Esercitazione 5

Statistica

Alfonso Iodice D'Enza
iodicede@unina.it

Università degli studi di Cassino

Outline

1. Correlazione

Outline

Outline

Misura del legame

Data una variabile doppia (X, Y), la misura del legame che caratterizza le componenti X ed Y si definisce

Misura del legame

Data una variabile doppia (X, Y), la misura del legame che caratterizza le componenti X ed Y si definisce

- connessione se X e Y sono mutabili
- correlazione se X e Y sono variabili

Misura del legame

Nel caso di variabili quantitative preferibile utilizzare una misura del legame che coinvolga, oltre le frequenze, anche le modalità (numeriche) delle variabili.
Le componenti della variabile doppia X e Y possono essere caratterizzate da diversa posizione e variabilità, risulta in genere che

$$\mu_x \neq \mu_y \text{ e } \sigma_x \neq \sigma_y$$

Misura del legame

Nel caso di variabili quantitative preferibile utilizzare una misura del legame che coinvolga, oltre le frequenze, anche le modalità (numeriche) delle variabili.
Le componenti della variabile doppia X e Y possono essere caratterizzate da diversa posizione e variabilità, risulta in genere che

$$\mu_x \neq \mu_y \text{ e } \sigma_x \neq \sigma_y$$

Volendo misurare le variazioni congiunte delle modalità di X ed Y, si fa riferimento alla versione standardizzata delle variabili, data da

$$Z_x = \frac{X - \mu_x}{\sigma_x} \text{ e } Z_y = \frac{Y - \mu_y}{\sigma_y}$$

questo per escludere dalla misura del legame gli effetti della differente media e varianza (essendo $\mu_x \neq \mu_y$ e $\sigma_x \neq \sigma_y$)

Il coefficiente di correlazione lineare di Pearson ρ

L'indice corrispondente alla media aritmetica del prodotto delle modalità standardizzate delle variabili si definisce coefficiente di correlazione lineare di Pearson ρ ed dato da

$$\rho_{xy} = \frac{1}{n}\sum_{i=1}^{n}(z_{x,i}z_{y,i}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}\times\frac{y_i-\mu_y}{\sigma_y}\right)$$

Il coefficiente di correlazione lineare di Pearson ρ

L'indice corrispondente alla media aritmetica del prodotto delle modalità standardizzate delle variabili si definisce coefficiente di correlazione lineare di Pearson ρ ed dato da

$$\rho_{xy} = \frac{1}{n}\sum_{i=1}^{n}(z_{x,i}z_{y,i}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}\times\frac{y_i-\mu_y}{\sigma_y}\right)$$

Con piccole trasformazioni si ottiene la presente formalizzazione

$$\rho_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

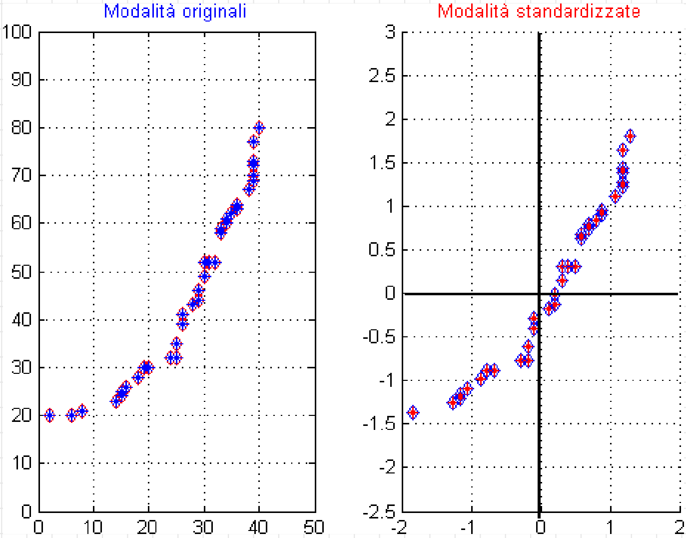
La quantità al numeratore si definisce covarianza: essa corrisponde alla media del prodotto degli scarti delle modalità di X e Y dalle rispettive medie. La covarianza misura la contenporanea variazione di X e Y con riferimento alle loro medie.

