Statistica Applicata
Corso di Laurea in Scienze Naturali
a. a. 2017/2018

prof. Federico Plazzi

25 Giugno 2018

Nome: _____

Cognome: _____

Matricola: _____

Alcune indicazioni:

- La prova è constituita da cinque 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

Un gruppo di ricercatori ha effettuato delle misure in campo su una particolare specie di pesce d'acqua dolce, misurando la lunghezza e il numero di scaglie lungo la linea laterale. Sono stati effettuati due campionamenti in due diversi corsi d'acqua, etichettati come "A" e "B": in ciascuno, sono stati catturati 15 individui. I risultati sono mostrati in tabella 1.

Tabella 1: Risultati delle misure effettuate.

Popolazione A		Popolazione B	
Lunghezza (cm)	Scaglie	Lunghezza (cm)	Scaglie
43.59	30	46.13	31
44.39	32	46.19	31
46.69	34	46.92	29
44.54	31	52.31	34
49.64	36	53.19	31
46.58	32	42.67	28
45.31	31	45.41	28
49.85	34	41.19	26
40.91	29	43.15	28
51.75	36	47.11	33
43.77	32	44.3	31
46.39	34	49.32	32
39.06	29	51.86	29
47.42	35	54.02	30
48.15	33	40.46	32

2 Esercizi

2.1 Statistiche di base

Calcolare media, varianza e deviazione standard della lunghezza degli esemplari della popolazione A.

Media	*Varianza*	*Deviazione standard*
$45,87$	$10,57$	$3,25$

2.2 Distribuzione dei dati

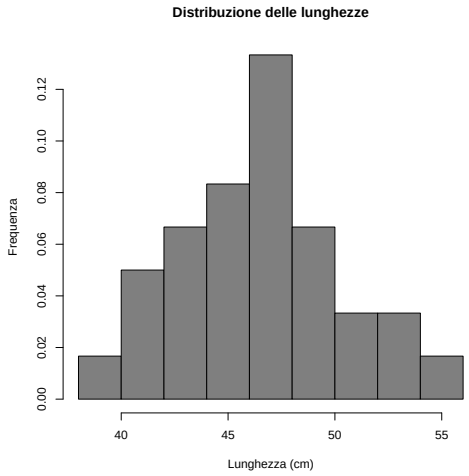

Figura 1: Distribuzione delle lunghezze di entrambe le popolazioni

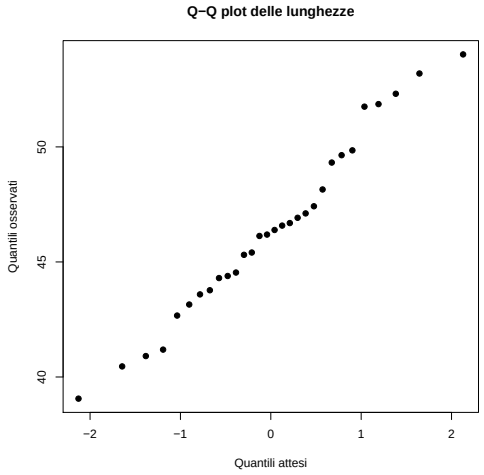

Figura 2: Q-Q Plot delle lunghezze di entrambe le popolazioni

La figura 1 mostra la distribuzione delle lunghezze di tutti e 30 gli individui campionati messe insieme. La figura 2 ne mostra il Q-Q plot. Se si va a verificare la correlazione lineare tra i punti mostrati nel Q-Q plot, quale dei seguenti è il risultato? Che cosa si può concludere a proposito della distribuzione dei dati?

