

Indici di dispersione

Consideriamo i tre insiemi di dati:

	$x_{(1)}$	$x_{(n)}$
X_1	30	40	50	60	70	80	90
X_2	30	40	50	60	70	80	<u>300</u>
X_3	<u>0</u>	40	50	60	70	80	90

Sono indici di dispersione:

- Lo scarto interquartile $Q_3 - Q_1$
- Il campo di variazione (Range) $x_{(n)} - x_{(1)}$

Osserviamo come cambiano i valori

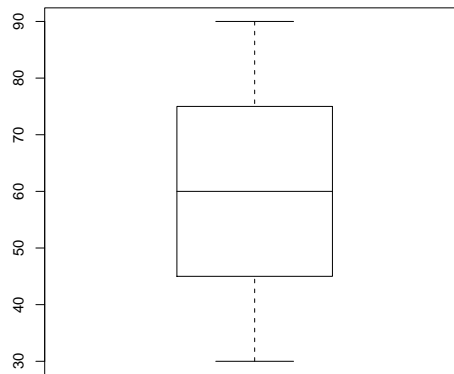
Insieme	Q_1	Me	Q_3	$Q_3 - Q_1$	M	Range
1	40	60	80	40	60	60
2	40	60	80	40	90	270
3	40	60	80	40	55.7	90

Il box-plot

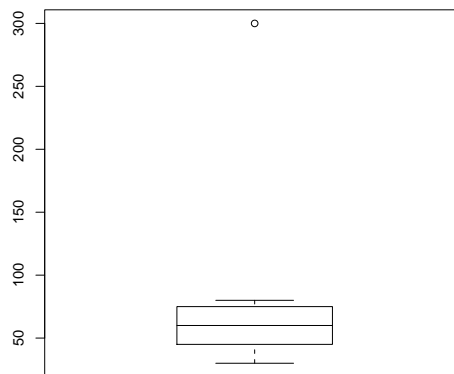
La distribuzione di una variabile statistica viene rappresentata come una scatola.

- gli estremi della scatola sono Q_1 e Q_3
- la scatola è tagliata dalla mediana
- baffo superiore: $Q_1 + 1.5 \cdot (Q_3 - Q_1)$
- baffo inferiore: $Q_3 - 1.5 \cdot (Q_3 - Q_1)$
- se non ci sono valori in corrispondenza dei baffi questi si accorciano al dato più vicino
- tutti i valori fuori dai baffi si segnano come punti isolati

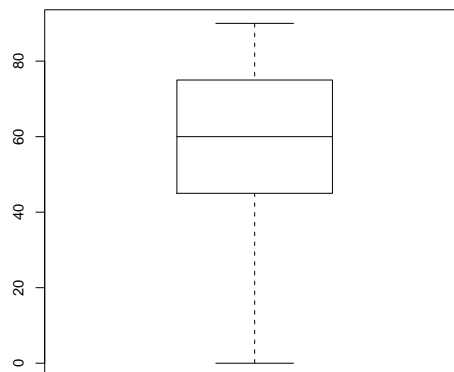
Box-plot della distribuzione di X_1



Box-plot della distribuzione di X_2



Box-plot della distribuzione di X_3

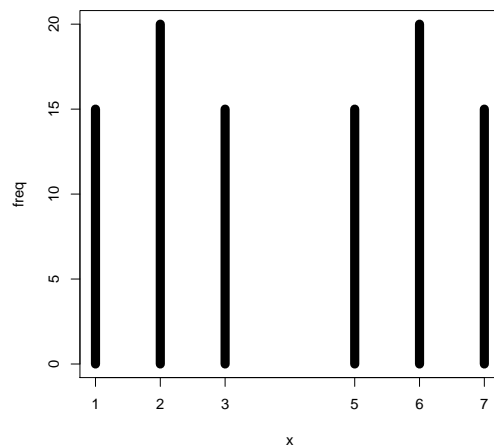
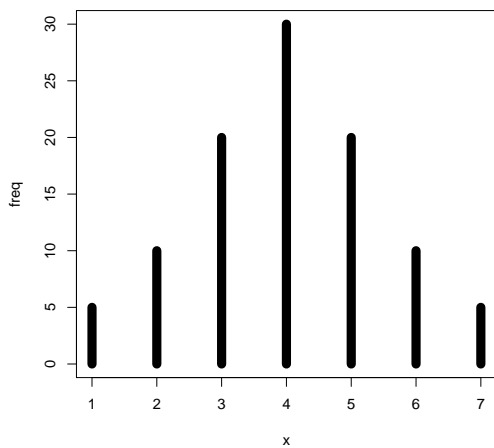


