Esercitazione 4

Statistica

Alfonso Iodice D'Enza
iodicede@gmail.com

Università degli studi di Cassino

Outline

Connessione in media

Si consideri una variabile Y quantitativa ed una variabile X qualitativa con ad esempio modalità (A, B, C). Siano n_a, n_b e n_c il numero di unità che presentano ciascuna delle modalità della variabile X, quindi $n = n_a + n_b + n_c$.
La media di Y e si ottiene considerando la distribuzione di Y è

$$\mu_y = \overline{y} = \frac{1}{n} \sum_{i=1}^{h} y_i$$

Le medie di Y condizionate a ciascuna delle modalità della variabile X è

$$\overline{y}_a = \overline{y}|A = \frac{1}{n_a} \sum_{i=1}^{n_a} y_i$$

$$\overline{y}_b = \overline{y}|B = \frac{1}{n_b} \sum_{i=1}^{n_b} y_i$$

$$\overline{y}_c = \overline{y}|C = \frac{1}{n_c} \sum_{i=1}^{n_c} y_i$$

Connessione in media

Si consideri una variabile Y quantitativa ed una variabile X qualitativa con ad esempio modalità (A, B, C). Siano n_a, n_b e n_c il numero di unità che presentano ciascuna delle modalità della variabile X, quindi $n = n_a + n_b + n_c$.
La media di Y e si ottiene considerando la distribuzione di Y è

$$\mu_y = \overline{y} = \frac{1}{n} \sum_{i=1}^{h} y_i$$

Le medie di Y condizionate a ciascuna delle modalità della variabile X è

$$\overline{y}_a = \overline{y}|A = \frac{1}{n_a} \sum_{i=1}^{n_a} y_i$$

$$\overline{y}_b = \overline{y}|B = \frac{1}{n_b} \sum_{i=1}^{n_b} y_i$$

$$\overline{y}_c = \overline{y}|C = \frac{1}{n_c} \sum_{i=1}^{n_c} y_i$$

Decomposizione della devianza

Ricordando che la devianza il numeratore della varianza... sia $j = 1, 2, 3$ se si fa riferimento alle modalità A, B, C rispettivamente.

$$
\begin{aligned}
Dev_y &= \sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_i - \overline{y})^2 = \\
&= \sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_i - \overline{y}_j + \overline{y}_j - \overline{y})^2 = \\
&= \sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_i - \overline{y}_i)^2 + \sum_{j=1}^{3}\sum_{i=1}^{n_j}(\overline{y}_i - \overline{y})^2 + \\
&+ 2\sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_j - \overline{y}_i)(\overline{y}_i - \overline{y})
\end{aligned}
$$

Decomposizione della devianza

$$= \sum_{j=1}^{3} \left[\sum_{i=1}^{n_j} (y_i - \overline{y}_j)^2 \right] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j +$$

$$+ 2 \sum_{j=1}^{3} (y_i - \overline{y}_j) \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y}) =$$

$$= \sum_{j=1}^{3} [Dev(Y \mid X = x_j)] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j =$$

$$= Devianza(interna) + Devianza(esterna)$$

Decomposizione della devianza

$$= \sum_{j=1}^{3} \left[\sum_{i=1}^{n_j} (y_i - \overline{y}_j)^2 \right] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j +$$

$$+ 2 \sum_{j=1}^{3} (y_i - \overline{y}_j) \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y}) =$$

$$= \sum_{j=1}^{3} [Dev(Y \mid X = x_j)] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j =$$

$$= Devianza(interna) + Devianza(esterna)$$

Decomposizione della devianza

$$= \sum_{j=1}^{3} \left[\sum_{i=1}^{n_j} (y_i - \overline{y}_j)^2 \right] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j +$$

$$+ 2 \sum_{j=1}^{3} (y_i - \overline{y}_j) \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y}) =$$

$$= \sum_{j=1}^{3} [Dev(Y \mid X = x_j)] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j =$$

$$= Devianza(interna) + Devianza(esterna)$$

Decomposizione della devianza

$$= \sum_{j=1}^{3} \left[\sum_{i=1}^{n_j} (y_i - \overline{y}_j)^2 \right] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j +$$

$$+ 2 \sum_{j=1}^{3} (y_i - \overline{y}_j) \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y}) =$$

