Statistica I

Unità P: la dipendenza tra variabili

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità P

Argomenti affrontati

- Dipendenza in media e altre modalità di dipendenza.
- Confronto tra indice η^2 ed indice χ^2.

Riferimenti al libro di testo

- §7.4

Descrizione del problema

- È noto che i **cuculi** depongono le proprie **uova** nei nidi di altri uccelli, a cui viene poi lasciato il compito della cova.
- I cuculi scelgono differenti specie di uccello (ad es: pettirossi o scriccioli) da utilizzare come nidi, a seconda del territorio.
- È presente una forma di adattamento dell'uovo del cuculo a quella dell'uccello "ospite".
- Per verificare quest'idea sono state misurate le **lunghezze** (in mm) di alcune uova di cuculo trovate in nidi di pettirossi e di scriccioli in due territori.

I dati grezzi

Pettirossi

```
[1] 21.05 21.85 22.05 22.05 22.05 22.25 22.45 22.45 22.65 23.05 23.05
[12] 23.05 23.05 23.05 23.25 23.85
```

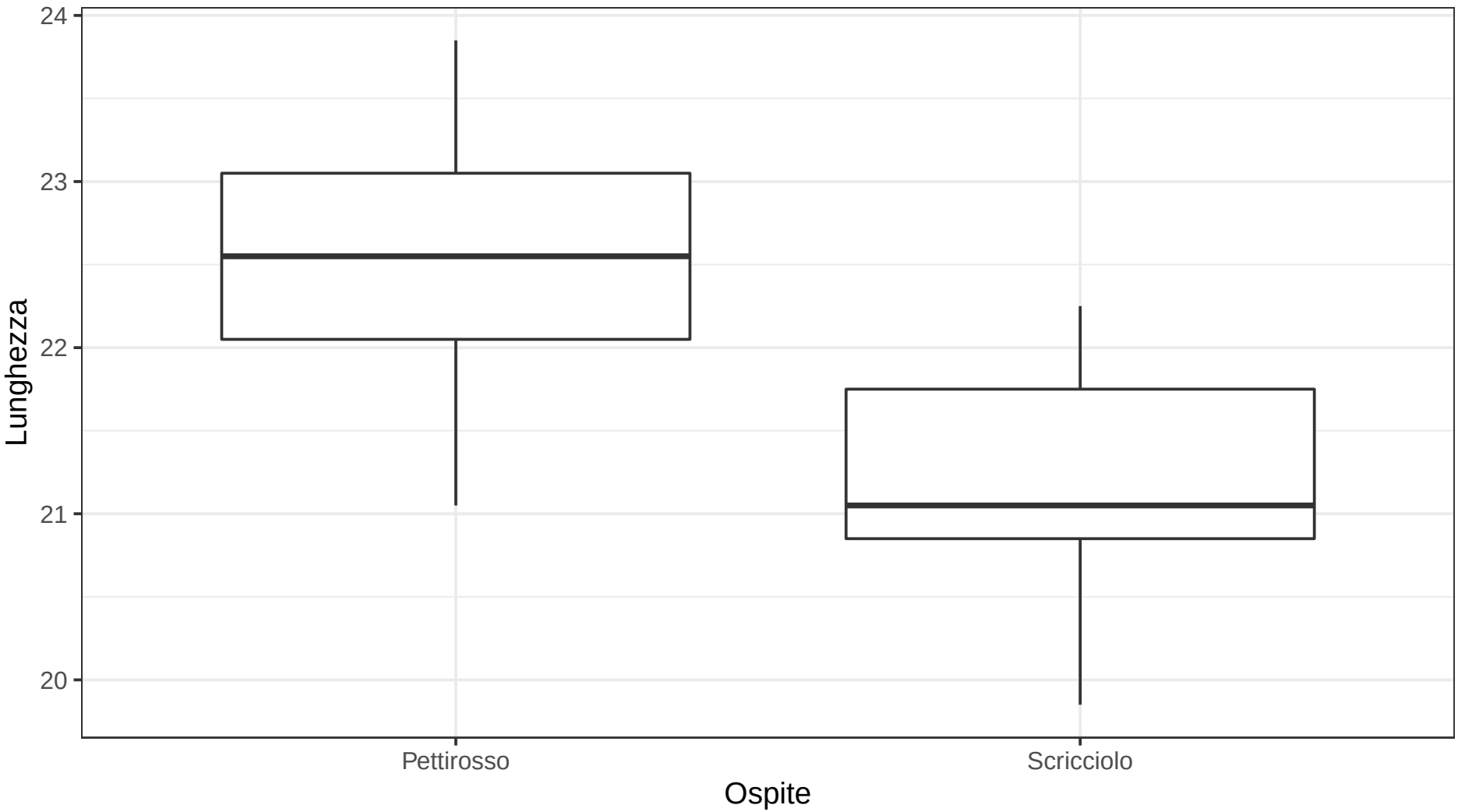
Scriccioli

```
 [1] 19.85 20.05 20.25 20.85 20.85 20.85 21.05 21.05 21.05 21.25 21.45
[12] 22.05 22.05 22.05 22.25
```