La varianza

Consideriamo le due distribuzioni

x_i	n_i	N_i
1	5	5
2	10	15
3	20	35
4	30	65
5	20	85
6	10	95
7	5	100
	100	

x_i	n_i	N_i
1	15	15
2	20	35
3	15	50
4	—	—
5	15	65
6	20	85
7	15	100
	100	



Calcoliamo media, moda e mediana delle due distribuzioni

(a) La moda è 4

La media è

$$\frac{1}{100} \sum_{i=1}^7 x_i n_i = \frac{400}{100} = 4$$

La mediana

$$(n + 1)/2 = 50.5 \quad n = 100$$

$$x_i \text{ di posizione } n/2 = 50 \quad x_{(50)} = 4$$

$$x_i \text{ di posizione } n/2 + 1 = 51 \quad x_{(51)} = 4$$

La mediana è 4

(b) Le mode sono 2 e 6 (Bimodale)

La media è 4

La mediana

$$(n + 1)/2 = 50.5 \quad n = 100$$

$$x_i \text{ di posizione } n/2 = 50 \quad x_{(50)} = 3$$

$$x_i \text{ di posizione } n/2 + 1 = 51 \quad x_{(51)} = 5$$

$$\text{La mediana è } \frac{3+5}{2} = 4$$

L'indice che misura la dispersione basato sugli scarti al quadrato dalla media:

La **varianza**

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

La **varianza campionaria**

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

Per il calcolo di tali indici si osservi che

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

e quindi

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

$$\bar{s}_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2 \right)$$

In presenza frequenze (assolute o relative)

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_n)^2 \cdot n_i = \sum_{i=1}^k (x_i - \bar{x}_n)^2 \cdot f_i$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^k x_i^2 \cdot n_i - (\bar{x}_n)^2 = \sum_{i=1}^k x_i^2 \cdot f_i - (\bar{x}_n)^2$$

Calcoliamo ora le varianze delle due distribuzioni

x_i	n_i	x_i^2	$x_i^2 \cdot n_i$		x_i	n_i	x_i^2	$x_i^2 \cdot n_i$
1	5	1	5		1	15	1	15
2	10	4	40		2	20	4	80
3	20	9	180		3	15	9	135
4	30	16	480		4	—	—	—
5	20	25	500		5	15	25	375
6	10	36	360		6	20	36	720
7	5	49	245		7	15	49	735
	100		1810			100		2060

$$\sigma_a^2 = \frac{1810}{100} - 4^2 = 2.1$$

mentre per quella (b)

$$\sigma_b^2 = \frac{2060}{100} - 4^2 = 4.6$$

Scarto quadratico medio

È espresso nella stessa unità di misura del fenomeno

$$\sigma = \sqrt{\sigma^2}$$

Regola dei 3-*sigma*

La maggior parte dei dati (il 89%) deve trovarsi nell'intervallo

$$[\bar{x}_n - 3 \cdot \sigma; \bar{x}_n + 3 \cdot \sigma]$$

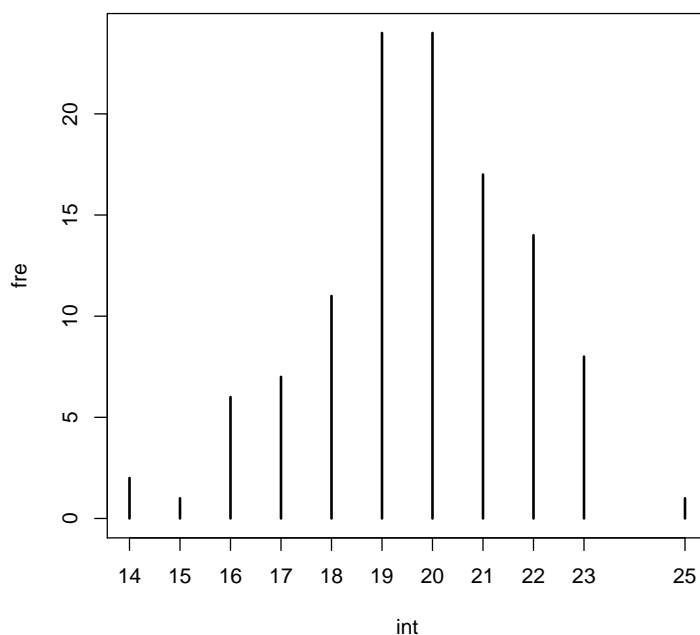
I dati esterni a tale intervallo sono detti *outlier*