Calcolo del coefficiente di correlazione

Si consideri l'esempio della variabile doppia reddito/grado di anzianità

Anzianità (X)	Reddito annuo (Y)	Z(X)	Z(Y)
2	20	-2,43	-1,42
6	20	-2,04	-1,42
8	21	-1,84	-1,36
14	23	-1,26	-1,26
15	24	-1,16	-1,2
15	24	-1,16	-1,2
15	24	-1,16	-1,2
15	25	-1,16	-1,15
16	26	-1,06	-1,1
18	28	-0,87	-0,99
19	30	-0,77	-0,88
20	30	-0,67	-0,88
24	32	-0,28	-0,77
25	32	-0,19	-0,77
25	35	-0,19	-0,61
26	39	-0,09	-0,4
26	41	-0,09	-0,29
26	41	-0,09	-0,29
28	43	0,11	-0,18
29	44	0,2	-0,13
29	46	0,2	-0,02
30	49	0,3	0,14
30	52	0,3	0,3
31	52	0,4	0,3
32	52	0,5	0,3
33	58	0,59	0,63
33	59	0,59	0,68
34	60	0,69	0,73
34	61	0,69	0,79
35	62	0,79	0,84
36	63	0,89	0,89
36	64	0,89	0,95
38	67	1,08	1,11
38	67	1,08	1,11
39	69	1,18	1,22
39	70	1,18	1,27
39	72	1,18	1,38
39	73	1,18	1,43
39	77	1,18	1,65
40	80	1,28	1,81

Calcolo del coefficiente di correlazione



Calcolo del coefficiente di correlazione

sia X = 'anni di anzianità' e Y = 'reddito annuo'

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2} = 10.13$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_y)^2} = 18.36$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) = 177.31$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{177.31}{10.13\times 18.36} = 0.95$$

Proprietà del coefficiente di correlazione

- se X e Y sono indipendenti, allora $\rho_{xy} = 0$ (NON vale il contrario)
- se $\rho_{xy} = 1$, allora $Y = \alpha + \beta X$ (ovvero Y una trasformazione lineare di X)
- se $\rho_{xy} = -1$, allora $Y = \alpha - \beta X$ (ovvero Y una trasformazione lineare di X)
- $\rho_{xy} = \rho_{yx}$
- $\rho_{xx} = 0$

Metodo alternativo per il calcolo di ρ

Da un punto di vista computazionale risulta conveniente l'utilizzo della seguente formulazione alternativa del coefficiente di correlazione lineare ρ basata sulle somme delle modalità delle componenti ($\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i$), sulle somme dei quadrati delle modalità delle componenti ($\sum_{i=1}^{n}(x_i)^2, \sum_{i=1}^{n} (y_i)^2$), sulla somma dei prodotti tra le modalità ($\sum_{i=1}^{n} x_i y_i$)

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n}(x_i)^2 - [\sum_{i=1}^{n} x_i]^2)(n\sum_{i=1}^{n}(y_i)^2 - [\sum_{i=1}^{n} y_i]^2)}}$$

