Esercitazione 7

Statistica

Alfonso Iodice D'Enza
iodicede@gmail.com

Università degli studi di Cassino

Outline

1. Relazioni tra variabili
2. Indipendenza
3. Indici di connessione
4. Il coefficiente di correlazione lineare
5. Dipendenza in variabili miste

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

Interdipendenza e dipendenza

Se le componenti di una variabile doppia (X, Y) oggetto di studio rivestono lo stesso ruolo ai fini dell'analisi si studia l'interdipendenza tra X e Y. Se si vuole studiare, invece, l'andamento della variabile Y rispetto ad X, si farà riferimento alla dipendenza di Y da X.

- Y si definisce variabile dipendente
- X si definisce variabile indipendente

Frequenze condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Distribuzione condizionata del carattere A rispetto alla j-sima modalità del carattere B

Frequenze condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Distribuzione condizionata del carattere B rispetto alla i-sima modalità del carattere A

Frequenze relative condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

9

La distribuzione delle frequenze relative condizionate della variabile A (k modalità) rispetto alla $j-$sima modalità della variabile B (h modalità) si ottiene dividendo ciascun elemento dell'$j-$ma colonna (frequenza assoluta) per il rispettivo totale di di colonna $n_{ij}/n_{.j}$ per $i = 1, \ldots, k$.

Frequenze relative condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{Kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

1

La distribuzione delle frequenze relative condizionate della variabile B (h modalità) rispetto alla $i-$sima modalità della variabile A (k modalità) si ottiene dividendo ciascun elemento dell'$i-$ma riga (frequenza assoluta) per il rispettivo totale di riga $n_{ij}/n_{i.}$ per $j = 1, \ldots, h$.

Esempio di tabella a doppia entrata

Si consideri di aver registrato il colore degli occhi e quello dei capelli di un collettivo di 592 persone. I risultati sono raccolti nella seguente tabella

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*	*Tot*
nero	68	119	26	7	220
azzurro	20	84	17	94	215
marrone	15	54	14	10	93
verde	5	29	14	16	64
Tot	108	286	71	127	592

Distribuzioni relative condizionate

Frequenze condizionate della variabile *capelli* rispetto alle modalità della variabile *occhi*

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*	*Tot*
nero	0.309	0.541	0.118	0.032	1
azzurro	0.093	0.391	0.079	0.437	1
marrone	0.161	0.581	0.151	0.108	1
verde	0.078	0.453	0.219	0.250	1

Frequenze condizionate della variabile *occhi* rispetto alle modalità della variabile *capelli*

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*
nero	0.630	0.416	0.366	0.055
azzurro	0.185	0.294	0.239	0.740
marrone	0.139	0.189	0.197	0.079
verde	0.046	0.101	0.197	0.126
Tot	1	1	1	1

Indipendenza

Indipendenza e distribuzioni condizionate

Le componenti di una variabile doppia (X, Y) sono indipendenti se le distribuzioni di frequenze relative condizionate $Y|X$ e $X|Y$ sono costanti.
Formalmente dovrà risultare per $Y|X$

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \frac{n_{i3}}{n_{.3}} = \ldots = \frac{n_{ih}}{n_{.h}}$$

e per $X|Y$

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \frac{n_{3j}}{n_{3.}} = \ldots = \frac{n_{kj}}{n_{k.}}$$

Indipendenza

Si supponga che nel precedente esempio sia stata osservata la seguente distribuzione doppia.

occhi/capelli	$neri$	$castani$	$rossi$	$biondi$	Tot
$nero$	40	106	26	47	220
$azzurro$	39	104	26	46	215
$marrone$	17	45	11	20	93
$verde$	12	31	8	14	64
Tot	108	286	71	127	592

Indipendenza

In questo caso le frequenze condizionate della variabile *capelli* rispetto alle modalità della variabile *occhi*

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*	*Tot*
nero	0.182	0.483	0.120	0.215	1
azzurro	0.182	0.483	0.120	0.215	1
marrone	0.182	0.483	0.120	0.215	1
verde	0.182	0.483	0.120	0.215	1
Tot	0.182	0.483	0.120	0.215	1

Mentre le frequenze condizionate della variabile *occhi* rispetto alle modalità della variabile *capelli*

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*	*Tot*
nero	0.372	0.372	0.372	0.372	0.372
azzurro	0.363	0.363	0.363	0.363	0.363
marrone	0.157	0.157	0.157	0.157	0.157
verde	0.108	0.108	0.108	0.108	0.108
Tot	1	1	1	1	1

