Regressione lineare → rappresentare con una retta due fenomeni X e Y quantitativi osservati contemporaneamente

Retta di regressione che spiega Y in funzione di x

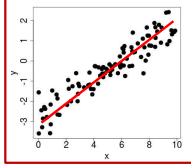
$\hat{Y} = \hat{a} + \hat{b}x$ *dove* $\hat{b} = \sigma_{xy}/\sigma_x^2$ *pendenza* della retta

$\hat{a} = \bar{y} - \hat{b}\bar{x}$ *intercetta* della retta con l'asse Y

Covarianza $\sigma_{xy} = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$

Correlazione lineare $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ *con* $-1 \le \rho_{xy} \le 1$

Correlazione lineare positiva

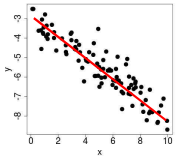


$\hat{b} > 0$

$\sigma_{xy} > 0$

$0 < \rho_{xy} \le 1$

Correlazione lineare negativa



$\hat{b} < 0$

$\sigma_{xy} < 0$

$-1 \le \rho_{xy} < 0$

Correlazione lineare $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ *con* $-1 \le \rho_{xy} \le 1$

$\rho_{xy} \approx -1$ *forte correlazione lineare negativa*

$\rho_{xy} \approx +1$ *forte correlazione lineare positiva*

} *regressione lineare* ***SI***

$\rho_{xy} \approx 0$ *assenza di correlazione lineare*

} *regressione lineare* ***NO***

Misura di ***bontà di adattamento*** *della retta di regressione ai dati*

$R^2 = \rho_{xy}{}^2$ *con* $0 \le R^2 \le 1$

$R^2 \approx +1$ *forte correlazione lineare*
buon adattamento della retta ai dati

} *regressione lineare* ***SI***

$\rho_{xy} \approx 0$ *assenza di correlazione lineare*
scarso adattamento della retta ai dati

} *regressione lineare* ***NO***

Test delle ipotesi di livello α
sul coefficiente b della retta di regressione $Y = a + bx + \varepsilon$

$$H_0: b = 0 \qquad H_1: b \neq 0$$

Il test controlla che il valore di b sia significativamente diverso da zero

Se $b = 0$ significa che Y non dipende da x → in tal caso la retta non è un buon modello per rappresentare i dati

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$

dove la Statistica Test è $t = \hat{b}/\hat{\sigma}_b$

$$\hat{\sigma}_b = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{s^2}{n\sigma_x{}^2}} \qquad s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

QUIZ

Un indice di correlazione lineare tra le variabili X e Y pari a -0.02 indica che

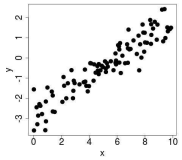
(a) X e Y sono correlate positivamente
(b) X e Y sono molto correlate
(c) X e Y sono molto connesse
(d) X e Y sono incorrelate

$-1 \le \rho_{xy} \le 1$

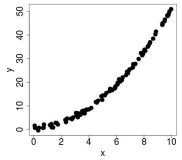
QUIZ

A quale grafico di dispersione corrisponde il coefficiente di correlazione lineare $\rho_{xy} = -0{,}92$?

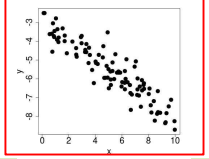
(a)



(b)



(c)



QUIZ

Una misura della bontà di adattamento della retta di regressione ai dati è

(a) la covarianza
(b) l'indice R^2
(c) l'indece χ^2
(d) l'intercetta $\hat{a}$

QUIZ

I valori assunti dal coefficiente di correlazione sono

(a) $-1 \leq \rho_{xy} \leq 1$
(b) $-1 \leq R^2 \leq 1$
(c) $0 \leq \rho_{xy} \leq 1$
(d) $0 \leq R^2 \leq 1$

QUIZ

In un modello di regressione lineare $Y = a + bx + \varepsilon$, la verifica della ipotesi $H_0: b = 0$ contro $H_1: b \neq 0$ fornisce p-value uguale a 0.03
Allora:

(a) si rifiuta H_0 a livello $\alpha = 0.01$
(b) non si rifiuta H_1 a livello $\alpha = 0.01$
(c) non si rifiuta H_0 a livello $\alpha = 0.01$
(d) la pendenza stimata è $\hat{b} = 0.03$

Si rifiuta H_0 se p-value $< \alpha$

QUIZ

Se il coefficiente di correlazione tra le variabili X e Y vale 0, allora

(a) X e Y sono correlate positivamente
(b) X e Y sono incorrelate
(c) X e Y sono indipendenti
(d) X e Y sono correlate negativamente

$-1 \leq \rho_{xy} \leq 1$

QUIZ

Due variabili X e Y sono legate dalla relazione $Y = -2X + 1$.
Allora:

(a) $\rho_{xy} = -2$
(b) $\rho_{xy} = -1$
(c) $R^2 = 0$
(d) $\rho_{xy} = 0$

(a) FALSO perché $-1 \leq \rho_{xy} \leq 1$

Nel testo si dichiara che tra X e Y c'è un perfetta relazione lineare, quindi:

(c)-(d) FALSO perché dovrebbe essere $R^2 = 1$ e $\rho_{xy} = \pm 1$

(b) VERO perché la pendenza $\hat{b} = -2$ indica una correlazione lineare negativa, quindi $\rho_{xy} = -1$

