazione negativa perfetta (tutti i punti su una retta: discordi)
- $r_{xy} > 0$: correlazione positiva
- $r_{xy} < 0$: correlazione negativa
- $r_{xy} \cong 0$: assenza di relazione lineare

Se $r_{xy} = \pm 1$ le variabili sono legate da una relazione lineare perfetta (diretta o inversa, rispettivamente). Si parla di relazione lineare in quanto r_{xy} misura se le coppie di valori (x_i, y_i) sono allineate lungo una retta del tipo $y = a + bx$.

Quando tra X e Y non vi è una relazione lineare o essa è estremamente debole, il valore dell'indice r_{xy} è zero o circa zero, e le variabili sono dette incorrelate.

ATTENZIONE: Il coefficiente di correlazione misura una associazione lineare. Il valore $r_{xy} = 0$ non indica tuttavia un'assenza di relazione tra le due variabili. Può esserci una relazione curvilinea.



Esempio: r_{xy} per la FIM

- Siano $Y =$ FIMPOST e $X =$ FIMPRE.
- Si ha

  $m_x = 109.3$
  $m_y = 114.6$
  $\sum(x_i - m_x)^2 = 8732.2$
  $\sum(y_i - m_y)^2 = 5433.6$
  $\sum(x_i - m_x)(y_i - m_y) = 5808.7$

- Allora:

$$r_{xy} = \frac{5808.7}{\sqrt{8732.2 \times 5433.6}} = 0.843$$

- Il valore 0.843 indica una correlazione positiva elevata tra la FIMPRE e la FIMPOST (come ci si aspettava dal grafico di dispersione).
- Con una relazione così, la FIMPOST potrebbe essere prevista in modo accurato conoscendo il valore della FIMPRE.

La regressione

La regressione

- Quando dall'analisi di un diagramma di dispersione emerge un particolare andamento della nuvola di punti di X e Y, è naturale chiedersi se esiste una qualche relazione statistica $Y = f(X)+$errore tra X e Y.
- Il problema è lo stesso di prima: si vuole studiare una relazione tra le variabili. La relazione non è più simmetrica!! Perchè si vuole comprendere come la variabile risposta Y sia influenzata dalla variabile esplicativa X.
- Se la relazione che emerge è di tipo lineare, si può esprimere la relazione statistica tra X e Y usando un modello molto semplice: l'equazione della retta.

Il modello è del tipo:

$$Y = a + bX + \text{errore}$$

con

$a =$ intercetta
$b =$ coefficiente angolare
errore $=$ la deviazione dalla retta dei punti osservati

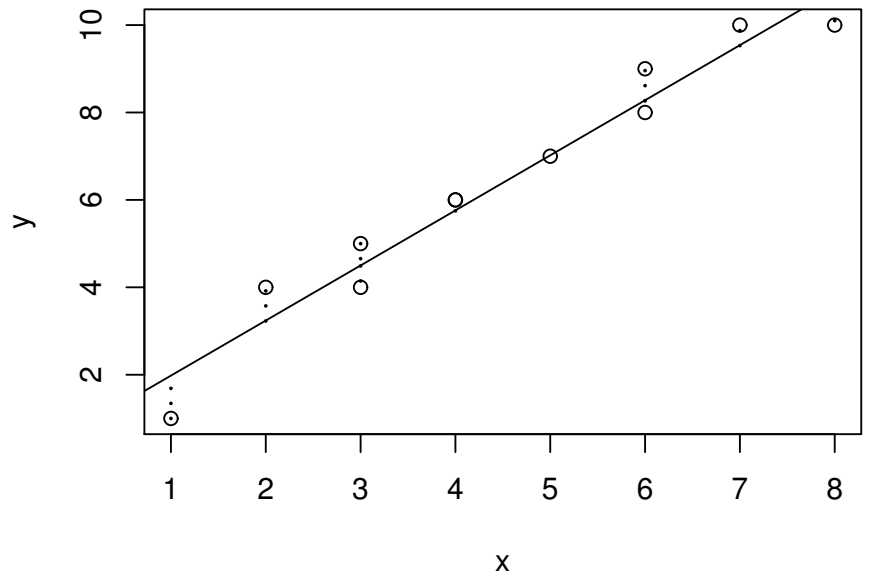
La regressione

- Se si calcolano "opportunamente" i valori di a e b, l'equazione può essere usata per prevedere il valore della Y a partire da un qualunque valore della X.
- PROBLEMA: come trovare la retta che si adatta nel modo migliore ai dati?
- Si devono determinare i valori di a e b che rendono la retta la più "vicina" possibile alle coppie osservate (x_i, y_i): la retta interpolante, cioè quella che passa tra i punti lasciando da essa scarti complessivamente minimi.
- I punti che stanno sulla retta sono le coppie di punti $(x_i, \hat{y}_i) = (x_i, a + bx_i)$, con $\hat{y}_i$ valori teorici o previsti, cioè i valori che la variabile Y dovrebbe assumere per $X = x_i$ se la relazione tra X e Y fosse esattamente quella ipotizzata $Y = a + bX$.
- r_{xy} misura quanto bene i dati sono allineati lungo tale retta. Come regola empirica, valori da 0.80 a 1 (o da -1 a -0.80) rivelano una accettabile relazione lineare di tipo diretto (o inverso). Ricordiamo che quando $r_{xy} = 0$ non è escluso che X e Y possono essere legate da altre relazioni, come $Y = \cos(X) + \exp(X^3)$, o altre "mostruosità" del genere.