Indipendenza

Se le componenti di una variabile doppia (X, Y) sono indipendenti (le distribuzioni di frequenze relative condizionate $Y|X$ e $X|Y$ sono costanti), allora vale la seguente relazione

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

con $i = 1, \ldots, k; j = 1, \ldots, h$

Pertanto, data una distribuzione doppia di frequenze, il legame tra le due componenti (mutabile) varierà tra una situazione di indipendenza (assenza di legame) e un qualche grado di connessione

Indice quadratico di connessione (X^2)

Gli indici per la misura della connessioni sono basati sulle differenze tra le frequenze osservate sul collettivo n_{ij} e le frequenze teoriche $\hat{n}_{ij}$, che si osserverebbero sul collettivo se le mutabili considerate fossero indipendenti.
Indice quadratico di connessione (X^2) è dato dalla seguente relazione

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- in caso di indipendenza, essendo $n_{ij} = \hat{n}_{ij}$, risulta $X^2 = 0$
- il massimo valore dell'indice è dato dalla seguente espressione: $n \times min(k-1, h-1)$

Indice quadratico di connessione (X^2)

Per calcolare l'indice quadratico di connessione che caratterizza le variabili $coloreocchi$ e $colorecapelli$, con distribuzione congiunta di frequenze

n_{ij} :

occhi/capelli	$neri$	$castani$	$rossi$	$biondi$	Tot
$nero$	68	119	26	7	220
$azzurro$	20	84	17	94	215
$marrone$	15	54	14	10	93
$verde$	5	29	14	16	64
Tot	108	286	71	127	592

si deve calcolare la distribuzione di frequenze che si osserverebbero in caso di indipendenza

$\hat{n}_{ij}$:

occhi/capelli	$neri$	$castani$	$rossi$	$biondi$	Tot
$nero$	40.135	106.284	26.385	47.196	220
$azzurro$	39.223	103.868	25.785	46.123	215
$marrone$	16.966	44.929	11.154	19.951	93
$verde$	11.676	30.919	7.676	13.730	64
Tot	108	286	71	127	592

Indice quadratico di connessione (X^2)

$\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$:

occhi/capelli	*neri*	*castani*	*rossi*	*biondi*
nero	19.346	1.521	0.006	34.234
azzurro	9.421	3.800	2.993	49.697
marrone	0.228	1.831	0.726	4.963
verde	3.817	0.119	5.211	0.375

L'indice X^2 è dato dunque dalla somma degli elementi in tabella

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = 19.346 + 1.521 + 0.006 + 34.234 + 9.421 + 3.800 + 2.993 +$$
$$+ 49.697 + 0.228 + 1.831 + 0.726 + 4.963 + 3.817 + 0.119 + 5.211 + 0.375 = 138.29$$

Indice ν di Cramer

avendo definito $n \times min(k-1, h-1)$ come valore massimo che X^2 può assumere, è possibile ottenere una versione normalizzata dell'indice di connessione. Viene definito indice ν di Cramer.

$$\nu = \sqrt{\frac{X^2}{n \times min(k-1, h-1)}}$$

con k e h numero di modalità delle componenti della mutabile doppia.
L'indice è normalizzato, quindi $0 \leq \nu \leq 1$.

Indice ν di Cramer

Con riferimento ai dati dell'esercizio, si ha che $X^2 = 138.29$, $n = 592$, $h = 4$ e $k = 4$

$$\nu = \sqrt{\frac{X^2}{n \times min(k-1, h-1)}} = \sqrt{\frac{138.29}{592 \times min(3,3)}} = 0.28$$

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

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

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.

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)
- $\rho_{xy} = \rho_{yx}$
- $\rho_{xx} = 0$

Il coefficiente di correlazione lineare di Pearson ρ

Esercizio

Si considerino i voti riportati da $n = 8$ studenti negli esami di $matematica$ e $statistica$.