Esercizio 1

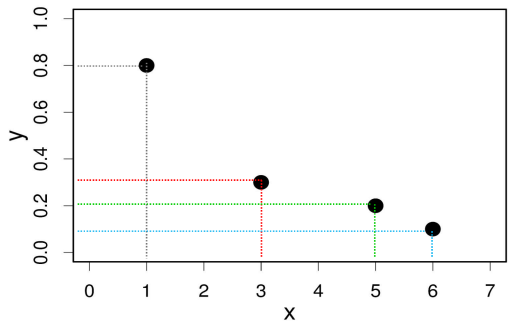
Si considerino le coppie di osservazioni da due variabili casuali X e Y in tabella:

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?
2) Stimare la retta di regressione che spiega Y in funzione di X e disegnarla sul grafico precedente
3) Verificare se la regressione è significativa a livello 10% .
4) Valutare la bontà di adattamento della retta stimata ai dati
5) Se possibile, prevedere il valore di Y dato $x = 2$ e quello di Y dato $x = 20$

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?

x_i	1	3	5	6	← in orizzontale
y_i	0.8	0.3	0.2	0.1	← in verticale



1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1

$n = 4$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+3+5+6}{4} = 3.75$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} =$$

$$= \frac{(1-3.75)^2 + (3-3.75)^2 + (5-3.75)^2 + (6-3.75)^2}{4} = 3.69$$

oppure

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} x_i{}^2}{n} - \bar{x}^2 = \frac{1^2+3^2+5^2+6^2}{4} - 3.75^2 = 3.69$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{3.69} = 1.92$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1

$n = 4$

$$\bar{y} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.8+0.3+0.2+0.1}{4} = 0.35$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} =$$

$$= \frac{(0.8-0.35)^2 + (0.3-0.35)^2 + (0.2-0.35)^2 + (0.1-0.35)^2}{4} = 0.07$$

oppure

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} y_i{}^2}{n} - \bar{y}^2 = \frac{0.8^2+0.3^2+0.2^2+0.1^2}{4} - 0.35^2 = 0.07$$

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{0.07} = 0.26$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1

$n = 4$

$\bar{x} = 3.75$ $\quad \sigma_x^2 = 3.69$ $\quad \sigma_x = 1.92$

$\bar{y} = 0.35$ $\quad \sigma_y^2 = 0.07$ $\quad \sigma_y = 0.26$

$$\sigma_{xy} = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} =$$

$$= \frac{1\times 0.8 + 3\times 0.3 + 5\times 0.2 + 6\times 0.1}{4} - 3.75\times 0.35 = -0.49$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-0.49}{1.92\times 0.26} = -0.98$$

Essendo il coefficiente di correlazione ρ_{xy} vicino a -1, le variabili X e Y mostrano una elevata **correlazione lineare negativa**

2) Stimare la retta di regressione che spiega Y in funzione di X e disegnarla sul grafico precedente

$\bar{x} = 3.75$ $\quad \sigma_x^2 = 3.69$ $\quad \sigma_x = 1.92$ $\quad \sigma_{xy} = -0.49$

$\bar{y} = 0.35$ $\quad \sigma_y^2 = 0.07$ $\quad \sigma_y = 0.26$ $\quad \rho_{xy} = -0.98$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-0.49}{3.69} = -0.13$$

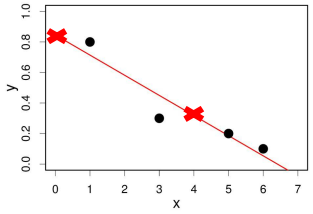
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 0.35 - (-0.13) \times 3.75 = 0.35 + 0.13 \times 3.75 = 0.84$$

Retta di regressione

$$\hat{Y} = 0.84 - 0.13x$$

Per disegnare una retta bastano due punti

- $(0,\ 0.84 - 0.13 \times 0) = (0,\ 0.84)$
- $(4,\ 0.84 - 0.13 \times 4) = (4,\ 0.32)$



3) Verificare se la regressione è significativa a livello 10% .

$\bar{x} = 3.75$ $\quad \sigma_x^2 = 3.69$ $\quad \sigma_x = 1.92$ $\quad \sigma_{xy} = -0.49$

$\bar{y} = 0.35$ $\quad \sigma_y^2 = 0.07$ $\quad \sigma_y = 0.26$ $\quad \rho_{xy} = -0.98$

$\hat{Y} = 0.84 - 0.13x$ $\quad n = 4$

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1

Test delle Ipotesi su b

$H_0: b = 0$ $\quad H_1: b \neq 0$ $\quad \alpha = 0.10$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $t = \hat{b}/\hat{\sigma}_b$

essendo $\hat{\sigma}_b = \sqrt{\frac{s^2}{n\sigma_x^2}}$ e $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

$\sigma_x^2 = 3.69$ $\quad \hat{Y} = 0.84 - 0.13x$ $\quad n = 4$

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1
$\hat{y}_i$	**0.71** $(0.84 - 0.13 \times 1)$	**0.45** $(0.84 - 0.13 \times 3)$	**0.19** $(0.84 - 0.13 \times 5)$	**0.06** $(0.84 - 0.13 \times 6)$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

$$= \frac{(0.8 - 0.71)^2 + (0.3 - 0.45)^2 + (0.2 - 0.19)^2 + (0.1 - 0.06)^2}{4-2} =$$

$$= \frac{0.0081 + 0.0225 + 0.0001 + 0.0016}{4-2} = 0.0161$$

$$\hat{\sigma}_b = \sqrt{\frac{s^2}{n\sigma_x^2}} = \sqrt{\frac{0.0161}{4 \times 3.69}} = 0.03 \qquad t = \frac{\hat{b}}{\hat{\sigma}_b} = \frac{-0.13}{0.03} = -4.33$$