Metodo alternativo per il calcolo di ρ

2	20	40	4	400
6	20	120	36	400
8	21	168	64	441
14	23	322	196	529
15	24	360	225	576
15	24	360	225	576
15	24	360	225	576
15	25	375	225	625
16	26	416	256	676
18	28	504	324	784
19	30	570	361	900
20	30	600	400	900
24	32	768	576	1024
25	32	800	625	1024
25	35	875	625	1225
26	39	1014	676	1521
26	41	1066	676	1681
26	41	1066	676	1681
28	43	1204	784	1849
29	44	1276	841	1936
29	46	1334	841	2116
30	49	1470	900	2401
30	52	1560	900	2704
31	52	1612	961	2704
32	52	1664	1024	2704
33	58	1914	1089	3364
33	59	1947	1089	3481
34	60	2040	1156	3600
34	61	2074	1156	3721
35	62	2170	1225	3844
36	63	2268	1296	3969
36	64	2304	1296	4096
38	67	2546	1444	4489
38	67	2546	1444	4489
39	69	2691	1521	4761
39	70	2730	1521	4900
39	72	2808	1521	5184
39	73	2847	1521	5329
39	77	3003	1521	5929
40	80	3200	1600	6400

I totali di colonna sono rispettivamente

- $\sum_{i=1}^{n} x_i = 1076$
- $\sum_{i=1}^{n} y_i = 1855$
- $\sum_{i=1}^{n} x_i y_i = 56992$
- $\sum_{i=1}^{n} (x_i)^2 = 33046$
- $\sum_{i=1}^{n} (y_i)^2 = 99509$

Metodo alternativo per il calcolo di ρ

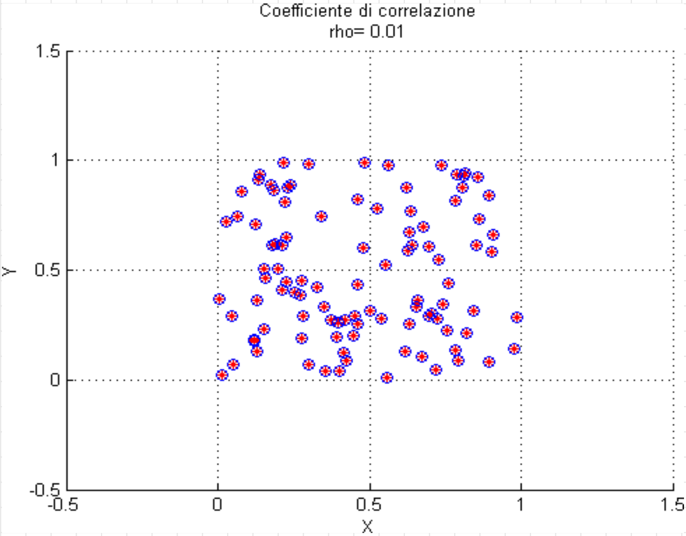
Si può dunque ricorrere alla seguente formula, sostituendo opportunamente le quantità trovate:

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n}(x_i)^2 - \left[\sum_{i=1}^{n} x_i\right]^2\right)\left(n\sum_{i=1}^{n}(y_i)^2 - \left[\sum_{i=1}^{n} y_i\right]^2\right)}}$$

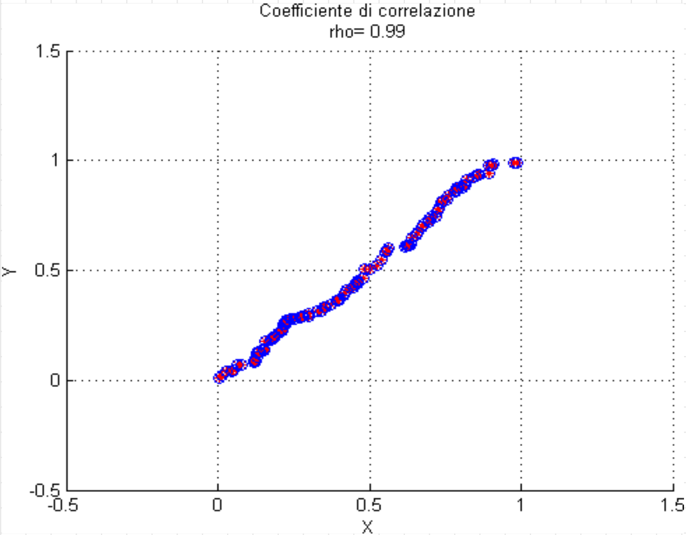
$$= \frac{40 \times 56992 - (1076)(1855)}{\sqrt{(40 \times 33046 - (1076)^2)(40 \times 99509 - (1855^2))}} =$$