- Queste osservazioni possono essere visti come un insieme di **dati bivariati**, organizzati come segue

Uovo	`ospite`	`lunghezza`
1	Pettirosso	21.05
2	Pettirosso	21.85
⋮	⋮	⋮
31	Scricciolo	22.25

I boxplot



Alcune analisi descrittive

- È evidente che le uova deposte nei nidi di scricciolo sono tendenzialmente più piccole.
- Questo è confermato anche dalla media e dalla mediana della variabile `lunghezza`.

Ospite	Numerosità	Media	Mediana	Deviazione standard
Pettirosso	16	22.575	22.55	0.663
Scricciolo	15	21.130	21.05	0.719

- È inoltre noto che la media complessiva dei dati è pari a $\bar{x} = 21.876$.
- Per valutare l'intensità di questa dipendenza in media, possiamo ottenere il **rapporto di correlazione** η^2 descritto nell'unità N.

Analisi della varianza

- Calcoliamo nel seguito le principali quantità di interesse per l'analisi della varianza.
- La **devianza tra i gruppi** è pari a

$$\mathscr{D}^2_{\text{tr}} = \sum_{j=1}^{n} n_j(\bar{x}_j - \bar{x})^2 = 16(22.575 - 21.876)^2 + 15(21.130 - 21.876)^2 = 16.16536.$$

- La **devianza entro i gruppi** è pari a

$$\mathscr{D}^2_{\text{en}} = d_1^2 + d_2^2 = (16 \times 0.4393750) + (15 \times 0.5162667) = 14.774,$$

per cui la devianza totale è $\mathscr{D}^2 = 16.16536 + 14.774 = 30.93936$.

- Infine, il **rapporto di correlazione** η^2 è pari a

$$\eta^2 = \frac{\mathscr{D}^2_{\text{tr}}}{\mathscr{D}^2} = \frac{16.16536}{30.93936} = 0.522,$$

evidenziando quindi una forte dipendenza in media.

Tabelle di contingenza

- È possibile analizzare questi dati utilizzando gli strumenti dell'unità O.
- In altri termini, possiamo costruire una **tabella di contingenza** per le variabili `ospite` e `lunghezza`.
- Si noti infatti che la variabile `lunghezza` è discreta: alcuni valori sono ripetuti.
- Calcolando ad esempio l'indice di connessione χ^2, è possibile stabilire se le due variabili sono o meno **indipendenti in distribuzione**.
- Si noti che l'indice di dipendenza in media η^2 era abbastanza grande, suggerendo che quindi anche l'indice χ^2 evidenzierà una forte dipendenza.

La distribuzione congiunta

Lunghezza	Pettirosso	Scricciolo	**Totale**
19.85	0	1	1
20.05	0	1	1
20.25	0	1	1
20.85	0	3	3
21.05	1	3	4
21.25	0	1	1
21.45	0	1	1
21.85	1	0	1
22.05	3	3	6
22.25	1	1	2
22.45	2	0	2
22.65	1	0	1
23.05	5	0	5
23.25	1	0	1
23.85	1	0	1
Totale	16	15	31

Le frequenze attese

Lunghezza	Pettirosso	Scricciolo	**Totale**
19.85	0.52	0.48	1
20.05	0.52	0.48	1
20.25	0.52	0.48	1
20.85	1.55	1.45	3
21.05	2.06	1.94	4
21.25	0.52	0.48	1
21.45	0.52	0.48	1
21.85	0.52	0.48	1
22.05	3.10	2.90	6
22.25	1.03	0.97	2
22.45	1.03	0.97	2
22.65	0.52	0.48	1
23.05	2.58	2.42	5
23.25	0.52	0.48	1
23.85	0.52	0.48	1
Totale	16	15	31

L'indice di connessione χ^2

- L'indice di **connessione** χ^2 è pari a

$$\chi^2 = \frac{(0-0.52)^2}{0.52} + \frac{(1-0.48)^2}{0.48} + \cdots + \frac{(0-0.48)^2}{0.48} = 19.98854.$$

- L'indice di **connessione** χ^2 **normalizzato** è pari invece a

$$\chi^2_{\text{norm}} = \frac{\chi^2}{n \min\{h-1, k-1\}} = \frac{19.98854}{31} = 0.6447917.$$

- Questo evidenza una forte **dipendenza in distribuzione** tra le due variabili.