$H_0: b = 0$ $\quad H_1: b \neq 0$ $\quad \alpha = 0.10$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $t = -4.33$

$$t_{1-\alpha/2}^{(n-2)} = t_{1-0.1/2}^{(4-2)} = t_{0.95}^{(2)} = 2.91999 \approx 2.92$$

È vero che $|t| > t_{1-\alpha/2}^{(n-2)}$:

$|-4.33| > 2.92$

→ Si rifiuta H_0 a livello 10% → La regressione è significativa a livello 10 %

4) Valutare la bontà di adattamento della retta stimata ai dati

$\bar{x} = 3.75$ $\sigma_x^2 = 3.69$ $\sigma_x = 1.92$ $\sigma_{xy} = -0.49$

$\bar{y} = 0.35$ $\sigma_y^2 = 0.07$ $\sigma_y = 0.26$ $\rho_{xy} = -0.98$

$\hat{Y} = 0.84 - 0.13x$ $n = 4$

$$R^2 = {\rho_{xy}}^2 = (-0.98)^2 = 0.96$$

Essendo il valore R^2 prossimo a 1, la retta di regressione ben si adatta ai dati

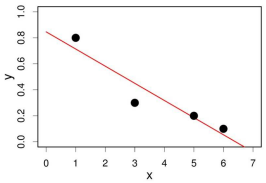
5) Se possibile, prevedere il valore di Y dato $x = 2$ e quello di Y dato $x = 20$

$\bar{x} = 3.75$ $\sigma_x^2 = 3.69$ $\sigma_x = 1.92$ $\sigma_{xy} = -0.49$

$\bar{y} = 0.35$ $\sigma_y^2 = 0.07$ $\sigma_y = 0.26$ $\rho_{xy} = -0.98$

$\hat{Y} = 0.84 - 0.13x$ $n = 4$

x_i	1	3	5	6
y_i	0.8	0.3	0.2	0.1



I dati di X variano da 1 a 6 → posso fare la previsione solo se x è tra 1 e 6

Per $x = 2$ si prevede $\hat{y} = 0.84 - 0.13 \times 2 = 0.58$

Per $x = 20$ non si può fare previsione

Esercizio 2

Si considerino le coppie di osservazioni da due variabili casuali X e Y in tabella:

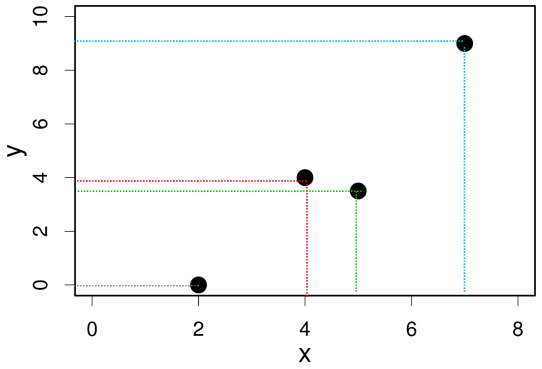
x_i	2	4	5	7
y_i	0	4	3.5	9

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?
2) Si calcoli la retta di regressione di Y su X e la si disegni sul grafico precedente
3) Verificare se la regressione è significativa a livello 5% .
4) Si valuti un opportuno indice di bontà della regressione
5) Se possibile, prevedere il valore di Y dato $x = 2.5$ e quello di Y dato $x = -2$

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?

x_i	2	4	5	7
y_i	0	4	3.5	9

← in orizzontale
← in verticale



1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	2	4	5	7
y_i	0	4	3.5	9

$n = 4$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+4+5+7}{4} = 4.5$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} =$$

$$= \frac{(2-4.5)^2 + (4-4.5)^2 + (5-4.5)^2 + (7-4.5)^2}{4} = 3.25$$

oppure

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} x_i{}^2}{n} - \bar{x}^2 = \frac{2^2+4^2+5^2+7^2}{4} - 4.5^2 = 3.25$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{3.25} = 1.8$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	2	4	5	7
y_i	0	4	3.5	9

$n = 4$

$$\bar{y} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+4+3.5+9}{4} = 4.12$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} =$$

$$= \frac{(0-4.12)^2 + (4-4.12)^2 + (3.5-4.12)^2 + (9-4.12)^2}{4} = 10.3$$

oppure

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} y_i{}^2}{n} - \bar{y}^2 = \frac{0^2+4^2+3.5^2+9^2}{4} - 4.12^2 = 10.3$$

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{10.3} = 3.21$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	2	4	5	7
y_i	0	4	3.5	9

$n = 4$

$\bar{x} = 4.5$ $\quad \sigma_x^2 = 3.25$ $\quad \sigma_x = 1.8$

$\bar{y} = 4.12$ $\quad \sigma_y^2 = 10.3$ $\quad \sigma_y = 3.21$

$$\sigma_{xy} = cov(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} =$$

$$= \frac{2\times 0 + 4\times 4 + 5\times 3.5 + 7\times 9}{4} - 4.5 \times 4.12 = 5.56$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{5.56}{1.8 \times 3.21} = 0.96$$