$$\frac{283700}{297470} = 0.95$$

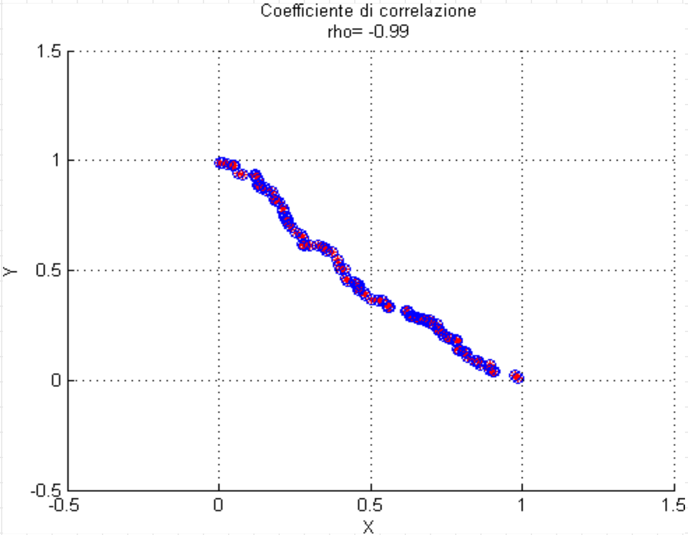
Coefficiente di correlazione: esempi di casi limite



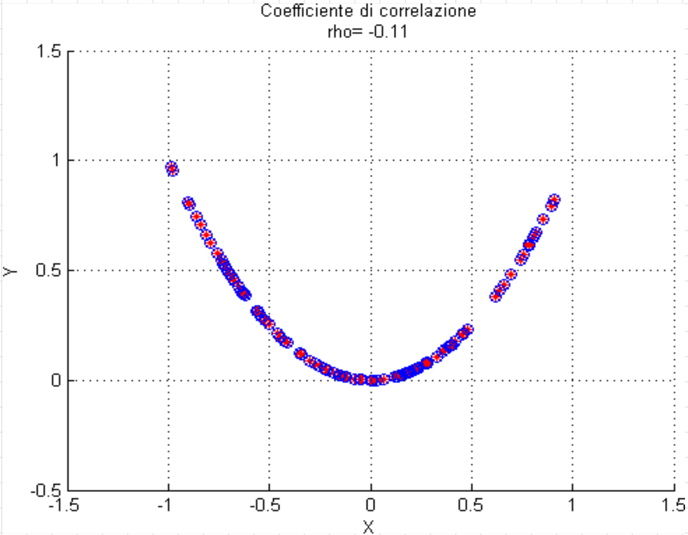
Coefficiente di correlazione: esempi di casi limite



Coefficiente di correlazione: esempi di casi limite



Coefficiente di correlazione: esempi di casi limite



Connessione in media

Data una distibuzione doppia di un carattere misto (X, Y), si dir che la componente Y indipendente in media da X se al variare delle modalità di X le medie condizionate di X rimangono costanti (vale il viceversa). Il fatto che Y sia indipendente in media da X non implica che sia vero il contrario (come invece accade per l'indipendenza in distribuzione).

Connessione in media

Data una distibuzione doppia di un carattere misto (X, Y), si dir che la componente Y indipendente in media da X se al variare delle modalità di X le medie condizionate di X rimangono costanti (vale il viceversa). Il fatto che Y sia indipendente in media da X non implica che sia vero il contrario (come invece accade per l'indipendenza in distribuzione).

$$\mu_y = \overline{y} = \frac{1}{n}\sum_{j=1}^{h} y_j n_{.j}$$

Rappresenta la media di Y e si ottiene considerando la distribuzione marginale di Y.

$$\overline{y}_i = \overline{y}|x_i = \frac{1}{n_{i.}}\sum_{j=1}^{h} y_j n_{ij}$$

Rappresenta la media di Y condizionata alla $i-ma$ modalità della variabile X.