Minimi quadrati

- Come cerchiamo la retta interpolante? Si noti che le quantità $e_i = y_i - \hat{y}_i$ misurano la distanza o scarto tra i valori di Y osservati e quelli teorici. In particolare, prendiamo la distanza quadratica, data da $(y_i - \hat{y}_i)^2$. Ne consegue che la distanza totale tra i valori osservati e teorici è

$$d(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \ .$$



La retta dei minimi quadrati

- La somma dei quadrati $d(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ dipende dalle incognite a e b, mentre y_i e x_i sono numeri osservati.
- La retta interpolante è quella i cui valori di a e di b che rendono minima $d(a,b)$, che viene detta retta dei minimi quadrati.

Si mostra che i valori a e b che minimizzano $d(a,b)$ sono dati da

$$\hat{b} = \frac{S_{xy}}{S_x^2} \qquad \hat{a} = m_y - \hat{b}\, m_x$$

- I calcoli richiesti sono gli stessi che servono per determinare il coefficiente di correlazione . . . non serve molto lavoro in più.
- Sia r_{xy} sia $\hat{b}$ dipendono al numeratore dalla covarianza S_{xy}. Essendo le quantità al denominatore sempre positive, è evidente che i segni di r_{xy} e di $\hat{b}$ sono concordi con il segno di S_{xy}.

Dimostrazione

Posto $y_i^* = y_i - bx_i$, $i = 1, \ldots, n$, la somma dei quadrati $d(a, b)$ può essere riscritta come $\sum_{i=1}^n (y_i^* - a)^2$. Quindi, per la proprietà dei minimi quadrati della media aritmetica, la quantità $\sum_{i=1}^n (y_i^* - a)^2$ è minima per

$$\hat{a} = \frac{1}{n}\sum_{i=1}^n y_i^* = \frac{1}{n}\sum_{i=1}^n (y_i - bx_i) = \frac{1}{n}\sum_{i=1}^n y_i - b\,\frac{1}{n}\sum_{i=1}^n x_i = m_y - b\,m_x \ .$$

Sostituendo tale valore in $d(a, b)$ si ottiene

$$\begin{aligned}
\sum_{i=1}^n (y_i - m_y - bx_i + bm_x)^2 &= \sum_{i=1}^n \left[(y_i - m_y) - b(x_i - m_x)\right]^2 \\
&= \sum_{i=1}^n (y_i - m_y)^2 + b^2 \sum_{i=1}^n (x_i - m_x)^2 - 2b \sum_{i=1}^n (y_i - m_y)(x_i - m_y) \\
&= nb^2 S_x^2 - 2nbS_{xy} + nS_y^2
\end{aligned}$$

Come funzione di b, si tratta di una funzione quadratica, il cui grafico è una parabola con concavità rivolta verso l'alto. Il minimo si ha in corrispondenza del vertice, ossia per

$$\hat{b} = \frac{-(-2nS_{xy})}{2nS_x^2} = \frac{S_{xy}}{S_x^2}$$

Esempio: FIM

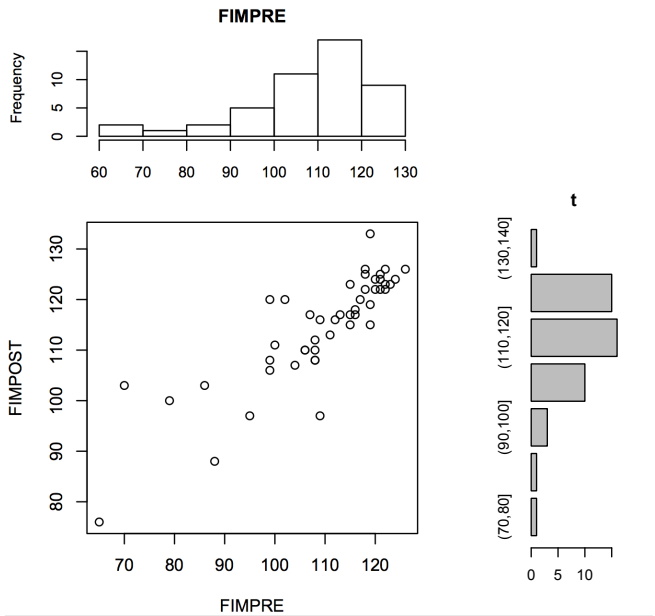
Nell'esempio dalla FIMPOST (Y) e FIMPRE (X) si trovano i seguenti valori di $\hat{a}$ e $\hat{b}$:

$$\hat{b} = 5808.702/8732.213 = 0.67 \qquad \hat{a} = 114.6 - 0.67 \times 109.3 = 41.37$$

La retta di regressione per questi dati è:

$$\hat{Y} = 41.37 + 0.67\, X = 41.37 + 0.67\, \text{FIMPRE}$$

Abbiamo il risultato: ma come interpretarlo e usarlo? La retta è UTILE per fare previsioni sulla variabile risposta. Ad esempio per $X = 90$, si trova $Y = 41.37 + 0.67 \times 90 = 101.67$.



Bontà dell'adattamento della retta ai dati

- Come possiamo valutare se la retta si adatta bene ai dati? Abbiamo bisogno di un indice capace di riassumere l'adattamento globale e la capacità esplicativa complessiva del modello in rapporto ai dati osservati.
- Si può utilizzare ancora il coefficiente di correlazione r_{xy}. E poiché non ha importanza se la correlazione è positiva o negativa, si eleva r_{xy} al quadrato $\Rightarrow$ COEFFICIENTE DI DETERMINAZIONE:

$$R^2 = r_{xy}^2$$

NOTA:

Se $R^2 = 1$: adattamento perfetto (tutti i punti sulla retta)
Se $R^2 = 0$: la retta non ha nulla da vedere con i dati
Se $R^2 = 0.8$: "buon livello" di adattamento

- ESEMPIO: $r_{xy} = 0.84^2 \Rightarrow R^2 = 0.71$, ossia la retta di regressione spiega discretamente la relazione.