Esercizio In uno studio per valutare la resistenza cubica a compressione di una particolare miscela di calcestruzzo si hanno disposizione i dati relativi a 115 campioni, i cui dati sono riportati nella seguente tabella (la resistenza è misurata in N/mm^2)

resistenza	14	15	16	17	18	19	20	21	22	23	25
frequenze	2	1	6	7	11	24	24	17	14	8	1

Calcolare la mediana, i due quartili e il quinto e il novantacinquesimo percentile, la varianza, la varianza campionaria e lo s.q.m., l'intervallo determinato dalla regole dei 3 σ e il box-plot

La Figura riporta il grafico delle frequenze assolute.



x_i	n_i	N_i	F_i
14	2	2	0.017
15	1	3	0.026
16	6	9	0.078
17	7	16	0.139
18	11	27	0.235
19	24	51	0.443
20	24	75	0.652
21	17	92	0.800
22	14	106	0.922
23	8	114	0.991
25	1	115	1.000

Da questa tabella, in particolare dall'ultima colonna, possiamo dedurre $Me_X = 20$, $Q_1 = 19$, $Q_3 = 21$. inoltre il quinto percentile $C_5 = 16$ mentre il novantacinquesimo $C_{95} = 23$.

Inoltre la media aritmetica risulta $\bar{x} = 19.7$, la varianza è $\sigma^2 = 4.21$, mentre $s^2 = 4.25$. Lo s.q.m. risulta $\sigma = 2.05$ e l'intervallo ottenuto con la regola dei 3 sigma

$$(\bar{x} - 3\sigma, \bar{x} + 3\sigma) = (13.55, 25.86)$$

Nessun dato è fuori da questo intervallo.

Disegniamo il boxplot della distribuzione.

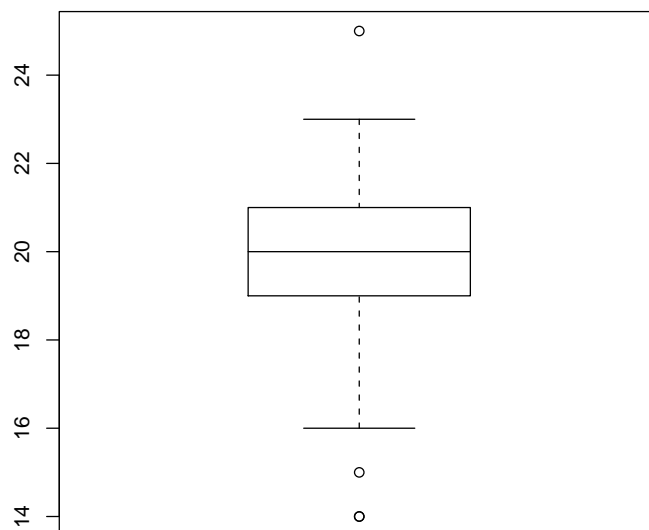
Calcoliamo i baffi:

$$Q_1 - 1.5(Q_3 - Q_1) = 19 - 1.5(21 - 19) = 16$$

Il baffo inferiore è 16

$$Q_3 + 1.5(Q_3 - Q_1) = 21 + 1.5(21 - 19) = 24$$

Il baffo superiore è 23 poiché non c'è il dato 24



Coefficiente di variazione

- se si vuole eliminare l'influenza dovuta all'unità di misura;
- se si vuole confrontare la variabilità di due diversi fenomeni;
- se si vuole eliminare l'influenza dovuta all'ampiezza campionaria n .

Si calcola il Coefficiente di variazione

$$CV = \frac{\sigma}{|\bar{x}_n|} \geq 0$$

Si tratta di un indice di dispersione che è un numero puro

Esercizio Calcolare la varianza, la varianza campionaria e lo s.q.m. per le distribuzioni Z e W . ($\sigma_Z^2 = 2.27$, $\sigma_W^2 = 216.9$, $s_Z^2 = 2.16$, $s_W^2 = 206.1$, $\sigma_Z = 1.47$, $\sigma_W = 14.36$). Calcolare il coefficiente di variazione per le distribuzioni Z e W e stabilire quale risulta più variabile. $\bar{Z} = 1.8$, $\bar{W} = 60.4$, $CV_Z = 0.81$, $CV_W = 0.22$.