Decomposizione della devianza

Ricordando che la devianza il numeratore della varianza...

$$
\begin{aligned}
Dev_y = & \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y})^2 n_{ij} = \\
& = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i+\overline{y}_i-\overline{y})^2 n_{ij} = \\
& = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{h}(\overline{y}_i-\overline{y})^2 n_{ij} + \\
& + 2\sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i)(\overline{y}_i-\overline{y}) n_{ij}
\end{aligned}
$$

Decomposizione della devianza

$$= \sum_{i=1}^{k} \left[\sum_{j=1}^{h} (y_j - \overline{y}_i)^2 n_{ij} \right] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} +$$

$$+ 2 \sum_{i=1}^{k} (y_j - \overline{y}_i) \sum_{j=1}^{h} (\overline{y}_i - \overline{y}) n_{ij} =$$

$$= \sum_{i=1}^{k} [Dev(Y \mid X = x_i)] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} =$$

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

Decomposizione della devianza

$$= \sum_{i=1}^{k}\left[\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij}\right]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}+$$

$$+\,2\sum_{i=1}^{k}(y_j-\overline{y}_i)\sum_{j=1}^{h}(\overline{y}_i-\overline{y})n_{ij} =$$

$$= \sum_{i=1}^{k}\left[Dev(Y \mid X=x_i)\right]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.} =$$

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

Decomposizione della devianza

$$= \sum_{i=1}^{k} \left[\sum_{j=1}^{h} (y_j - \overline{y}_i)^2 n_{ij} \right] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} +$$

$$+ 2 \sum_{i=1}^{k} (y_j - \overline{y}_i) \sum_{j=1}^{h} (\overline{y}_i - \overline{y}) n_{ij} =$$

$$= \sum_{i=1}^{k} [Dev(Y \mid X = x_i)] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} =$$

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

Decomposizione della devianza

$$= \sum_{i=1}^{k} \left[\sum_{j=1}^{h} (y_j - \overline{y}_i)^2 n_{ij} \right] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} +$$

$$+ 2 \sum_{i=1}^{k} (y_j - \overline{y}_i) \sum_{j=1}^{h} (\overline{y}_i - \overline{y}) n_{ij} =$$

$$= \sum_{i=1}^{k} [Dev(Y \mid X = x_i)] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} =$$

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

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.

Se Y indipendente in media da X, allora le medie condizionate $\overline{y}_i$ saranno tutte costanti, la variabilità ad esse associate sar uguale a zero. In particolare risulter $Dev(B) = 0$ quindi

$$Dev(Y) = Dev(W) + 0$$

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.
Se Y indipendente in media da X, allora le medie condizionate $\overline{y}_i$ saranno tutte costanti, la variabilità ad esse associate sar uguale a zero. In particolare risulter $Dev(B) = 0$ quindi

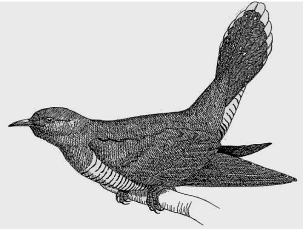
$$Dev(Y) = Dev(W) + 0$$

Quindi, per quantificare la dipendenza in media di Y da X occorre un indice basato su $Dev(B)$.

$$\eta^2 = \frac{Dev(B)}{Dev(Y)}$$

Il nido del cuculo

Il cuculo è un uccello caratterizzato da una particolare abitudine: depone le uova nei nidi di altri uccelli, e lascia dunque che siano altre specie a covarle. Ovviamente, il tutto funziona se la dimensione delle uova nel nido ospite sono compatibili con quelle del nido ospitante. In alcuni territori, il cuculo depone le uova in nidi di scricciolo, in altri sceglie nidi di pettirosso.