Interpretazione di R^2 come proporzione di varianza spiegata

- Siano $\hat{y}_i = \hat{a} + \hat{b}x_i$, $i = 1, \ldots, n$, i valori calcolati sulla retta dei minimi quadrati.
- La somma dei residui $y_i - \hat{y}_i$ vale zero.
  Infatti, $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i) = \sum_{i=1}^n (y_i - m_y + \hat{b}m_x - \hat{b}x_i) =$
  $\sum_{i=1}^n (y_i - m_y) - \hat{b}\sum_{i=1}^n (x_i - m_x) = 0$ (proprietà di baricentro).
- Inoltre, $\sum_{i=1}^n (y_i - \hat{y}_i)x_i = \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - m_x) =$
  $\sum_{i=1}^n (y_i - m_y + \hat{b}m_x - \hat{b}x_i)(x_i - m_x) = nS_{xy} - \hat{b}nS_x^2 = 0\,.$
- Allora, dall'identità $\sum_{i=1}^n (y_i - m_y)^2 = \sum_{i=1}^n (y_i \pm \hat{y}_i - m_y)^2 =$
  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - m_y)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - m_y)$, usando le due relazioni precedenti, si vede facilmente che l'ultima sommatoria vale zero.
  Dunque $\frac{1}{n}\sum_{i=1}^n (y_i - m_y)^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - m_y)^2$
  ossia

  VARIANZA TOTALE=VARIANZA RESIDUA+VARIANZA SPIEGATA
- Si vede infine che R^2 =VARIANZA SPIEGATA/VARIANZA TOTALE.

  Infatti, $\sum_{i=1}^n (\hat{y}_i - m_y)^2 = \sum_{i=1}^n (m_y - \hat{b}m_x + \hat{b}x_i - m_y)^2 = n\hat{b}^2 S_x^2 = nS_{xy}^2/S_x^2$. E quindi

$$\frac{\sum_{i=1}^n (\hat{y}_i - m_y)^2}{\sum_{i=1}^n (y_i - m_y)^2} = \frac{nS_{xy}^2}{S_x^2 nS_y^2} = R^2\,.$$

Esempio: Tensione, corrente e resistenza

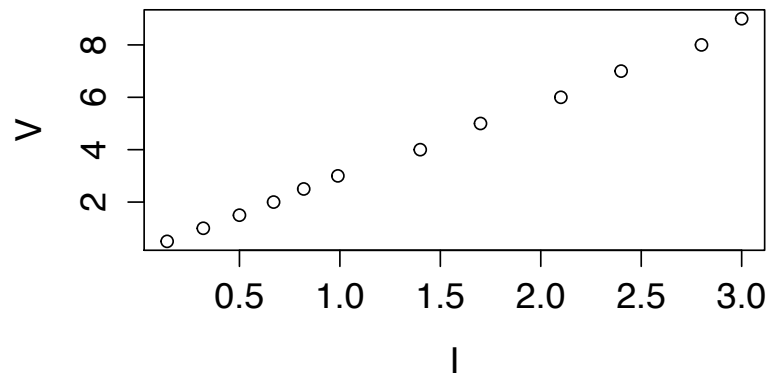
I seguenti dati riportano $n = 12$ misurazioni della tensione (`V`) e della corrente (`I`):

`V = (0.5,1,1.5,2,2.5,3,4,5,6,7,8,9)` in *volt*
`I = (0.14,0.32,0.50,0.67,0.82,0.99,1.4,1.7,2.1,2.4,2.8,3)` in *ampere*

La relazione lineare tra le due variabili è esprimibile come

$$V = a + bI + \text{errore}$$

e ci si attende dal modello teorico $a \doteq 0$ *volt*, $b = Res$ *volt/ampere*, dove Res è una costante di proporzionalità che misura la resistenza, e un valore di R^2 estremamente elevato.



Posto $X = I$ e $Y = V$, si ha:

$m_x = 1.403$ e $m_y = 4.125$

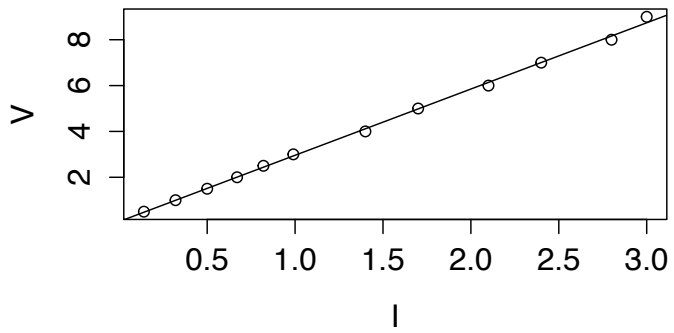
$S_x^2 = 0.892$, $S_y^2 = 7.463$ e $S_{xy} = 2.578$

$\rightarrow \hat{b} = 2.578/0.892 = 2.89$ *volt/ampere* e $\hat{a} = 4.125 - 2.89 \times 1.403 = 0.07$ *volt*.