Essendo il coefficiente di correlazione ρ_{xy} vicino a $+1$, le variabili X e Y mostrano una elevata **correlazione lineare positiva**

2) Si calcoli la retta di regressione di Y su X e la si disegni sul grafico precedente

$\bar{x} = 4.5$ $\quad \sigma_x^2 = 3.25$ $\quad \sigma_x = 1.8$ $\quad \sigma_{xy} = 5.56$

$\bar{y} = 4.12$ $\quad \sigma_y^2 = 10.3$ $\quad \sigma_y = 3.21$ $\quad \rho_{xy} = 0.96$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{5.56}{3.25} = 1.71$$

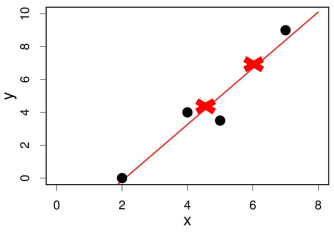
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 4.12 - 1.71 \times 4.5 = -3.58$$

Retta di regressione

$$\hat{Y} = -3.58 + 1.71x$$

Per disegnare una retta bastano due punti

- $(\bar{x},\ \bar{y}) = (4.5,\ 4.12)$
- $(6,\ -3.58 + 1.71 \times 6) = (6,\ 6.68)$



3) Verificare se la regressione è significativa a livello 5% .

$\bar{x} = 4.5$ $\quad \sigma_x^2 = 3.25$ $\quad \sigma_x = 1.8$ $\quad \sigma_{xy} = 5.56$

$\bar{y} = 4.12$ $\quad \sigma_y^2 = 10.3$ $\quad \sigma_y = 3.21$ $\quad \rho_{xy} = 0.96$

$\hat{Y} = -3.58 + 1.71x$ $\quad n = 4$

x_i	2	4	5	7
y_i	0	4	3.5	9

Test delle Ipotesi su b

$H_0: b = 0$ $\quad H_1: b \neq 0$ $\quad \alpha = 0.05$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $t = \hat{b}/\hat{\sigma}_b$

essendo $\hat{\sigma}_b = \sqrt{\frac{s^2}{n\sigma_x{}^2}}$ e $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

$\sigma_x^2 = 3.25$ $\quad \hat{Y} = -3.58 + 1.71x$ $\quad n = 4$

x_i	2	4	5	7
y_i	0	4	3.5	9
$\hat{y}_i$	**-0.16** $(-3.58 + 1.71 \times 2)$	**3.26** $(-3.58 + 1.71 \times 4)$	**4.97** $(-3.58 + 1.71 \times 5)$	**8.39** $(-3.58 + 1.71 \times 7)$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{(\mathbf{0 + 0.16})^2 + (4 - 3.26)^2 + (3.5 - 4.97)^2 + (9 - 8.39)^2}{4-2} =$$

$$= \frac{0.0256 + 0.5476 + 2.1609 + 0.3721}{4-2} = 1.5531$$

$$\hat{\sigma}_b = \sqrt{\frac{s^2}{n\sigma_x{}^2}} = \sqrt{\frac{1.5531}{4 \times 3.25}} = 0.35 \qquad \mathbf{t = \frac{\hat{b}}{\hat{\sigma}_b} = \frac{1.71}{0.35} = 4.89}$$

$H_0: b = 0$ $\quad H_1: b \neq 0$ $\quad \alpha = 0.05$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $\mathbf{t = 4.89}$

$$\mathbf{t_{1-\alpha/2}^{(n-2)} = t_{1-0.05/2}^{(4-2)} = t_{0.975}^{(2)} = 4.30265 \approx 4.30}$$

È vero che $|t| > t_{1-\alpha/2}^{(n-2)}$:

$|4.89| > 4.30$

→ Si rifiuta H_0 a livello 5% → La regressione è significativa a livello 5 %

4) Si valuti un opportuno indice di bontà della regressione

$\bar{x} = 4.5$ $\quad \sigma_x^2 = 3.25$ $\quad \sigma_x = 1.8$ $\quad \sigma_{xy} = 5.56$

$\bar{y} = 4.12$ $\quad \sigma_y^2 = 10.3$ $\quad \sigma_y = 3.21$ $\quad \rho_{xy} = 0.96$

$\hat{Y} = -3.58 + 1.71x$ $\quad n = 4$

$$R^2 = \rho_{xy}{}^2 = (0.96)^2 = 0.92$$

Essendo il valore R^2 prossimo a 1, la retta di regressione ben si adatta ai dati

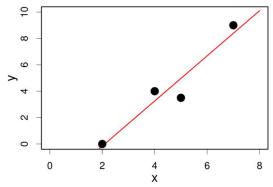
5) Se possibile, prevedere il valore di Y dato $x = 2.5$ e quello di Y dato $x = -2$

$\bar{x} = 4.5$ $\quad \sigma_x^2 = 3.25$ $\quad \sigma_x = 1.8$ $\quad \sigma_{xy} = 5.56$

$\bar{y} = 4.12$ $\quad \sigma_y^2 = 10.3$ $\quad \sigma_y = 3.21$ $\quad \rho_{xy} = 0.96$

$\hat{Y} = -3.58 + 1.71x$ $\quad n = 4$

x_i	2	4	5	7
y_i	0	4	3.5	9



I dati di X variano da 2 a 7 → posso fare la previsione solo se x è tra 2 e 7

Per $x = 2.5$ si prevede $\hat{y} = -3.58 + 1.71 \times 2.5 = 0.69$

Per $x = -2$ non si può fare previsione

Esercizio 3

Si considerino le coppie di osservazioni da due variabili casuali X e Y in tabella:

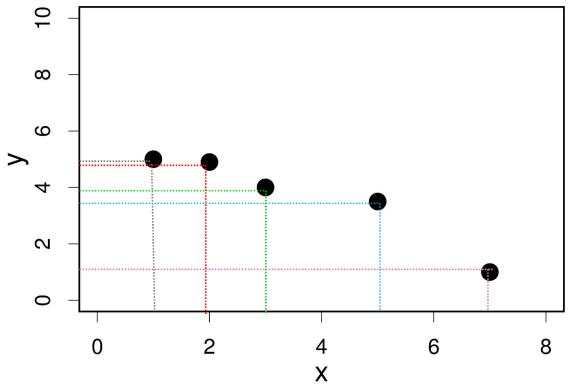
x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?
2) Calcolare la retta di regressione di Y in funzione di X e disegnarla sul grafico precedente
3) Verificare se la regressione è significativa a livello 2% .
4) Valutare la bontà della regressione con un indice opportuno
5) Qualora la regressione fosse risultata significativa, prevedere il valore di Y in $x = 0$ e quello di Y in $x = 4$

1) Rappresentare graficamente i dati: i due fenomeni sono correlati? Come e quanto?

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

← in orizzontale
← in verticale



1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

$n = 5$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+5+7}{5} = 3.6$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} =$$

$$= \frac{(1-3.6)^2 + (2-3.6)^2 + (3-3.6)^2 + (5-3.6)^2 + (7-3.6)^2}{5} = 4.64$$

oppure

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} x_i^{\,2}}{n} - \bar{x}^2 = \frac{1^2 + 2^2 + 3^2 + 5^2 + 7^2}{5} - 3.6^2 = 4.64$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{4.64} = 2.15$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

$n = 5$

$$\bar{y} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5 + 4.9 + 4 + 3.5 + 1}{5} = 3.68$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

$$= \frac{(5 - 3.68)^2 + (4.9 - 3.68)^2 + (4 - 3.68)^2 + (3.5 - 3.68)^2 + (1 - 3.68)^2}{5} = 2.11$$

oppure

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} y_i{}^2}{n} - \bar{y}^2 = \frac{5^2 + 4.9^2 + 4^2 + 3.5^2 + 1^2}{5} - (3.68)^2 = 2.11$$

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{2.11} = 1.45$$

1) Rappresentare graficamente i dati: **i due fenomeni sono correlati? Come e quanto?**

Si deve calcolare il coefficiente di correlazione ρ_{xy}

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

$n = 5$

$\bar{x} = 3.6$ $\quad \sigma_x^2 = 4.64$ $\quad \sigma_x = 2.15$

$\bar{y} = 3.68$ $\quad \sigma_y^2 = 2.11$ $\quad \sigma_y = 1.45$

$$\sigma_{xy} = cov(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} =$$

$$= \frac{1 \times 5 + 2 \times 4.9 + 3 \times 4 + 5 \times 3.5 + 7 \times 1}{5} - 3.6 \times 3.68 = -2.99$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-2.99}{2.15 \times 1.45} = -0.96$$

Essendo il coefficiente di correlazione ρ_{xy} vicino a -1, le variabili X e Y mostrano una elevata **correlazione lineare negativa**

2) Calcolare la retta di regressione di Y in funzione di X e disegnarla sul grafico precedente

$\bar{x} = 3.6$ $\quad \sigma_x^2 = 4.64$ $\quad \sigma_x = 2.15$ $\quad \sigma_{xy} = -2.99$

$\bar{y} = 3.68$ $\quad \sigma_y^2 = 2.11$ $\quad \sigma_y = 1.45$ $\quad \rho_{xy} = -0.96$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-2.99}{4.64} = -0.64$$

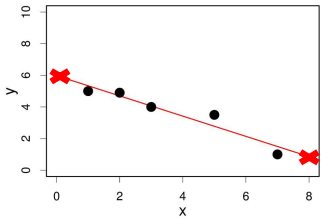
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 3.68 - (-0.64) \times 3.6 = 3.68 + 0.64 \times 3.6 = 5.98$$

Retta di regressione

$$\hat{Y} = 5.98 - 0.64x$$

Per disegnare una retta bastano due punti

- $(0,\ 5.98 - 0.64 \times 0) = (0,\ 5.98)$
- $(8,\ 5.98 - 0.64 \times 8) = (8,\ 0.86)$



3) Verificare se la regressione è significativa a livello 2% .

