# Statistica Applicata
# Corso di Laurea in Scienze Naturali
# a. a. 2016/2017

prof. Federico Plazzi

5 Luglio 2017

Nome: _____

Cognome: _____

Matricola: _____

**Alcune indicazioni:**

- La prova è constituita da quattro esercizi; dopo ogni esercizio c'è lo spazio in cui scrivere la risposta o le risposte. In caso questo spazio non sia sufficiente, si può continuare a rispondere sul retro del foglio, avendo cura di indicare il numero dell'esercizio a fianco della continuazione della risposta.
- Alcuni esercizi richiedono semplici calcoli, per i quali è consentito l'uso di una calcolatrice ed eventualmente la consultazione di una o più delle tabelle allegate.
- Altri esercizi richiedono invece la lettura dei dati: verrà valutata in questo caso l'argomentazione che giustifica l'interpretazione fornita.
- La durata massima della prova è di 60 minuti.
- Si prega di non scrivere nulla sulle tabelle allegate.

# 1 Dati

Alcuni entomologi stanno studiando una certa specie di libellula (Odonata: Anisoptera), distribuita su un ampio areale. Vengono effettuati campionamenti in 6 diverse località, indicate con le lettere da A a F, corrispondenti ad altrettante popolazioni, e in ciascuna vengono catturati 6 individui, indicati con i numeri da 1 a 6. Per ciascuno vengono misurate la lunghezza dell'ala (destra) anteriore, la lunghezza dell'ala (destra) posteriore e la lunghezza dell'addome. I rilevamenti sono riassunti in Tabella 1.

# 2 Esercizi

## 2.1 Statistiche di base

### 2.1.1 Apertura alare

Quale popolazione ha le ali anteriori più lunghe, considerando la media aritmetica della lunghezza dell'ala anteriore che, negli anisotteri, è la più lunga delle due?

```
> tapply(forewing,population,mean)
```

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 36.13167 | 36.85000 | 31.55667 | 41.67000 | 35.89167 | 31.63333 |

### 2.1.2 Variabilità interna alla popolazione D

Calcola la deviazione standard della lunghezza dell'addome nella popolazione D.

```
> standard.deviation(abdomen[population == "D"])
[1] 0.6086985
```

Tabella 1: Misurazioni relative a 36 libellule.

| Individuo | Ala anteriore (mm) | Ala posteriore (mm) | Addome (mm) |
|---|---|---|---|
| A1 | 40, 02 | 35, 21 | 35, 55 |
| A2 | 33, 37 | 29, 44 | 35, 62 |
| A3 | 36, 87 | 26, 17 | 33, 57 |
| A4 | 36, 56 | 28, 76 | 37, 15 |
| A5 | 32, 33 | 31, 78 | 33, 31 |
| A6 | 37, 64 | 33, 4 | 39, 61 |
| B1 | 35, 19 | 33, 96 | 35, 17 |
| B2 | 36, 68 | 34, 7 | 37, 54 |
| B3 | 42, 17 | 29, 03 | 36, 85 |
| B4 | 36, 59 | 32, 96 | 36, 47 |
| B5 | 34, 83 | 32, 75 | 34, 49 |
| B6 | 35, 64 | 33, 56 | 35, 15 |
| C1 | 23, 21 | 22, 33 | 34, 06 |
| C2 | 34, 83 | 31, 25 | 30, 33 |
| C3 | 31, 93 | 25, 17 | 30, 29 |
| C4 | 31, 22 | 27, 73 | 25, 46 |
| C5 | 30, 45 | 26, 78 | 30, 27 |
| C6 | 37, 7 | 28, 64 | 29, 7 |
| D1 | 42, 89 | 31, 53 | 41, 23 |
| D2 | 42, 15 | 32, 07 | 41, 21 |
| D3 | 41, 31 | 34, 02 | 42, 8 |
| D4 | 41, 42 | 32, 5 | 42, 04 |
| D5 | 40, 68 | 35, 56 | 41, 14 |
| D6 | 41, 57 | 31, 01 | 42, 07 |
| E1 | 35, 81 | 32, 11 | 36, 61 |
| E2 | 35, 75 | 30, 64 | 35, 82 |
| E3 | 35, 35 | 31, 45 | 35, 25 |
| E4 | 35, 93 | 34, 14 | 35, 99 |
| E5 | 36, 81 | 30, 89 | 35, 22 |
| E6 | 35, 7 | 31, 28 | 35, 41 |
| F1 | 29, 52 | 28, 22 | 30, 09 |
| F2 | 32, 37 | 26, 61 | 31, 09 |
| F3 | 30, 72 | 26, 42 | 34, 51 |
| F4 | 30, 79 | 23, 6 | 33, 03 |
| F5 | 31, 41 | 24, 75 | 28, 25 |
| F6 | 34, 99 | 25, 38 | 31, 56 |

## 2.2 Distribuzione dei risultati

La Figura 1 mostra, in alto, le distribuzioni delle tre variabili osservate e, in basso, i Q-Q plot delle stesse variabili. Di seguito (Tabella 2) sono anche riportati i risultati del test di Shapiro e Wilk effettuato su ciascuna variabile.