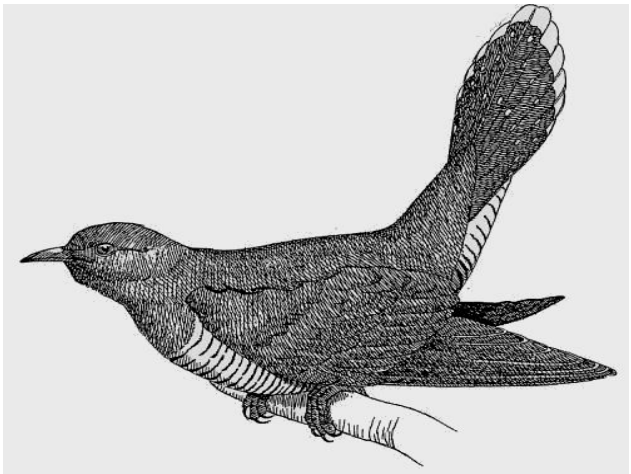
$$= \sum_{j=1}^{3} [Dev(Y \mid X = x_j)] + \sum_{j=1}^{3} (\overline{y}_j - \overline{y})^2 n_j =$$

$$= Devianza(interna) + Devianza(esterna)$$

Esempio di calcolo della decomposizione della devianza

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 deposte cambia in media a seconda del tipo di nido in cui vengono deposte.

Esempio di calcolo della decomposizione della devianza

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

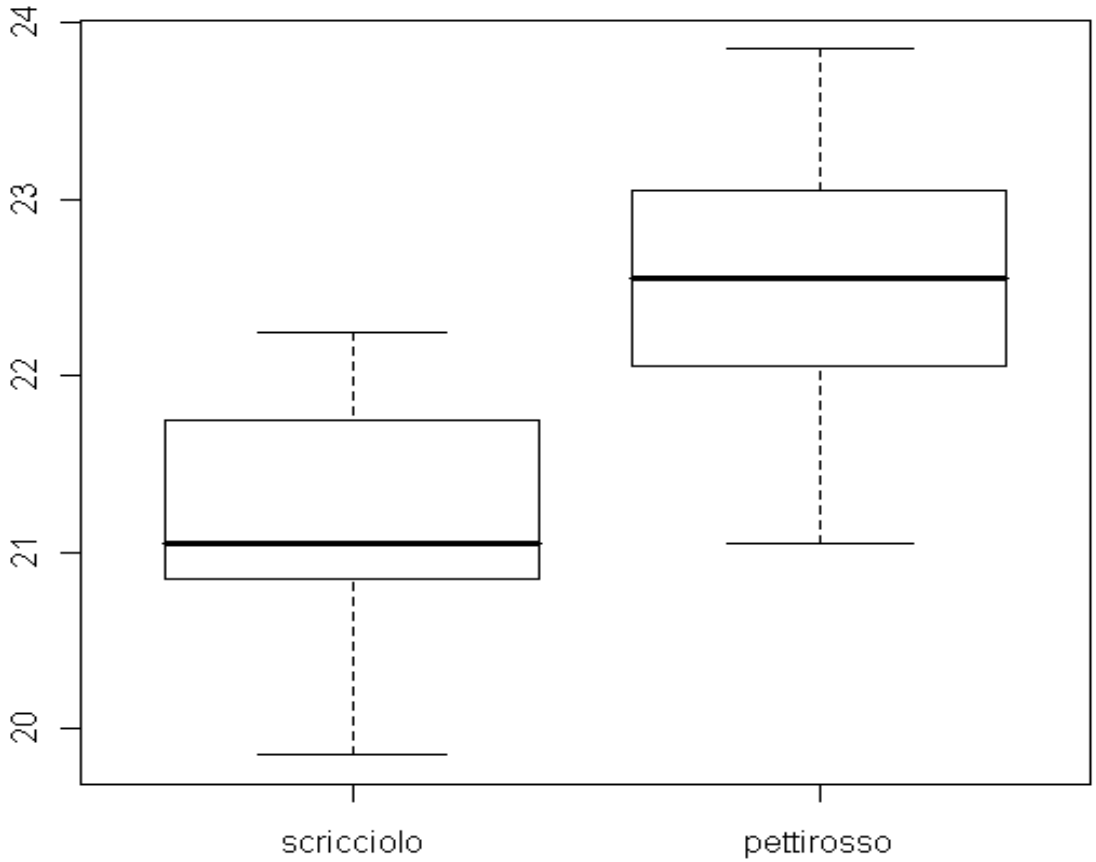
```
> P
      [,1]
 [1,] 21.05
 [2,] 21.85
 [3,] 22.05
 [4,] 22.05
 [5,] 22.05
 [6,] 22.25
 [7,] 22.45
 [8,] 22.45
 [9,] 22.65
[10,] 23.05
[11,] 23.05
[12,] 23.05
[13,] 23.05
[14,] 23.05
[15,] 23.25
[16,] 23.85
```

```
> summary(pettirosso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.05   22.05   22.55   22.57   23.05   23.85
```