	$matematica(x_i)$	$statistica(y_i)$
1	24	23
2	27	28
3	30	30
4	26	27
5	29	30
6	18	20
7	21	20
8	22	25

- Si misuri il legame lineare che caratterizza le due variabili

Svolgimento

É necessario calcolare le medie aritmetiche μ e gli scarti quadratici medi σ

- Il voto medio ottenuto dagli studenti all'esame di matematica è
  $\mu_m = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{197}{8} = 24.625$
- Il voto medio ottenuto dagli studenti all'esame di statistica è $\mu_s = \frac{\sum_{i=1}^{8} y_i}{n} = \frac{203}{8} = 25.375$

	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$
1	24	23	-0.62	-2.38	0.39	5.64
2	27	28	2.38	2.62	5.64	6.89
3	30	30	5.38	4.62	28.89	21.39
4	26	27	1.38	1.62	1.89	2.64
5	29	30	4.38	4.62	19.14	21.39
6	18	20	-6.62	-5.38	43.89	28.89
7	21	20	-3.62	-5.38	13.14	28.89
8	22	25	-2.62	-0.38	6.89	0.14
Tot	197	203			119.875	115.875

scarti quadratici medi: $\sigma_m = \sqrt{\frac{\sum_{i=1}^{8} (x_i - \mu_m)^2}{n}} = \sqrt{\frac{119.875}{8}} = 3.87$

$\sigma_s = \sqrt{\frac{\sum_{i=1}^{8} (y_i - \mu_s)^2}{n}} = \sqrt{\frac{115.875}{8}} = 3.805$

Svolgimento

Per calcolare il coefficiente di correlazione lineare resta da calcolare la covarianza, ovvero la media aritmetica del prodotto degli scarti dalla media.

	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x) \times (y_i - \mu_y)$
1	24.00	23.00	-0.62	-2.38	1.48
2	27.00	28.00	2.38	2.62	6.23
3	30.00	30.00	5.38	4.62	24.86
4	26.00	27.00	1.38	1.62	2.23
5	29.00	30.00	4.38	4.62	20.23
6	18.00	20.00	-6.62	-5.38	35.61
7	21.00	20.00	-3.62	-5.38	19.48
8	22.00	25.00	-2.62	-0.38	0.98
Tot	197	203			111.125

La covarianza è

$$\sigma_{ms} = \frac{\sum_{i=1}^{8} (x_i - \mu_m)(y_i - \mu_s)}{n} = \frac{111.125}{8} = 13.89$$

É ora possibile calcolare il coefficiente di correlazione dato da

$$\rho_{ms} = \frac{\sigma_{ms}}{\sigma_m \sigma_s} = \frac{13.89}{3.87 \times 3.805} = 0.943$$