Medie e varianze per gruppi

Vogliamo risalire alla media e alla varianza totale delle retribuzioni in una particolare azienda conoscendo i dati relativi alle donne e agli uomini

n_U	n_D	\bar{x}_U	\bar{x}_D	σ_U^2	σ_D^2
42500	12000	50.72	44.02	275.78	163.23

La media generale \bar{x}_n si ottiene come la *media delle medie*

$$\begin{aligned}\bar{x}_n &= \frac{\bar{x}_d \cdot n_d + \bar{x}_u \cdot n_u}{n_d + n_u} \\ &= \frac{50.72 \cdot 42500 + 44.02 \cdot 12000}{54500} = 49.24\end{aligned}$$

In generale se vi sono k gruppi, di media \bar{x}_i e numerosità n_i

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^k \bar{x}_i \cdot n_i$$

La varianza generale σ^2 si ottiene come somma di due indici

$$\sigma^2 = \sigma_B^2 + \sigma_W^2$$

dove σ_B^2 è detta *varianza between*, cioè *varianza tra i gruppi* e si calcola come la *varianza delle medie* dei gruppi

$$\sigma_B^2 = \frac{(\bar{x}_d - \bar{x}_n)^2 \cdot n_d + (\bar{x}_u - \bar{x}_n)^2 \cdot n_u}{n_u + n_d}$$

σ_W^2 è detta *varianza within*, cioè *varianza nei gruppi* e si calcola come la *media delle varianze* dei gruppi

$$\sigma_W^2 = \frac{\sigma_d^2 \cdot n_d + \sigma_u^2 \cdot n_u}{n_u + n_d}$$

In generale, indicata con σ_i^2 , la varianza dei gruppi i

$$\sigma_B^2 = \frac{1}{n} \sum_{i=1}^k (\bar{x}_i - \bar{x}_n)^2 \cdot n_i \quad (\text{varianza delle medie})$$

$$\sigma_W^2 = \frac{1}{n} \sum_{i=1}^k \sigma_i^2 \cdot n_i \quad (\text{media delle varianze})$$

Calcoliamo la varianza totale delle retribuzioni:

$$\begin{aligned}\sigma_B^2 &= \frac{(\bar{x}_d - \bar{x}_n)^2 \cdot n_d + (\bar{x}_u - \bar{x}_n)^2 \cdot n_u}{n_u + n_d} = \\ &= \frac{(44.02 - 49.24)^2 \cdot 12000 + (50.72 - 49.24)^2 \cdot 42500}{54500} \\ &= 7.71\end{aligned}$$

$$\begin{aligned}\sigma_W^2 &= \frac{\sigma_d^2 \cdot n_d + \sigma_u^2 \cdot n_u}{n_u + n_d} \\ &= \frac{163.23 \cdot 12000 + 275.78 \cdot 42500}{54500} \\ &= 251.00\end{aligned}$$

Quindi

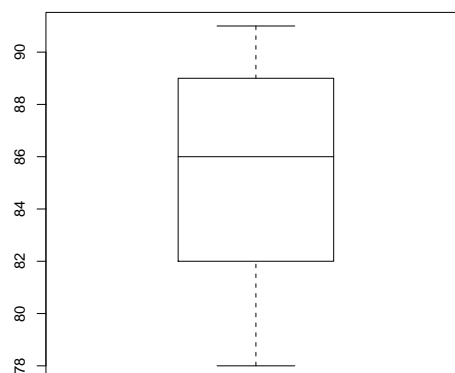
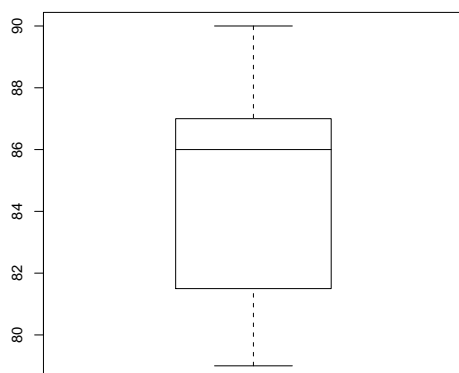
$$\sigma^2 = \sigma_B^2 + \sigma_W^2 = 7.71 + 251.00 = 258.71$$

Esercizio: La seguente tabella riporta i dati relativi alla produzione di bobine di un filato di alta qualità prodotto da due macchine per la filatura, le lunghezze sono espresse in m . Per la macchina denominata A si hanno a disposizione 7 campioni prodotti, per la macchina denominata B si hanno a disposizione 9 campioni.