$\bar{x} = 3.6$ $\quad \sigma_x^2 = 4.64$ $\quad \sigma_x = 2.15$ $\quad \sigma_{xy} = -2.99$

$\bar{y} = 3.68$ $\quad \sigma_y^2 = 2.11$ $\quad \sigma_y = 1.45$ $\quad \rho_{xy} = -0.96$

$\hat{Y} = 5.98 - 0.64x$ $\quad n = 5$

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1

Test delle Ipotesi su *b*

$$H_0: b = 0 \qquad H_1: b \neq 0 \qquad \alpha = 0.02$$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $t = \hat{b}/\hat{\sigma}_b$

essendo $\hat{\sigma}_b = \sqrt{\frac{s^2}{n\sigma_x{}^2}}$ e $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

$\sigma_x^2 = 4.64$ $\quad \hat{Y} = 5.98 - 0.64x$ $\quad n = 5$

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1
$\hat{y}_i$	**5.34** $(5.98 - 0.64 \times 1)$	**4.70** $(5.98 - 0.64 \times 2)$	**4.06** $(5.98 - 0.64 \times 3)$	**2.78** $(5.98 - 0.64 \times 5)$	**1.50** $(5.98 - 0.64 \times 7)$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

$$= \frac{(5-5.34)^2 + (4.9-4.70)^2 + (4-4.06)^2 + (3.5-2.78)^2 + (1-1.50)^2}{5-2} =$$

$$= \frac{0.1156 + 0.0400 + 0.0036 + 0.5184 + 0.2500}{5-2} = 0.3092$$

$$\hat{\sigma}_b = \sqrt{\frac{s^2}{n{\sigma_x}^2}} = \sqrt{\frac{0.3092}{5 \times 4.64}} = 0.12 \qquad \boldsymbol{t = \frac{\hat{b}}{\hat{\sigma}_b} = \frac{-0.64}{0.12} = -5.33}$$

$H_0: b = 0$ $\quad H_1: b \neq 0$ $\quad \alpha = 0.02$

Si rifiuta H_0 se $|t| > t_{1-\alpha/2}^{(n-2)}$ dove la statistica test è $\boldsymbol{t = -5.33}$

$$\boldsymbol{t_{1-\alpha/2}^{(n-2)} = t_{1-0.02/2}^{(5-2)} = t_{0.990}^{(3)} = 4.54070 \approx 4.54}$$

È vero che $|t| > t_{1-\alpha/2}^{(n-2)}$:

$$|-5.33| > 4.54$$

→ Si rifiuta H_0 a livello 2% → La regressione è significativa a livello 2 %

4) Valutare la bontà della regressione con un indice opportuno

$\bar{x} = 3.6$ $\quad \sigma_x^2 = 4.64$ $\quad \sigma_x = 2.15$ $\quad \sigma_{xy} = -2.99$

$\bar{y} = 3.68$ $\quad \sigma_y^2 = 2.11$ $\quad \sigma_y = 1.45$ $\quad \rho_{xy} = -0.96$

$\hat{Y} = 5.98 - 0.64x$ $\quad n = 5$

$$R^2 = {\rho_{xy}}^2 = (-0.96)^2 = 0.92$$

Essendo il valore R^2 prossimo a 1, la retta di regressione ben si adatta ai dati

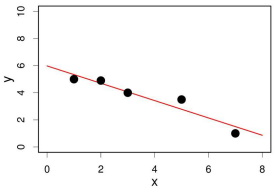
5) Qualora la regressione fosse risultata significativa, prevedere il valore di Y in $x = 0$ e quello di Y in $x = 4$

$\bar{x} = 3.6$ $\quad \sigma_x^2 = 4.64$ $\quad \sigma_x = 2.15$ $\quad \sigma_{xy} = -2.99$

$\bar{y} = 3.68$ $\quad \sigma_y^2 = 2.11$ $\quad \sigma_y = 1.45$ $\quad \rho_{xy} = -0.96$

$\hat{Y} = 5.98 - 0.64x$ $\quad n = 5$

x_i	1	2	3	5	7
y_i	5	4.9	4	3.5	1



I dati di X variano da 1 a 7 → posso fare la previsione solo se x è tra 1 e 7

Per $x = 0$ non si può fare previsione

Per $x = 4$ si prevede $\hat{y} = 5.98 - 0.64 \times 4 = 3.42$

Tabella di contingenza (a doppia entrata)

per rappresentare due fenomeni X e Y osservati contemporaneamente

X \ Y	y_1	y_2	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\cdot}$
⋮	⋮	⋮	…	⋮	⋮
x_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$	n

- r = n. righe
- c = n. colonne
- n_{ij} = n. di volte che la coppia (x_i, y_j) è stata osservata
- n = n. totale di coppie osservate
- $n_{i\cdot}$ = n. marginale riga i
- $n_{\cdot j}$ = n. marginale colonna j

$n^*_{ij} = n_{i\cdot}n_{\cdot j}/n$ = n. di volte che la coppia (x_i, y_j) sarebbe osservata se X e Y fossero indipendenti

Indice di connessione

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Valore minimo = 0 → indipendenza tra X e Y

Valore massimo = $n \cdot \min(r-1, c-1)$ → massima connessione tra X e Y

Indice di connessione

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Valore minimo = 0 → indipendenza tra X e Y

Valore massimo = $n \cdot \min(r-1, c-1)$ → massima connessione tra X e Y

Test chi-quadro di indipedenza

$H_0: \chi^2 = 0$ → indipendenza tra X e Y

$H_1: \chi^2 > 0$ → connessione tra X e Y

Rifiuto H_0 a livello α se $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$

Indice relativo di connessione

$$\tilde{\chi}^2 = \frac{\chi^2}{n \cdot \min(r-1, c-1)}$$

Valore minimo = 0 → indipendenza tra X e Y

Valore massimo = 1 → massima connessione tra X e Y

Esercizio 4

La tabella riporta il numero (in milioni) di contribuenti, suddiviso per fasce di età, che lo scorso anno ha effettuato la dichiarazione dei redditi autonomamente tramite il servizio web dell'Agenzia delle Entrate:

età \ web	SI	NO
<40	0,8	10,3
40 ⊢ 60	1,2	12,5
> 60	0,4	15,6

Verificare a livello 1% se c'è associazione tra le due variabili considerate?