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 $(\sum_{i=1}^{n} x_i y_i)$

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

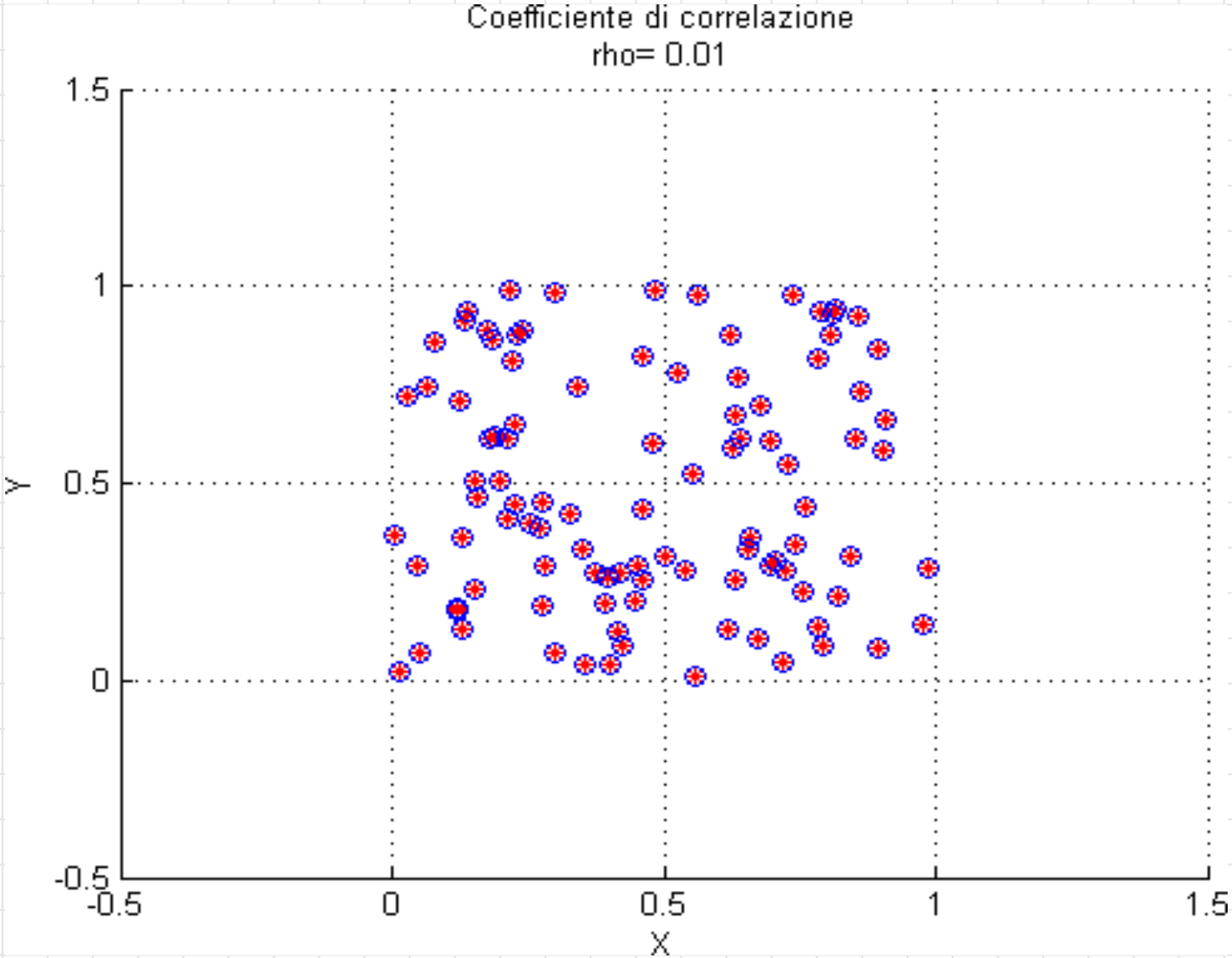
Metodo alternativo per il calcolo di ρ

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	24	23	576	529	552
2	27	28	729	784	756
3	30	30	900	900	900
4	26	27	676	729	702
5	29	30	841	900	870
6	18	20	324	400	360
7	21	20	441	400	420
8	22	25	484	625	550
	$\sum x = 197$	$\sum y = 203$	$\sum x^2 = 4971$	$\sum y^2 = 5267$	$\sum xy = 5110$

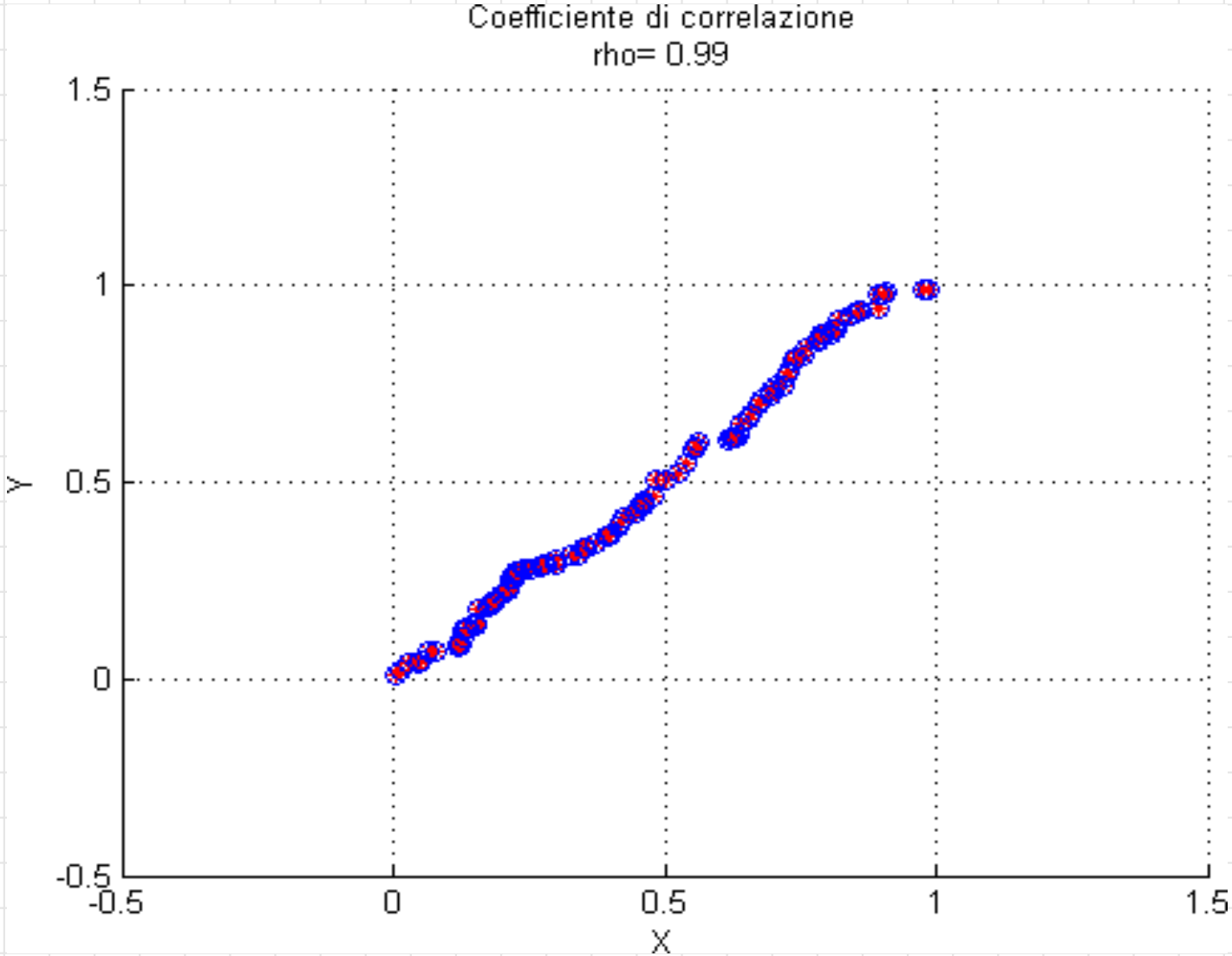
$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n}(x_i)^2 - \left[\sum_{i=1}^{n} x_i\right]^2)(n\sum_{i=1}^{n}(y_i)^2 - \left[\sum_{i=1}^{n} y_i\right]^2)}} =$$