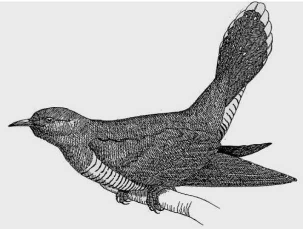


Si consideri di aver osservato la lunghezza di $n_1 = 15$ uova di cuculo ritrovate in nidi di scricciolo e $n_2 = 16$ uova di cuculo ritrovate in nidi di pettirosso. Si vuole verificare se la lunghezza delle uova dipende in media dal tipo di nido in cui vengono deposte.

Calcolo del rapporto di correlazione

Il nido del cuculo

Il cuculo è un uccello caratterizzato da una particolare abitudine: depone le uova nei nidi di altri uccelli, e lascia dunque che siano altre specie a covarle. Ovviamente, il tutto funziona se la dimensione delle uova nel nido ospite sono compatibili con quelle del nido ospitante. In alcuni territori, il cuculo depone le uova in nidi di scricciolo, in altri sceglie nidi di pettirosso.



Si consideri di aver osservato la lunghezza di $n_1 = 15$ uova di cuculo ritrovate in nidi di scricciolo e $n_2 = 16$ uova di cuculo ritrovate in nidi di pettirosso. Si vuole verificare se la lunghezza delle uova dipende in media dal tipo di nido in cui vengono deposte.

Calcolo del rapporto di correlazione

Scricciolo

Sia S la lunghezza delle uova di cuculo nei nidi di scricciolo

```
> S
      [,1]
 [1,] 19.85
 [2,] 20.05
 [3,] 20.25
 [4,] 20.85
 [5,] 20.85
 [6,] 20.85
 [7,] 21.05
 [8,] 21.05
 [9,] 21.05
[10,] 21.25
[11,] 21.45
[12,] 22.05
[13,] 22.05
[14,] 22.05
[15,] 22.25
> summary(scricciolo)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.85   20.85   21.05   21.13   21.75   22.25
```

Pettirosso

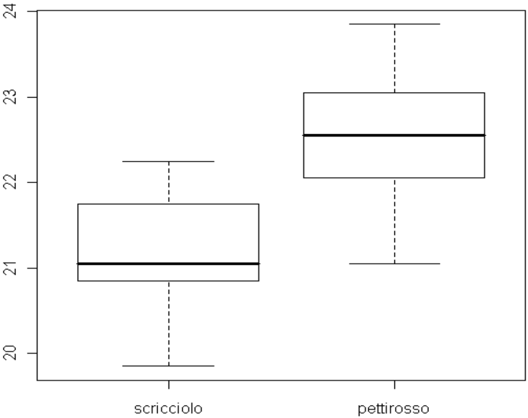
Sia P la lunghezza delle uova di cuculo nei nidi di pettirosso

```
> P
      [,1]
 [1,] 21.05
 [2,] 21.85
 [3,] 22.05
 [4,] 22.05
 [5,] 22.05
 [6,] 22.25
 [7,] 22.45
 [8,] 22.45
 [9,] 22.65
[10,] 23.05
[11,] 23.05
[12,] 23.05
[13,] 23.05
[14,] 23.05
[15,] 23.25
[16,] 23.85
> summary(pettirosso)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.05   22.05   22.55   22.57   23.05   23.85
```

Calcolo del rapporto di correlazione

Confronto tra le distribuzioni

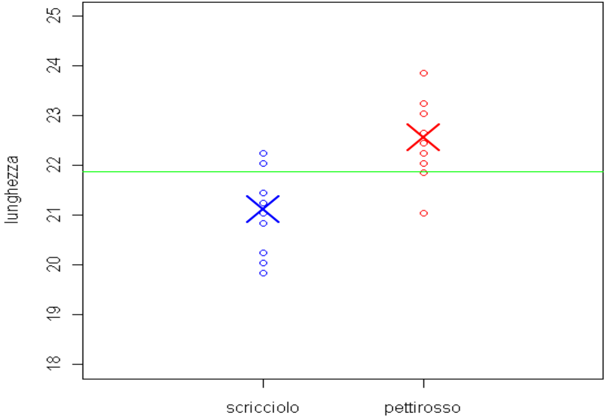
Un primo confronto grafico via box plot tra le due distribuzioni mostra che le uova deposte in nidi di pettirosso hanno una lunghezza maggiore di quelle deposte in nidi di scricciolo.