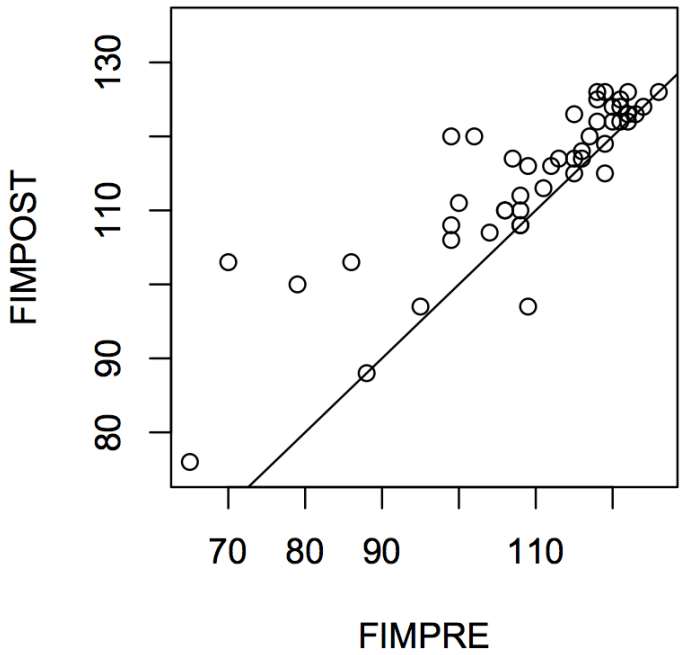
La retta di regressione per questi dati è:

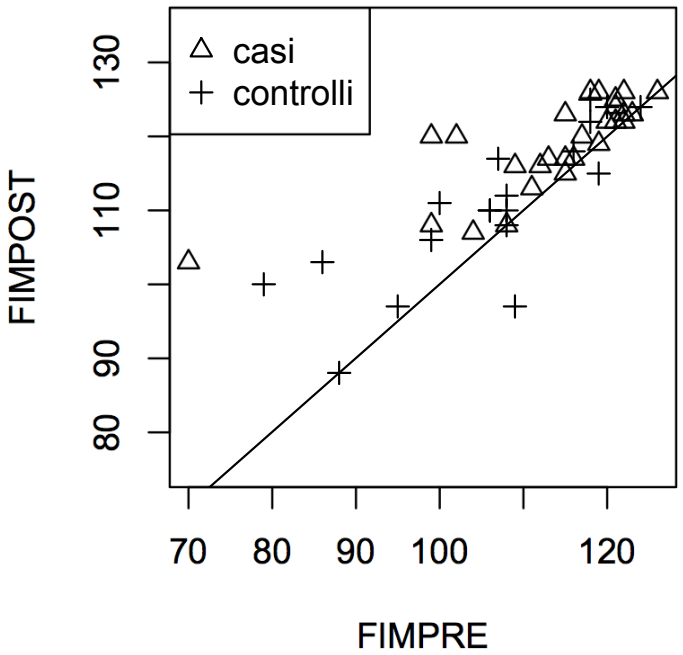
$$\hat{Y} = 0.07 + 2.89\, X$$

Con correlazione $r_{xy} = 0.999$ ($R^2 = 0.9985$), tale modello evidenza una relazione lineare tra le due variabili. Inoltre, $a \doteq 0$ *volt* come ci si aspettava dal modello teorico, mentre $Res = 2.89$ *volt/ampere*.



Esempio: FIM per TRATTAMENTO





- Posto Y_{R} = FIMPOST con realtà virtuale e X_{R} = FIMPRE con realtà virtuale, si ha:
  $m_{x_R} = 113.29$ e $m_{y_R} = 118.93$
  $S^2_{x_R} = 129.75$, $S^2_{y_R} = 46.38$ e $S_{xy_R} = 58.09$

  La retta di regressione per questi dati è:

$$\hat{Y}_{R} = 68.18 + 0.45\, X_{R}$$

  La correlazione è $r_{xy_R} \doteq 0.75$.

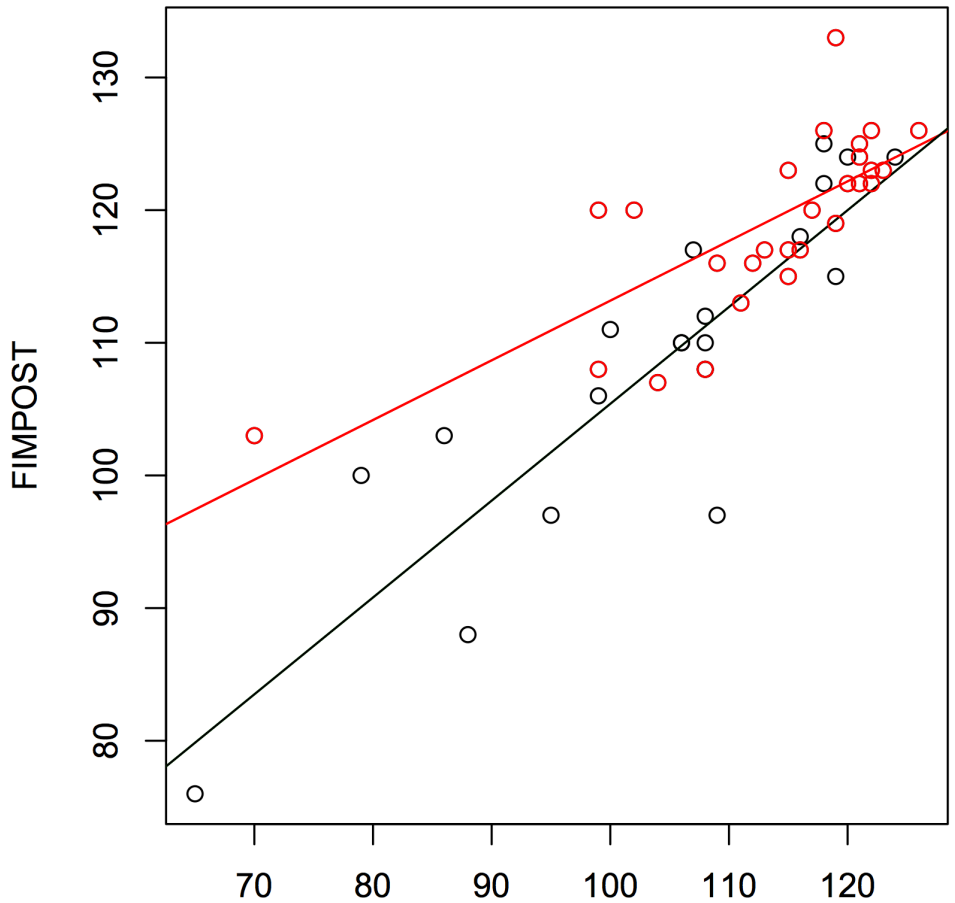
- Posto Y_{F} = FIMPOST con fisioterapia e X_{F} = FIMPRE con fisioterapia, si ha:
  $m_{x_F} = 103.95$ e $m_{y_F} = 108.65$
  $S^2_{x_F} = 229.21$, $S^2_{y_F} = 158.66$ e $S_{xy_F} = 168.14$