Esempio di calcolo della decomposizione della devianza

Confronto tra le distribuzioni

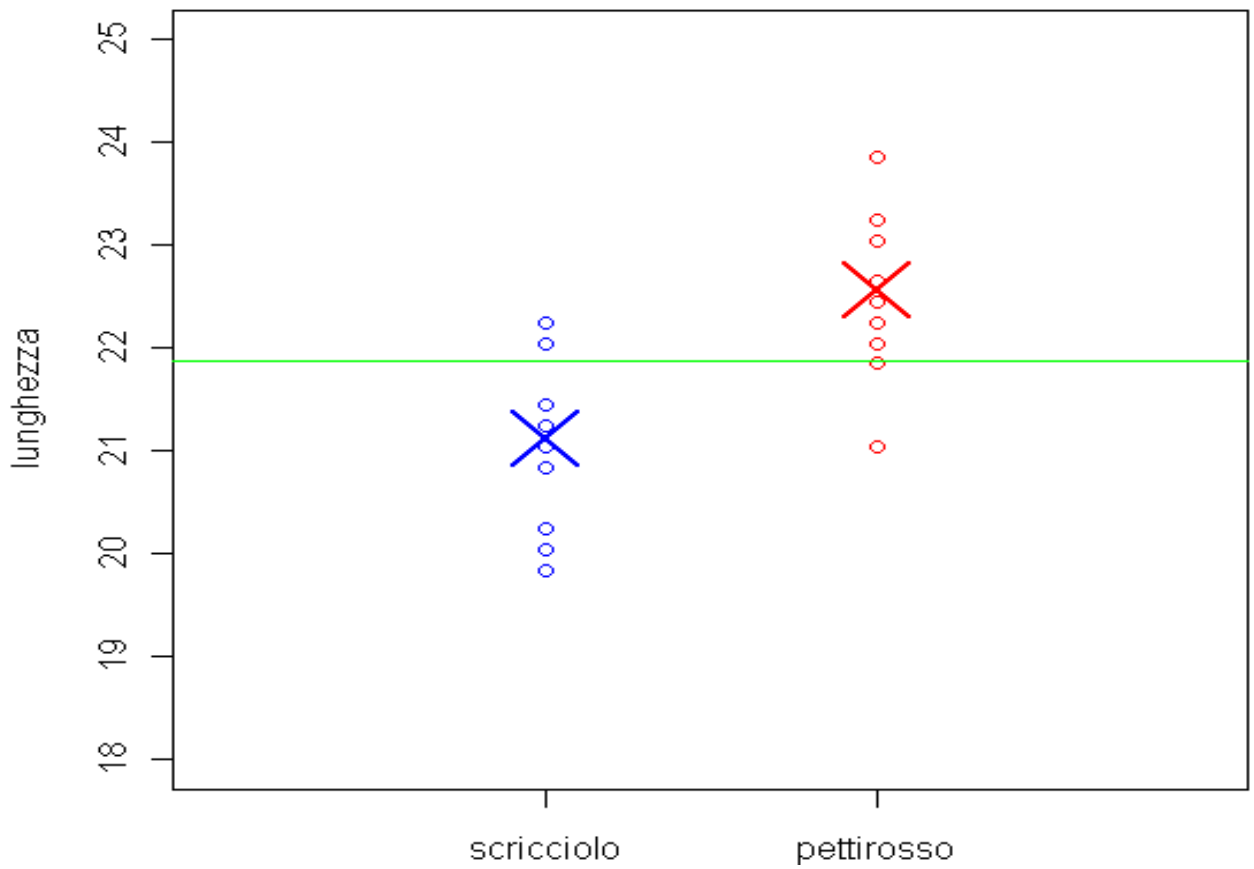
Un primo confronto grafico via box plot tra le due distribuzioni mostra che le uova deposte in nidi di pettirosso hanno una lunghezza maggiore di quelle deposte in nidi di scricciolo.



Esempio di calcolo della decomposizione della devianza

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.



Esempio di calcolo della decomposizione della devianza

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

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