$$= \frac{8 \times 5110 - (197 \times 203)}{\sqrt{(8 \times 4971 - (197)^2) \times (8 \times 5267 - (203)^2)}} = 0.943$$

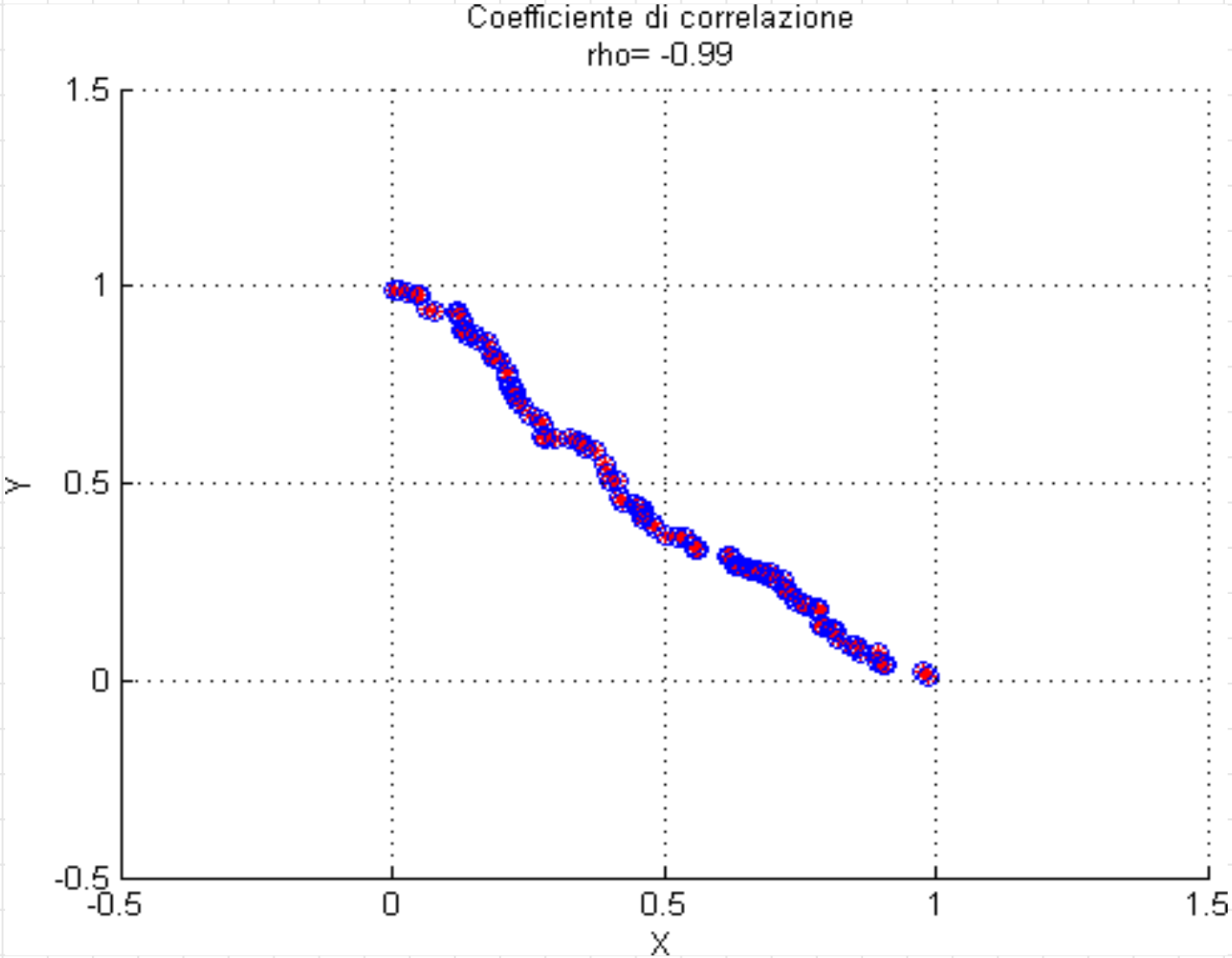
Coefficiente di correlazione: esempi di casi limite