  La retta di regressione per questi dati è:

$$\hat{Y}_{F} = 32.40 + 0.73\, X_{F}$$

  La correlazione è $r_{xy_F} \doteq 0.88$.

Esempio: FIM per trattamento



Appendice: proprietà della media e della varianza

Media

- Linearità: $m_{a+bx} = a + bm_x$, con $a, b \in \mathbb{R}$
- Combinazione lineare: $m_{ax+by} = am_x + bm_y$, con $a, b \in \mathbb{R}$

Varianza

- Invarianza rispetto a traslazioni: $S^2_{a+x} = S^2_x$, con $a \in \mathbb{R}$
- Omogeneità (di secondo grado): $S^2_{bx} = b^2 S^2_x$, con $b \in \mathbb{R}$

  $\rightarrow S^2_{a+bx} = b^2 S^2_x$, con $a, b \in \mathbb{R}$
- Combinazione lineare: $S^2_{ax+by} = a^2 S^2_x + b^2 m^2_y + 2abS_{xy}$, con $a, b \in \mathbb{R}$ e $S^2_{ax-by} = a^2 S^2_x + b^2 m^2_y - 2abS_{xy}$, con $a, b \in \mathbb{R}$

Esercizi

(1) La gascromotografia è una tecnica per analizzare miscele di gas. I dati che seguono mostrano la quantità di una certa sostanza (Y) e la corrispondente misura ottenuta da un gascromatografo (X):

quantità	0.25	0.25	0.25	1	1	1	5	5	5	20	20	20
misura	6.55	7.98	6.54	29.7	30	30.1	211	204	212	929	905	922

1) Disegnare il diagramma di dispersione dei dati
2) Calcolare la quantità media di sostanza
3) Calcolare la retta di regressione che permette di prevedere la quantità di sostanza come funzione della misura ottenuta dal gascromatografo
4) Calcolare l'indice di correlazione
5) Per una quantità di sostanza pari a 2, il gascromatografo ha fornito una misura pari a?

(2) La seguente tabella mostra per vari anni il numero di incidenti stradali in una certa regione:

Anno	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Incidenti	5413	6122	6705	6824	7790	7698	8571	8688	9422	9904

1) Si calcoli il numero medio di incidenti in un anno.
2) Si fornire una rappresentazione grafica dei dati opportuna.
3) Si calcoli la retta di regressione che permette di prevedere il numero di incidenti come funzione dell'anno.
4) Si calcoli il coefficiente di correlazione.
5) Si fornisca una previsione per il numero di incidenti per il 2001.

Ancora dati bivariati:

Associazione

Caso di studio: Efficacia del casco protettivo

Nella tabella 2×2 che segue sono riportati i dati che illustrano i risultati di uno studio sull'efficacia dei caschi protettivi per bicicletta nella prevenzione dei traumi cranici (su $n = 793$ soggetti coinvolti in incidenti).

Trauma cranico	Casco SI	Casco NO	totale
SI	17	218	235
NO	130	428	558
totale	147	646	793

- Il modo più comune per rappresentare sinteticamente i dati categoriali sono le Tabelle di contingenza (distribuzioni di frequenza doppie).
- Esse costituiscono l'organizzazione in formato tabulare delle frequenze per variabili qualitative bivariate.
- Le tabelle di contingenza possono essere anche uno strumento idoneo per indagare le relazioni esistenti tra le modalità di due caratteri quantitativi suddivisi in classi.

Rappresentazione generale di una tabella di contingenza

Distribuzione di frequenza doppia per X e Y:

			X			
Y	x_1	$\dots$	x_j	$\dots$	x_J	totale
y_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_I	n_{I1}	$\dots$	n_{Ij}	$\dots$	n_{IJ}	$n_{I.}$
totale	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.J}$	n

- n_{ij} $(i = 1, \dots, I, j = 1, \dots, J)$ distribuzione di frequenza congiunta
- $n_{i.}$ $(i = 1, \dots, I)$ distribuzione di frequenza marginale della Y
- $n_{.j}$ $(j = 1, \dots, J)$ distribuzione di frequenza marginale della X
- $n_{ij}/n_{i.}$ $(j = 1, \dots, J)$ distribuzione di frequenza condizionata della X data $Y = y_i$
- $n_{ij}/n_{.j}$ $(i = 1, \dots, I)$ distribuzione di frequenza condizionata della Y data $X = x_j$

Caso di studio: Efficacia del casco protettivo

Trauma cranico	Casco SI	Casco NO	totale
SI	17	218	235
NO	130	428	558
totale	147	646	793

	X		
Y	x_1	x_2	totale
y_1	n_{11}	n_{12}	$n_{1.}$
y_2	n_{21}	n_{22}	$n_{2.}$
totale	$n_{.1}$	$n_{.2}$	n

PROBLEMA: Per esaminare l'efficacia del casco protettivo si vuole valutare se esiste un'associazione (relazione) tra traumi cranici (Y) ed uso del casco (X) tra i soggetti coinvolti in un incidente.