X=«età» è quantitativa in classi e Y=«utilizzo del web» è qualitativa

→ **Test chi-quadro per verificare se c'è indipendenza o connessione tra X e Y**

$H_0: \chi^2 = 0$ (indipendenza) $H_1: \chi^2 > 0$ (connessione)

Rifiuto H_0 a livello $\alpha = 0.01$ se $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$ dove

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} \qquad n^*_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

$H_0: \chi^2 = 0$ (indipendenza) $H_1: \chi^2 > 0$ (connessione)

Rifiuto H_0 a livello $\alpha = 0.01$ se $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$ dove

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} \qquad n^*_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

- N. righe $r = 3$, N. colonne $c = 2$
- Prima calcolo le marginali $n_{i\cdot}$ e $n_{\cdot j}$ (ultima riga/colonna)
- Poi calcolo delle frequenze n^*_{ij} (tra parentesi)

età \ web	SI	NO	
<40	0,8 (11,1 × 2,4/40,8 = 0,65)	10,3 (11,1 × 38,4/40,8 = 10,55)	0,8 + 10,3 = 11,1
40 ⊢ 60	1,2 (13,7 × 2,4/40,8 = 0,81)	12,5 (13,7 × 38,4/40,8 = 12,89)	1,2 + 12,5 = 13,7
> 60	0,4 (16,0 × 2,4/40,8 = 0,94)	15,6 (16,0 × 38,4/40,8 = 15,06)	0,4 + 15,6 = 16,0
	0,8 + 1,2 + 0,4 = 2,4	10,3 + 12,5 + 15,6 = 38,4	40,8

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

$$= \frac{(0,8-0,65)^2}{0,65} + \frac{(1,2-0,81)^2}{0,81} + \frac{(0,4-0,94)^2}{0,94} +$$

$$+ \frac{(10,3-10,45)^2}{10,45} + \frac{(12,5-12,89)^2}{12,89} + \frac{(15,6-15,06)^2}{15,06} = \mathbf{0,57}$$

età \ web	SI	NO	
<40	**0,8** (11,1 × 2,4/40,8 = **0,65**)	**10,3** (11,1 × 38,4/40,8 = **10,45**)	0,8 + 10,3 = **11,1**
40 ⊢ 60	**1,2** (13,7 × 2,4/40,8 = **0,81**)	**12,5** (13,7 × 38,4/40,8 = **12,89**)	1,2 + 12,5 = **13,7**
> 60	**0,4** (16,0 × 2,4/40,8 = **0,94**)	**15,6** (16,0 × 38,4/40,8 = **15,06**)	0,4 + 15,6 = **16,0**
	0,8 + 1,2 + 0,4 = **2,4**	10,3 + 12,5 + 15,6 = **38,4**	**40,8**

$H_0: \chi^2 = 0$ (indipendenza) $H_1: \chi^2 > 0$ (connessione)

Rifiuto H_0 a livello $\alpha = 0,01$ se $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*} = 0,57 \qquad r = 3 \qquad c = 2$$

$$\alpha = 1\% = 0,01 \rightarrow \chi^{(r-1)(c-1)}_{1-\alpha} = \chi^{(3-1)\cdot(2-1)}_{1-0,01} = \chi^{(2)}_{0,99} = 9,21034$$

Non vale che $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$

$$0,57 < 9,21034$$

Non si rifiuta H_0 a livello $\alpha = 0,01$
Non c'è connessione tra le variabili «età» e «utilizzo del web», ossia le due variabili possono essere considerate indipendenti

Il valore prossimo a zero dell'**indice relativo di connessione** conferma il risultato del test :

$$\tilde{\chi}^2 = \frac{\chi^2}{n \cdot \min(r-1, c-1)} = \frac{0,57}{40,8 \cdot min(3-1,\ 2-1)} = 0,014$$

QUIZ

La distribuzione di probabilità più adatta a verificare se due variabili osservate congiuntamente sono indipendenti è

(a) chi-quadro
(b) t-Student
(c) normale
(d) binomiale

QUIZ

Se l'**indice relativo di connessione** tra X e Y risulta uguale a 0, allora tra X e Y c'è

(a) massima connessione
(b) minima connessione
(c) dipendenza
(d) correlazione

$0 \leq \tilde{\chi}^2 \leq 1$

QUIZ

Se l'indice relativo di connessione di X e Y è uguale a 1, allora

(a) $\rho_{XY} = 1$
(b) le variabili sono indipendenti
(c) le variabili non sono indipendenti
(d) $R^2 = 1$

$0 \leq \tilde{\chi}^2 \leq 1$

QUIZ

L'indice χ^2 per due caratteri X e Y risulta 0 se

(a) X e Y sono connessi
(b) X e Y sono correlati
(c) X e Y sono indipendenti
(d) nessuna delle precedenti