Macchina A	87	79	81	86	90	87	82			
Macchina B	82	79	78	84	88	91	89	90	86	

Calcolare la varianza tra i gruppi e la varianza nei gruppi. Dire in quale gruppo c'è maggiore variabilità

Disegniamo il boxplot delle due distribuzioni



La media delle due distribuzioni è

$$\bar{x}_A = 84.57, \quad \bar{x}_B = 85.22$$

La media TOTALE è

$$\bar{x} = \frac{1}{16} (84.57 * 7 + 85.22 * 9) = 84.94$$

La varianza delle due distribuzioni è

$$\sigma_A^2 = 13.39, \quad \sigma_B^2 = 20.17$$

La varianza TRA i gruppi è

$$\sigma_B^2 = \frac{1}{16} \left((\bar{x}_A - \bar{x})^2 n_A + (\bar{x}_B - \bar{x})^2 n_B \right)$$

da cui

$$\sigma_B^2 = \frac{1}{16} (0.13 * 7 + 0.08 * 9) = 0.10$$

La varianza NEI gruppi è

$$\sigma_W^2 = \frac{1}{16} (\sigma_A^2 \cdot n_A + \sigma_B^2 \cdot n_B)$$

da cui

$$\sigma_W^2 = \frac{1}{16} (13.39 * 7 + 20.17 * 9) = 17.20$$

La varianza TOTALE è

$$\sigma^2 = \sigma_B^2 + \sigma_W^2 = 17.31$$

Infine

$$CV_A = \frac{\sqrt{13.39}}{84.57} = 0.04 \quad CV_B = \frac{\sqrt{20.17}}{84.94} = 0.05$$

I cuculi e Darwin

- I cuculi depongono le proprie uova nei nidi di altri uccelli.
- In certi territori i cuculi sembrano preferire una specie di uccello come “ospite”, in altri un'altra.
- Darwin: ci si aspetta quindi una qualche forma di adattamento dell'uovo del cuculo a quella dell'uccello “ospite”.
- Per verificare questa idea sono state misurate le lunghezze (in *mm*) di alcune uova di cuculo trovate in nidi di pettirossi e di scriccioli in due territori, uno in cui i cuculi “preferiscono” i pettirossi, l'altro in cui “preferiscono” gli scriccioli.

Pettirossi

21.05 21.85 22.05 22.05 22.05 22.25 22.45 22.45 22.65
23.05 23.05 23.05 23.05 23.05 23.25 23.85

Scriccioli

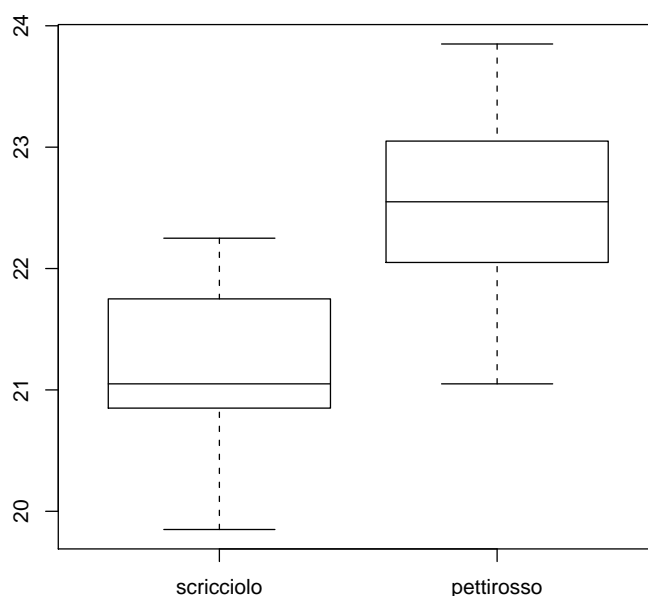
19.85 20.05 20.25 20.85 20.85 20.85 21.05 21.05 21.05
21.25 21.45 22.05 22.05 22.05 22.25

E' evidente che le uova deposte nei nidi di scricciolo sono tendenzialmente più piccole. Poichè le uova degli scriccioli sono più piccole di quelle dei pettirossi sembra che i cuculi diano ragione a Darwin.