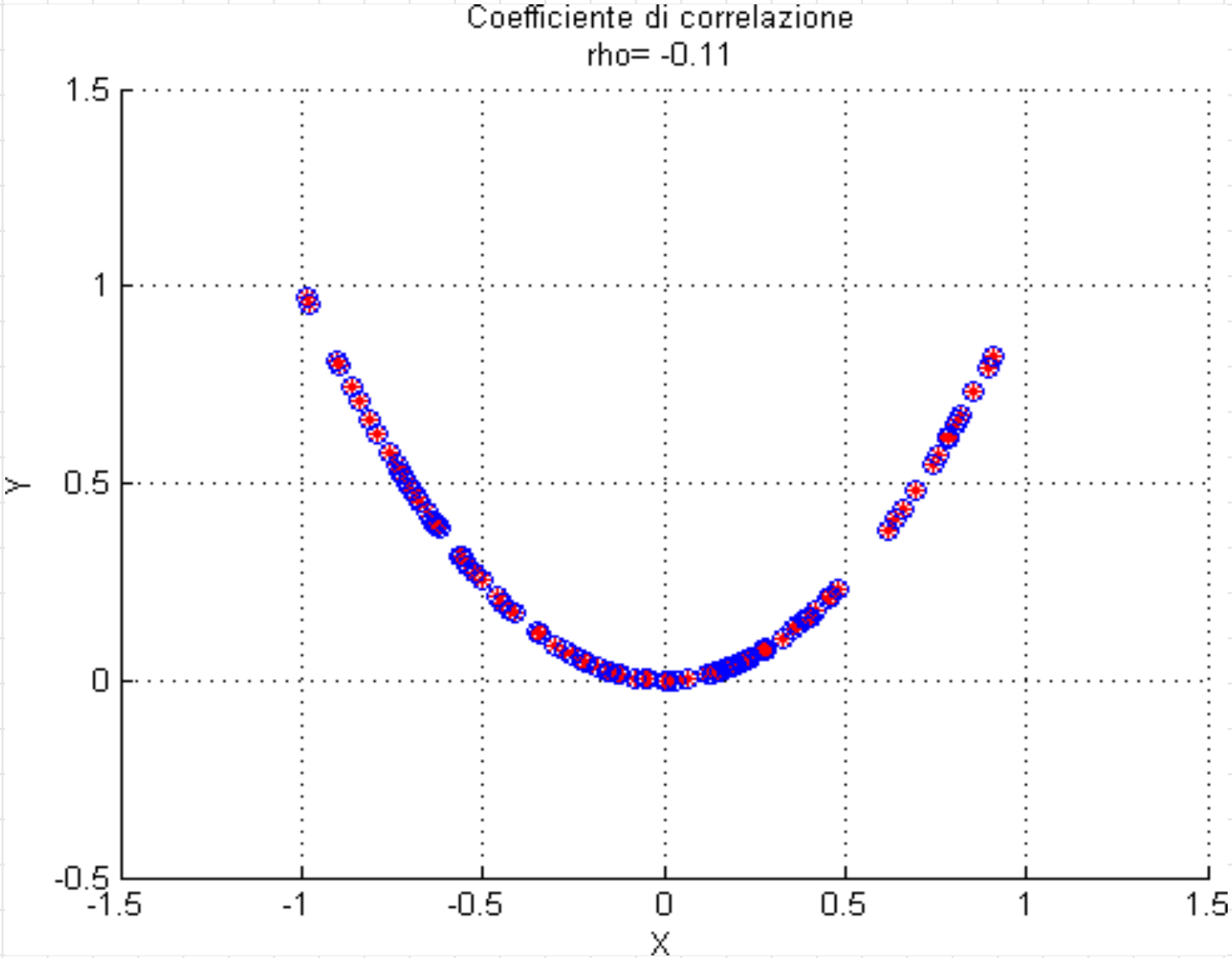
Coefficiente di correlazione: esempi di casi limite



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...