	Stima	p-value	r	R^2
Intercetta	$46,41$			
Pendenza	$3,85$	$< 2 \times 10^{-16}$	$0,99$	$0,98$

	Stima	p-value	r	R^2
Intercetta	$55,29$			
Pendenza	$-3,04$	$3,21 \times 10^{-11}$	$0,97$	$0,95$

	Stima	p-value	r	R^2
Intercetta	$40,84$			
Pendenza	$4,94$	$7,54 \times 10^{-8}$	$0,12$	$0,23$

Il risultato più verosimile è il primo: dalla figura 2 si vede chiaramente che la correlazione deve essere positiva, per cui il secondo risultato non è possibile, indicando un coefficiente angolare negativo. Il terzo risultato non può essere corretto per l'intercetta (che dal grafico risulta circa 45*) e per il fatto che il p-value elevato (*$7,54 \times 10^{-8}$*) è poco compatibile con il modesto coefficiente r per questo numero di punti; il valore di R^2, infine, non è il quadrato di quello di r. Considerando quindi la correlazione lineare significativa tra i punti del Q-Q plot (*$p < 2 \times 10^{-16}$*) e la forma del grafico della distribuzione, si deve conlcudere che le lunghezze si distribuiscono in modo normale.*

2.3 Differenze in lunghezze tra le due popolazioni

Per verificare se esistano differenze in lunghezza tra le due popolazioni, si eseguono diversi test statistici. Qual è quello corretto? Cosa si può concludere?

- Test t a campioni appaiati:

```
Paired t-test

t = -0.73127, df = 14, p-value = 0.0213
alternative hypothesis:  true difference in means is not equal
to 0
```

- Test t a campioni non appaiati:

```
Two Sample t-test

t = -0.75902, df = 26.312, p-value = 0.4546
alternative hypothesis:  true difference in means is not equal
to 0
```

- Test di Wilcoxon:

```
Wilcoxon signed rank test

V = 45, p-value = 0.4212
alternative hypothesis:  true location shift is not equal to 0
```

- Test di Mann e Whitney:

```
Wilcoxon rank sum test

W = 101, p-value = 0.0004
alternative hypothesis:  true location shift is not equal to 0
```

I campioni non sono appaiati (non c'è ragione di confrontare il primo pesce della popolazione A con il primo della popolazione B, il secondo con il secondo e così via) e, dall'esercizio precedente, si sa che la distribuzione della lunghezza è normale: il test corretto è perciò il secondo, il test t a campioni non appaiati.

2.4 Scaglie lungo la linea laterale

Il numero di scaglie lungo la linea laterale è legato in qualche modo alla lunghezza dell'esemplare? Le tabelle 2 e 3 mostrano il risultato delle ricerca di un modello di correlazione lineare tra lunghezza e numero di scaglie lungo la linea laterale nelle due popolazioni. Cosa si può concludere?

Tabella 2: Popolazione A

	Stima	p-value	r	R^2
Intercetta	$4,20$			
Pendenza	$0,62$	$3,35 \times 10^{-6}$	$0,91$	$0,82$

Tabella 3: Popolazione B

	Stima	p-value	r	R^2
Intercetta	$20,83$			
Pendenza	$0,20$	$0,14$	$0,40$	$0,16$

Sebbene non siano significativamente diverse dal punto di vista della lunghezza, apparentemente le due popolazioni differiscono però per quanto riguarda il numero di scaglie lungo la linea laterale. Mentre infatti la correlazione tra lunghezza e numero di scaglie lungo la linea laterale è significativa e abbastanza forte nella popolazione A ($p = 3,35 \times 10^{-6}$, $R^2 = 0,82$), nella popolazione B la correlazione non è significativa ($p = 0,14$).

2.5 Individuo "aberrante"

Qualche giorno dopo i campionamenti, viene portato ai ricercatori da alcuni pescatori un individuo, pescato nel corso d'acqua A, definito "aberrante" per la sua dimensione: $39,11$ cm appena. È vero?

È vero. Si sa dal secondo esercizio che la distribuzione delle lunghezze è normale, per cui si può procedere a calcolare la deviata normale per questo individuo:

$$Z = \frac{X - \mu}{\sigma} = \frac{39,11 - 45,87}{3,25} = -2,08$$

Dalla tabella si ottiene che questo valore di Z è associato a un'area sotto la curva normale di $1,0000 - 0,9812 = 0,0192$, il che significa che valori così o più piccoli corrispondono a una quota di $1,92\%$ della distributione totale, che è significativamente piccola in quanto inferiore al 5% (e anche al $2,5\%$ nel caso di un test a due code).