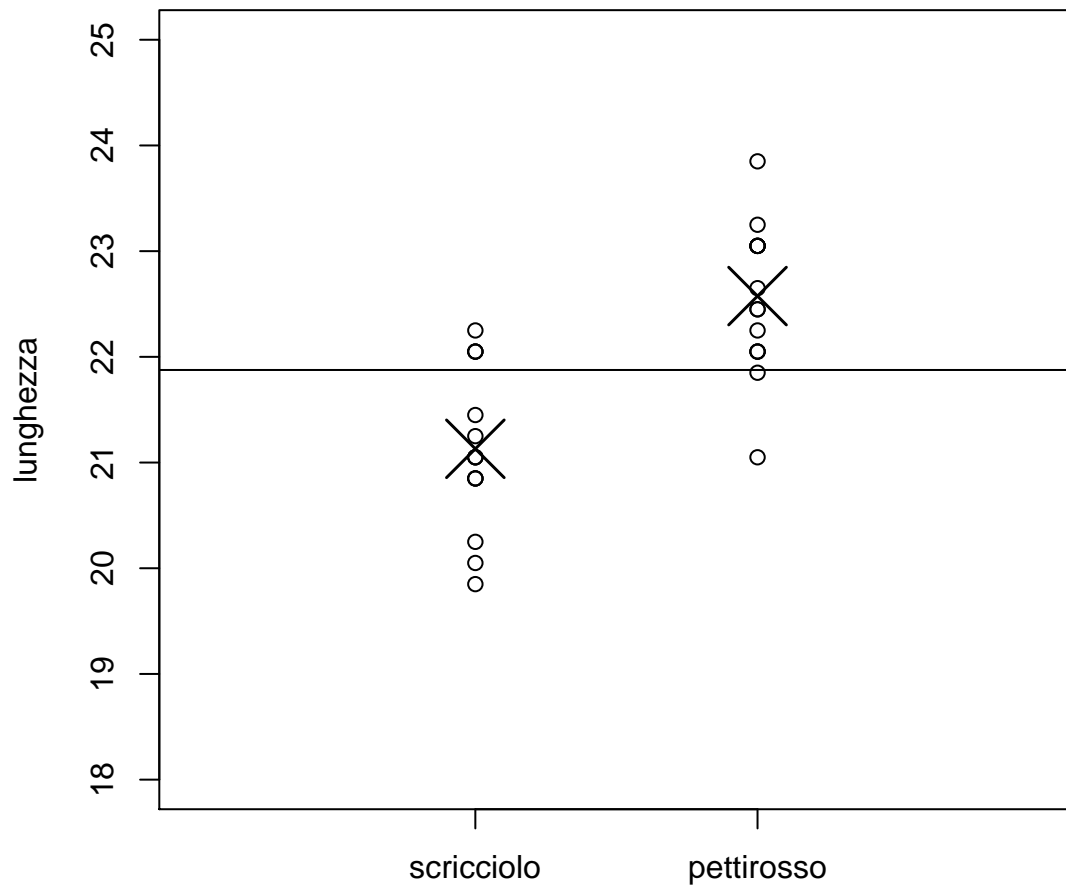
Alcune misure di sintesi per le lunghezze

ospite	n	\bar{x}	Me	σ	Q_1	Q_3
pettirosso	16	22.57	22.55	0.66	22.05	23.05
scricciolo	15	21.13	21.05	0.72	20.85	21.75

Il boxplot ci da un'idea di come sono diverse le distribuzioni



Calcoliamo la media totale



$$\bar{x}_n = \frac{22.57 \cdot 16 + 21.13 \cdot 15}{31} = 21.87$$

$$\sigma_B^2 = \frac{(22.57 - 21.87)^2 \cdot 16 + (21.13 - 21.87)^2 \cdot 15}{31} = 0.52$$

$$\sigma_W^2 = \frac{0.66^2 \cdot 16 + 0.72^2 \cdot 15}{31} = 0.48$$

La varianza totale

$$\sigma^2 = \sigma_B^2 + \sigma_W^2 = 0.52 + 0.48 = 1.00$$

Vediamo quale delle due distribuzioni è più variabile

$$CV_S = \frac{0.72}{21.13} = 0.03, \quad CV_P = \frac{0.66}{22.57} = 0.03$$