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

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

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

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

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

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

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

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

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

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

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

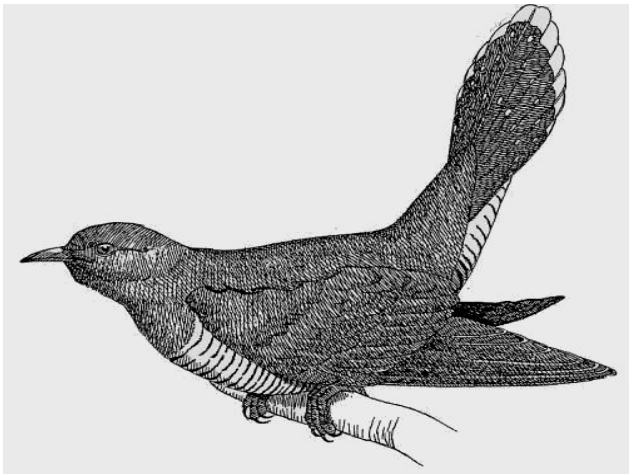
Quindi, per quantificare la dipendenza in media di Y da X occorre un indice basato su $Dev(B)$.

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

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

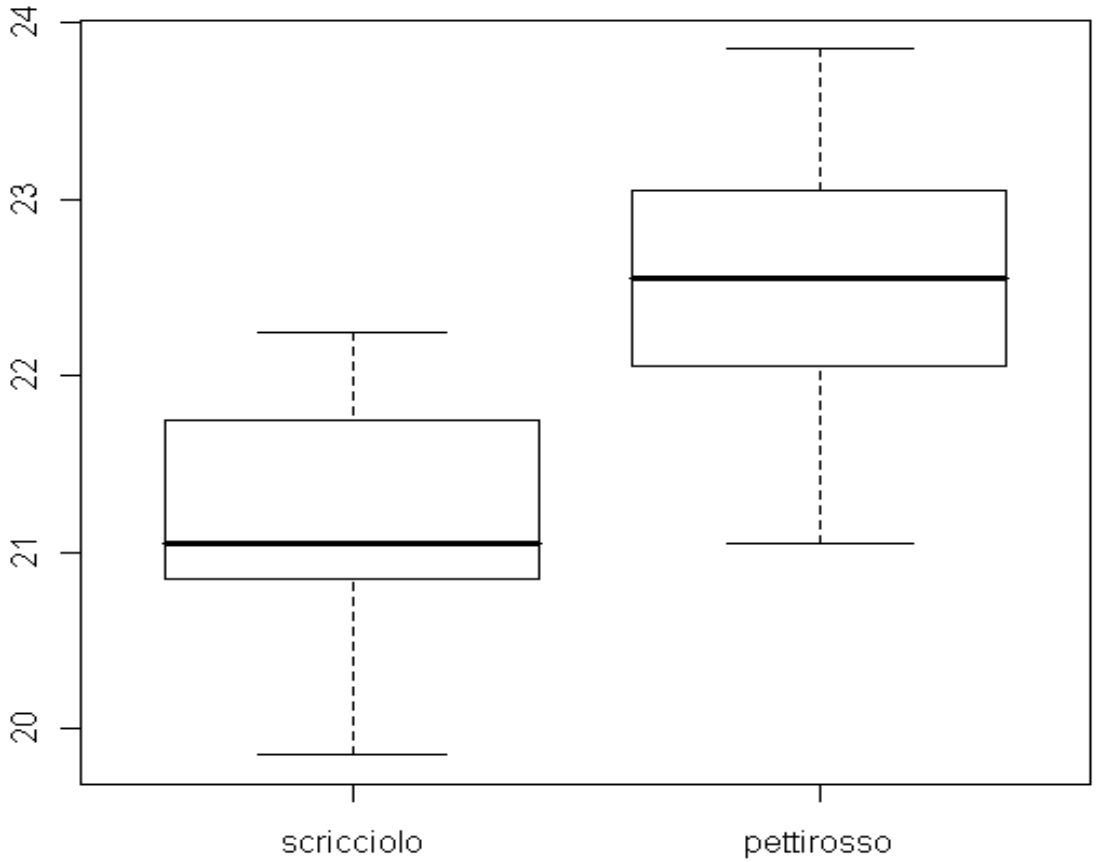