Dipendenza in media, dipendenza in distribuzione

- Le due distribuzioni della `lunghezza` condizionate all'`ospite` sono diverse, ovvero esiste **dipendenza in distribuzione**.
- Inoltre, le due medie erano diverse: tra le due variabili esiste **dipendenza in media**.
- In generale, una variabile numerica y è dipendente (indipendente) in media da un'altra variabile x se le medie delle distribuzioni condizionate sono diverse (uguali) tra loro.
- Queste due forme di dipendenza, seppur legate tra loro, non sono uguali.
- Altre forme di dipendenza sono possibili: dipendenza in mediana, dipendenza in varianza, etc.

Dipendenza in distribuzione

- Si noti che questi concetti di indipendenza (in media, mediana, etc.) sono **più deboli** di quello di indipendenza in distribuzione.

- In altri termini, l'indipendenza in distribuzione implica necessariamente indipendenza in media, mediana, etc. Ad esempio, avremo che

$$\chi^2 = 0 \implies \eta^2 = 0.$$

- Il viceversa non è vero: indipendenza in media, mediana, etc., non necessariamente implicano l'indipendenza in distribuzione. Ad esempio:

$$\eta^2 = 0 \not\implies \chi^2 = 0,$$

  ovvero è possibile che $\eta^2 = 0$ e che $\chi^2 > 0$.

- Discutiamo questo punto con riferimento alla sola indipendenza in media.

Indipendenza in distribuzione $\Longrightarrow$ indipendenza in media

- Richiamiamo la definizione di tabella di contingenza e **supponiamo che la variabile *y* sia numerica**, ovvero che $d_1, \ldots, d_k$ siano dei numeri.

Variabile x	**Variabile y** d_1	$\ldots$	d_j	$\ldots$	d_k	**Totale**
c_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1k}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{ik}	n_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_h	n_{h1}	$\ldots$	n_{hj}	$\ldots$	n_{hk}	n_{h+}
Totale	n_{+1}	$\ldots$	n_{+j}	$\ldots$	n_{+k}	n

Indipendenza in distribuzione $\Longrightarrow$ indipendenza in media

- Le **medie condizionate** $\bar{y}_1, \ldots, \bar{y}_h$, condizionate a $x = c_i$, per $i = 1, \ldots, h$, sono dunque pari a:

$$\bar{y}_i = \frac{1}{n_{i+}} \sum_{j=1}^{k} n_{ij} d_j, \qquad i = 1, \ldots, h.$$

- In caso di **indipendenza in distribuzione** tra y e x, dall'unità O sappiamo che le quantità $n_{ij}/n_{i+} = f_{+j}$ sono uguali, pertanto

$$\bar{y}_i = \sum_{j=1}^{k} f_{+j} d_j, \qquad i = 1, \ldots, h,$$

  che ovviamente implica che $\bar{y}_1 = \cdots = \bar{y}_h$.

- È immediato quindi far vedere che se $\chi^2 = 0$ allora anche $\eta^2 = 0$.

Indipendenza in media $\not\Longrightarrow$ indipendenza in distribuzione

- Per quel che riguarda la seconda affermazione, possiamo procedere con un esempio.
- È facile infatti costruire una tabella di contingenza in cui esiste indipendenza in media ma non indipendenza in distribuzione.
- Si verifichi che questo è quello che accade con la seguente distribuzione congiunta.

	Variabile y					
Variabile x	3	4	5	6	7	**Totale**
c_1	0	1	1	1	0	3
c_2	1	0	1	0	1	3
Totale	1	1	1	1	1	6

- Infatti, si noti che le medie condizionate sono uguali $\bar{y}_1 = \bar{y}_2 = 5$, pertanto $\eta^2 = 0$. Tuttavia, è anche facile verificare che $\chi^2 > 0$.