Calcolo del rapporto di correlazione

Confronto tra le distribuzioni

Un ulteriore confronto grafico tra le due distribuzioni consiste in un diagramma per punti: sono riportate graficamente le medie condizionate, mentre la media generale è rappresentata dalla linea orizzontale.



Calcolo del rapporto di correlazione

Si indica con $\mu_X = 21.875$ la lunghezza media delle $n = n_1 + n_2$ uova complessivamente considerate. Le medie condizionate al nido in cui le uova sono state deposte sono rispettivamente $\mu_{X|S} = 21.13$ e $\mu_{X|P} = 22.57$. La devianza delle medie condizionate rispetto alla media generale è dunque

$$dev_b = (21.13-21.875)^2 \times 15 + (22.57-21.875)^2 \times 16 = 16.165$$

mentre la devianza complessiva è data da

$$dev_{tot} = (19.85 - 21.875)^2 + (20.05 - 21.875)^2 + \\ + \ldots + (23.25 - 21.875)^2 + (23.85 - 21.875)^2 = 30.94$$

$$\eta^2 = \frac{dev_b}{dev_{tot}} = \frac{16.165}{30.94} = 0.522$$

Calcolo del rapporto di correlazione: valori in classi

Si consideri l'esempio della variabile doppia reddito/grado di anzianità

Zona\reddito(in migliaia)	[10,15]	[15, 20[	[20,25[	[25,30[	Totale
Nord	0	7	34	5	46
Centro	1	18	5	1	25
Sud	31	1	0	0	32
Totale	32	26	39	6	103

Calcolo del rapporto di correlazione

Ai fini del calcolo del rapporto di correlazione necessario calcolare la devianza totale della variabile $Dev(Y)$ e la devianza tra le classi $Dev(B)$ (ovvero la devianza tra le medie condizionate $Y \mid X = x_i, i = 1, 2, \dots, k$ e la media globale). Dunque

$$\mu(Y) = \frac{1}{103}(12.5 \times 32) + (17.5 \times 26) + \\ + (22.5 \times 39) + (27.5 \times 6) = 14.9$$

Calcolo del rapporto di correlazione

$$\mu(Y \mid x_i = Nord) = \frac{1}{46}(12.5 \times 0) + (17.5 \times 7) + \\ + (22.5 \times 34) + (27.5 \times 5) = 22.28$$

$$\mu(Y \mid x_i = Centro) = \frac{1}{25}(12.5 \times 1) + (17.5 \times 18) + \\ + (22.5 \times 5) + (27.5 \times 1) = 18.7$$

$$\mu(Y \mid x_i = Sud) = \frac{1}{32}(12.5 \times 31) + (17.5 \times 1) + \\ + (22.5 \times 0) + (27.5 \times 0) = 12.66$$

Calcolo del rapporto di correlazione

$$dev(Y) = (12.5 - 14.9)^2 \times 32 + (17.5 - 14.9)^2 \times 26 + \\ + (22.5 - 14.9)^2 \times 39 + (27.5 - 14.9)^2 \times 6 = 3565.3$$

$$dev(B) = (22.28 - 14.9)^2 \times 46 + (18.7 - 14.9)^2 \times 25 + \\ + (12.66 - 14.9)^2 \times 32 = 3026.9$$

$$\eta^2 = \frac{dev(B)}{dev(Y)} = \frac{3026.9}{3565.3} = 0.849$$