Alcuni dei risultati mostrati in Figura 1 potrebbero sembrare in disaccordo

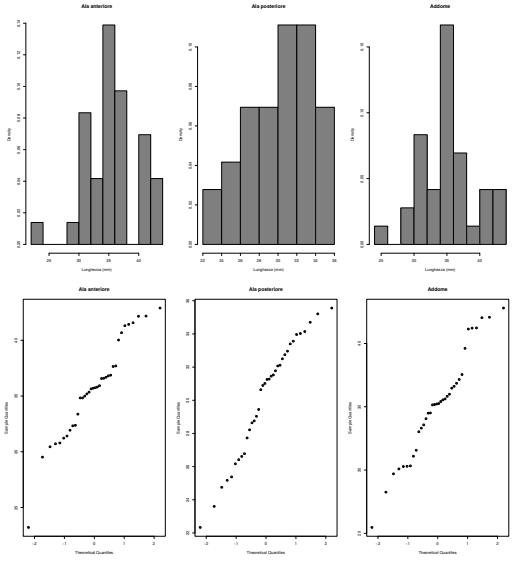


Figura 1: Istogrammi e Q-Q Plot.

Tabella 2: Test di Shapiro e Wilk.

| Variabile | W | p-value |
|---|---|---|
| Ala anteriore | $0,95350$ | $0,1347$ |
| Ala posteriore | $0,95537$ | $0,1543$ |
| Addome | $0,96568$ | $0,3195$ |

con i risultati del test di Shapiro e Wilk. A che cosa potrebbe essere dovuta questa discrepanza? In conclusione, cosa possiamo affermare sulla distribuzione dei dati? C'è qualche variabile con una distribuzione particolare?

*Le variabili sono tutte a distribuzione normale, come si evince dai test di Shapiro e Wilk, che hanno tutti un p-value superiore a 0.05. Alcuni istogrammi non approssimano molto bene la distribuzione normale e soprattutto alcuni Q-Q plot non mostrano la linearità attesa dei punti, ma questo potrebbe essere dovuto, per gli istogrammi, a una scelta sbagliata del numero di categorie da utilizzare e, per tutti e due i tipi di grafico, al fatto che i campioni sono abbastanza pochi (soltanto 36).*

## 2.3 Correlazione ala-addome

La tabella 3 mostra i risultati ottenuti dal calcolo di un modello di correlazione lineare tra lunghezza dell'ala anteriore e lunghezza dell'addome. Cosa possiamo concludere?

Tabella 3: Correlazione tra lunghezza di ala anteriore e addome.

| | Stima | p-value | $r$ | $R^2$ |
|---|---|---|---|---|
| Intercetta | $10,1395$ | | | |
| Pendenza | $0,7009$ | $7,576 \times 10^{-7}$ | $0,7196151$ | $0,5178459$ |

*La correlazione è positiva ($r > 0$) e significativa ($p <<< 0,05$): all'aumento di lunghezza dell'ala anteriore corrisponde l'aumento di lunghezza dell'addome, anche se la forza della correlazione non è grandissima ($R^2 = 0,5178459$).*

## 2.4 One-Way ANOVA

Gli entomologi che conducono lo studio sono interessati a vedere se esistano differenze significative tra le sei popolazioni, basandosi sulla lunghezza dell'ala anteriore. A questo scopo, calcolano la devianza *entro* gruppi e la devianza *tra* gruppi, usando come gruppi le sei popolazioni da A a F. I risultati sono trascritti in Tabella 4.

Tabella 4: One-Way ANOVA. D, devianza; $\sigma^2$, varianza; g.l., gradi di libertà.

| | D | $\sigma^2$ | g.l. | $F$ | p-value |
|---|---|---|---|---|---|
| *tra* | $425,131000$ | | | | |
| *entro* | $218,474800$ | | | | |

1. Per quale motivo l'ANOVA è un approccio valido in questo caso?
2. Completa la tabella 4 calcolando le due varianze e indicando i gradi di libertà.
3. Calcola di valore di $F$: è significativo? Cosa possiamo concludere?
4. L'analisi potrebbe procedere ulteriormente? Perché? Come?

*L'ANOVA è indicata perché la variabile è a distribuzione normale; oltretutto i campioni sono tutti della stessa dimensione.*
*I gradi di libertà sono 5 per la varianza tra gruppi (perché i gruppi in tutto sono 6) e 30 per la varianza entro gruppi (perché ci sono 36 osservazioni e 6 gruppi). Le varianze si ottengono dividendo le rispettive devianze per i gradi di libertà; il valore di F si ottiene dividendo la varianza* tra *gruppi per la varianza* entro *gruppi. I risultati sono mostrati in Tabella 5. Consultando le tabelle allegate si*

Tabella 5: One-Way ANOVA. D, devianza; $\sigma^2$, varianza; g.l., gradi di libertà.

| | D | $\sigma^2$ | g.l. | $F$ | p-value |
|---|---|---|---|---|---|
| *tra* | $425,131000$ | $85,026200$ | 5 | | |
| *entro* | $218,474800$ | $7,282494$ | 30 | $11,675420$ | $< 0,001$ |

*vede che F è altamente significativo: il valore esatto sarebbe* $2,54 \times 10^{-6}$.
*Possiamo quindi concludere che esiste una strutturazione nella specie e che le sei popolazioni non sono tutte uguali per quanto riguarda la lunghezza dell'ala anteriore. Il passo successivo potrebbe essere il test di Tukey per verificare se alcuni gruppi si discostano significativamente dagli altri.*