Date le due variabili categoriali, si vuole valutare se X e Y sono dipendenti.

L'indipendenza

- Nella tabella si possono considerare le distribuzioni condizionate di Y (Trauma Cranico) dato $X = x$ (Uso del Casco), nonché la distribuzione marginale di Y, considerate come distribuzioni di frequenza relativa, in modo da ovviare alle diverse numerosità.

- Una situazione estrema si ha quando le distribuzioni condizionate sono tutte uguali: in tale caso è inutile tenere sotto controllo X per evidenziare una fonte sistematica di variabilità dei valori di Y.

- Nell'esempio si ha:

Trauma cranico	Casco SI	Casco NO	totale
SI	$17/147 = 0.12$	$218/646 = 0.34$	$235/793 = 0.29$
NO	$130/147 = 0.88$	$428/646 = 0.66$	$558/793 = 0.71$
totale	1	1	1

da cui si nota che le distribuzioni non sono somiglianti.

Indipendenza

- Si parla di indipendenza statistica quando la conoscenza della modalità di una delle due variabili in esame non migliora la "previsione" della modalità dell'altra.
- Se le distribuzioni condizionate sono tutte somiglianti, allora Y è indipendente da X.
- Condizione necessaria e sufficiente affinché Y sia indipendente da X è che valga, per ogni $i = 1, \ldots, I$ e $j = 1, \ldots, J$, il seguente risultato.

Se X e Y sono indipendenti, la generica frequenza assoluta corrispondente alla i-esima modalità di X e alla j-esima modalità di Y deve essere uguale a

$$a_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

ossia le frequenze a_{ij} attese (teoriche) in ipotesi di indipendenza sono il prodotto tra totale della riga e totale della colonna diviso per n.

Dimostrazione

In base alla definizione, Y è indipendente da X se, per ogni j, le distribuzioni condizionate di Y dato $X = x_j$ sono tutte uguali, ossia si ha

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{iJ}}{n_{.J}} = p_i^*$$

Ma allora

$$n_{i.} = \sum_{j=1}^{J} n_{ij} = \sum_{j=1}^{J} p_i^* n_{.j} = p_i^* \sum_{j=1}^{J} n_{.j} = n p_i^*$$

Pertanto deve valere l'identità

$$n_{i.} = n p_i^* = n \frac{n_{ij}}{n_{.j}}$$

da cui si ottiene

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

nell'ipotesi di indipendenza.

Indice χ^2 di Pearson

La statistica χ^2 di Pearson è basata sul confronto tra le frequenze osservate e quelle attese in ipotesi di indipendenza.

La formula per il calcolo della statistica χ^2 è

$$\chi^2 = \sum \frac{(\text{osservate - attese})^2}{\text{attese}} = \sum \frac{(n_{ij} - a_{ij})^2}{a_{ij}}$$

Il calcolo è fatto confrontando le frequenze attese e quelle osservate per ogni cella della tabella, e poi i risultati sono sommati.

frequenze osservate

Trauma cranico	Casco SI	Casco NO	totale
SI	17	218	235
NO	130	428	558
totale	147	646	793

frequenze attese

Trauma cranico	Casco SI	Casco NO	totale
SI	43.56	191.44	235
NO	103.44	454.56	558
totale	147	646	793

Si trova $\chi^2 = 27.20$.

Interpretazione

Rimane da interpretare il valore calcolato per la statistica χ^2.

Per renderci conto se il valore trovato è "grande" o "piccolo" potrebbe essere utile sapere che

$$0 \leq \chi^2 \leq \max(\chi^2) = n \min\left((I-1),(J-1)\right)$$

Si ha
– $\chi^2 = 0$ nel caso di indipendenza tra X e Y ($n_{ij} = a_{ij}$)
– $\chi^2 = \max(\chi^2)$ nel caso di dipendenza perfetta tra X e Y (ad ogni modalità di X corrisponde sempre una sola modalità di Y)
– si avvicina sempre più a $\max(\chi^2)$ quanto più forte è il legame tra le due variabili studiate ($n_{ij} - a_{ij}$ grandi e quindi χ^2 grande)

Nell'esempio sull'efficacia del casco protettivo si ha:

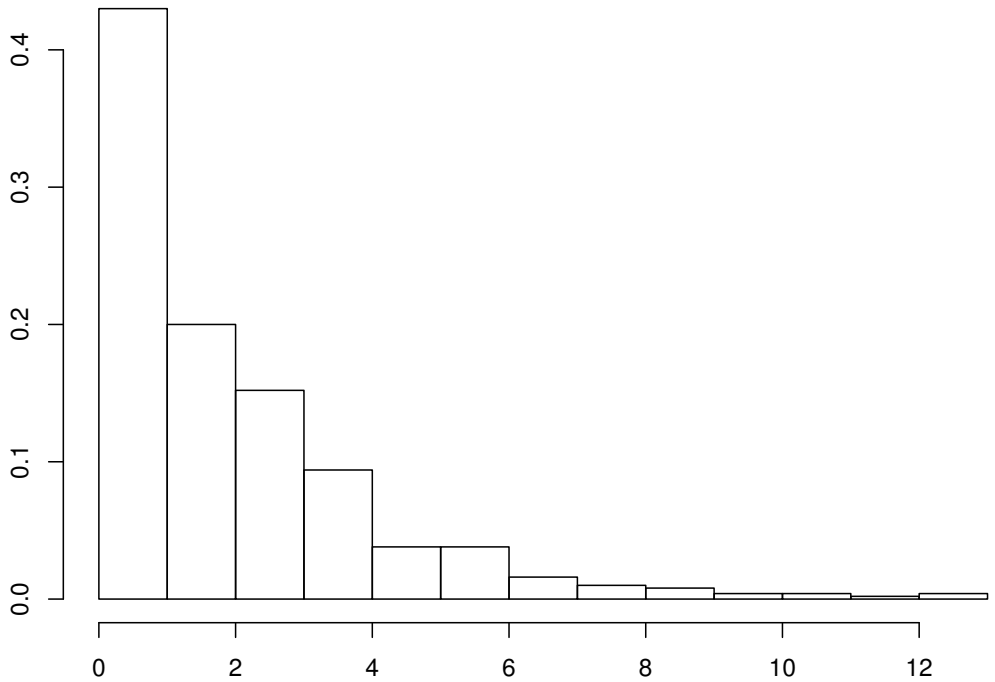
$$\chi^2 = 27.20 \quad n = 793 \quad I = J = 2$$

e $\max(\chi^2) = n \min((r-1),(c-1)) = 793 \min(1,1) = 793$. E quindi?

Interpretazione

- Se X e Y sono indipendenti, ci si aspetta un valore osservato della statistica χ^2 "piccolo".
- Viceversa, se X e Y sono dipendenti, ci si aspetta un valore osservato della statistica χ^2 "grande".
- Per interpretare il valore osservato della statistica χ^2 si può usare un riassunto "probabilistico" dell'evidenza contro l'ipotesi di indipendenza.
- Per capire bene questo serve il Calcolo delle Probabilità. Comunque, intuitivamente, pensiamo che la tabella sia ottenuta effettuando un campionamento casuale da una popolazione in cui c'è effettivamente indipendenza.
- Ipotizzando di ripetere il campionamento casuale molte volte, si calcola la proporzione di tabelle osservate che danno un valore della statistica χ^2 maggiore o uguale a quello osservato nei dati. Un valore piccolo di questa proporzione (p-value) indica che è difficile avere una tabella come quella osservata pescando da una popolazione dove c'è effettivamente indipendenza e dunque indica una evidenza contro l'ipotesi di indipendenza.

Interpretazione



La proporzione (frequenza relativa) di valori maggiori o uguali di 27.2 è praticamente zero.

Esempio

Nella tabella che segue viene mostrata una classificazione di $n = 141$ pesci predati e non predati (X) da parte di uccelli, secondo il livello di infestazione (Y) da parte di particolari vermi (trematodi).

	predati	non predati	totale
non infestati	1	49	50
poco infestati	10	35	45
molto infestati	37	9	46
totale	48	93	141

Essendo le due variabili qualitative, un indice appropriato per lo studio della relazione tra X e Y è la statistica χ^2 di Pearson.

Calcoliamo le frequenze attese nell'ipotesi di indipendenza:

	predati	non predati	totale
non infestati	17	33	50
poco infestati	15.3	29.7	45
molto infestati	15.7	30.3	46
totale	48	93	141

L'indice χ^2 di Pearson è allora:

$$\chi^2 = (1-17)^2/17 + \ldots + (9-30.3)^2/30.3 = 69.5$$

con $\max(\chi^2) = 141 \min((2-1),(3-1)) = 141$ e p-value$\doteq 0$.

Il valore trovato indica che i dati a disposizione evidenziano relazione tra pesci predati e non predati da parte di uccelli e il livello di infestazione da parte di trematodi.

Approfondiamo una particolare tabella:
I Test Diagnostici

Caso di studio: Test di screening

- $n = 1000$ soggetti, di cui è nota la presenza/assenza di una particolare patologia, sono stati sottoposti a un nuovo test di screening (poco costoso e poco invasivo). Lo scopo dello screening è diagnosticare precocemente la malattia, quando è ancora curabile.
- Si vogliono studiare le proprietà diagnostiche del test di screening.

	Paziente Malato	Paziente Sano	Totale
Test Positivo	291	7	298
Test Negativo	9	693	702
Totale	300	700	1000

Test diagnostico

- Una delle ragioni principali per effettuare misurazioni cliniche è fornire uno strumento di supporto alle diagnosi.
- La misurazione clinica fornisce un test diagnostico, che consente di classificare i soggetti in due gruppi:
  - gruppo dei pazienti sani (M^-)
  - gruppo dei pazienti malati (M^+)
- Il test è positivo (T^+) se segnala la presenza della malattia ed è negativo (T^-) se non la segnala.

Ma in che modo la statistica è utile al test diagnostico?

La matrice di confusione

- Il test diagnostico purtroppo non è infallibile.
- In genere con le misurazioni del test diagnostico nei due gruppi (pazienti M^+ e pazienti M^-) si possono ottenere:
  - dei pazienti malati correttamente classificati come positivi (TP = *True Positive* o Veri Positivi)
  - dei pazienti malati classificati come negativi (FN = *False Negative* o Falsi Negativi)
  - dei pazienti sani correttamente classificati come negativi (TN = *True Negative* o Veri Negativi)
  - dei pazienti sani classificati come positivi (FP = *False Positive* o Falsi Positivi)

ERRORE DI I TIPO
(FALSO POSITIVO)

ERRORE DI II TIPO
(FALSO NEGATIVO)

La matrice di confusione

- I quattro valori TP, FN, TN e FP possono essere rappresentati in una tabella a doppia entrata (chiamata matrice di confusione o tabella di errata classificazione) che conta il numero di casi classificati correttamente o meno:

	Paziente Malato (M^+)	Paziente Sano (M^-)	Totale
Test Positivo (T^+) Test Negativo (T^-)	TP (Veri Positivi) FN (Falsi Negativi)	FP (Falsi Positivi) TN (Veri Negativi)	TP + FP FN + TN
Totale	TP + FN	FP + TN	

L'accuratezza del test diagnostico

	Paziente Malato (M^+)	Paziente Sano (M^-)	Totale
Test Positivo (T^+) Test Negativo (T^-)	TP (Veri Positivi) FN (Falsi Negativi)	FP (Falsi Positivi) TN (Veri Negativi)	TP + FP FN + TN
Totale	TP + FN	FP + TN	

- La validità del test può essere misurata tramite la corretta classificazione dei pazienti sani e malati.
- L'accuratezza del test è definita come

$$\text{accuratezza} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}$$

- Ma vogliamo tenere conto anche di FP e FN.

Ma cosa chiediamo a un test?

- A partire dalla classificazione, si possono ottenere due importanti indici della qualità del test: la sensibilità e la specificità.
- La sensibilità (*sensitivity*) è definita come

$$\text{sensibilità} = \frac{\text{TP}}{\text{TP+FN}}$$

ed esprime la proporzione di Veri Positivi (TP) rispetto al numero totale di positivi effettivi, ossia di pazienti malati (TP+FN). Un test diagnostico è sensibile al 100% quando tutti i malati risultano positivi.

- La specificità (*specificity*) è definita come

$$\text{specificità} = \frac{\text{TN}}{\text{FP+TN}}$$

e misura la proporzione di Veri Negativi (TN) rispetto al numero totale di negativi effettivi, ossia di pazienti sani (FP+TN). Un test diagnostico è specifico al 100% quando tutti i sani risultano negativi.

Sensibilità e Specificità

- È chiaro che un test diagnostico sensibile e specifico al 100% non lascerebbe dubbi.
- Un test specifico ha alta capacità di classificare i SANI come NEGATIVI al test (basso rischio di Falsi Positivi).
- Un test sensibile ha alta capacità di classificare i MALATI come POSTIVI al test (basso rischio di Falsi Negativi).
- Elevata sensibilità e bassa specificità o viceversa?
- Se la malattia è a grave rischio e richiede un intervento immediato, è preferibile un test molto sensibile.
- Se la malattia ha conseguenze non gravi, è meglio un test molto specifico.

Caso di studio: Test di screening

	Paziente Malato (M^+)	Paziente Sano (M^-)	Totale
Test Positivo (T^+)	291	7	298
Test Negativo (T^-)	9	693	702
Totale	300	700	1000

- Prevalenza $= 300/1000 = 0.30$
- Sensibilità $= 291/300 = 0.97$
- Specificità $= 693/700 = 0.99$

Per concludere

Esempio da prove Invalsi per la classe seconda superiore

Una scuola è costituita da due piani e i 900 alunni che la frequentano sono così distribuiti:

	biennio	triennio	totale
I piano	180	360	540
II piano	140	220	360
totale	320	580	900

Quali fra le seguenti affermazioni è falsa?

(A) Il 40% degli alunni della scuola si trova al II piano. (R. 360/900=0.4)
(B) I 2/3 degli alunni del I piano frequentano il triennio. (R. 360/540=0.667)
(C) Gli alunni del triennio costituiscono il 70% del totale. (R. 580/900=0.64)
(D) Il 20% degli alunni della scuola frequenta il biennio in un'aula del I piano. (R. 180/900=0.2)

Esempio da prove Invalsi per la classe seconda superiore

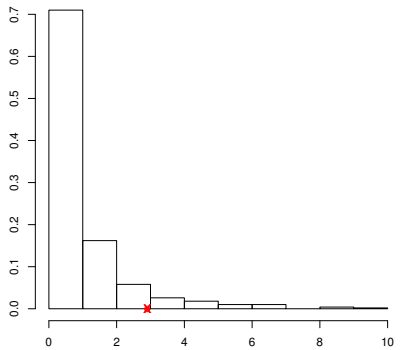
In più ... la tabella nell'ipotesi di indipendenza è:

	biennio	triennio	totale
I piano	192	348	540
II piano	128	232	360
totale	320	580	900

L'indice χ^2 di Pearson è allora:

$$\chi^2 = (180 - 192)^2/192 + \ldots + (220 - 232)^2/232 = 2.91$$

con $\max(\chi^2) = 900 \min((2-1),(2-1)) = 900$ e p-value$= 0.10$. Il valore trovato indica che i dati a disposizione non evidenziano una relazione tra X e Y.



Esercizi

(1) La seguente tabella mostra come 319 studenti universitari si distribuiscono sulla base delle due variabili $X =$ tipo di maturità e $Y =$ numero di esami superati durante il primo anno.

Maturità	Esami superati		
	0-1	2-5	> 5
classica	10	67	31
scientifica	4	52	36
altre	14	65	40

Si calcoli la statistica χ^2 di Pearson.

(2) In un'indagine sulle preferenze alimentari si sono svolte 139 interviste e si è chiesto di indicare la preferenza tra tre alimenti liquidi (caffè-thè-succo) e tre alimenti solidi (biscotto-pane-brioche) da consumare a colazione. La tabella è tuttavia disponibile con alcuni dati mancanti (NA).

liquidi	solidi			
	biscotto	pane	brioche	tot
caffè	45	NA	5	58
thè	NA	NA	31	NA
succo	5	27	6	NA

1) Sapendo che 40 intervistati hanno risposto pane tra gli alimenti solidi, si completi la tabella.
2) Si calcoli la statistica χ^2 di Pearson.