Ricordiamo che $0 \leq \chi^2 \leq n \cdot min(r-1, c-1)$

QUIZ

I valori assunti dall'indice di connessione sono

(a) $-1 \leq \rho_{XY} \leq 1$

(b) $0 \leq \tilde{\chi}^2 \leq \mathbf{1}$

(c) $0 \leq \chi^2 \leq n \cdot min(r-1, c-1)$

(d) $0 \leq \chi^2 \leq 1$

QUIZ

I valori assunti dall'indice relativo di connessione sono

(a) $-1 \leq \rho_{XY} \leq 1$

(b) $0 \leq \tilde{\chi}^2 \leq \mathbf{1}$

(c) $0 \leq \chi^2 \leq n \cdot min(r-1, c-1)$

(d) $0 \leq \chi^2 \leq 1$

Esercizio 5

Alcuni pazienti affetti da una certa patologia sono stati classificati per durata (in giorni) dello stato febbrile e per tipo di trattamento somministrato:

durata / cura	$1 \vdash 5$	$5 \vdash 10$	$10 \vdash 15$	
A	0	47		59
B		0	0	25
C	1		18	
		64		

1) Completare la tabella di contingenza
2) Verificare a livello 5% se le due variabili considerate sono indipendenti?

durata / cura	$1 \vdash 5$	$5 \vdash 10$	$10 \vdash 15$	
A	0	47	12	59
B	25	0	0	25
C	1	17	18	36
	26	64	30	120

riga 1 → $59 - 0 - 47 = 12$
riga 2 → $25 - 0 - 0 = 25$
colonna 2 → $64 - 47 - 0 = 17$

colonna 1 → $0 + 25 + 1 = 26$
colonna 3 → $12 + 0 + 18 = 30$
Riga 3 → $1 + 17 + 18 = 36$

Totale pazienti → $26 + 64 + 30 = 120$

- N. righe $r = 3$, N. colonne $c = 3$
- Prima calcolo le marginali $n_{i\cdot}$ e $n_{\cdot j}$ (ultima riga/colonna)
- Poi calcolo delle frequenze n^*_{ij} (tra parentesi)

durata / cura	$1 \vdash 5$	$5 \vdash 10$	$10 \vdash 15$	
A	0 $(59 \times 26/120)$ $(\mathbf{12,8})$	47 $(59 \times 64/120)$ $(\mathbf{31,5})$	12 $(59 \times 30/120)$ $(\mathbf{14,8})$	$0 + 47 + 12 = \mathbf{59}$
B	25 $(25 \times 26/120)$ $(\mathbf{5,4})$	0 $(25 \times 64/120)$ $(\mathbf{13,3})$	0 $(25 \times 30/120)$ $(\mathbf{6,2})$	$25 + 0 + 0 = \mathbf{25}$
C	1 $(36 \times 26/120)$ $(\mathbf{7,8})$	17 $(36 \times 64/120)$ $(\mathbf{19,2})$	18 $(36 \times 30/120)$ $(\mathbf{9})$	$1 + 17 + 18 = \mathbf{36}$
	$0 + 25 + 1 =$ **26**	$47 + 0 + 17 =$ **64**	$12 + 0 + 18 =$ **30**	**120**

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} =$$

$$= \frac{(0-12{,}8)^2}{12{,}8} + \frac{(25-5{,}4)^2}{5{,}4} + \frac{(1-7{,}8)^2}{7{,}8} + \ \ldots + \frac{(18-9)^2}{9} =$$

$$= \mathbf{126{,}8}$$

durata / cura	$\mathbf{1 \vdash 5}$	$\mathbf{5 \vdash 10}$	$\mathbf{10 \vdash 15}$	
A	**0** $(59 \times 26/120)$ $\mathbf{(12,8)}$	**47** $(59 \times 64/120)$ $\mathbf{(31,5)}$	**12** $(59 \times 30/120)$ $\mathbf{(14,8)}$	$0 + 47 + 12 = \mathbf{59}$
B	**25** $(25 \times 26/120)$ $\mathbf{(5,4)}$	**0** $(25 \times 64/120)$ $\mathbf{(13,3)}$	**0** $(25 \times 30/120)$ $\mathbf{(6,2)}$	$25 + 0 + 0 = \mathbf{25}$
C	**1** $(36 \times 26/120)$ $\mathbf{(7,8)}$	**17** $(36 \times 64/120)$ $\mathbf{(19,2)}$	**18** $(36 \times 30/120)$ $\mathbf{(9)}$	$1 + 17 + 18 = \mathbf{36}$
	$0 + 25 + 1 =$ **26**	$47 + 0 + 17 =$ **64**	$12 + 0 + 18 =$ **30**	**120**

$\boldsymbol{H_0: \chi^2 = 0}$ (indipendenza) $\quad \boldsymbol{H_1: \chi^2 > 0}$ (connessione)

Rifiuto H_0 a livello $\alpha = 0{,}05$ se $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} = 126{,}8 \qquad r = 3 \qquad c = 3$$

$\alpha = 5\% = 0{,}05 \quad \rightarrow \quad \chi^{(r-1)(c-1)}_{1-\alpha} = \chi^{(3-1)\cdot(3-1)}_{1-0{,}05} = \chi^{(4)}_{0{,}95} = 9{,}48773$

Vale che $\chi^2 > \chi^{(r-1)(c-1)}_{1-\alpha}$

$126{,}8 > 9{,}48773$

Si rifiuta H_0 a livello $\alpha = 0{,}05$

C'è connessione tra «cura» e «durata della febbre»

Il valore dell'**indice relativo di connessione** conferma il risultato del test

$$\tilde{\chi}^2 = \frac{126.8}{120 \times \min(3-1,\ 3-1)} = 0{,}53$$

indica una connessione tra «cura» e «durata della febbre» abbastanza forte.