```
> P
      [,1]
 [1,] 21.05
 [2,] 21.85
 [3,] 22.05
 [4,] 22.05
 [5,] 22.05
 [6,] 22.25
 [7,] 22.45
 [8,] 22.45
 [9,] 22.65
[10,] 23.05
[11,] 23.05
[12,] 23.05
[13,] 23.05
[14,] 23.05
[15,] 23.25
[16,] 23.85
```

```
> 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

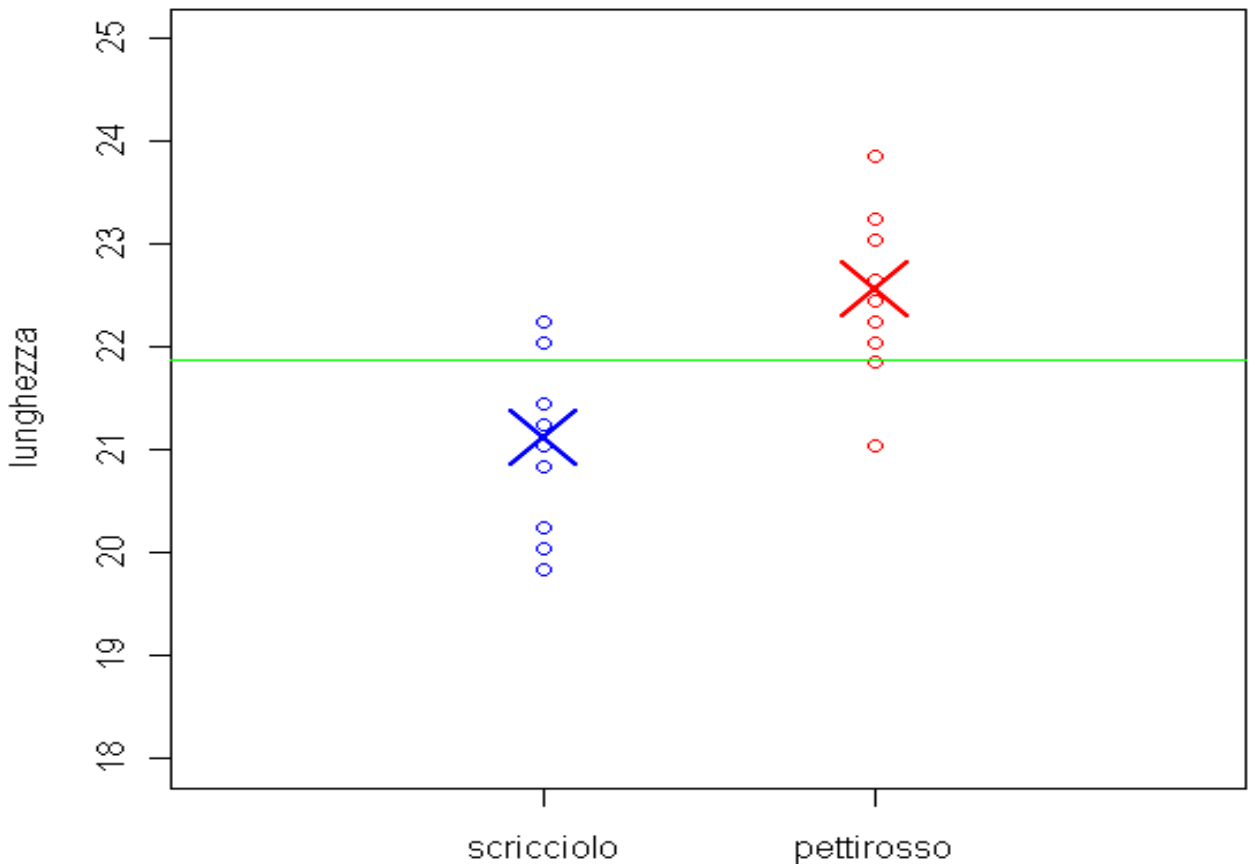
Un primo confronto grafico via box plot tra le due distribuzioni mostra che le uova deposte in nidi di pettirosso hanno una lunghezza maggiore di quelle deposte in nidi di scricciolo.



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 ï¿ $\frac{1}{2}$ 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 $Y \mid X = x_i, i = 1, 2, \ldots, k$ 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

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

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

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